APPENDIX A

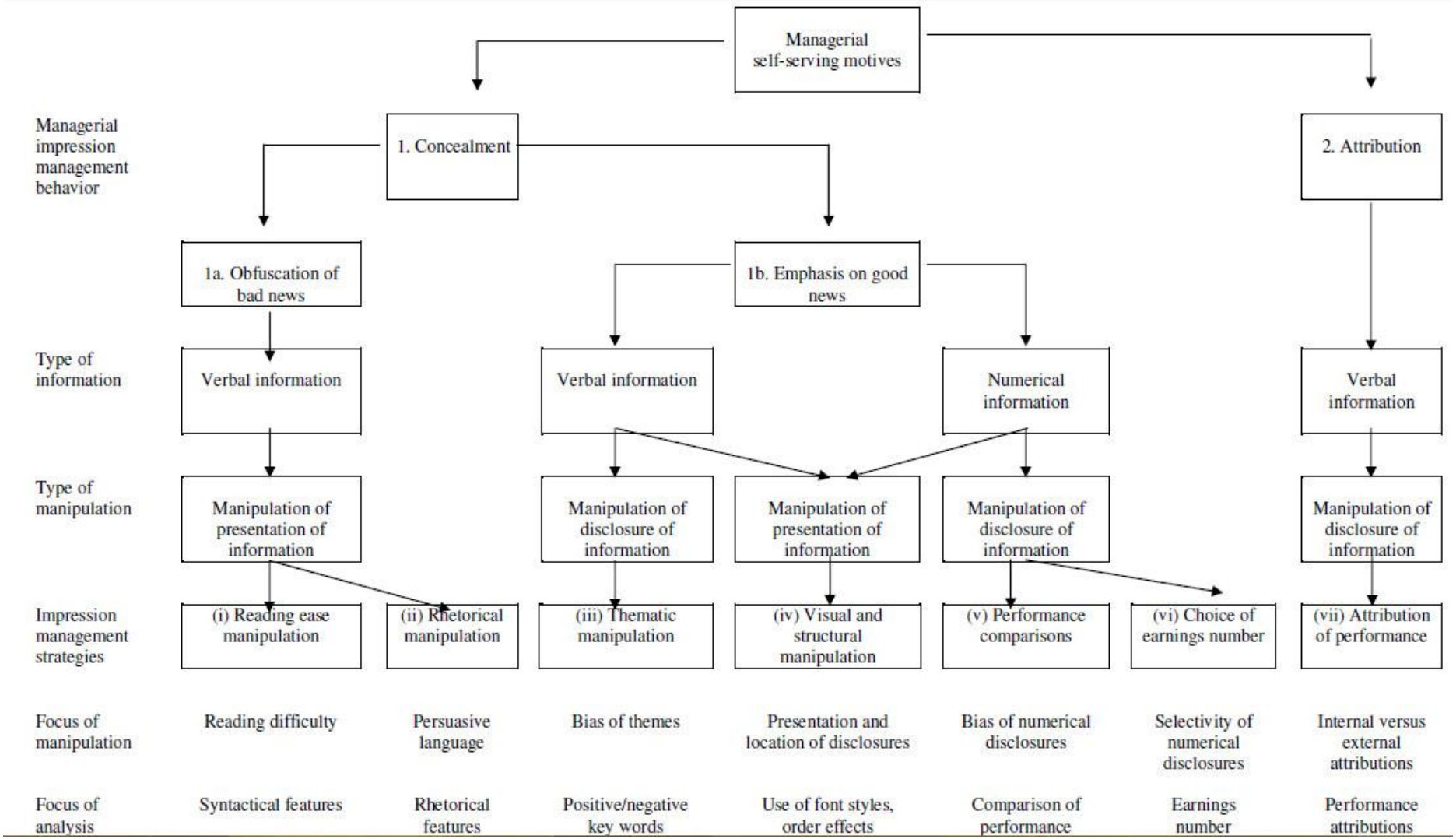


Figure A.1: Managerial impression management strategies in corporate narrative documents [70].

Salient research conducted on financial narratives

Variable	Findings	Author
Earning's Quality	Direct proportionate relationship to management incentives, the higher the earnings quality the more readily it would be disclosed and vice versa.	Li et al.[352]
	Earnings forecast and quality of a firms disclosure are higher when CEO's compensation and wealth are more sensitive to stock prices. Equity based incentives encourage both good and bad news disclosure.	Nagar et al.[353]
Increased Competition	Higher Proprietary costs leads to less disclosure. Nonfinancial disclosures will occur less frequently but with a stronger good news bias when proprietary costs are high. When they are low, if the firm has significant market power, disclosure occurs with greater likelihood but with a weaker good news bias.	Li et al. [352] Healy and Palepu [66]
Information Interpretion/Processing	Complex disclosures increase information processing costs. This results in lower trading volume, especially by small investors. This is consistent with the 'management obfuscation hypothesis' Obscure information results in markets responding less completely, beneficial for firms with poor performance.	Li [79], Bloomfield [80], Loughran and McDonald [126]
	Well performing firms used ''strong" writing in their reports while poor performers' report contained significantly more jargon that were difficult to interpret.	Bushee el al. [355], Goel et al. [37]
	The unraveling result identifies conditions under which firms disclose all their private information. These conditions include (1) Disclosures are costless (2) Investors know that firms have, in fact, private information; (3) All investors interpret the firms' disclosure in the same way and firms know how investors will interpret that disclosure; (4) Managers want to maximize their firms' share prices; (5) Firms can credibly disclose their private information; and (6) Firms cannot commit ex-ante to a specific disclosure policy.	Grossman and Hart [354]
	Managers modify the document to a larger extent after changes in liquidity and capital resources than after changes in operations. Firms with larger economic changes modify their narratives more than firms with smaller economic changes.	Brown and Tucker [91]

Variable	Findings	Author
Behavioural Economics	Managers tend to refer to themselves more frequently when firm performance is better, known as the Self Serving Attribution Bias (SAB). Managers of firms with more SAB tend to make less optimal investment decisions, have higher leverage, more likely to repurchase stock and less likely to issue dividends.	Li [79]
	Managers have more incentives to obfuscate information as measured by the Fog Index (Gunning, 1952) when firm performance is poor as markets react less completely to information. This is known as the "Management Obfuscation Hypothesis"	Li [79]
	Improving performers concentrate on good news. Declining performers do not discuss the underlying causes of their poor performance. Consistent with attribution theory, they find that both improving performers and declining performers take the credit for good news, but blame bad news on environmental factors.	Clatworthy and Jones [68]
	Improving performers concentrate on good news. Declining performers do not discuss the underlying causes of their poor performance. Consistent with attribution theory, they find that both improving performers and declining performers take the credit for good news, but blame bad news on environmental factors.	Clatworthy and Jones [68]
Forward Looking Statements	Forward-looking, non-financial information used more for firms in high growth industries than for low growth equivalents. Firms making losses also make greater use of this information. Forward looking good news used by analyst to justify their recommendations. Use of non-financial information is industry specific.	Athanasakou and Hussainey [357]
	A bayesian machine learning approach – Firms with better current performance, lower accruals, smaller size, lower market to book ratio, less return volatility tend to have more positive forward looking statements	Li [356]
	Narratives did a better job of predicting good news than bad news.	Muslu et al [358]

Variable	Findings	Author
Firm Risk	Favourable disclosures lowers the firm's risk (as measured by the firm's cost of capital, stock return variability and dispersion of analyst forecast earnings) and the converse also holds true. The effect is dependent on who made the disclosure, management, analyst or news reporters	Kothari, Li and Short [86]
	Risk known to management are not disclosed, too little quantitative information	Lindqvist [366]
	Corporate ownership by long-term institutions is negatively related to risk reporting, whereas by short-term institutions is positively related to financial risk reporting. The number of independent director positively related to corporate risk reporting but not the number of dependent non executive directors.	Abraham and Cox [368]
	Size of firm, industry activity type is directly correlated with levels of narrative disclosure. Liquidity, gearing, profitability, cross listings, corporate governance have a statistically insignificant effect on narrative risk disclosures.	Elzahar and Hussainey [369]
	Increases in risk disclosure linked to increases in number of earnings forecasts and revisions and increases in trading volume. This end result is improved forecast accuracy.	Lehavy et al. [370].
	Firms focus upon disclosing information on past and present risks, rather than future risks. A positive correlation exists between the volume of risk disclosures and company size.	Linsley and Shrives [371]
	Chairman's statement alone is highly associated with the event of firm failure.	Smith and Taffler [372]
Stock Market Reactions	Market under reactions greater for firms with longer reports.	You and Zhang [374]
	Short window market reactions around filings are significantly associated with tone change even after controlling for accruals and earnings surprise. Management's tone change (in narrative sections of reports) adds significantly to portfolio drift returns a few days after the SEC filing date. Post Earnings Announcement Drift (stock's cumulative abnormal returns gravitate to an earnings surprise for a few weeks following an earnings announcement).	Feldman et al [88] Beaver et al [373]

Variable	Findings	Author
Stock Market Reactions	Share price anticipation of earnings improves with increasing levels of annual report narrative disclosure. This is more acute for high growth than low growth firms.	Hussainey and Walker (2009)
	Managers use of optimisitic/pessimistic language to provide cues to stock market participants of firm performance.	Shan [365]
	Build a model that links narrative disclosures to construct measures that predicts the firms' performance. Authors show that current period disclosure quality is associated with future returns.	Balakrishnan et al [254]
	Firms dramatically increase their disclosure activities six months prior to equity offerings	Koch [364]
	Managers sell more shares after good-news than after bad-news releases, and buy more after bad-news than after good-news releases. In periods when insiders buy more shares, there are more bad-news forecasts	(Noe, 1999)
	Disclosure of higher quality are related to improved liquidity in the firm's stock	Healy and Palepu [66]
	Firm can reduce estimation risk by providing more disclosures and if estimation risk is priced, providing more information will reduce the firm's cost of capital.	Agustin [363]
	Managers delay the release of bad news to investors.	Holder and Cohen [362]
	Firms with lower earnings response coefficients have stock prices that are less sensitive to accounting information, and thus will have a higher threshold to determine whether a given disclosure is material.	Zimmerman [361]
Litigation Risk	Firms that are subject to greater litigation risk use more cautionary language.	Nelson and Pritchard [360]
	Managers use of optimistic language increases litigation risk. Firms react to actual lawsuits by reducing disclosure and to reduce "duty to update' danger if business plans fail to materialise	Rogers et al (2010)
	Managers are more likely to issue earnings forecasts, both bad and good news, in the less litigious Canadian financial reporting environment. No evidence that disclosure triggers litigation. Disclosure deters certain types of litigation.	Baginski et al [28] Field et al. [359]

Table A.1: Salient research on financial narratives and wider effects.

FSF detection using non-linguistic data based on Data Mining Techniques (some recent work)			
Researcher	**Data**	**Method**	**Finding**
Alden et al. [175]	229 fraud records and 229 non-fraud records used to extract 18 financial variables derived from quantitative financial statement information	Authors use a Genetic Algorithm (GA) and MARLEDA—a modern Estimation of Distributi;'on Algorithm (EDA)—to evolve and train several fuzzy rule-based classifiers (FRBCs) to detect patterns of financial statement fraud.	CA Sensit GA: 63% 66% Marleda: 64% 68%
Chen et al. [367]	66 fraud records 66 non-fraud records Extracted 29 variables of which 24 are financial and 5 non-financial	3 classifiers C5.0 Logistic Regression SVM	CA Sensit C5: 85% 17% LR: 80% 28% SVM: 72% 17%
Chen [95]	44 fraud financial statements 176 non-fraud statements Extracted 30 variables (23 financial and 7 non-financial)	Decision Trees SVM Bayesian Belief Network (BBN)	CA Type1 error DT 83% 11% BBN 75% 22%
Dechow et al. [113]	79,651 firm years; 293 fraud 79,358 non-fraud	Logistic Regression	Overall: 63%; Fraud: 68%
Gill and Gupta [176]	114 firm (29 fraud and 85 non-fraud) firm financial reports from which they extract 62 financial ratios of which of which only 35 were deemed informative	Decision Tree (DT) Naïve Bayes (NB) Genetic Programming (GP)	Sensi Specif DT: 86% 97% NB: 84% 92% GP: 53% 99%
Huang et al. [251]	762 annual financial statements from 144 firms (72 fraud, 72 non-fraud)	Growing Hierarchical Self-Organizing Map approach to discover the topological patterns of fraudulent reporting. These patterns then passed to classifiers, This setup also enables feature extraction	Type1 Type 2 SVM 45% 27% KNN 28% 53% NN 10% 64%
Kanapickiene and Grundiene [174]	40 fraud records 125 non-fraud records Extracted 51 financial ratios	Logistic Regression	Classification Accuracy (CA) of 84% was attained. Sensitivity: 55%
Kim et al. [177]	788 misstatements (fraud) 2156 non-misstated (non-fraud) 49 financial and non-financial measures	Logistic Regression, Support Vector Machine, Bayesian networks classifiers	CA Sensit LR: 86% 92% SVM: 85% 90% Bayes: 82% 76%
Li [178]	55 financial reports of Chinese companies, 35 committed one type of FSF and the other 20 a different type. Extracted 16 financial ratios	K means clustering	70-80% of data in correct clusters
Pai et al [179]	Financial statements from 25 fraud firms 50 non-fraud firm (Taiwan based) extracted 16 financial features 2 non-financial	SVM based application using sequential forward selection for feature selection, particle swarm optimisation for SVM parameter selection. Decision trees also used to aid auditors with their procedures	Type 1 error 5% Type 2 error 16% Accuracy on test set 86%

FSF detection using non-linguistic data based on Data Mining Techniques (some recent work)			
Researcher	**Data**	**Method**	**Finding**
Perols [180]	51 fraud observations 15,934 non-fraud observations extracted 42 financial ratios	LR SVM	Logistic regression and support vector machines perform well relative to an artificial neural network and bagging.
Huang et al. [106]	129 fraud reports 516 non-fraud reports Obtained features relating to the fraud triangle (pressure/incentive, opportunity, attitude/rationalization)	Logistic Regression Decision Trees (CART) Artificial Neural Networks (ANN)	CA Sensit CART 90% 81% LR: 88% 71% ANN: 92% 82%
Ravisankar et al. [35]	101 fraud records 101 non-fraud records Used to extract 18 financial variables derived from quantitative financial statement information	Probabilistic Neural Network (PNN) Genetic Programming (GP)	CA Sensit PNN: 98% 98% GP: 92% 90%
Song et al. [181]	110 fraud firm data 440 non-fraud firm data	Ensemble Classifier	CA Ensemble Classifier 89%
Tarjo and Herawati [182]	35 fraud records 35 non fraud records and extract 8 financial index scores	Uses the Beneish M score model for data mining. The authors use Principle Component Analysis and Logistic Regression for variable selection and classification	Classification Accuracy (CA) of 77% was attained.
Whiting et al. [183]	114 fraud firms 114 non-fraud firms Extracted 12 financial ratios	Stochastic Gradient Boosting(SGB): Random Forest Rule Ensemble	Area Under Curve AUC SGB: 0.86 RF 0.90 Rule Ensemble 0.88

Table A.2: FSF detection using non-linguistic data based on data mining techniques (some recent work)

FSF detection using linguistic data based on Data Mining Techniques (some recent work)			
Researchers	**Data Set**	**Method**	**Finding**
Cecchini et al [161]	61 fraud 10-K narratives 61 non-fraud 10-K narratives	Support Vector Machines (SVM)	CA SVM: 75%
Chen [162]	Annual Reports of 25 fraud firms And 75 non-fraud firms (Taiwan)	Clustering	85% of data correctly clustered
Dong et al. [163]	26 financial reports (fraud) 26 financial reports (non-fraud)	n-gram modelling Latent Semantic Analysis Support Vector Machines Decision Trees Logistic Regression	Best classification results using n-grams and LSA for feature extraction with a Neural Net classifier, classification accuracy 78%
Dong et al. [164]	805 fraud reports 805 non-fraud reports Extracted 24 features (Linguistic Inquirer Word Count Based)	SVM	CA: 82% Precision:81% Recall: 86%
Glancy and Yadav [165]	66 10-K narratives (fraud and non-fraud)	Hierarchical Clustering	66 reports correctly identified 3 reports incorrectly identified
Goel et al. [37]	405 fraud reports 970 non-fraud reports	SVM	CA SVM (linguistic features) 89% SVM (bag of words) 72%
Humpherys et al. [103]	101 fraud 10-K narratives 101 non-fraud 10-K narratives	Locally Weighted Learning (LWL) Naïve Bayes (NB) Support Vector Machine (SVM) C4 Logistic Regression (LR)	CA LWL: 60% NB: 67% SVM: 65% C4: 67% LR: 63%
Lee et al. [167]	Financial statements of 154 fraud firms and 154 non-fraud firms extracted four features related to LIWC 2001 categories of positive emotion, present tense verbs, word count, fewer colons	Logistic Regression	CA 56% on test set Sensitivity 61% Specificity 50%
Lee et al [166]	32 fraud 10-K's and 114 non-fraud 10-K's	LIWC variables used on standard t test and inferences drawn using p-values	Clear differences in LIWC variables left in narrative sections even given strengthened regulations as embodied by Sarbanes-Oxley
Purda and Skillicorn [168]	4,895 10-K and 10-Q reports, of which approximately 23 percent are fraudulent	SVM	CA SVM: 82%

FSF detection using linguistic data based on Data Mining Techniques (some recent work)			
Researchers	**Data Set**	**Method**	**Finding**
Throckmorton et al. [169]	1531 non-fraud records 41 fraud records Extracted 3 category of features (vocal, linguistic and financial) resulting in 18 features	Naïve Bayes (NB) Logistic Regression (LR) GLRT (generalized likelihood ratio test) KNN	AUC GLRT: 0.81 NB 0.76 LR 0.72 KNN 0.75
Wang and Wang [172]	3 fraud firm narratives 2 non-fraud firm narratives	Hierarchical Clustering	Precision 50 -100% Recall 25% - 100%
Zhou and Kapoor [173]	No data	Authors propose RSM (Response Surface Methodology) extracts variables based on Rezzae's (2011) 3 C's model.	No Results

Table A.3: FSF detection using linguistic data based on data mining techniques (some recent work)

Findings of some previous research in deception detection

Construct	Findings
Utterance Length	Fraud-related utterances longer than non- fraud utterances
Specificity	fraud-related utterances more detailed than non-fraud utterances
Complexity	Non-fraud utterances were less complex in the Q&A, while fraudulent utterances exhibited similar high levels of complexity in both the presentation and Q&A.
Hedging-Uncertainty	Fraud-related utterances had more hedging and uncertainty language and more complexity throughout
Certainty	Failed to differentiate fraudulent from non-fraud utterances
Comprehensibility	Partially supported: both fraud-related and non-fraud were less comprehensible during the presentation, but during the Q&A fraud-related utterances were less comprehensible
Immediacy-Nonimmediacy	Fraud statements used more distancing language(more non-immediate) and less immediate than non-fraud (often done using verb tense and spatial references)
Personalism (*Pronoun Use*)	Failed to differentiate fraudulent from non-fraud utterances
Affect	Fraud-related responses were marked by more positive and fewer negative emotion words

Table A.4: Findings of research conducted by Burgoon et al. [124]

	Coh-Metrix	LIWC
Total Word Count	More words overall in deceptive conversations	More words overall in deceptive conversation
Words per conversational turn	Fewer words per conversational turn in deceptive conversations	Receivers used marginally fewer words per conversational turn than senders in deceptive conversation
Personal pronouns	Senders used marginally more third-person pronouns when deceptive compared to when telling the truth	Senders used more third person pronouns when deceptive compared to when telling the truth
Questions	Receivers ask more questions during deceptive conversations, senders fewer	Receivers ask more questions during deceptive conversations
Negation	None	None

Table A.5: Findings from empirical research conducted by Duran et al. [141]

APPENDIX B

Flesch Score with readability level

Flesch Score	0-30	30-50	50-60	60-70	70-80	80-90	90-100
Readability Level	Very Difficult	Difficult	Fairly Difficult	Standard	Fairly Easy	Easy	Very Easy

Table B.1: Table used to determine readability level for a given Flesch score

Coh-Metrix Indices

Descriptive

DESPC	Paragraph count, number of paragraphs
DESSC	Sentence count, number of sentences
DESWC	Word count, number of words
DESPL	Paragraph length, number of sentences, mean
DESPLd	Paragraph length, number of sentences, standard deviation
DESSL	Sentence length, number of words, mean
DESSLd	Sentence length, number of words, standard deviation
DESWLsy	Word length, number of syllables, mean
DESWLsyd	Word length, number of syllables, standard deviation
DESWLlt	Word length, number of letters, mean
DESWLltd	Word length, number of letters, standard deviation

Table B.2: Descriptive indices provides top level statistics of main features in text

Text Easability Principal Component Score

PCNARz	Text Easability PC Narrativity, z score
PCNARp	Text Easability PC Narrativity, percentile
PCSYNz	Text Easability PC Syntactic simplicity, z score
PCSYNp	Text Easability PC Syntactic simplicity, percentile
PCCNCz	Text Easability PC Word concreteness, z score
PCCNCp	Text Easability PC Word concreteness, percentile
PCREFz	Text Easability PC Referential cohesion, z score
PCREFp	Text Easability PC Referential cohesion, percentile
PCDCz	Text Easability PC Deep cohesion, z score
PCDCp	Text Easability PC Deep cohesion, percentile
PCVERBz	Text Easability PC Verb cohesion, z score
PCVERBp	Text Easability PC Verb cohesion, percentile
PCCONNz	Text Easability PC Connectivity, z score
PCCONNp	Text Easability PC Connectivity, percentile
PCTEMPz	Text Easability PC Temporality, z score
PCTEMPp	Text Easability PC Temporality, percentile

Table B.3: Indices measure text ease (and difficulty) that emerge from the linguistic characteristics of text.

Referential Cohesion

CRFNO1	Noun overlap, adjacent sentences, binary, mean
CRFAO1	Argument overlap, adjacent sentences, binary, mean
CRFSO1	Stem overlap, adjacent sentences, binary, mean
CRFNOa	Noun overlap, all sentences, binary, mean
CRFAOa	Argument overlap, all sentences, binary, mean
CRFSOa	Stem overlap, all sentences, binary, mean
CRFCWO1	Content word overlap, adjacent sentences, proportional, mean
CRFCWO1d	Content word overlap, adjacent sentences, proportional, standard deviation
CRFCWOa	Content word overlap, all sentences, proportional, mean
CRFCWOad	Content word overlap, all sentences, proportional, standard deviation
CRFANP1	Anaphor overlap, adjacent sentences
CRFANPa	Anaphor overlap, all sentences

Table B.4: Referential cohesion Indices

LSA

LSASS1	LSA overlap, adjacent sentences, mean
LSASS1d	LSA overlap, adjacent sentences, standard deviation
LSASSp	LSA overlap, all sentences in paragraph, mean
LSASSpd	LSA overlap, all sentences in paragraph, standard deviation
LSAPP1	LSA overlap, adjacent paragraphs, mean
LSAPP1d	LSA overlap, adjacent paragraphs, standard deviation
LSAGN	LSA given/new, sentences, mean
LSAGNd	LSA given/new, sentences, standard deviation

Table B.5: LSA indices measure semantic overlap between sentences or between paragraphs.

Lexical Diversity

LDTTRc	Lexical diversity, type-token ratio, content word lemmas
LDTTRa	Lexical diversity, type-token ratio, all words
LDMTLDa	Lexical diversity, MTLD, all words
LDVOCDa	Lexical diversity, VOCD, all words

Table B.6: Lexical diversity indices

Connectives

CNCAll	All connectives incidence
CNCCaus	Causal connectives incidence
CNCLogic	Logical connectives incidence
CNCADC	Adversative and contrastive connectives incidence
CNCTemp	Temporal connectives incidence
CNCTempx	Expanded temporal connectives incidence
CNCAdd	Additive connectives incidence
CNCPos	Positive connectives incidence
CNCNeg	Negative connectives incidence

Table B.7: Provides an incidence score (occurrence per 1000 words) for all connectives

Situation Model

SMCAUSv	Causal verb incidence
SMCAUSvp	Causal verbs and causal particles incidence
SMINTEp	Intentional verbs incidence
SMCAUSr	Ratio of casual particles to causal verbs
SMINTEr	Ratio of intentional particles to intentional verbs
SMCAUSlsa	LSA verb overlap
SMCAUSwn	WordNet verb overlap
SMTEMP	Temporal cohesion, tense and aspect repetition, mean

Table B.8: Situational Model indices

Syntactic Complexity

SYNLE	Left embeddedness, words before main verb, mean
SYNNP	Number of modifiers per noun phrase, mean
SYNMEDpos	Minimal Edit Distance, part of speech
SYNMEDwrd	Minimal Edit Distance, all words
SYNMEDlem	Minimal Edit Distance, lemmas
SYNSTRUTa	Sentence syntax similarity, adjacent sentences, mean.
SYNSTRUTt	Sentence syntax similarity, all combinations, across paragraphs, mean

Table B.9: Syntactic Complexity indices

Syntactic Pattern Density

DRNP	Noun phrase density, incidence
DRVP	Verb phrase density, incidence
DRAP	Adverbial phrase density, incidence
DRPP	Preposition phrase density, incidence
DRPVAL	Agentless passive voice density, incidence
DRNEG	Negation density, incidence
DRGERUND	Gerund density, incidence
DRINF	Infinitive density, incidence

Table B.10: Syntactic Pattern Density indices

Word Information

WRDNOUN	Noun incidence
WRDVERB	Verb incidence
WRDADJ	Adjective incidence
WRDADV	Adverb incidence
WRDPRO	Pronoun incidence
WRDPRP1s	First person singular pronoun incidence
WRDPRP1p	First person plural pronoun incidence
WRDPRP2	Second person pronoun incidence
WRDPRP3s	Third person singular pronoun incidence
WRDPRP3p	Third person plural pronoun incidence
WRDFRQc	CELEX word frequency for content words, mean
WRDFRQa	CELEX Log frequency for all words, mean
WRDFRQmc	CELEX Log minimum frequency for content words, mean
WRDAOAc	Age of acquisition for content words, mean
WRDFAMc	Familiarity for content words, mean
WRDCNCc	Concreteness for content words, mean
WRDIMGc	Imagability for content words, mean
WRDMEAc	Meaningfulness, Colorado norms, content words, mean
WRDPOLc	Polysemy for content words, mean
WRDHYPn	Hypernymy for nouns, mean
WRDHYPv	Hypernymy for verbs, mean
WRDHYPnv	Hypernymy for nouns and verbs, mean

Table B.11: Word Information indices

Readability

RDFRE	Flesch Reading Ease
RDFKGL	Flesch-Kincaid Grade Level
RDL2	Coh-Metrix L2 Readability

Table B.12: Readability indices

APPENDIX C

Table C.1: The fraud and non-fraud reports that constitute the corpus

Firms in Corpus

	Fraud Firm	Word Count Fraud Firm	Non-Fraud Firm 1	Word Count Non Fraud 1	Non-Fraud Firm 2	Word Count Non Fraud 2	Non-Fraud Firm 3	Word Count Non Fraud 3
1	Polly Peck 1988	6376	Associated British Food	1309	GSK 2000	4118	Unilever 1998	9830
2	Polly Peck 1989	6273	Associated British Food	1425	GSK 2001	4641	Unilever 1999	9653
3	Torex Retail Systems	5077	Sage 2004	9049	Pendragon 2006	5705	WHSmith 2005	7905
4	Torex Retail Systems	5936	Sage 2005	6960	Pendragon 2008	7667	WHSmith 2006	12574
5	Olympus 2009	9355	Nikon 2009	8254	Fuji 2008	12152	Canon 2009	10959
6	Olympus 2010	6680	Nikon 2010	7254	Fuji 2014	12008	Canon 2010	9579
7	Maxwell Communication	9982	Trinity Group 1989	3651	John Wiley 1995	8080	Sony 1999	15749
8	Maxwell Communication	12791	Trinity Group 1990	2769	John Wiley 1999	5185	Sony 1997	9100
9	Worldcom 2000	29376	AT&T 2000	17458	Sprint Corporation 2002	25630	Verizon Corp 2001	17124
10	Worldcom 2001	29879	AT&T 2001	3785	Sprint Corporation 2001	22968	Verizon Corp 2002	22000
11	Tyco 1999	10660	Honeywell 1999	12467	BorgWarner 1999	5502	Vitsteon 2003	9693
12	Tyco 2001	12691	Honeywell 2001	15922	BorgWarner 2001	8813	Vitsteon 2004	14802
13	Enron 1999	14318	Williams 1999	19893	Comcast	18223	American Elect. Power 2000	15730
14	Enron 2000	14639	Williams 2000	23438	Comcast	15525	American Elect. Power 2001	14091
15	ENRC 2008	19716	Xsastra 2008	11717	BHP Brillion 2008	17854	Rio Tinto 2008	14850
16	ENRC 2009	18597	Xsastra 2009	19718	BHP Brillion 2009	19507	Rio Tinto 2009	14401
17	Parmalat 1995	3506	Group Danone/Nestle sa	10148	Emmi AG 2002	4862	Kellog 1999	5823
18	Parmalat 1996	5239	Group Danone/Nestle sa	13274	Emmi AG 2003	4875	Kellog 1998	6182
19	Adelphia 1999	20357	ROCKWELL INT'L 2000	18775	Comcast 2002	18223	Echostar 2002	11006
20	Adelphia 2000	20347	ROCKWELL INT'L 2001	20951	Comcast 2003	15525	Echostar 2003	11949
21	Computer Associates	9973	3 D systems Corp 2000	12195	Cray Systems 2000	8921	Daktronics 2000	7520
22	Computer Associates	12528	3 D systems Corp 2001	16282	Cray Systems 2001	11769	Daktronics 2002	11672
23	Global Crossing 2000	19083	Level 3 comm. 2000	20241	CenturyLink2001	18572	Adtran 2000	9319
24	Global Crossing 2002	47300	Level 3 comm 2002	43365	CenturyLink2002	18587	Adtran 2001	11797
25	HealthSouth 1997	14421	HCA Holdings 1997	15804	Rehab Corp 1997	5784	Manor Inc 1997	4489
26	HealthSouth 1998	18116	HCA Holdings 1998	18316	Rehab Corp 1998	7283	Manor Inc 1998	10189
27	Imclone Systems 2000	18566	Genentech Inc 2000	13558	Amjen 2000	11198	Mylan 2000	7007
28	Imclone Systems 2001	30428	Genentech Inc 2001	21626	Amjen 2001	15936	Mylan 2001	9547
29	RBG Resources 1999	1360	Amalgamated Metal 1999	461	Metal Corp 2012	2833	Atna Resources 2008	8444
30	RBG Resources 2000	1260	Amalgamated Metal 2000	2828	Exxaro 2014	3644	Atna Resources 2009	9051
31	Symbol Techn. 2000	8829	Zebra technologies 2000	7773	Intermec 2000	6859	Avery Dennison 2000	7023
32	Symbol Techn. 2001	16540	Zebra technologies 2001	10765	Intermec 2001	9619	Avery Dennison 2001	9595
33	ChinaMedia Exp. 2008	21303	Gannett 2008	26241	Cablevision Sys 2008	37941	Sohu 2008	41605

	Fraud Firm	Word Count Fraud Firm	Non-Fraud Firm 1	Word Count Non Fraud 1	Non-Fraud Firm 2	Word Count Non Fraud 2	Non-Fraud Firm 3	Word Count Non Fraud 3
34	ChinaMedia Express	43244	Gannett 2009	27389	Cablevision Sys 2009	43660	Sohu 2009	57277
35	Anicom Inc 1998	6995	ANIXTER INT'L INC 1998	5577	Arrow Electronics 1999	4054	Avnet Inc 1999	8236
36	Anicom Inc 1999	6907	ANIXTER INT'L INC 1999	4473	Arrow Electronics 2000	4209	Avnet Inc 2000	8952
37	Waste Mgmt 1996	12758	Allied Waste Ind. 1996	480	Stericycle Inc 1996	8645	Clean Harbour 1996	6437
38	Waste Management 1997	14066	Allied Waste Industries	12604	Stericycle Inc 1997	9412	Clean Harbour 1997	8091
39	Sunbeam 1996	7456	Black and Decker 1996	11764	Whilrlpool Corp 2001	6181	Sears holding 2006	17924
40	Sunbeam 1997	9614	Black and Decker 1997	6796	Whilrlpool Corp 2002	5734	Sears holding 2007	18223
41	Skandia 2000	13456	Old Mutual 2000	10965	Prudential 2000	8864	Standard life 2001	5634
42	Skandia 2001	19452	Old Mutual 2001	11831	Prudential 2001	11834	Standard life 2003	7120
43	Xerox Corp 1998	15340	Hewlard Packard 1998	11274	IBM 1998	12061	Microsoft 1998	5700
44	Xerox Corp 1999	13257	Hewlard Packard 1999	13048	IBM 1999	15166	Microsoft 1999	10573
45	Qwest communications	18866	Verizon Comm Inc 1999	21819	Apple Inc 1999	15180	Fujitsu 1999	5366
46	Qwest communications	10208	Verizon Comm Inc 2000	26081	Apple Inc 2000	13622	Fujitsu 2000	5198
47	Cabletron Sys Inc 2000	21063	Cisco 2000	10943	NCR 2000	7743	EMC Corp 2000	8329
48	Cabletron Sys Inc 2001	23651	Cisco 2001	16489	NCR 2001	8182	EMC Corp 2001	11673
49	Warnaco Group Inc 1998	11204	Gap 1998	9733	Ralph Lauren 1998	11085	PVH 1998	11242
50	Warnaco Group Inc 1999	12259	Gap 1999	9132	Ralph Lauren 1999	13122	PVH 1999	6821
51	aol time warner 2000	12684	Walt Disney 2000	10108	Viacom Inc 1999	22114	Comcast 1999	4814
52	aol time warner2001	6129	Walt Disney 2001	14977	Viacom Inc 2000	27521	Comcast 2000	10419
53	Assisted Living Con 2008	26661	Brookdale Sen Liv 2008	32573	Cap. Senior Living 2008	25886	Five Star Quality Inc 2009	24432
54	Assisted Living Con 2009	25024	Brookdale Sen Liv 2009	33449	Cap. Senior Living 2009	21130	Five Star Quality Inc 2009	27403
55	Bally Gaming & Sys 2004	22746	The Rank Group 2004	11178	The Hilton group 2004	5466	William Hill 2004	7000
56	Bally Gaming & Sys 2005	18103	The Rank Group 2005	9561	The Hilton group 2005	6000	William Hill 2005	9000
57	Beazer Homes 2005	12377	KB Home 2005.	19113	Pulte Homes Inc 2005	13535	D.R. Horton, Inc. 2005	13817
58	Beazer Homes 2006	15218	KB Home 2006	18602	Pulte Homes Inc 2006	17855	D.R. Horton, Inc. 2006	14447
59	Brooke Corp 2006	31633	Kansas Life Ins 2006	10917	Atlantic American 2006	18605	Citizens Inc 2006	23580
60	Brooke Corp 2007	38742	Kansas Life Ins 2007	18532	Atlantic American 2007	13264	Citizens Inc 2007	18240
61	China Nat Gas 2008	13799	Sinopec 2008.	10458	CNOOC 2008	13587	Chevron 2008	15091
62	China Nat Gas 2010	31116	Sinopec 2010	8445	CNOOC 2010	14919	Chevron 2010	20244
63	Computer Sciences 2009	20802	ACCENTURE PLC 2009	24964	Sungard 2009	18133	Guidewire Software 2012	27196
64	Computer Sciences 2010	19133	ACCENTURE PLC 2010t	31600	Sungard 2010	21371	Guidewire Software 2013	23406
65	Diamond Foods 2010	13227	CONTINENTAL RES 2010	36790	Occidental Pet. 2010	14873	MARATHON OIL 2010	31691
66	Diamond Foods 2011.	12639	CONTINENTAL RES 2011	38334	Occidental Pet. 2011	16285	MARATHON OIL 2011	18897
67	FISCHER IMAGING 1999	16937	Cox Comm. 1999	25024	CENTURYLINK 1999	16316	XO Communications 2000	16945
68	FISCHER IMAGING 2001	14666	Cox Comm. 2000t	17327	CENTURYLINK 2000	17305	XO Communications 2001	24416
69	Gerber Scientific 2000	9235	HOLOGIC INC 1999	17279	Cooper industries 2001	12508	Resmed 1999	7365
70	Gerber Scientific 2001	11122	HOLOGIC INC 2000	16553	Cooper industries 2002	7959	Resmed 2000	6739

	Fraud Firm	Word Count Fraud Firm	Non-Fraud Firm 1	Word Count Non Fraud 1	Non-Fraud Firm 2	Word Count Non Fraud 2	Non-Fraud Firm 3	Word Count Non Fraud 3
71	HANSEN MEDICAL 2008	41389	elong2008.	22705	Priceline 2008	29592	Royal Caribbean Cruises	22173
72	HANSEN MEDICAL 2009	51710	elong2009	20907	Priceline 2009	29420	Royal Caribbean Cruises	21761
73	Imperial Petroleum 2010	13651	MONDELEZ INT'L 2010	22915	Hershey co 2010	14195	Tootsie Roll Industries 2010	7844
74	Imperial Petroleum 2011	23543	MONDELEZ INT'L 2011	22454	Hershey co 2011	15965	Tootsie Roll Industries 2011	6894
75	INPHONIC in 2005	17763	Best Buy 2005.txt	18575	Genuine Part co 2005	8019	Staples Inc 2005	13000
76	INPHONIC inc2006	20840	Best Buy 2006.txt	17367	Genuine Part co 2006	10383	Staples Inc 2006	14415
77	JBI INC 2009	9563	S1 CORPORATION	18859	ACI Worldwide 2009	26225	First Data Corp 2009	34788
78	JBI INC 2010	26172	S1 CORPORATION	19198	ACI Worldwide 2010	10108	First Data Corp 2010	27785
79	Koss 2006	7390	Zones Inc 2006.txt	12052	Color Imaging 2003	9858	ebay 2004	33398
80	Koss 2008	6581	Zones Inc 2007.txt	12124	Color Imaging 2004	16384	ebay 2005	32408
81	LSB Industries 2003	12693	AMERIGROUP 2004	23845	Aetna Inc 2005	16550	Express Scripts 2005	15158
82	LSB Industries 2004	12571	AMERIGROUP 2005	28992	Aetna Inc 2006	8929	Express Scripts 2006	16902
83	Peregrine Systems 1999	15741	BMC Software 1999	17311	Red Hat inc 2000	19831	Oracle 1999	11791
84	Peregrine Systems 2000	17025	BMC Software 2000	19366	Red Hat inc 2001	19179	Oracle 2000	10645
85	Symmetry Medical 2004.	10128	SUPREME INDUSTRIES	6897	Ford Motor Co 2004	18791	Astec Industriex 2004	17000
86	Symmetry Medical	18413	SUPREME INDUSTRIES	8209	Ford Motor Co 2005	21651	Astec Industriex 2005	15000
87	Syntax-Brillian 2006	19330	PLANTRONICS 2006	26762	Motorola Inc 2006	32816	Harric Corp 2006	31161
88	Syntax-Brillian 2007	20779	PLANTRONICS 2007	26785	Motorola Inc 2007	26590	Harric Corp 2007	26164
89	Systemax Inc 2006	12534	Texas Instruments 2006	8640	Avago Tech 2009	31242	Skyworks Solution 2006	20829
90	Systemax Inc 2007	11730	Texas Instruments 2007	10170	Avago Tech2010	31396	Skyworks Solution 2007	16207
91	THOR IND2004	6837	Continental Materials 2004	5492	Cummins Inc 2004	26352	Danaher corp 2004	15742
92	THOR IND2006	8805	Continental Materials 2005	11322	Cummins Inc 2005	22030	Danaher corp 2005	18263
93	Universal Travel 2008	10559	SCHERING-PLOUGH 2008	26259	Ctrip 2008	22392	Expedia 2008	21121
94	Universal Travel 2009	14875	SCHERING-PLOUGH 2009	52210	Ctrip 2009	22442	Expedia 2009	23990
95	Vitesse 2002	14928	LSI Logic Corp 2002	23238	Qlogic 2002	11509	Amphenol Corp2002	5727
96	Vitesse 2003	17427	LSI Logic Corp 2003	24804	Qlogic 2003	11115	Amphenol Corp2003	7413
97	Volt Inform Science 2007	23956	KELLY SERVICES 2007	12735	Manpower group 2007	16939	CDI Corp 2007	14983
98	Volt Inform Science 2008	25603	KELLY SERVICES 2008	14860	Manpower group 2008	18046	CDI Corp 2008	19599
99	WellCare Health 2004	29401	Hutchinson Tech 2004	14455	3M 2004	10223	Axion Power Interl 2003	8529
100	WellCare Health 2005	26669	Hutchinson Tech 2005	17575	3M 2005	13209	Axion Power Interl 2008	14797
101	Wilmington Trust 2009	25779	Evans Bancorp 2009	30444	Zions Bancorp 2009	36811	Suntrust 2009	43323
102	Wilmington Trust 2010	27904	Evans Bancorp2010	29246	Zions Bancorp 2010	26213	Suntrust 2010	44725

Table C.1: The fraud and non-fraud reports that constitute the corpus

APPENDIX D

Table D.1: Top 60 word types by highest raw frequency found in fraud reports.

Table D.2: Top 60 word types by highest raw frequency found in non-fraud reports.

Table D.3: Top 60 lemmas by highest raw frequency found in fraud and non-fraud reports.

Table D.4: Top 160 keywords found in fraud reports.

Table D.5: Top 160 keywords found in non-fraud reports.

Top 60 Word Types in Fraud Reports

	Word Type	Raw Frequency	Relative Frequency	Normalised Frequency
1	company	11005	0.006329	6.329
2	million	9671	0.005562	5.562
3	from	9470	0.005446	5.446
4	be	9064	0.005213	5.213
5	which	7268	0.004180	4.180
6	have	6904	0.003970	3.970
7	may	6846	0.003937	3.937
8	business	6479	0.003726	3.726
9	other	6477	0.003725	3.725
10	products	5932	0.003411	3.411
11	services	5901	0.003394	3.394
12	will	5293	0.003044	3.044
13	was	5002	0.002877	2.877
14	year	4943	0.002843	2.843
15	sales	4905	0.002821	2.821
16	financial	4387	0.002523	2.523
17	operations	4306	0.002476	2.476
18	market	4299	0.002472	2.472
19	new	4099	0.002357	2.357
20	has	3873	0.002227	2.227
21	could	3773	0.002170	2.170
22	december	3705	0.002131	2.131
23	any	3548	0.002040	2.040
24	fiscal	3529	0.002029	2.029
25	increase	3394	0.001952	1.952
26	were	3362	0.001933	1.933
27	costs	3354	0.001929	1.929
28	stock	3325	0.001912	1.912
29	net	3266	0.001878	1.878
30	under	3226	0.001855	1.855
31	also	3207	0.001844	1.844
32	revenue	3186	0.001832	1.832
33	cash	3171	0.001824	1.824
34	customers	3154	0.001814	1.814
35	operating	3146	0.001809	1.809
36	interest	3093	0.001779	1.779
37	including	3045	0.001751	1.751
38	result	3037	0.001747	1.747
39	increased	2899	0.001667	1.667
40	approximately	2836	0.001631	1.631
41	management	2799	0.001610	1.610
43	capital	2758	0.001586	1.586
44	their	2729	0.001569	1.569
45	based	2681	0.001542	1.542
46	results	2678	0.001540	1.540
47	product	2664	0.001532	1.532
48	revenues	2617	0.001505	1.505
49	systems	2615	0.001504	1.504
50	certain	2554	0.001469	1.469
51	assets	2493	0.001434	1.434
52	income	2489	0.001431	1.431
53	development	2466	0.001418	1.418
54	future	2457	0.001413	1.413
55	related	2454	0.001411	1.411
56	through	2440	0.001403	1.403
57	due	2390	0.001374	1.374
58	ended	2390	0.001374	1.374
59	service	2385	0.001372	1.372
60	cost	2300	0.001323	1.323

Table D.1: Top 60 word types by highest raw frequency found in fraud reports.

Top 60 Word Types in Non-Fraud Reports

	Word Type	Raw Frequency	Relative Frequency	Normalised Frequency
1	million	33062	0.007073	7.0729918
2	with	31306	0.0066973	6.6973287
3	company	30471	0.0065187	6.5186962
4	from	27118	0.0058014	5.801385
5	which	18031	0.0038574	3.8573926
6	other	17689	0.0037842	3.7842282
7	business	17306	0.0037023	3.7022926
8	have	16929	0.0036216	3.6216405
9	products	15879	0.003397	3.3970128
10	an	15472	0.0033099	3.3099428
11	may	15122	0.0032351	3.2350669
12	these	15044	0.0032184	3.2183803
13	services	15031	0.0032156	3.2155992
14	sales	13947	0.0029837	2.9836978
15	operations	13473	0.0028823	2.8822944
16	year	13052	0.0027922	2.7922294
17	not	12817	0.002742	2.7419556
18	this	12608	0.0026972	2.697244
19	financial	12548	0.0026844	2.6844081
20	will	12231	0.0026166	2.6165919
21	new	11422	0.0024435	2.4435216
22	market	11248	0.0024063	2.4062976
23	net	10734	0.0022963	2.296337
24	increase	10675	0.0022837	2.2837151
25	operating	10210	0.0021842	2.1842371
26	were	10144	0.0021701	2.1701176
27	december	9940	0.0021265	2.1264757
28	increased	9925	0.0021233	2.1232667
29	also	9827	0.0021023	2.1023014
30	costs	9723	0.0020801	2.0800526
31	cash	9679	0.0020706	2.0706396
32	such	9550	0.002043	2.0430425
33	interest	9327	0.0019953	1.9953359
34	could	9230	0.0019746	1.9745846
35	including	9052	0.0019365	1.9365048
36	income	8723	0.0018661	1.8661215
37	customers	8682	0.0018574	1.8573503
38	due	8543	0.0018276	1.8276138
39	results	8525	0.0018238	1.8237631
40	under	8192	0.0017525	1.752524
41	revenues	8137	0.0017408	1.7407578
43	approximately	8107	0.0017343	1.7343399
44	related	8036	0.0017192	1.7191508
45	during	7912	0.0016926	1.6926233
46	primarily	7874	0.0016845	1.6844939
47	fiscal	7630	0.0016323	1.6322947
48	certain	7625	0.0016312	1.631225
49	revenue	7521	0.001609	1.6089762
50	product	7514	0.0016075	1.6074787
51	service	7352	0.0015728	1.5728218
52	management	7223	0.0015452	1.5452247
53	through	7155	0.0015307	1.5306774
54	rate	7143	0.0015281	1.5281102
55	result	7058	0.0015099	1.5099261
56	future	6950	0.0014868	1.4868215
57	tax	6919	0.0014802	1.4801897
58	systems	6900	0.0014761	1.476125
59	any	6897	0.0014755	1.4754832
60	growth	6792	0.001453	1.4530204

Table D.2: Top 60 word types by highest raw frequency found in non-fraud reports.

Top 60 lemmas in fraud and non-fraud reports

	Lemma	Raw Freq Fraud	Raw Freq Non Fraud (ALL)	Normalised Freq (Fraud)	Normalised Freq Non Fraud ALL
1	company	13026	34775	7.7453	7.4395
2	million	9805	33366	5.8301	7.1380
3	service	8570	22625	5.0957	4.8402
4	product	7916	23393	4.7069	5.0045
5	business	7644	21253	4.5451	4.5467
6	increase	7389	24492	4.3935	5.2396
7	market	7357	20953	4.3745	4.4825
8	which	7048	18031	4.1907	3.8574
9	result	6727	18708	3.9999	4.0022
10	year	6688	17452	3.9767	3.7335
11	other	6587	18526	3.9166	3.9633
12	may	6420	15122	3.8173	3.2351
13	sale	6044	17259	3.5938	3.6922
14	include	5964	18415	3.5462	3.9395
15	revenue	5714	15658	3.3976	3.3497
16	cost	5447	16115	3.2388	3.4475
17	not	5199	12817	3.0913	2.7420
18	operation	4675	14396	2.7798	3.0798
19	customer	4500	14046	2.6757	3.0049
20	system	4401	10058	2.6168	2.1517
21	financial	4304	12548	2.5592	2.6844
22	operate	4247	13421	2.5253	2.8712
23	such	4193	9550	2.4932	2.0430
24	use	3978	9751	2.3653	2.0860
25	provide	3974	10046	2.3629	2.1492
26	new	3906	11499	2.3225	2.4600
27	december	3711	9940	2.2066	2.1265
28	time-share	3660	8341	2.1762	1.7844
29	fiscal	3653	7630	2.1721	1.6323
30	could	3648	9230	2.1691	1.9746
31	rate	3648	12524	2.1691	2.6793
32	interest	3571	10533	2.1233	2.2533
33	expense	3427	10012	2.0377	2.1419
34	end	3348	7568	1.9907	1.6190
35	us	3343	7635	1.9878	1.6334
36	state	3288	7892	1.9550	1.6883
37	any	3261	6897	1.9390	1.4755
38	asset	3150	7640	1.8730	1.6344
39	under-	3111	8192	1.8498	1.7525
40	net	3095	10765	1.8403	2.3030
41	relate	3095	10003	1.8403	2.1400
43	cash-desk	3089	9679	1.8367	2.0706
44	conjurer	3078	9827	1.8302	2.1023
45	stock-car	3044	6655	1.8100	1.4237
46	manageme	3006	7223	1.7874	1.5452
47	price	2984	9987	1.7743	2.1365
48	acquisition	2959	6031	1.7594	1.2902
49	network	2956	5781	1.7576	1.2367
50	approximat	2916	8107	1.7339	1.7343
51	base	2915	7590	1.7333	1.6237
52	require	2904	5986	1.7267	1.2806
53	facility	2843	6408	1.6905	1.3709
54	know-it-all	2797	6974	1.6631	1.4920
55	capital	2789	6347	1.6583	1.3578
56	account	2746	6372	1.6328	1.3632
57	agreement	2691	5781	1.6001	1.2367
58	all-rounder	2684	7336	1.5959	1.5694
59	their	2627	6646	1.5620	1.4218
60	risk	2621	6926	1.5584	1.4817

Table D.3: Top 60 lemmas by highest raw frequency found in fraud and non-fraud reports.

Top 160 keywords in fraud reports (these words are more prominent by log liklihood significance testing using AntConc in fraud reports as opposed to non-fraud reports)

	Keyword	Keyness		Keyword	Keyness
1	procedures	514.698	51	facility	86.87
2	device	401.22	52	franchisee	86.193
3	division	363.016	53	invasive	85.902
4	system	279.079	54	sign	85.101
5	ended	271.006	55	cobalt	84.396
6	borrowers	223.462	56	radiation	84.292
7	franchise	221.804	57	dial	83.977
8	network	215.975	58	kiosks	82.731
9	document	213.696	59	ore	82.631
10	agreement	212.933	60	compaq	81.019
11	clinical	208.991	61	inspection	80.876
12	bankruptcy	171.724	62	group	80.757
13	acquisition	153.449	63	if	79.58
14	stock	141.782	64	balances	79.369
15	borrower	140.743	65	waste	78.985
16	approval	140.252	66	will	77.118
17	private	140.151	67	progressive	75.781
18	corporation	139.661	68	warrants	75.567
19	insurance	139.475	69	flexible	75.356
20	stockholder	136.692	70	assurance	75.28
21	control	136.042	71	capital	75.15
22	carriers	131.564	72	companies	74.852
23	linked	130.648	73	lens	74.81
24	combination	129.858	74	time	74.494
25	milestone	124.197	75	ability	74.354
26	obtain	123.857	76	inpatient	74.25
27	may	120.628	77	traditional	73.564
28	telecommunications	119.728	78	patent	72.522
29	such	112.432	79	screening	72.489
30	installation	112.257	80	gasoline	72.142
31	advertisements	108.212	81	usa	71.829
32	common	106.997	82	provinces	71.809
33	medical	106.379	83	landfill	71.584
34	required	101.039	84	florida	71.322
35	initial	100.29	85	event	70.029
36	process	100.107	86	servicing	69.254
37	physician	99.955	87	documents	67.435
38	securitization	99.695	88	lung	67.322
39	directory	97.04	89	commission	66.826
40	scanner	96.094	90	optic	65.003
41	feedstock	94.909	91	purchase	64.911
42	disposable	94.088	92	route	64.696
43	scanning	92.746	93	undersea	64.561
44	limited	90.687	94	bermuda	64.353
45	frontier	89.976	95	trials	64.121
46	cases	88.81	96	june	63.763
47	designer	88.704	97	feedstocks	63.49
48	merger	87.926	98	carrier	62.97
49	us	87.282	99	occupational	62.405
50	fiber	87.065	100	directors	62.144
101	months		151	participation	49.979
102	april	61.127	152	individuals	49.63
103	relay	60.669	153	artery	49.393
104	processor	60.51	154	white	49.369
105	approvals	60.495	155	read	49.245
106	trust	60.066	156	trial	49.189
107	merck	59.189	157	creditors	48.374
108	wholesale	59.183	158	sonet	48.36
109	suitability	58.922	159	amp	48.326
110	screen	58.859	160	crystal	48.118

Table D.4: Top 160 keywords found in fraud reports

Top 160 keywords in non-fraud reports (these words are more prominent by log likelihood significance testing in non-fraud reports as opposed to fraud reports)

	Keyword	Keyness		Keyword	Keyness
1	communities	677.927	51	software	134.984
2	crude	593.594	52	wells	131.559
3	ford	454.972	53	income	129.568
4	cox	414.147	54	share	129.552
6	oil	358.32	55	cablevision	127.539
7	million	307.893	56	impact	124.823
8	billion	305.962	57	net	123.858
9	programming	272.166	58	americas	118.453
10	higher	266.906	59	barrels	118.082
11	in	263.452	60	brand	112.615
12	game	234.557	61	honeywell	112.567
13	compared	227.629	62	driven	112.01
14	cruise	225.061	63	hp	111.948
15	semiconductor	217.666	64	blockbuster	111.762
16	percent	212.432	65	debit	108.437
17	proved	210.812	66	dairy	106.945
18	exploration	200.101	67	offset	106.107
19	sears	199.969	68	rate	105.902
20	automotive	199.558	69	undeveloped	104.559
21	cellular	198.908	70	pharmacy	103.659
22	com	194.168	71	broadcasting	102.355
23	home	192.393	72	drive	100.442
24	hotel	191.891	73	celebrity	99.867
25	analog	189.888	74	quarter	99.24
26	red	181.821	75	bmc	98.493
27	drilling	175.103	76	refined	98.108
28	tinto	169.391	77	online	97.582
29	old	166.094	78	travel	97.094
30	merchant	165.535	79	during	96.552
31	pharmaceutical	164.553	80	pharmacies	96.16
32	segment	164.322	81	ships	95.779
33	car	163.297	82	motorola	94.668
34	primarily	162.688	83	gains	94.595
35	earnings	161.09	84	garden	94.528
36	suspension	160.001	85	increased	92.874
37	generic	159.309	86	engine	92.822
38	hat	156.631	87	greenhouse	92.659
39	linux	152.942	88	pretax	91.852
40	lower	152.431	89	gold	91.511
41	assemblies	151.103	90	harris	90.655
42	notes	151.075	91	video	89.505
43	tax	150.887	92	emea	89.359
44	community	150.153	93	betting	87.52
45	news	148.529	95	cnooc	86.44
46	due	143.664	96	living	85.934
47	senior	142.454	97	reserves	85.089
48	amgen	140.446	98	homebuilding	84.743
49	fourth	139.006	99	level	84.522
50	broadcast	137.579	100	cummins	84.471
101	program	84.369	151	optimum	70.848
102	reflecting	84.345	152	strong	70.217
103	see	84.261	153	australian	70.018
104	investments	84.227	154	players	69.849
105	declined	83.78	155	newspaper	69.544
106	semiconductors	82.83	156	spending	69.473
107	africa	82.377	157	housing	69.208
108	diluted	82.307	158	partially	69.196
109	pharmaceuticals	81.22	159	vacation	68.92
110	newspapers	80.719	160	mobile	68.888

Table D.5: Top 160 keywords found in non-fraud reports

APPENDIX E

.

Financial Performance

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
Loss	1227	3348	0.30	YES	0.03	Overuse in Fraud set as compared to Non Fraud set
Margin	498	1800	28.25	YES	-0.38	Underuse in Fraud set as compared to Non Fraud set
Revenue	2903	7521	10.24	YES	0.10	Overuse in Fraud set as compared to Non Fraud set
Sale	1249	7521	764.51	YES	-1.12	Underuse in Fraud set as compared to Non Fraud set
turnover	67	195	0.11	NO	-0.07	Underuse in Fraud set as compared to Non Fraud set
Cost	2166	6392	5.87	YES	-0.09	Underuse in Fraud set as compared to Non Fraud set
Performance	963	3640	76.26	YES	-0.44	Underuse in Fraud set as compared to Non Fraud set
Profit	686	2417	31.37	YES	-0.34	Underuse in Fraud set as compared to Non Fraud set
Result	2970	7058	50.17	YES	0.23	Overuse in Fraud set as compared to Non Fraud set

Table E.1: Results obtained from using Log Likelihood Calculator [198, 204] on words under ‘Financial Performance’ category from [199].

Financial Position

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
asset	733	1702	16.16	YES	0.26	Overuse in Fraud set as compared to Non Fraud set
borrowing	157	399	0.89	YES	0.13	Overuse in Fraud set as compared to Non Fraud set
debt	1572	5008	22.80	YES	-0.20	Underuse in Fraud set as compared to Non Fraud set
Cash	3089	9679	34.34	YES	-0.17	Underuse in Fraud set as compared to Non Fraud set
liability	628	1415	18.63	YES	0.30	Overuse in Fraud set as compared to Non Fraud set

Table E.2: Results obtained from using Log Likelihood Calculator [198, 204] on words under ‘Financial Position’ category from [199].

Self Reference

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
company	11059	30471	0.61	NO	0.01	Overuse in Fraud set as compared to Non Fraud set
division	735	742	363.02	YES	1.46	Overuse in Fraud set as compared to Non Fraud set

Table E.3: Results obtained from using Log Likelihood Calculator [198, 204] on words under 'Self- Reference' category from [199]

General financial terms

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
currency	873	3069	39.14	YES	-0.34	Underuse in Fraud set as compared to Non Fraud set
exchange	1260	3727	3.69	YES	-0.09	Underuse in Fraud set as compared to Non Fraud set
expenditure	117	317	0.06	NO	0.04	Overuse in Fraud set as compared to Non Fraud set
fixed	598	1696	0.18	NO	-0.03	Underuse in Fraud set as compared to Non Fraud set
interest	3075	9327	17.91	YES	-0.13	Underuse in Fraud set as compared to Non Fraud set
investment	1062	3638	37.28	YES	-0.30	Underuse in Fraud set as compared to Non Fraud set
net	3091	10734	123.86	YES	-0.32	Underuse in Fraud set as compared to Non Fraud set
risk	1668	4448	2.07	NO	0.06	Overuse in Fraud set as compared to Non Fraud set
sterling	89	142	15.98	YES	0.80	Overuse in Fraud set as compared to Non Fraud set
capital	2787	6344	75.15	YES	0.29	Overuse in Fraud set as compared to Non Fraud set
financial	4304	12548	7.37	YES	-0.07	Underuse in Fraud set as compared to Non Fraud set
share	1164	4639	129.55	YES	-0.52	Underuse in Fraud set as compared to Non Fraud set
tax	1818	6919	150.89	YES	-0.45	Underuse in Fraud set as compared to Non Fraud set

Table E.4: Results obtained from using Log Likelihood Calculator [198, 204] on words under 'Self- Reference' category from [199]

Comparison

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
rate	1992	7143	105.90	YES	-0.37	Underuse in Fraud set as compared to Non Fraud set
reduce	694	2033	1.44	NO	-0.08	Underuse in Fraud set as compared to Non Fraud set
up	1001	2755	0.07	NO	0.01	Overuse in Fraud set as compared to Non Fraud set
compare	32	51	5.77	YES	0.80	Overuse in Fraud set as compared to Non Fraud set
decrease	1064	3569	29.95	YES	-0.27	Underuse in Fraud set as compared to Non Fraud set
growth	1965	6792	75.32	YES	-0.31	Underuse in Fraud set as compared to Non Fraud set
high	1965	3527	230.02	YES	0.63	Overuse in Fraud set as compared to Non Fraud set
increase	3288	10675	62.34	YES	-0.22	Underuse in Fraud set as compared to Non Fraud set
increasingly	151	521	5.69	YES	-0.31	Underuse in Fraud set as compared to Non Fraud set
level	625	2577	84.52	YES	-0.57	Underuse in Fraud set as compared to Non Fraud set
lower	894	3853	152.43	YES	-0.63	Underuse in Fraud set as compared to Non Fraud set
more	2215	6109	0.10	YES	0.01	Underuse in Fraud set as compared to Non Fraud set
overall	450	1521	13.76	YES	-0.28	Underuse in Fraud set as compared to Non Fraud set
than	2158	5708	3.82	YES	0.07	Overuse in Fraud set as compared to Non Fraud set

Table E.5: Results obtained from using Log Likelihood Calculator [198, 204] on words under ‘Comparison’ category from [199].

Up words

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
grow	234	625	0.27	YES	0.06	Underuse in Fraud set as compared to Non Fraud set
higher	938	4532	266	YES	-0.80	Underuse in Fraud set as compared to Non Fraud set
increase	3288	10675	62.34	YES	-0.22	Underuse in Fraud set as compared to Non Fraud set
more	2215	6109	0.10	YES	0.01	Overuse in Fraud set as compared to Non Fraud set
overall	450	1521	13.76	YES	-0.28	Underuse in Fraud set as compared to Non Fraud set

Table E.6: Results obtained from using Log Likelihood Calculator [198, 204] on words under ‘Up’ category from [199].

Down words

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
decrease	1064	3569	29.95	YES	-0.27	Underuse in Fraud set as compared to Non Fraud set
lower	894	3853	152.43	YES	-0.63	Underuse in Fraud set as compared to Non Fraud set
reduce	694	2033	1.44	YES	-0.08	Underuse in Fraud set as compared to Non Fraud set

Table E.7: Results obtained from using Log Likelihood Calculator [198, 204] on words under 'Down' category from [199].

Temporal

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
before	539	4400	732.87	YES	-1.55	Underuse in Fraud set as compared to Non Fraud set
completed	639	1594	5.25	YES	0.16	Overuse in Fraud set as compared to Non Fraud set
continued	579	2292	61.87	YES	-0.51	Underuse in Fraud set as compared to Non Fraud set
during	2263	7912	96.55	YES	-0.33	Underuse in Fraud set as compared to Non Fraud set
end	872	3056	38.02	YES	-0.33	Underuse in Fraud set as compared to Non Fraud set
last	327	848	1.12	NO	0.10	Overuse in Fraud set as compared to Non Fraud set
new	3874	11422	10.17	YES	-0.09	Underuse in Fraud set as compared to Non Fraud set
now	326	991	2.00	NO	-0.13	Underuse in Fraud set as compared to Non Fraud set
previous	163	630	14.99	YES	-0.48	Underuse in Fraud set as compared to Non Fraud set

Table E.8: Results obtained from using Log Likelihood Calculator [198, 204] on words under 'Temporal' category from [199]

General Words

Lemma	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
activity	274	808	0.72	NO	-0.09	Underuse in Fraud set as compared to Non Fraud set
both	962	3263	30.52	YES	-0.29	Underuse in Fraud set as compared to Non Fraud set
business	6417	17306	4.23	YES	0.04	Overuse in Fraud set as compared to Non Fraud set
but	715	2456	25.79	YES	-0.31	Underuse in Fraud set as compared to Non Fraud set
cent	175	768	32.60	YES	-0.66	Underuse in Fraud set as compared to Non Fraud set
customer	1627	5364	37.49	YES	-0.25	Underuse in Fraud set as compared to Non Fraud set
development	2133	6377	8.60	YES	-0.11	Underuse in Fraud set as compared to Non Fraud set
due	2340	8543	143.66	YES	-0.39	Underuse in Fraud set as compared to Non Fraud set
exceptional	51	236	12.11	YES	-0.74	Underuse in Fraud set as compared to Non Fraud set
facility	1535	3176	86.87	YES	0.43	Overuse in Fraud set as compared to Non Fraud set
include	1432	4430	12.64	YES	-0.15	Underuse in Fraud set as compared to Non Fraud set
management	3005	7223	43.86	YES	0.21	Overuse in Fraud set as compared to Non Fraud set
market	4151	11248	1.95	NO	0.04	Underuse in Fraud set as compared to Non Fraud set
network	1986	3627	215.98	YES	0.61	Overuse in Fraud set as compared to Non Fraud set
not	5199	12817	52.29	YES	1.13	Overuse in Fraud set as compared to Non Fraud set
number	1450	4011	0.02	0.01	0.01	Overuse in Fraud set as compared to Non Fraud set
operating	3131	10210	62.89	YES	-0.23	Underuse in Fraud set as compared to Non Fraud set
operations	4264	13473	54.44	YES	-0.19	Underuse in Fraud set as compared to Non Fraud set
per	1406	4987	68.17	YES	-0.35	Underuse in Fraud set as compared to Non Fraud set
products	5469	15879	7.81	YES	-0.06	Underuse in Fraud set as compared to Non Fraud set
programme	94	130	25.15	YES	1.01	Overuse in Fraud set as compared to Non Fraud set
property	835	2205	1.57	NO	0.07	Overuse in Fraud set as compared to Non Fraud set
Retail	601	1894	7.35	YES	-0.18	Underuse in Fraud set as compared to Non Fraud set
services	5926	15031	35.08	YES	0.13	Overuse in Fraud set as compared to Non Fraud set
significant	2099	5652	1.53	NO	0.05	Overuse in Fraud set as compared to Non Fraud set
store	164	914	78.59	YES	-1.00	Underuse in Fraud set as compared to Non Fraud set
strong	402	1741	70.22	YES	-0.64	Underuse in Fraud set as compared to Non Fraud set
systems	2528	6900	0.61	NO	0.03	Overuse in Fraud set as compared to Non Fraud set
total	1543	4863	18.90	YES	-0.18	Underuse in Fraud set as compared to Non Fraud set
trading	312	1270	39.27	YES	-0.55	Underuse in Fraud set as compared to Non Fraud set
years	1601	4400	0.15	NO	0.02	Overuse in Fraud set as compared to Non Fraud set

Table E.9: Results obtained from using Log Likelihood Calculator [198, 204] on words under ‘General Words’ category from [199].

Words from Negative word list [126,143]	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
under	3110	8192	6.45	YES	0.08	Overuse in Fraud
risk, riskier, risks, risky	2629	6949	4.78	YES	0.07	Overuse in Fraud
loss/lower/losses/less/low decrease/decreases/decreasing/decline/down/declined/declines/declining	16547	24087	3882	YES	0.93	Overuse in Fraud
adverse, adversely, against, unable	3385	8314	36.13	YES	0.18	Overuse in Fraud
downgrade, downgraded, downgrades, downgrading, downsize, downsized, downsizing	72	181	0.51	NO	0.14	Overuse in Fraud
Downtime, down, downturn, downturns, downward, downwards, dampen, dampened, deteriorate, deteriorated, deteriorates, deteriorating, deterioration, devalue, devalued, devaluing	542	1705	6.44	YES	-0.18	Underuse in Fraud
Impair, impaired, impairing, impairment, impairments, impairs, impede, impeded, impediment, impediments, impeding	804	2664	19.67	YES	-0.25	Underuse in Fraud
problem, problematic, problems	349	521	76.43	YES	0.90	Overuse in fraud
Fail, failed, failing, fails, failure, failures	889	2038	22.37	YES	0.28	Overuse in fraud
Difficult, difficulties, difficulty	570	1640	0.51	NO	-0.05	Underuse in Fraud
Bankrupt, bankruptcies, bankruptcy	291	236	194	YES	1.78	Overuse in fraud
Restructure, restructured, restructures, restructuring, restructurings	578	1584	0.08	NO	0.02	Overuse in Fraud
Litigants, litigate, litigated, litigating, litigation, litigations	400	1072	0.38	NO	0.05	Overuse in Fraud
Negative, negatively	454	1668	29.26	YES	-0.40	Underuse in Fraud
Terminate, terminated, terminates, terminating, termination, terminations	548	1575	0.46	NO	-0.05	Underuse in Fraud
Delay, delayed, delaying, delays	488	1151	9.03	YES	0.24	Overuse in Fraud
Harm, harmed, harmful, harming, harms	283	579	17.18	YES	0.44	Overuse in Fraud
Damage, damaged, damages, damaging	284	729	1.28	NO	0.11	Overuse in Fraud
Concern, concerned, concerns	143	359	1.04	NO	0.15	Overuse in Fraud
shortage, shortages, shortfall, shortfalls	128	285	4.22	YES	0.32	Overuse in Fraud
Challenge, challenged, challenges, challenging	225	830	14.25	YES	-0.41	Underuse in Fraud
Unfavourable, unfavourably, unfavourable	89	450	30.25	YES	-0.86	Underuse in Fraud
Uncertain, uncertainty	278	693	2.31	NO	0.16	Overuse in Fraud
Disrupt, disrupted, disrupting, disruption, disruptions, disruptive	180	742	24.31	YES	-0.57	Underuse in Fraud

Words from Negative word list [126, 143]	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
Weak, weaken, weakened, weakening, weakens, weaker, weakest, weakly, weakness, weaknesses	225	668	0.74	NO	-0.10	Underuse in Fraud
severe	73	199	0.02	NO	0.03	Overuse in Fraud
claims	797	1737	31.29	YES	0.35	Overuse in Fraud
unable	533	1052	39.78	YES	0.49	Overuse in Fraud
force	309	587	28.05	YES	0.55	Overuse in Fraud
Penalty, penalties, penalize. penalized	217	485	6.92	YES	0.31	Overuse in Fraud
Volatile, volatility	248	859	9..70	YES	-0.32	Underuse in Fraud
inability	188	327	24.95	YES	0.68	Overuse in Fraud
able	927	1771	81.86	YES	0.54	Overuse in Fraud

Table E.10 Some words in negative word list derived by [126], log likelihood scores displayed.

Words from Positive word list [126,143]	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
Increase, increased, increases, increasing	7389	21412	9.63	YES	-0.06	Underuse in Fraud Set
good, great, greater, greatest, greatly, grew, grow, growing, grown, grows, growth	3412	10109	10.52	YES	-0.09	Underuse in Fraud Set
High, higher, highest	2074	6464	21.00	YES	-0.17	Underuse in Fraud Set
More, most	3120	7092	85.27	YES	0.29	Overuse in Fraud set
Profitability, profitable, profitably	645	1417	23.77	YES	0.34	Overuse in Fraud set
Expand, expanded, expanding, expands, expansion	1314	3330	7.93	YES	0.13	Overuse in Fraud set
Strength, strengthen, strengthened, strengthening, strengthens, strengths, strong, stronger, strongest	747	3059	97.26	YES	-0.56	Underuse in Fraud set
Gain, gained, gaining, gains, exceed, exceeded, exceeding, exceeds	961	3462	53.07	YES	-0.37	Underuse in Fraud set
Improve, improved, improvement, improvements, improves, improving	1144	4597	133.68	YES	-0.53	Underuse in Fraud set
Advantage, advantaged, advantageous, advantageously, advantages	458	862	43.47	YES	0.56	Overuse in Fraud Set
Enhance, enhanced, enhancement, enhancements, enhances, enhancing	543	1855	18.54	YES	-0.30	Underuse in Fraud set
Effective, efficiencies, efficiency, efficient, efficiently	1472	3605	16.43	YES	0.18	Overuse in Fraud Set
Beneficial, beneficially, benefit, benefited, benefiting, benefitted	673	2168	11.47	YES	-0.21	Underuse in Fraud set
Satisfaction, satisfactorily, satisfactory, satisfied, satisfies, satisfy, satisfying	365	866	6.29	YES	0.23	Overuse in Fraud set
Excellence, excellent, excels, exceptional exceptionally	146	515	6.75	YES	-0.34	Underuse in Fraud set
Collaborate, collaborated, collaborates, collaborating, collaboration, collaborations, collaborative, collaborator, collaborators	118	261	4.11	YES	0.33	Overuse in Fraud set
Solid	180	369	10.79	YES	0.44	Overuse in Fraud Set
Attractive, attractiveness	143	288	9.53	YES	0.46	Overuse in Fraud Set

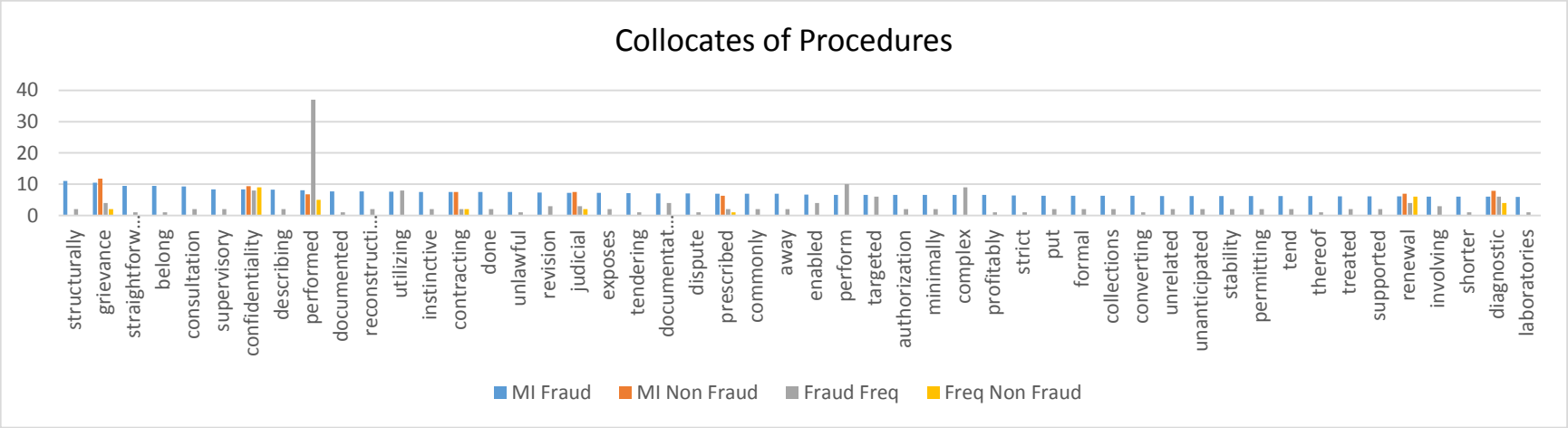
Table E.11 Some words in positive word list derived by [126], log likelihood scores displayed.

Words from Uncertainty word list [126, 143]	RF Fraud	RF NF	L-L Score	Significant p < 0.05	Log Ratio	Comment
less	625	1540	6.34	YES	0.17	Overuse in Fraud
may, maybe, might	6609	155598	121.33	YES	0.24	Overuse in Fraud
can	1771	4145	35.79	YES	0.25	Overuse in Fraud
could	3648	9230	22.77	YES	0.14	Overuse in Fraud
believe, believed, believes	1869	4302	45.14	YES	0.27	Overuse in Fraud
About	630	1483	11.92	YES	0.24	Overuse in Fraud
Overall	450	1521	13.76	YES	-0.28	Underuse in Fraud
general, generally	2533	6417	15.39	YES	0.13	Overuse in Fraud
Fluctuate, fluctuated, fluctuates, fluctuating, fluctuation, fluctuations	561	1790	8.33	YES	-0.20	Underuse in Fraud
Likely, liklihood	390	911	8.06	YES	0.25	Overuse in Fraud
Differ, differed, differing, differs	170	596	7.44	YES	-0.34	Underuse in fraud
Exposure, exposures	411	1403	13.93	YES	-0.30	Underuse in fraud
Volatile, volatilities, volatility	251	874	10.32	YES	-0.33	Underuse in fraud
Uncertain, uncertainties, uncertainty	449	1390	4.02	YES	-0.16	Underuse in fraud
Revise, revised	81	293	4.62	YES	-0.38	Underuse in fraud
Sometime, sometimes, somewhat	106	227	4.75	YES	-0.38	Overuse in Fraud
rather	150	347	3.44	YES	0.26	Overuse in Fraud
pending	194	429	6.77	YES	0.33	Overuse in Fraud
Nearly	64	274	10.55	YES	-0.62	Underuse in Fraud

Table E.12 Some words in uncertainty word list derived by [126], log likelihood scores displayed.

APPENDIX F

1. Collocates of Procedures



Example Key Words in Context

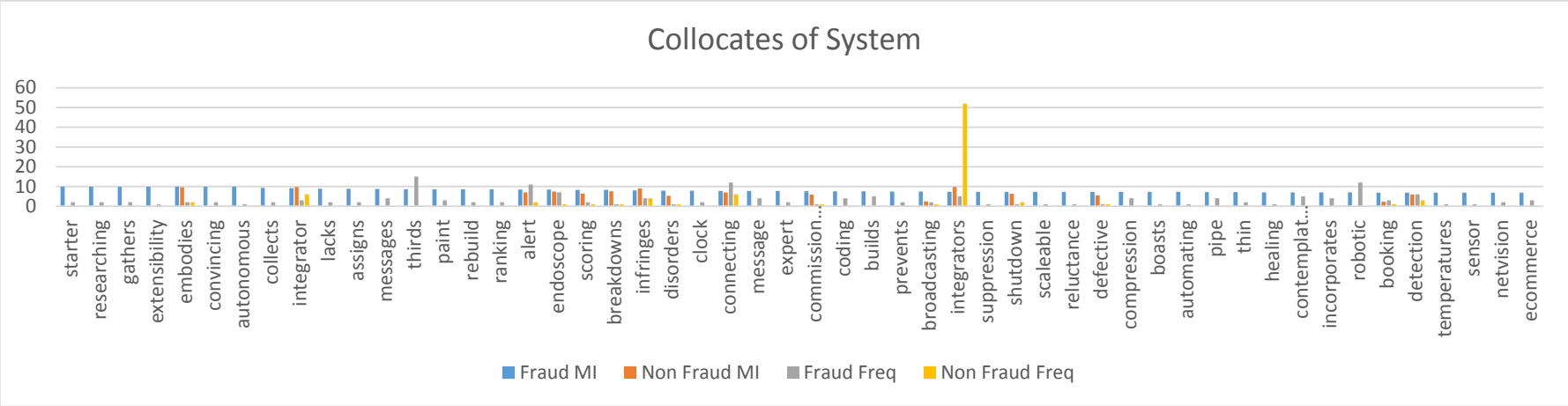
"…one of the most frequently performed procedures, involves x-ray imaging guidance (fraud report)"

"…mammography and other imaging and diagnostic procedures performed by our products (non-fraud reports)"

Procedures	MI Fraud	MI Non Fraud
Mean	3.73	1.19
Variance	5.47	5.20
Observations	241	241
P(T<=t) two-tail	1.43383E-29	

Figure F.1: Keyword 'procedures' - collocates, concordance and significance.

2 Collocates of System



Example Key Words in Context

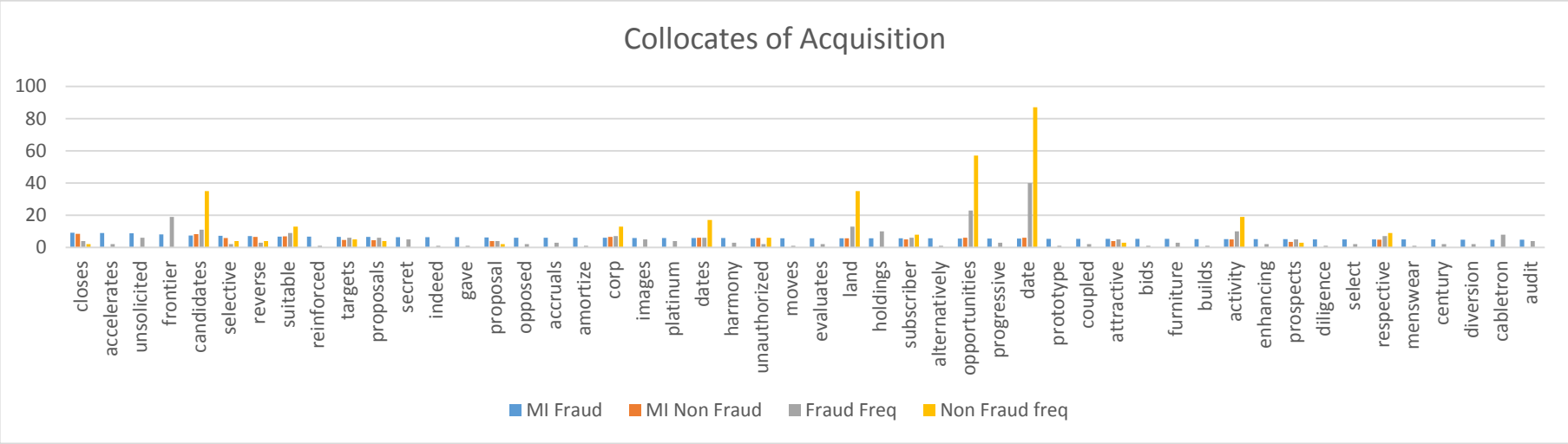
"…adopted rules implementing an emergency alert system (fraud report)"

"…marking and lighting; and - emergency alert system requirements. The FCC may enforce" non-fraud report)"

System	Fraud MI	Non Fraud MI
Mean	3.260102364	1.375517758
Variance	5.973719499	4.635025897
Observations	571	571
P(T<=t) two-tail	2.89962E-40	

Figure F.2: Keyword 'system' - collocates, concordance and significance.

3 Collocates of Acquisition



Example Key Words in Context

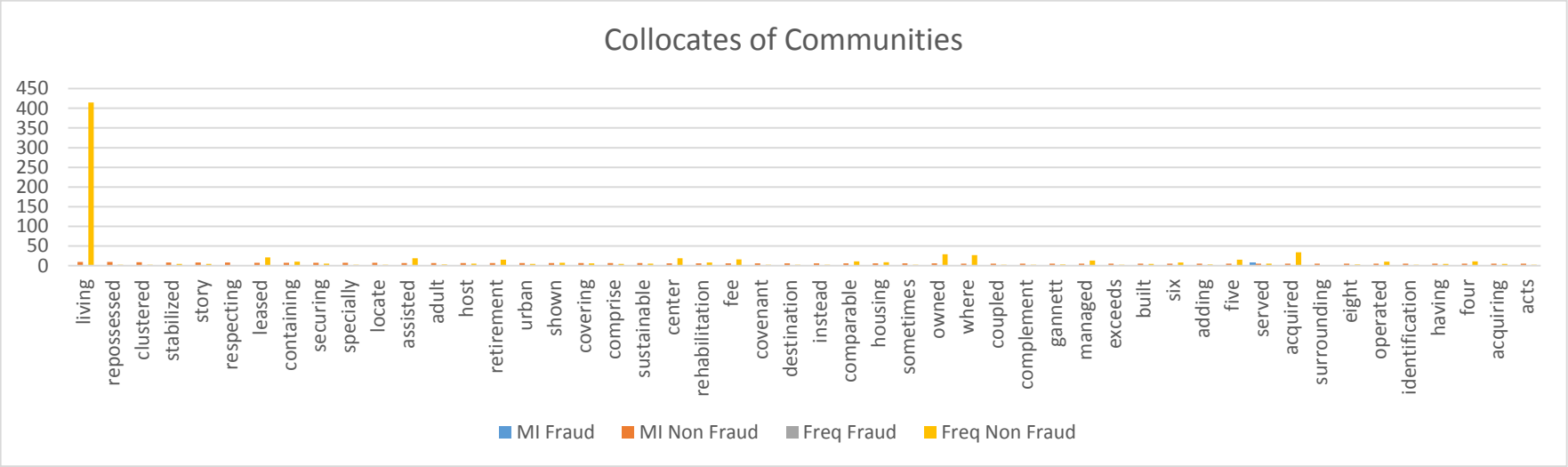
"…Increased competition for acquisition candidates may increase purchase price (fraud report)"

"…because there should be more complementary acquisition candidates available for us to consider (non-fraud report)"

Acquisition	MI Fraud	MI Non Fraud
Mean	2.34	0.85
Variance	5.02	3.40
Observations	343	343
P(T<=t) two-tail	2.76767E-20	

Figure F.3: Keyword 'Acquisition' - collocates, concordance and significance.

4 Collocates of Communities



Example Key Words in Context

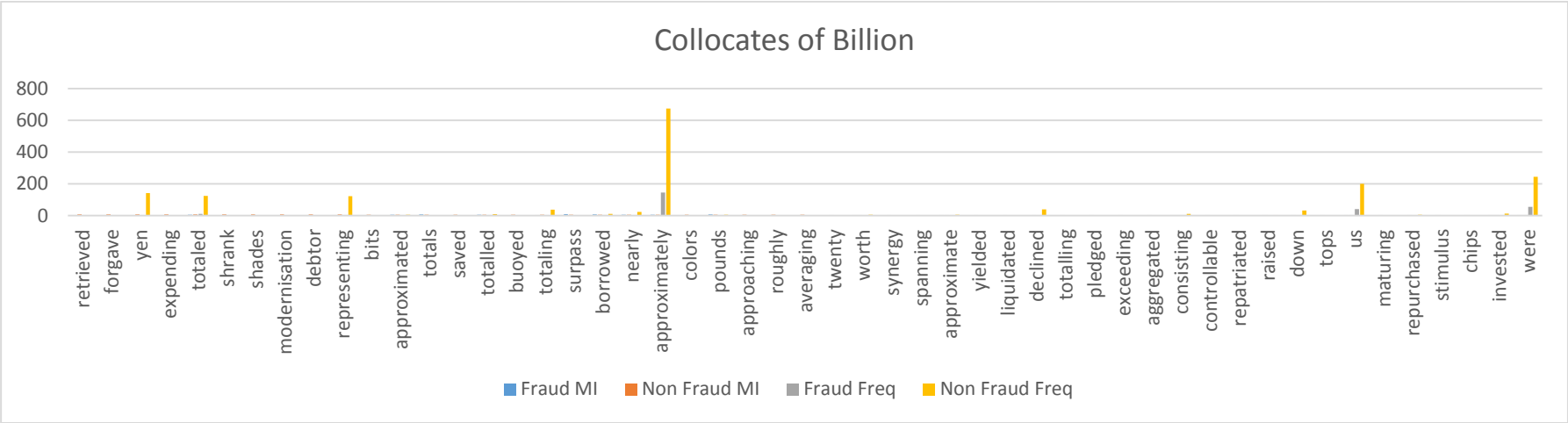
"..We also provide information about our active communities and mortgage financing through (fraud reports)"

"…Our senior living communities offer residents a supportive" (non-fraud report)

Communities	MI Fraud	MI Non-Fraud
Mean	0.30	10.59
Variance	2.47	1244.07
Observations	323	323
P(T<=t) two-tail	2.95E-07	

Figure F.4: Keyword 'Communities' - collocates, concordance and significance.

5. Collocates of Billion



Example Key Words in Context

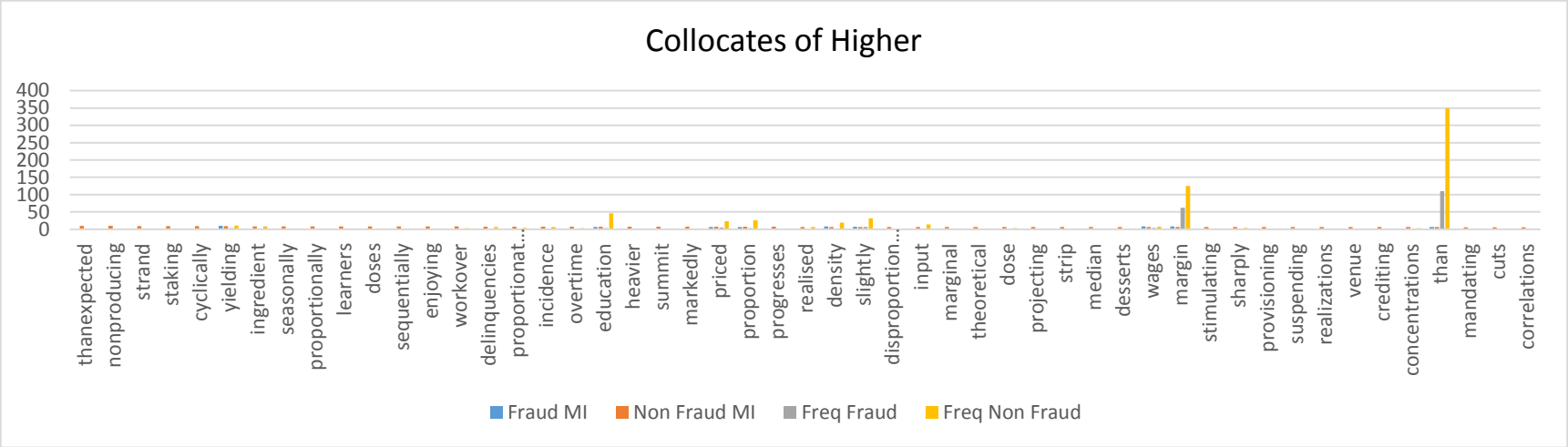
"…fiscal year 2001 represents a decrease of $2.26 billion over fiscal year 2000. Excluding a special" (fraud report)

"…We forecast these savings to total $2 billion over the next decade. Process improvements"(non-fraud report)

Billion	Fraud MI	Non-Fraud MI
Mean	0.76	1.28
Variance	2.91	4.75
Observations	592	592
P(T<=t) two-tail	7.02469E-06	

Figure F.5: Keyword 'Billion' - collocates, concordance and significance.

6 Collocates of Higher



Example Key Words in Context

"...a result of the decrease in higher margin calling card and dial" (Fraud Report)

"…through deliberate efforts to grow new higher margin business and to cycle out" (Non Fraud Report)

Higher	MI Fraud	MI Non Fraud
Mean	0.862	2.312
Variance	3.110	5.069
Observations	894	894
P(T<=t) two-tail	9.3087E-49	

Figure F.6: Keyword 'Higher' - collocates, concordance and significance.

Million

Fraud			*Non-Fraud*	
collocate	**MI**	***Example of Concordances***	**MI**	***Example of Concordances***
tourists	7.490	*with over eight million tourists annually*		*Not Observed*
inhabitants	6.905	*million inhabitants with a good con*		*Not Observed*
tonnes	6.747	*million tonnes of measured and million tonnes of stock was added*	6.403	*million tonnes capacity in second half*
gallons	6.727	*million gallons in the month*		*Not Observed*
pertained	6.490	*million, pertained to the sales*		*Not Observed*
deduction	6.490	*million deduction related to*		*Not Observed*
approximately	5.989	*approximately $1 million to the landlord in*	-2.401	*Approximately 275 company-operated*
totalled	5.753	*the Imaging Systems Business totalled*		*The Company totaled RMB 51.0 billion*
respectively	4.905	*and $1.8 million, respectively. The timing*		*Not Observed*
decreased	4.612	*Net sales decreased $15.3 million, or*	-1.710	*Other income decreased $66 million in 2000*
approx	6.905	*approx. 440 million dollars*	6.558	*(approx. $52.8 million) of products*
reverted		*Not observed*	6.14	*million reverted to the Company*
aggregated	3.717	*acquired companies aggregated $1,592.3 million.*	4.909	*Companies aggregated $476 million*
awarding		*Not observed*	4.684	*awarding $35 million in compensatory*
decreased	4.612	*administrative expenses decreased*	4.536	*working capital decreased*
invested	3.927	*invested $90.0 million for an approximate*	4.226	*We have invested $100 million toward*
declined	3.954	*segment revenue declined*	4.144	*Accrued liabilities declined $74.9 million,*
repayable		*Not observed*	4.143	*million repayable in three years*
yielded		*Not observed*	4.084	*yielded $163.8 million of net proceeds,*
contributed	4.528	*contributed $110 million of revenue*	4.043	*contributed $49.7 million in revenues*

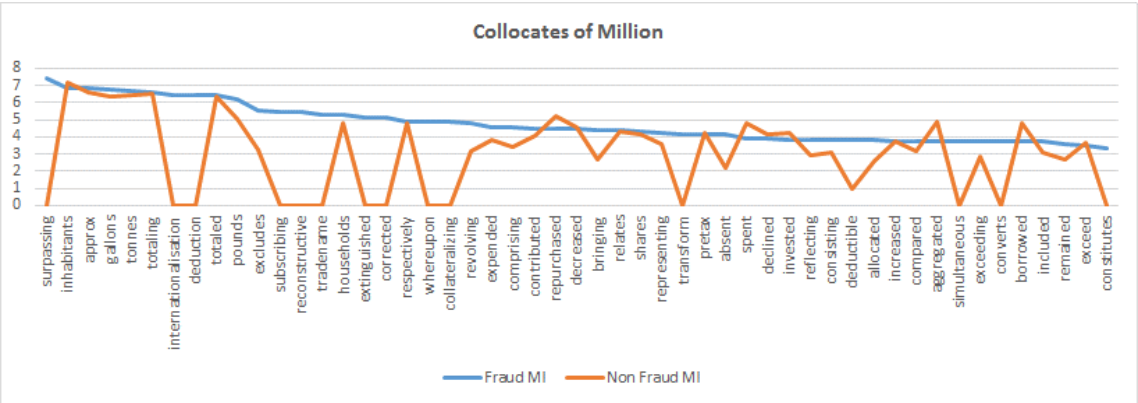


Figure F.7: Collates and Concordances for ‘Million’.

Business

Fraud				*Non-Fraud*		
Collocate	**MI**	***Freq***	***Example of Concordances***	**MI**	***Freq***	***Example of Concordances***
savvy	8.06	*1*	*will think and act with business savvy.*			*Not Observed*
combinations	7.61	*82*	*assets and liabilities in business combinations, valuation*	7.536	*132*	*purchase method for business combinations and, accordingly*
oldest	8.06	*1*	*And our oldest business, the interstate*			*Not Observed*
doing	7.09	*33*	*cost-effective way of doing business, mapping the growth*	7.588	*171*	*regions with lower costs of doing business, may be able to prohibitions on doing business and damage t*
transcends	7.06	*23*	*production system that transcends business segments*			*Not Observed*
Continuity	6.98	*8*	*disaster recovery and business continuity plan*	7.205	*59*	*Our business continuity and disaster*
revoking	6.74	*2*	*levying fines, revoking business and other licenses*			*Not Observed*
exploiting	6.74	*2*	*new technologies or exploiting business opportunities;*			*Not Observed*
Endeavours	6.48	*1*	*several other business endeavours and are not obligated*			*Not Observed*
Conducts	6.39	*11*	*Skandia conducts business throughout*	5.604	*20*	*management conducts business development activities*
skilfully			*Not observed*	8.077	*1*	*our UK business skilfully to reach*
Transact(s)	7.14	*10*	*we may in the future transact business in other currencies*	7.854	*6*	*The company transacts business in various foreign*
Solicits			*Not observed*	7.492	*2*	*American Southern also solicits business from government*
Rationale			*Not observed*	7.077	*2*	*The business rationale for an institutional pharmacy*
coinciding			*Not observed*	7.077	*1*	*strategic review of the business, coinciding with the move*
Bottomed			*Not observed*	7.077	*1*	*The Company's overall business bottomed in 2001*
awaited			*Not observed*	7.077	*1*	*traffic safety systems business awaited a new highway*
suffers			*Not observed*	6.492	*2*	*suffers business disruptions*
agility			*Not observed*	6.492	*2*	*production requires overall business agility. We must operate efficiently.*
topped			*Not observed*	6.077	*1*	*embedded value of new business topped 374 million dollars.*

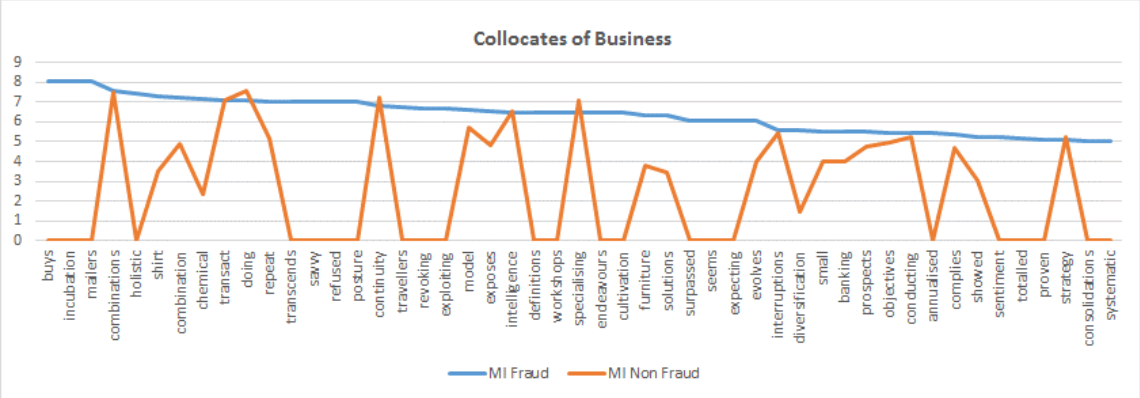


Figure F.8: Collates and Concordances for ‘Business’.

Products

Fraud				*Non-Fraud*		
Collocate	**MI**	***Freq***	***Example of Concordances***	**MI**	***Freq***	***Example of Concordances***
Recalling	8.195	*2*	*Or approvals recalling products, ceasing product*			*Not Observed*
Viable	7.592	*27*	*into commercially viable products principally related*	6.091	*19*	*developing commercially viable products from the purchased*
noncompetitive	6.87	*2*	*render these products noncompetitive. Therefore,*	4.742	*1*	*permitted to sell other, noncompetitive products only*
Saleable	6.75	*14*	*the Iron Ore Division's principal saleable products for the years ended*	5.742	*2*	*in the production of our saleable products. The three*
Defective	6.610	*2*	*also lead to liability for defective products as a result of lawsuits*	4.786	*3*	*consumer can return defective products during the*
Upcoming	6.025	*2*	*The company's upcoming products will meet market*			*Not Observed*
Finished	5.680	*14*	*Components and finished products. Our products*	6.023	*40*	*materials related to finished products. The mailing*
Competing	5.668	*17*	*Is superior to competing products would have a*	5.382	*35*	*new technological advancements and competing products entering*
Adhere	5.494	*2*	*To which products adhere to these technical*			*Not Observed*
Diversifying	5.338	*3*	*efforts in this direction, diversifying products and packaging*			*Not Observed*
Plunged			*Not Observed*	8.201	*2*	*Demand for refined oil products plunged after experienced*
Inspecting			*Not Observed*	8.201	*1*	*products, inspecting plants and*
Noncompliant			*Not Observed*	7.201	*1*	*noncompliant products to compliant products*
Plummeted			*Not Observed*	6.616	*1*	*prices of petrochemical products plummeted harshly.*
expeditiously			*Not Observed*	6.616	*2*	*products expeditiously will be a significant*
exported	5.19	*1*	*products exported by local producers.*			*Revenue in fiscal 2006 from products exported from the*
unsold	4.73	*1*	*Outdoor products unsold from the 1995*	6.086	*19*	*have rights to return unsold products. We may purchase*
fabricate	6.87	*2*	*Our third-party foundries fabricate products for other companies*	6.031	*4*	*These foundries fabricate products for other companies*
differentiated				5.699	*9*	*Provide differentiated products and service*

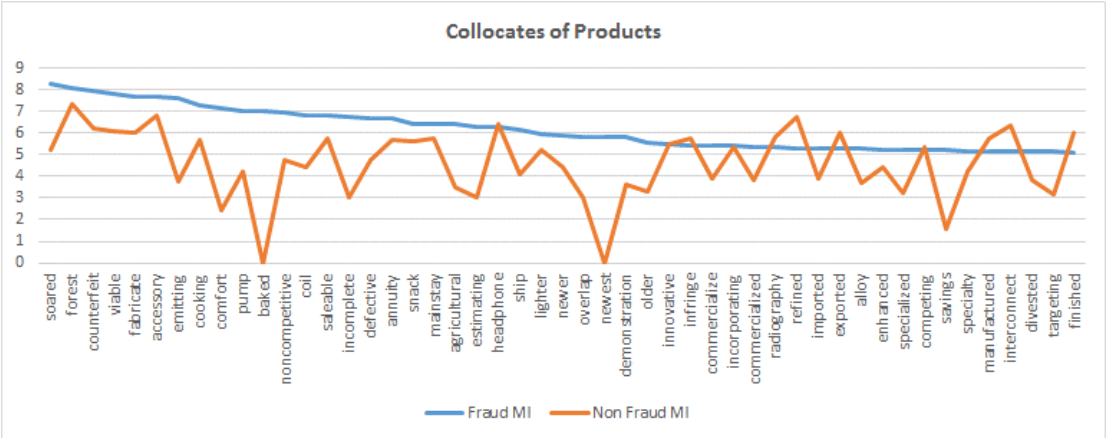


Figure F.9: Collates and Concordances for 'Products'.

Sales

Fraud				*Non-Fraud*		
Collocate	**MI**	***Freq***	***Example of Concordances***	**MI**	***Freq***	***Example of Concordances***
Sacrificed	8.469	*1*	*Excluding the impact of sacrificed sales in December and*			*Not Observed*
upswing	7.469	*1*	*the ability to meet a future sales upswing. Skandia*			*Not Observed*
Outpacing	7.469	*1*	*In net sales outpacing required expenses*			*Not Observed*
Force(s)	7.455	*149*	*In the sales force in the original market*	6.62	*26*	*EM Americas reorganization, the sales forces of Hamilton*
Cycle	6.764	*23*	*delays inherent in our lengthy sales cycle increase the risk* *Due to this lengthy sales cycle, we may experience*	5.162	*31*	*In the sales cycle by engineers a* *Our sales cycle is lengthy and variable,*
representatives	6.711	*34*	*with third-party sales representatives to strengthen*	6.45	*88*	*domestically through direct sales representatives and a dealer*
successes	6.595	*3*	*Skandia's sales successes in the American market*	4.344	*2*	*were also direct to sales successes. Global Television*
export	6.354	*15*	*The Company's export sales have generally*	5.160	*38*	*US export sales, raw materials are a* *US export sales are not a significant*
Direct	6.156	*105*	*network of direct sales and service offices* *augmented by a direct sales force working*	5.950	*322*	*Through both direct sales and dealer sales*
Volumes	5.919	*42*	*Some contractual sales volumes have been reduced*	6.33	*246*	*attempt to increase unitsales volumes and our market share*
tradeshows			*Not observed*	8.388	*2*	*materials and tradeshows. Sales and marketing*
slipped			*Not observed*	7.803	*2*	*In Japan sales slipped 19.1% to*
Deflated			*Not observed*	7.388	*1*	*These external events deflated sales by approximately*
intensifies			*Not observed*	7.066	*2*	*pattern of quarterly sales, intensifies the risk that a supplier*
shrinking			*Not observed*	6.388	*1*	*In response to shrinking sales volume and surging silver prices,*
dropped			*Not observed*	6.218	*6*	*Domestic sales dropped slightly due to*
store	4.120	*8*	*Same Store Sales Revenue generation,*	6.068	*183*	*Comparable store sales gains, which could* *Comparable store sales gain benefited*
steadied			*Not observed*	6.066	*1*	*Bagley steadied sales and improved margin*

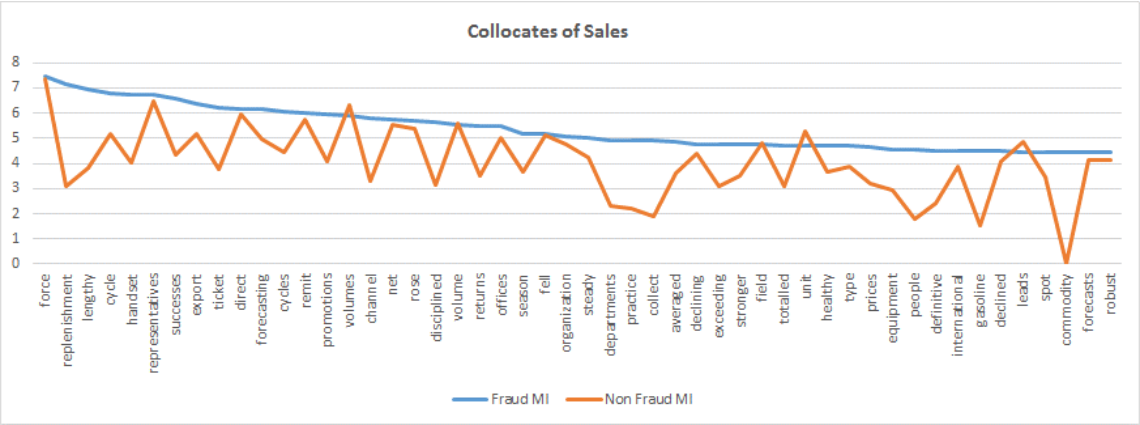


Figure F.10: Collates and Concordances for 'Sales'.

Financial

Collocate	MI	*Freq*	*Example of Concordances*	MI	*Freq*	*Example of Concordances*
wherewithal	8.63071	*1*	*Who lack the financial wherewithal to follow suit.*	**8.54118**	*1*	*Had the financial wherewithal, either alone*
consolidated	7.56669	*606*	*Consolidated Financial Statements included in Item 8*	7.44	*1843*	*Consolidated Financial Statements under the*
statements	8.16598	*1013*	*Adelphia Consolidated Financial Statements included in Item*	8.021	*2573*	*our Consolidated Financial Statements and related*
reporting	7.0724	*215*	*Internal Control financial reporting is effective based*	6.566	*368*	*The accuracy of our financial reporting is dependent on*
collapse	7.045	*1*	*A bankruptcy or financial collapse of one of these fiber*	5.293	*2*	*A bankruptcy or financial collapse of one of the*
instruments	6.704	*136*	*Other derivative financial instruments for the purpose*	6.95	*494*	*our derivative and financial instrument positions.*
Position	6.562	*139*	*Material impact on our financial position and results of operations*	6.782	*519*	*our consolidated financial position, results*
greater	5.986	*4*	*As well as greater financial resources and lower costs*	5.380	*156*	*our competitors may have greater financial, marketing or*
covenant	5.886	*40*	*Breach of such financial covenant, and is discussing*	5.055	*10*	*was in compliance with the financial covenant requirements*
results	4.8804	*199*	*Accordingly the financial results of the acquired systems*	4.961	*713*	*our business and financial results may*
condition	8.48	*760*	*the Company's financial condition and results of*	8.415	*2165*	*"Management's Discussion and Analysis of Financial Condition and Results*
institutions	7.229	*53*	*originated by other financial institutions. However, increase*	7.783	*440*	*or sold to financial institutions engaged in the*
meltdown			*Not observed*	7.541	*1*	*Given the financial meltdown in September*
strength	6.143	*23*	*continued solvency and financial strength of the*	6.326	*106*	*Factor such as financial strength, stability,*
Stability	6.65	*14*	*Assure the financial stability of corporate gaming*	6.1488	*20*	*Experience and financial stability of the vendor.*
Turmoil	4.72	*1*	*Recent financial turmoil may be affected. I*	6.1006	*7*	*As the recent financial turmoil may be affected.*
Misleading			*Not observed*	5.8407	*2*	*Inaccurate or misleading financial statements*

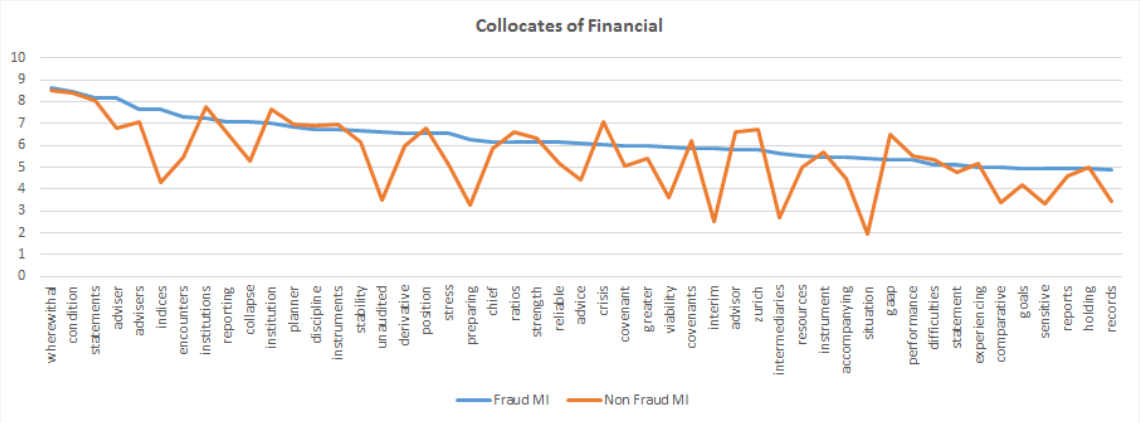


Figure F.11: Collates and Concordances for 'Financial'.

Operations

Fraud				Non-Fraud		
Collocate	**MI**	***Freq***	***Example of Concordances***	**MI**	***Freq***	***Example of Concordances***
unify	9.67	*2*	*Each of them to unify operations in existing service*			*Not Observed*
discontinued	8.14	*167*	*and are presented as discontinued operations.*	7.789	*415*	*The results from discontinued operations, along with the*
Continuing	6.88	*122*	*Income from all continuing operations before income*	7.163	*513*	*To support continuing operations, capital*
route	6.62	*22*	*Including our route operations in Nevada and*	2.152	*116*	*development of new operations, including complementary*
fundamental	6.198	*6*	*Limitations on liens, operations, fundamental changes,*			*Not observed*
divested	5.902	*4*	*divested operations. As of December*	5.06	*8*	*Included within divested operations. The impairment*
overview	5.457	*16*	*Operations Overview Applied*	5.725	*63*	*Operations Operations Overview We believe that s*
collection	5.602	*29*	*number of independent collection operations to enhance f*	3.017	*14*	*acquire landfills and collection operations. Accordingly,*
began	5.309	*22*	*14 markets which began operations in 1996*	4.105	*31*	*We began operations in Florida as a result of*
profitably	5.1304	*2*	*Our PDP operations profitably, and our failure*	0.210	*1*	*build significant and profitable operations over the long term.*
International	5.131	*83*	*Our business. International operations, unfavorable*	5.291	*390*	*Discontinued operations. International sales increased*
regrettably			*Not Observed*	8.438	*1*	*Business operations. Regrettably, we did not*
disappointingly			*Not Observed*	8.438	*1*	*Pellet operations. Disappointingly, this slowdown*
Crippled			*Not Observed*	8.438	*1*	*An explosion crippled operations. Mike Julian drove*
Seating			*Not Observed*	7.49	*14*	*Visteon in 2003 was the exit from ourseating operations in Chesterfield, Mic*
Centralizing			*Not Observed*	6.631	*4*	*Centralizing operations and production facilities*
flourished			*Not Observed*	6.438	*1*	*South African operations flourished, with share increases*
summarize			*Not Observed*	5.853	*1*	*consolidated statements of operations summarize operating*

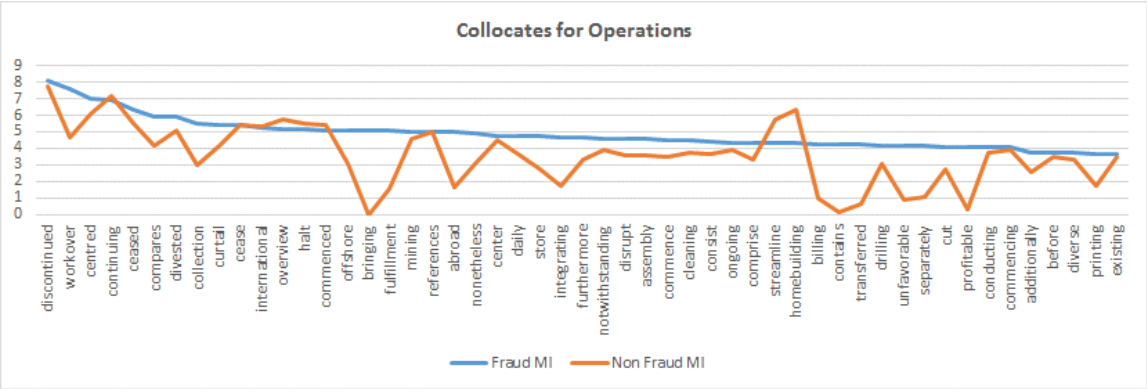


Figure F.12: Collates and Concordances for ‘Operations’.

Revenue

Fraud				*Non-Fraud*		
Collocate	**MI**	***Freq***	***Example of Concordances***	**MI**	***Freq***	***Example of Concordances***
unfortunately	7.507	*1*	*Of our total revenue. Unfortunately, these positives*			*Not Observed*
correlates	7.09	*1*	*Retirement services revenue correlates with an increase*			*Not Observed*
persistent	9.092	*1*	*Affect our revenue. Persistent operational*			*Not Observed*
lifetime	6.507	*1*	*Three times the lifetime revenue of standalone long*			*Not Observed*
recognize	7.043	*29*	*We do not recognize revenue until the customer*	6.105	*39*	*methodology used to recognize revenue during the past three*
generating	6.400	*13*	*Our historical revenue-generating assets completed,*	6.517	*37*	*Cause us to lose revenue-generating opportunities an*
Growth	5.581	*170*	*Enhance revenue growth and better leverage*	5.741	*586*	*Our full fiscal revenue growth was approximately*
total	5.405	*119*	*as a percentage of total revenue were as follows*	5.31	*312*	*As a percentage of total revenue, including revenue*
deferred	5.00	*30*	*Due dates deferred revenue balances increased*	4.94	*65*	*Which included an increase in deferred revenue of approximately*
generated	4.889	*24*	*we pay a percentage of the revenue generated by the machines.*	4.53	*41*	*With little to no revenue generated from operations.*
shortfall	5.81	*3*	*This revenue shortfall would have a*	7.65	*13*	*Unexpected revenue shortfall, which may harm our*
comparisons			*Not Observed*	7.051	*27*	*Year-over-year revenue comparisons were favorably*
Recognition	7.63	*99*	*In our revenue recognition which would*	7.60	*256*	*revenue recognition guidance for arrangement*
unearned			*Not Observed*	6.60	*5*	*Million in unearned revenue due to the early*
grew	3.77	*3*	*Carrier voice revenue grew due to completion o*	6.13	*55*	*No wonder AT&T Wireless grew revenue by 37 percent last year*

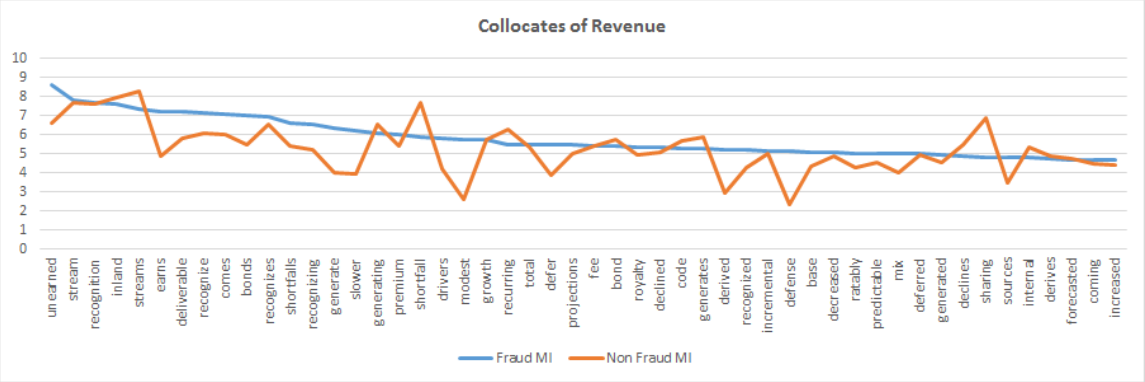


Figure F.13: Collates and Concordances for ‘Revenue’.

Cost

Collocate	MI	*Freq*	*Example of Concordances*	MI	*Freq*	*Example of Concordances*
Prohibitive	8.56	*1*	*Will be cost-prohibitive for stockholders*	**8.70**	*4*	*And cost prohibitive, particularly for cert*
overrruns	9.56	*5*	*experience substantial cost overruns in completing d*	9.33	*38*	*Substantial cost overruns in manufacturing*
unmanaged	8.56	*3*	*Low cost, unmanaged and managed*	5.92	*1*	*Low cost, unmanaged devices to*
Low	7.62	*104*	*Standard is a low-cost wire-replacement*	7.48	*251*	*has developed its own low-cost chip sets*
structure	7.07	*42*	*Our cost structure during periods*	6.67	*130*	*Our cost structure and better align our*
reductions	7.14	*51*	*The cost reductions consisted primarily*	6.96	*149*	*And cost reductions in mergers and*
effective	6.80	*151*	*With a cost-effective and comprehensive*	6.22	*261*	*High quality cost-effective solutions to clients u*
justified	6.75	*1*	*When cost-justified. The Company'*			*Not Observed*
competitiveness	6.62	*3*	*Maintaining cost competitiveness Information*	6.76	*19*	*Overall cost competitiveness of its mainframe*
Discipline(d)	6.31	*2*	*To exercise cost discipline, with careful*	4.22	*1*	*disciplined cost controls and investing*
containment	9.469	*15*	*more efficiencies and cost containment in 2000*	9.33	*69*	*excellence cost containment; and corporate functions,*
consciousness			*Not Observed*	8.51	*1*	*Promote the cost consciousness among both ou*
savings	6.98	*63*	*Such as cost savings and revenue enhancements*	8.13	*301*	
prohibitive	8.56	*1*	*Will be cost-prohibitive for stockholders*	8.71	*4*	*Obtain and cost prohibitive, particularly*
disadvantages			*Not Observed*	7.90	*4*	*These cost disadvantages may adversely affect*
amortized	5.97	*9*	*balance sheet at amortized cost, which approximates*	6.59	*44*	*Recorded at amortized cost or fair value*
efficiencies	5.56	*7*	*Maximize cost efficiencies and expand*	5.90	*26*	*To realize cost efficiencies in the purchasing*
advantages	4.58	*5*	*To create cost advantages through successful*	5.60	*19*	*environmental and cost advantages, shot up in price*
escalation			*Not Observed*	5.46	*2*	*Industry wide cost escalation for energy and fuel*
strict			*Not Observed*	5.70	*5*	*Through strict cost control measures*

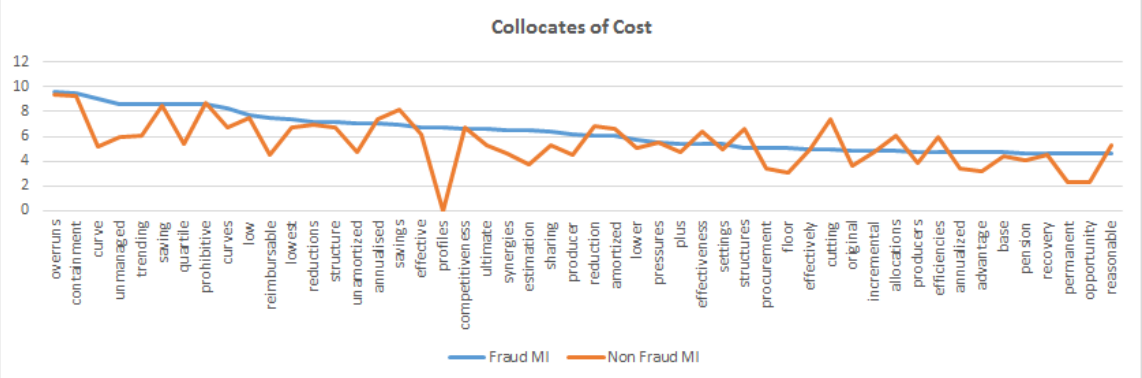


Figure F.14: Collates and Concordances for ‘Cost’.

Company

Fraud			*Non-Fraud*	
Collocate	**MI**	***Example of Concordances***	**MI**	***Example of Concordances***
uncovered	7.303	*The Company uncovered inefficiencies in*		*Not observed*
transacts	7.303	*The Company transacts certain of its business*	6.038	*The Company transacts business*
teamed	7.303	*Company teamed up with a distributor in*		*Not observed*
reevaluates	7.303	*The Company re-evaluates and*		*Not observed*
redefined	7.303	*the Company redefined its core product*		*Not observed*
proceeded	7.303	*the Company proceeded with a strategy*		*Not observed*
prefers	7.303	*The Company prefers a rigorous defense*		*Not observed*
cheque	7.303	*a public blank cheque company and has been*		*Not observed*
bolsters	7.303	*Owning a large finance company bolsters*		*Not observed*
amortizes	7.303	*The Company amortizes acquired technology*	5.676	*The company amortizes the capitalized*
seals	6.888	*relation to company seals and the proposed T*		*Not observed*
consulted	6.718	*The company consulted with the American*		*Not observed*
cautions	6.718	*The Company cautions investors that any*	6.939	*The Company cautions readers that*
learned	6.566	*The Company learned that a*		*Not observed*
undertakes	6.455	*The Company undertakes hedging activities* *The Company undertakes to correct defects*	6.322	*The Company undertakes no obligation* *The Company undertakes no duty*
derives	6.455	*The Company derives revenues from the* *The Company derives from these customers*	4.642	*The Company derives its revenues*
inspects	6.303	*The Company inspects and tests its manufacture*		*Not observed*
discloses	6.303	*The Company discloses the lowest level input*	4.676	*The Company discloses results of operation*
classifies	6.303	*The Company classifies as available*	6.261	*The Company classifies outstanding*
achieves	6.303	*actual results that the Company achieves may*	5.261	*The Company achieves quality control* *The Company achieves greater visibility*
considers	6.239	*The Company considers technological* *The Company considers that such patents* *Company considers its relationship with*	5.654	*The Company considers appropriate* *The Company considers the requirement*
participates	6.133	*The Company participates in a highly dynamic* *The Company participates in advertising*	4.891	*The Company participates in a joint venture*
believes	6.099	*The Company believes this combination* *The Company believes that the*		*The Company believes it can minimize* *The Company believes that the costs*
undertook	6.040	*The Company undertook its fiscal* *The Company undertook a strategic*	5.609	*The company undertook reductions* *The company undertook numerous*
expects	6.014	*The Company expects to implement* *The Company expects its gross profit* *The Company expects to receive fees*	6.359	*The Company expects that financial* *The Company expects to be required*
recognises	5.981	*The Company recognises its responsibility*	5.261	*The Company recognises that the contribution*
possesses	5.981	*The Company possesses a Consent Order*		*Not observed*
notifies	5.981	*The Company notifies the franchising*		*Not observed*
consolidates	5.981	*The Company consolidates the financial results*		*Not observed*
tackled		*Not Observed*	7.261	*The Company tackled a variety*
revisits		*Not Observed*	7.261	*The Company revisits its assessment*
renounced		*Not Observed*	7.261	*The Company renounced its rights*
overestimates		*Not Observed*	7.261	*The Company overestimates the demand*

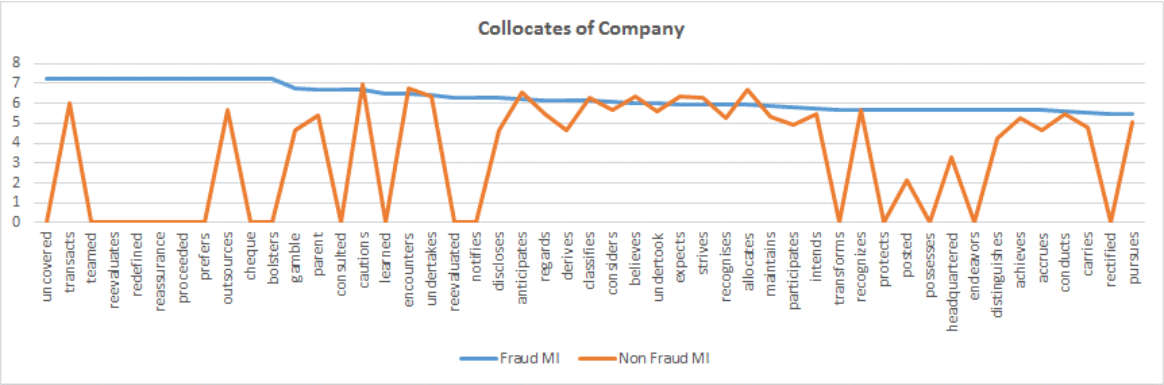


Figure F.15: Collates and Concordances for 'Company'

Significance Testing over Node words

Company	Fraud	Non Fraud
Mean	1.765	1.271
Variance	7.197	8.176
Observations	1102	1778
Hypothesized Mean Difference	0	
df	2446	
t Stat	4.68154235	
P(T<=t) one-tail	1.50139E-06	
t Critical one-tail	1.645476828	
P(T<=t) two-tail	3.00279E-06	
t Critical two-tail	1.960934314	

Business	Fraud	Non Fraud
Mean	2.085	1.866
Variance	5.193	5.2903
Observations	1208	2402
Hypothesized Mean Difference	0	
df	2439	
t Stat	2.725337808	
P(T<=t) one-tail	0.003234727	
t Critical one-tail	1.645478617	
P(T<=t) two-tail	0.006469454	
t Critical two-tail	1.9609371	

Million	Fraud	Non Fraud
Mean	1.253	0.441
Variance	6.169	6.028
Observations	851	1493
Hypothesized Mean Difference	0	
df	1751	
t Stat	7.646589236	
P(T<=t) one-tail	1.69083E-14	
t Critical one-tail	1.645724318	
P(T<=t) two-tail	3.38165E-14	
t Critical two -tail	1.961319715	

Operations	Fraud	Non Fraud
Mean	2.012	1.596
Variance	4.838	5.457
Observations	828	1585
Hypothesized Mean Difference	0	
df	1768	
t Stat	4.317876563	
P(T<=t) one-tail	8.31355E-06	
t Critical one-tail	1.645715942	
P(T<=t) two-tail	1.66271E-05	
t Critical two-tail	1.96130667	

Figure F.16 Significance testing on the mean of MI scores for node words.

Significance Testing over Node words

Sales	*Fraud*	*Non Fraud*
Mean	2.111	1.719
Variance	4.768	5.0669
Observations	969	1758
Hypothesized Mean Difference	0	
df	2047	
t Stat	4.439533584	
P(T<=t) one-tail	4.74757E-06	
t Critical one-tail	1.645598357	
P(T<=t) two-tail	9.49514E-06	
t Critical two-tail	1.96112356	

Cost	*Fraud*	*Non Fraud*
Mean	2.6913	1.8781
Variance	5.9865	5.9069
Observations	531	1030
Hypothesized Mean Difference	0	
df	1064	
t Stat	6.235392888	
P(T<=t) one-tail	3.24422E-10	
t Critical one-tail	1.646286996	
P(T<=t) two-tail	6.48845E-10	
t Critical two-tail	1.962196058	

Services	Fraud	Non Fraud
Mean	2.4251	1.9749
Variance	5.6254	6.0432
Observations	1090	1825
Hypothesized Mean Difference	0	
df	2356	
t Stat	4.890871294	
P(T<=t) one-tail	5.35814E-07	
t Critical one-tail	1.645500644	
P(T<=t) two-tail	1.07163E-06	
t Critical two-tail	1.9609714	

financial	Fraud	Non Fraud
Mean	2.126	1.519
Variance	6.667	7.640
Observations	584	1129
Hypothesized Mean Difference	0	
df	1252	
t Stat	4.502263	
P(T<=t) one-tail	3.67E-06	
t Critical one-tail	1.646072	
P(T<=t) two-tail	7.35E-06	
t Critical two-tail	1.961861	

Figure F.16 Significance testing on the mean of MI scores for node words.

Products	Fraud	Non Fraud
Mean	2.5915	2.1957
Variance	5.3121	6.5310
Observations	1098	2212
Hypothesized Mean Difference	0	
df	2401	
t Stat	4.4841441	
P(T<=t) one-tail	3.83252E-06	
t Critical one-tail	1.645488513	
P(T<=t) two-tail	7.66504E-06	
t Critical two-tail	1.96095	

Revenue	Fraud	Non Fraud
Mean	2.4135	1.9314
Variance	4.8245	5.5445
Observations	599	1056
Hypothesized Mean Difference	0	
df	1315	
t Stat	4.177463476	
P(T<=t) one-tail	1.57107E-05	
t Critical one-tail	1.646013208	
P(T<=t) two-tail	3.14215E-05	
t Critical two-tail	1.961769626	

Figure F.16 Significance testing on the mean of MI scores for node words.

APPENDIX G

Matched Pair: Top 60 stems by greatest difference in tf-Idf scores (summed) in fraud (f) reports compared with stems in non-fraud (nf) reports

	Stems in 'f' reports	Total tf-Idf 'f' reports	Total Tf-Idf 'nf' reports	Difference
1	devic	0.10571945	0	0.10571945
2	client	0.10205208	0	0.10205208
3	fuel	0.08011972	0	0.08011972
4	vehicl	0.10228369	0.03399129	0.0682924
5	fiscal	0.20144456	0.13568874	0.06575582
6	optic	0.06535638	0	0.06535638
7	deal	0.06182158	0	0.06182158
8	hospit	0.05693346	0	0.05693346
9	retail	0.09604561	0.04243907	0.05360654
10	voic	0.05262212	0	0.05262212
11	switch	0.05110227	0	0.05110227
12	stockhold	0.05014303	0	0.05014303
13	food	0.04988683	0	0.04988683
14	china	0.10180196	0.05754793	0.04425403
15	wholesal	0.04419466	0	0.04419466
16	warranti	0.04054094	0	0.04054094
17	specialti	0.03946635	0	0.03946635
18	residenti	0.03927225	0	0.03927225
19	system	0.03835399	0	0.03835399
20	tool	0.03807631	0	0.03807631
21	reimburs	0.03746532	0	0.03746532
22	royalti	0.03732733	0	0.03732733
23	document	0.03598047	0	0.03598047
24	occup	0.03509495	0	0.03509495
25	websit	0.03497561	0	0.03497561
26	procedur	0.03419688	0	0.03419688
27	link	0.03416518	0	0.03416518
28	carrier	0.08282036	0.05027803	0.03254233
29	telecommun	0.08397591	0.05230843	0.03166748
30	medic	0.09941493	0.06875234	0.03066259
31	loan	0.0921735	0.06385731	0.02831619
32	energi	0.07765938	0.05096052	0.02669886
33	percent	0.11667412	0.09036769	0.02630643
34	divis	0.07825816	0.052072	0.02618616
35	enterpris	0.0521168	0.03235899	0.01975781
36	solut	0.09247053	0.07302385	0.01944668
37	environment	0.05340908	0.03762084	0.01578824
38	manufactur	0.11592739	0.10393374	0.01199365
39	decemb	0.04770741	0.03616494	0.01154247
40	publish	0.06165727	0.05031125	0.01134602
41	center	0.04717133	0.03630083	0.0108705
43	channel	0.05084303	0.04011509	0.01072794
44	communic	0.05428149	0.04368988	0.01059161
45	merger	0.04529034	0.03596125	0.00932909
46	claim	0.04442223	0.03550325	0.00891898
47	digit	0.09512914	0.08635222	0.00877692
48	raw	0.04645117	0.03771192	0.00873925
49	imag	0.10347502	0.09525592	0.0082191
50	trademark	0.04388392	0.03622218	0.00766174
51	billion	0.13410229	0.12678599	0.0073163
52	research	0.04535514	0.03815322	0.00720192
53	sharehold	0.04455997	0.03771069	0.00684928
54	construct	0.05974625	0.05404985	0.0056964
55	network	0.06085778	0.05633255	0.00452523
56	outsourc	0.04468736	0.04100987	0.00367749
57	fee	0.03420919	0.03062013	0.00358906
58	brand	0.05246696	0.04897183	0.00349513
59	repurchas	0.03580121	0.03321732	0.00258389
60	goodwil	0.04108503	0.03932869	0.00175634

Table G.1: tf-idf scores for most prominent stems in 'f' reports with corresponding stem in 'nf' reports.

Matched Pair: Top 60 stems by greatest difference in Tf-Idf scores (summed) in non-fraud (nf) reports compared with stems in fraud (f) reports

	Stems in 'nf' reports	Total tf-Idf 'nf' reports	Total Tf-Idf 'f' reports	Difference
1	cabl	0.15077548	0	0.15077548
2	wireless	0.0929783	0	0.0929783
3	store	0.13085247	0.0560792	0.07477327
4	leas	0.06947488	0	0.06947488
5	assembl	0.10021699	0.03098164	0.06923535
6	wast	0.06772747	0	0.06772747
7	communiti	0.06234306	0	0.06234306
8	expens	0.05345611	0	0.05345611
9	decreas	0.05322742	0	0.05322742
10	south	0.0526138	0	0.0526138
11	chemic	0.05253087	0	0.05253087
12	product	0.04937928	0	0.04937928
13	hard	0.04873959	0	0.04873959
14	may	0.04803749	0	0.04803749
15	solid	0.04761886	0	0.04761886
16	franc	0.04672747	0	0.04672747
17	foreign	0.04647612	0	0.04647612
18	drive	0.04562224	0	0.04562224
19	week	0.04497014	0	0.04497014
20	award	0.04408618	0	0.04408618
21	euro	0.04358548	0	0.04358548
22	shipment	0.04318433	0	0.04318433
23	travel	0.04313244	0	0.04313244
24	propos	0.04285249	0	0.04285249
25	estim	0.04274031	0	0.04274031
26	rose	0.04248418	0	0.04248418
27	launch	0.04171794	0	0.04171794
28	affili	0.04008933	0	0.04008933
29	primarili	0.03954743	0	0.03954743
30	revenu	0.03931838	0	0.03931838
31	explor	0.03882063	0	0.03882063
32	televis	0.10894325	0.07079475	0.0381485
33	storag	0.06708623	0.02915954	0.03792669
34	million	0.09529364	0.06159713	0.03369651
35	company'	0.1557651	0.12729291	0.02847219
36	commod	0.07116521	0.04329829	0.02786692
37	print	0.07976996	0.05261669	0.02715327
38	home	0.07551101	0.05055429	0.02495672
39	softwar	0.10110049	0.07731477	0.02378572
40	ordinari	0.05562832	0.03338776	0.02224056
41	video	0.06233066	0.04026289	0.02206777
43	insur	0.06929475	0.04737336	0.02192139
44	segment	0.06612925	0.04462123	0.02150802
45	inventori	0.0735101	0.05243677	0.02107333
46	portfolio	0.04762091	0.02772011	0.0199008
47	group	0.05120245	0.03130343	0.01989902
48	impair	0.04386717	0.02571111	0.01815606
49	page	0.05735848	0.03936613	0.01799235
50	licens	0.05857893	0.04248332	0.01609561
51	air	0.05678988	0.04239465	0.01439523
52	restructur	0.05545096	0.04148029	0.01397067
53	gross	0.04427801	0.03038908	0.01388893
54	premium	0.04919909	0.03532201	0.01387708
55	dividend	0.04150993	0.02774273	0.0137672
56	reserv	0.03976878	0.02653349	0.01323529
57	health	0.05020451	0.0369799	0.01322461
58	telephon	0.04625594	0.03315529	0.01310065
59	ventur	0.0441226	0.0314009	0.0127217
60	field	0.04531608	0.03480748	0.0105086

Table G.2: tf-idf scores for most prominent stems in 'nf' reports with corresponding stem in 'f' reports.

Peer Set: Top 60 stems by greatest difference in tf-Idf scores (averaged) in fraud (f) reports compared with stems in non-fraud (nf) reports

	Stems in 'nf' reports	Total tf-Idf 'f' reports	Total Tf-Idf nf' reports	Difference
1	client	0.001000511	0	0.001000511
2	optic	0.000640749	0	0.000640749
3	deal	0.000606094	0	0.000606094
4	hospit	0.000558171	0	0.000558171
5	voic	0.000515903	0	0.000515903
6	devic	0.001036465	0.000535075	0.00050139
7	stockhold	0.000491598	0	0.000491598
8	food	0.000489087	0	0.000489087
9	commod	0.000424493	0	0.000424493
10	telecommun	0.000823293	0.000417015	0.000406278
11	fuel	0.000785487	0.000384821	0.000400667
12	video	0.000394734	0	0.000394734
13	residenti	0.000385022	0	0.000385022
14	system	0.00037602	0	0.00037602
15	divis	0.000767237	0.000397899	0.000369338
16	reimburs	0.000367307	0	0.000367307
17	document	0.00035275	0	0.00035275
18	occup	0.000344068	0	0.000344068
19	code	0.000335605	0	0.000335605
20	procedur	0.000335264	0	0.000335264
21	link	0.000334953	0	0.000334953
22	china	0.000998058	0.000672199	0.00032586
23	billion	0.001314728	0.00099378	0.000320948
24	vehicl	0.001002781	0.000692945	0.000309837
25	energi	0.000761366	0.000460941	0.000300426
26	carrier	0.000811964	0.000521595	0.000290369
27	loan	0.000903662	0.000623428	0.000280234
28	retail	0.000941624	0.00070442	0.000237203
29	imag	0.001014461	0.000790969	0.000223492
30	switch	0.000501003	0.000292322	0.000208681
31	solut	0.000906574	0.000704523	0.000202051
32	enterpris	0.000510949	0.000326497	0.000184452
33	publish	0.000604483	0.000448635	0.000155848
34	fiscal	0.001974947	0.001825555	0.000149392
35	center	0.000462464	0.000324731	0.000137733
36	environment	0.000523618	0.000386959	0.000136659
37	patent	0.000825918	0.00069426	0.000131659
38	construct	0.000585748	0.000470892	0.000114855
39	digit	0.000932639	0.000826179	0.000106459
40	technolog	0.000439417	0.000335152	0.000104265
41	warranti	0.00039746	0.000293703	0.000103757
43	goodwil	0.000402794	0.000302222	0.000100572
44	medic	0.000974656	0.000876623	9.80336E-05
45	tool	0.000373297	0.0002772	9.60971E-05
46	royalti	0.000365954	0.000271438	9.45167E-05
47	sharehold	0.000436862	0.000351968	8.48944E-05
48	network	0.000596645	0.00051235	8.42945E-05
49	manufactur	0.001136543	0.001054159	8.23843E-05
50	outsourc	0.000438111	0.000358646	7.94654E-05
51	wholesal	0.000433281	0.000354216	7.90652E-05
52	channel	0.000498461	0.000426632	7.18294E-05
53	raw	0.000455404	0.0003914	6.40039E-05
54	specialti	0.000386925	0.000332629	5.42962E-05
55	claim	0.000435512	0.000385906	4.96061E-05
56	merger	0.000444023	0.000397541	4.64821E-05
57	trademark	0.000430235	0.000404722	2.55124E-05
58	transport	0.000382595	0.000357709	2.48853E-05
59	bank	0.000407668	0.000384595	2.30735E-05
60	air	0.000415634	0.000400073	1.55606E-05

Table G.3: Most prominent stems in 'f' reports with corresponding stem in 'nf' reports (avg. for PS).

Peer Set: Top 60 stems by greatest difference in tf-Idf scores (averaged) in non-fraud (nf) reports compared with stems in fraud (f) reports

	Stems in 'nf' reports	Total tf-Idf 'nf' reports	Total Tf-Idf 'f' reports	Difference
1	wireless	0.001092132	0	0.001092132
2	communiti	0.000727655	0	0.000727655
3	display	0.000661381	0	0.000661381
4	pension	0.00057437	0	0.00057437
5	company'	0.001813805	0.00124797	0.000565835
6	store	0.001079832	0.000549796	0.000530036
7	explor	0.000499361	0	0.000499361
8	foreign	0.000494257	0	0.000494257
9	cooper	0.000481643	0	0.000481643
10	revenu	0.000477756	0	0.000477756
11	card	0.000474286	0	0.000474286
12	leas	0.000468271	0	0.000468271
13	care	0.000810821	0.000356971	0.00045385
14	chemic	0.000430662	0	0.000430662
15	launch	0.000425313	0	0.000425313
16	supplier	0.000422245	0	0.000422245
17	allianc	0.000417964	0	0.000417964
18	home	0.000899095	0.00049563	0.000403465
19	transmiss	0.000399633	0	0.000399633
20	south	0.000385471	0	0.000385471
21	expens	0.000378333	0	0.000378333
22	assembl	0.000619876	0.000303742	0.000316134
23	insur	0.000725942	0.000464445	0.000261497
24	onlin	0.000624168	0.000364344	0.000259824
25	reserv	0.000515857	0.000260132	0.000255725
26	treatment	0.000504409	0.000267256	0.000237153
27	telephon	0.000557897	0.000325052	0.000232845
28	comput	0.000477248	0.000249084	0.000228164
29	segment	0.000660213	0.000437463	0.00022275
30	storag	0.00050512	0.000285878	0.000219242
31	engin	0.000543312	0.000329996	0.000213315
32	impair	0.000439659	0.00025207	0.000187589
33	million	0.000777136	0.000603893	0.000173243
34	distributor	0.000506467	0.000342863	0.000163605
35	softwar	0.000915543	0.000757988	0.000157555
36	intellectu	0.000462028	0.000318044	0.000143983
37	gross	0.000440605	0.000297932	0.000142673
38	brand	0.00063938	0.000514382	0.000124998
39	currenc	0.000576693	0.000455685	0.000121007
40	licens	0.000536907	0.000416503	0.000120404
41	global	0.000456716	0.000337363	0.000119353
43	book	0.000401473	0.000289851	0.000111622
44	field	0.000449997	0.00034125	0.000108747
45	health	0.00046653	0.000362548	0.000103982
46	live	0.000439782	0.000337854	0.000101928
47	premium	0.000447657	0.000346294	0.000101363
48	electr	0.000444014	0.000354502	8.95122E-05
49	ventur	0.000394063	0.000307852	8.62108E-05
50	backlog	0.000411511	0.000331657	7.98534E-05
51	internet	0.000595127	0.000523624	7.1503E-05
52	percent	0.001213817	0.001143864	6.99534E-05
53	europ	0.000468786	0.000403873	6.49129E-05
54	televis	0.000749943	0.000694066	5.5877E-05
55	restructur	0.000462446	0.00040667	5.5777E-05
56	decemb	0.00052068	0.00046772	5.29605E-05
57	regul	0.000509258	0.00046323	4.60272E-05
58	hardwar	0.000373763	0.000329876	4.38872E-05
59	fee	0.000376729	0.000335384	4.13443E-05
60	communic	0.000570509	0.000532171	3.83375E-05

Table G.4: Most prominent stems in 'nf' reports with corresponding stem in 'f' reports (avg. for PS).

Matched Pair (MP): Top 60 bigrams by greatest difference in tf-Idf scores (summed) in ‘f’ reports compared with corresponding bigrams in ‘nf’ reports

	Bigrams	Tf-Idf score 'f' reports (summed)	Tf-Idf score 'nf' reports (summed)	Difference
1	the companys	0.26920602	0.15788694	0.11131908
2	in fiscal	0.15437456	0.05798495	0.09638961
3	ended	0.06990432	0.04138838	0.02851594
4	or of	0.02513291	0	0.02513291
5	its products	0.02505703	0	0.02505703
6	for fiscal	0.05676202	0.03185406	0.02490796
7	preferred stock	0.02363506	0	0.02363506
8	fiscal year	0.05307819	0.02950745	0.02357074
9	to for	0.02287157	0	0.02287157
10	total revenues	0.02163292	0	0.02163292
11	we may	0.05451014	0.03316604	0.0213441
12	may not	0.02110074	0	0.02110074
13	revenues for	0.02077227	0	0.02077227
14	to us	0.02063725	0	0.02063725
15	income before	0.02024862	0	0.02024862
16	our revenues	0.02023257	0	0.02023257
17	our company	0.01974245	0	0.01974245
18	years ended	0.01953352	0	0.01953352
19	raw materials	0.01933823	0	0.01933823
20	year ended	0.03957503	0.02042666	0.01914837
21	a business	0.01911813	0	0.01911813
22	its customers	0.01888301	0	0.01888301
23	gross margin	0.01883883	0	0.01883883
24	accounts	0.01876468	0	0.01876468
25	may be	0.01862782	0	0.01862782
26	our common	0.03806802	0.02098387	0.01708415
27	during fiscal	0.03460379	0.01898402	0.01561977
28	our products	0.04859054	0.03414481	0.01444573
29	the company	0.08351705	0.07041034	0.01310671
30	common stock	0.03295773	0.02018166	0.01277607
31	if we	0.03557546	0.02364682	0.01192864
32	the merger	0.02719579	0.0158704	0.01132539
33	we are	0.04486292	0.03378286	0.01108006
34	we have	0.04818549	0.03822722	0.00995827
35	we will	0.02800981	0.02005468	0.00795513
36	the fiscal	0.02832004	0.0204285	0.00789154
37	our ability	0.03404345	0.02620109	0.00784236
38	us to	0.0263767	0.01886421	0.00751249
39	company	0.03047248	0.02309757	0.00737491
40	credit facility	0.02883274	0.02191646	0.00691628
41	that we	0.03107877	0.0242964	0.00678237
43	believes that	0.0286001	0.02231605	0.00628405
44	north america	0.02240713	0.01633588	0.00607125
45	company also	0.02220928	0.01628171	0.00592757
46	our business	0.04060909	0.0348118	0.00579729
47	of revenues	0.02185608	0.01606407	0.00579201
48	adversely affect	0.02133027	0.01574572	0.00558455
49	intellectual	0.02448551	0.0193394	0.00514611
50	which we	0.02114953	0.01667698	0.00447255
51	affect our	0.02888853	0.02460273	0.0042858
52	of our	0.09628796	0.09230711	0.00398085
53	we do	0.02211373	0.0184657	0.00364803
54	to our	0.03247811	0.0290265	0.00345161
55	of fiscal	0.03157896	0.02815014	0.00342882
56	we believe	0.03053446	0.02722426	0.0033102
57	and we	0.01904268	0.01575974	0.00328294
58	and our	0.02116487	0.01796535	0.00319952
59	million for	0.02022032	0.01705569	0.00316463
60	december and	0.01998516	0.01694224	0.00304292

Table G.5: Most prominent bigrams in ‘f’ reports with corresponding bigram in ‘nf’ reports for (MP).

Matched Pair (MP): Top 60 bigrams by greatest difference in tf-idf scores (summed) in 'nf' reports compared with corresponding bigrams in 'f' reports

	Bigrams	Tf-Idf score 'nf' reports (summed)	Tf-Idf score 'f' reports (summed)	Difference
1	operating income	0.04668362	0	0.04668362
2	in from	0.04039078	0	0.04039078
3	in compared	0.03647118	0	0.03647118
4	or per	0.03389952	0	0.03389952
5	continuing	0.0317575	0	0.0317575
6	the company's	0.10949466	0.07796875	0.03152591
7	health care	0.02933305	0	0.02933305
8	the notes	0.02859745	0	0.02859745
9	compared with	0.05399808	0.02668797	0.02731011
10	shares in	0.026841	0	0.026841
11	increased by	0.02667788	0	0.02667788
12	of billion	0.02585372	0	0.02585372
13	a percent	0.02543926	0	0.02543926
14	primarily due	0.02472482	0	0.02472482
15	notes to	0.02461461	0	0.02461461
16	to consolidated	0.02444606	0	0.02444606
17	approximately	0.02341146	0	0.02341146
18	latin america	0.0229722	0	0.0229722
19	to billion	0.02250064	0	0.02250064
20	real estate	0.02225821	0	0.02225821
21	revenue growth	0.02217364	0	0.02217364
22	see note	0.02210685	0	0.02210685
23	by million	0.021819	0	0.021819
24	from in	0.0217452	0	0.0217452
25	this was	0.02141295	0	0.02141295
26	in was	0.02100834	0	0.02100834
27	united kingdom	0.02099677	0	0.02099677
28	million at	0.02056334	0	0.02056334
29	in we	0.02027967	0	0.02027967
30	million compared	0.02006517	0	0.02006517
31	percent of	0.03322951	0.01579095	0.01743856
32	billion in	0.03935853	0.02192287	0.01743566
33	at december	0.03783919	0.02322388	0.01461531
34	million in	0.0422534	0.02795533	0.01429807
35	to million	0.03224134	0.02054101	0.01170033
36	from million	0.02883059	0.01962633	0.00920426
37	million of	0.02879198	0.01965047	0.00914151
38	million or	0.04285469	0.03455762	0.00829707
39	partially offset	0.0225553	0.01431725	0.00823805
40	increased million	0.02848806	0.02081004	0.00767802
41	operating	0.02132477	0.01426159	0.00706318
43	company had	0.02096279	0.01439236	0.00657043
44	fair value	0.02217414	0.0157116	0.00646254
45	and million	0.02419018	0.01843117	0.00575901
46	our customers	0.03081409	0.02524903	0.00556506
47	we also	0.02282897	0.01736369	0.00546528
48	of million	0.04509887	0.03981205	0.00528682
49	joint venture	0.02250015	0.01788409	0.00461606
50	was million	0.02008821	0.01576733	0.00432088
51	intangible assets	0.02090273	0.01684948	0.00405325
52	the internet	0.02040979	0.01639237	0.00401742
53	of revenue	0.02333859	0.01933173	0.00400686
54	foreign currency	0.02162029	0.01768813	0.00393216
55	research and	0.02972025	0.0260999	0.00362035
56	net sales	0.08076056	0.07751267	0.00324789
57	and development	0.02115419	0.01806311	0.00309108
58	our operating	0.02026717	0.01737853	0.00288864
59	of december	0.03007105	0.02742609	0.00264496
60	new products	0.0204415	0.01806364	0.00237786

Table G.6: Most prominent bigrams in 'nf' reports with corresponding bigram in 'f' reports for (MP).

Peer Set: Top 60 bigrams by greatest difference in tf-Idf scores (averaged) in ‘f’ reports compared with bigrams in ‘nf’ reports

	Bigrams	Tf-Idf score 'f' reports (avg)	Tf-Idf score 'nf' reports (avg)	Difference
1	the companys	0.003248121	0.002164575	0.001083546
2	in fiscal	0.001718545	0.001042306	0.000676239
3	fiscal year	0.00065665	0.00032042	0.00033623
4	for fiscal	0.000627958	0.000325853	0.000302105
5	or of	0.000283477	0	0.000283477
6	total revenues	0.000245985	0	0.000245985
7	to for	0.000222738	0	0.000222738
8	the state	0.000214898	0	0.000214898
9	our revenue	0.000212102	0	0.000212102
10	to us	0.000210296	0	0.000210296
11	may not	0.000200549	0	0.000200549
12	revenues for	0.000196907	0	0.000196907
13	income before	0.000195775	0	0.000195775
14	our company	0.000192264	0	0.000192264
15	shares of	0.00019125	0	0.00019125
16	our revenues	0.000188058	0	0.000188058
17	accounts	0.000187972	0	0.000187972
18	year ended	0.000488616	0.000301003	0.000187614
19	our sales	0.000186436	0	0.000186436
20	years ended	0.000186065	0	0.000186065
21	our common	0.000401976	0.000218264	0.000183712
22	a business	0.000181284	0	0.000181284
23	we may	0.000549962	0.00038806	0.000161902
24	the fiscal	0.000328306	0.000175535	0.000152772
25	ended december	0.00063151	0.000489013	0.000142497
26	common stock	0.000334752	0.000220669	0.000114083
27	preferred stock	0.000254383	0.000146209	0.000108174
28	company has	0.0004217	0.000313568	0.000108132
29	if we	0.000327246	0.00022618	0.000101066
30	our products	0.00048359	0.000404563	7.90274E-05
31	credit facility	0.000293316	0.000218694	7.46214E-05
32	the company	0.0007249	0.000657153	6.77471E-05
33	its products	0.000214968	0.000158148	5.68207E-05
34	our ability	0.000321067	0.000265109	5.59571E-05
35	selling general	0.000231371	0.000175773	5.55984E-05
36	north america	0.000237732	0.000184195	5.35368E-05
37	we believe	0.000302107	0.000252386	4.9721E-05
38	the merger	0.000283822	0.000235547	4.82753E-05
39	affect our	0.000279179	0.000231988	4.71905E-05
40	that we	0.00028517	0.000238033	4.71364E-05
41	adversely affect	0.000221906	0.000175062	4.68445E-05
43	we are	0.000421943	0.0003762	4.57427E-05
44	in north	0.000187935	0.000143207	4.47287E-05
45	we do	0.00021028	0.000168068	4.22114E-05
46	revenues in	0.000191295	0.000150129	4.11662E-05
47	of revenues	0.000227545	0.000189434	3.81113E-05
48	which we	0.000211007	0.000175482	3.55245E-05
49	of fiscal	0.00035729	0.000323824	3.34653E-05
50	sales for	0.000217889	0.000184623	3.32661E-05
51	net sales	0.000739444	0.000706343	3.31012E-05
52	sfas no	0.000184799	0.000152273	3.25252E-05
53	us to	0.000244202	0.000212519	3.16834E-05
54	we have	0.000462494	0.000431253	3.12407E-05
55	intangible assets	0.000194536	0.000166173	2.83628E-05
56	believes that	0.000272388	0.000244303	2.80847E-05
57	the previous	0.000230475	0.000202487	2.79879E-05
58	of revenue	0.000213141	0.000191448	2.16927E-05
59	of net	0.00019647	0.000178943	1.75268E-05
60	services and	0.000191502	0.000175769	1.57327E-05

Table G.7: Most prominent bigrams in ‘f’ reports with corresponding bigram in ‘nf’ reports for (PS).

Peer Set: Top 60 bigrams by greatest difference in tf-Idf scores (averaged) in 'nf' reports compared with bigrams in 'f' reports

	Bigrams	Tf-Idf score 'nf' reports (avg)	Tf-Idf score 'f' reports (avg)	Difference
1	the company's	0.001267829	0.000733505	0.00053432
2	health care	0.000462471	0	0.00046247
3	operating income	0.000362355	0	0.00036235
4	in compared	0.00032987	0	0.00032987
5	to billion	0.000285817	0	0.00028582
6	in from	0.000248597	0	0.0002486
7	of billion	0.000242965	0	0.00024296
8	continuing	0.000229548	0	0.00022955
9	increased by	0.000219252	0	0.00021925
10	million increase	0.000218603	0	0.0002186
11	sales growth	0.000217772	0	0.00021777
12	the notes	0.000217442	0	0.00021744
13	primarily due	0.000211467	0	0.00021147
14	a million	0.000201964	0	0.00020196
15	by million	0.000199542	0	0.00019954
16	note of	0.000198582	0	0.00019858
17	latin america	0.000196387	0	0.00019639
18	interest in	0.000193566	0	0.00019357
19	from in	0.00019152	0	0.00019152
20	sales increased	0.000190112	0	0.00019011
21	million million	0.000190099	0	0.0001901
22	to consolidated	0.000188388	0	0.00018839
23	partially offset	0.000187955	0	0.00018796
24	in we	0.000187588	0	0.00018759
25	earnings per	0.000185264	0	0.00018526
26	at december	0.000375748	0.000204053	0.00017169
27	compared with	0.000409701	0.000271233	0.00013847
28	percent of	0.000272246	0.000153785	0.00011846
29	billion in	0.000288747	0.0001864	0.00010235
30	of december	0.000312242	0.000229429	8.2813E-05
31	million in	0.000342279	0.000260124	8.2155E-05
32	joint venture	0.000238297	0.000156511	8.1785E-05
33	million of	0.000252557	0.000180774	7.1783E-05
34	gross margin	0.000250224	0.000183457	6.6767E-05
35	and million	0.00024025	0.000174362	6.5888E-05
36	of sales	0.000215584	0.000150663	6.4921E-05
37	of million	0.000425706	0.000368637	5.7069E-05
38	compared to	0.000255544	0.000198657	5.6887E-05
39	fair value	0.0002209	0.000164048	5.6852E-05
40	to million	0.000260633	0.000203835	5.6797E-05
41	we also	0.000217152	0.000167415	4.9737E-05
43	increased million	0.000235286	0.000187076	4.821E-05
44	from million	0.00023923	0.000192415	4.6815E-05
45	our customers	0.000278193	0.00023994	3.8254E-05
46	company believes	0.000330978	0.000294795	3.6184E-05
47	research and	0.000310458	0.000274313	3.6145E-05
48	in our	0.00034033	0.000306276	3.4054E-05
49	was million	0.000195508	0.00016198	3.3528E-05
50	and development	0.000230614	0.00020153	2.9084E-05
51	foreign currency	0.000209621	0.000180752	2.8869E-05
52	million or	0.000388002	0.000364605	2.3397E-05
53	the internet	0.000187035	0.00016517	2.1865E-05
54	company is	0.000215459	0.000195231	2.0228E-05
55	approximately	0.000272161	0.000253402	1.8759E-05
56	north american	0.0002231	0.000205357	1.7743E-05
57	as compared	0.000220337	0.000203312	1.7025E-05
58	on our	0.00036372	0.000347823	1.5897E-05
59	with our	0.000201309	0.000186139	1.517E-05
60	intellectual property	0.000264034	0.000251601	1.2433E-05

Table G.8: Most prominent bigrams in 'nf' reports with corresponding bigram in 'f' reports for (PS).

Matched Pair: Top 60 trigrams by greatest difference in tf-Idf scores (summed) in fraud reports compared with trigrams in non-fraud reports

	Trigrams	Tf-Idf score 'f' reports (summed)	Tf-Idf score 'nf' reports (summed)	Difference
1	to million in	0.01807225	0.03235652	-0.01428427
2	from million in	0.01380522	0.02551922	-0.011714
3	during the year	0.02172313	0.0313362	-0.00961307
4	in the company	0.01892631	0.02795737	-0.00903106
5	partially offset by	0.01431759	0.02250746	-0.00818987
6	the company had	0.0145302	0.02116857	-0.00663837
7	compared to million	0.01713201	0.02335586	-0.00622385
8	note to the	0.01609239	0.02022641	-0.00413402
9	research and	0.02958349	0.03334301	-0.00375952
10	of the company's	0.02989858	0.03320327	-0.00330469
11	revolving credit facility	0.01523948	0.01842736	-0.00318788
12	of million in	0.01425331	0.01721558	-0.00296227
13	results of operations	0.01509584	0.01776241	-0.00266657
14	as of december	0.02906583	0.03160208	-0.00253625
15	the company is	0.02563359	0.02654495	-0.00091136
16	our operating results	0.01346688	0.01426827	-0.00080139
17	could have a	0.01372059	0.01425769	-0.0005371
18	foreign currency	0.01348211	0.01369379	-0.00021168
19	effect on our	0.02096432	0.02115031	-0.00018599
20	and administrative	0.01585984	0.01581951	4.033E-05
21	cost of sales	0.01542456	0.01533142	9.314E-05
22	selling general and	0.0202206	0.01976273	0.00045787
23	that we will	0.01472083	0.01398272	0.00073811
24	the company to	0.01515216	0.01431803	0.00083413
25	products and services	0.01633713	0.01532626	0.00101087
26	for our products	0.01485901	0.01352491	0.0013341
27	the united states	0.01399988	0.01261173	0.00138815
28	the company expects	0.01391118	0.01248348	0.0014277
29	the fiscal year	0.02607164	0.02435679	0.00171485
30	adverse effect on	0.01553739	0.01357016	0.00196723
31	on the companys	0.02079395	0.01876806	0.00202589
32	general and	0.01523208	0.01318893	0.00204315
33	to the consolidated	0.01476313	0.01270628	0.00205685
34	million for the	0.01699607	0.0149248	0.00207127
35	are unable to	0.01435096	0.01204001	0.00231095
36	the financial	0.02423276	0.02190211	0.00233065
37	price of our	0.01493262	0.01255624	0.00237638
38	million and million	0.01989379	0.01736946	0.00252433
39	for the fiscal	0.01436241	0.0117969	0.00256551
40	the consolidated	0.01450072	0.01180414	0.00269658
41	we do not	0.01961712	0.0167579	0.00285922
43	of approximately	0.01487869	0.01200793	0.00287076
44	the company will	0.0184646	0.0155041	0.0029605
45	in which we	0.01438656	0.01138531	0.00300125
46	the company may	0.01497391	0.01170784	0.00326607
47	charge of million	0.01462649	0.01133881	0.00328768
48	and the company	0.01455737	0.0109379	0.00361947
49	to the company	0.01492419	0.01086029	0.0040639
50	december the	0.01926607	0.01498553	0.00428054
51	deferred tax assets	0.0152499	0.01090886	0.00434104
52	we believe that	0.02415392	0.01980059	0.00435333
53	as compared to	0.02039294	0.01597178	0.00442116
54	of the company	0.03084641	0.02628352	0.00456289
55	adversely affect our	0.02274149	0.01814673	0.00459476
56	that the company	0.01694867	0.01232162	0.00462705
57	the prior year	0.02181559	0.0166906	0.00512499
58	in north america	0.01773394	0.01209505	0.00563889
59	for the year	0.0233871	0.01753527	0.00585183
60	years ended december	0.01954028	0.01337235	0.00616793

Table G.9: Most prominent trigrams in 'f' reports with corresponding trigram in 'nf' reports (MP).

Matched Pair: Top 60 trigrams by greatest difference in tf-Idf scores (summed) in non-fraud reports compared with trigrams in fraud reports

	Trigrams	Tf-Idf score 'nf' reports (summed)	Tf-Idf score 'f' reports (summed)	Difference
1	in compared to	0.03082811	0	0.03082811
2	million in and	0.02534002	0	0.02534002
3	in and million	0.02447869	0	0.02447869
4	the notes to	0.0202322	0	0.0202322
5	million to million	0.02001287	0	0.02001287
6	at the end	0.01863603	0	0.01863603
7	end of the	0.01843567	0	0.01843567
8	of the year	0.02925679	0.01082255	0.01843424
9	was million in	0.01814285	0	0.01814285
10	the year and	0.01740834	0	0.01740834
11	report on form	0.01709927	0	0.01709927
12	expenses as a	0.01674503	0	0.01674503
13	in the us	0.01669696	0	0.01669696
14	due to higher	0.01661511	0	0.01661511
15	annual report on	0.01648455	0	0.01648455
16	in the fourth	0.01636731	0	0.01636731
17	the second half	0.01620866	0	0.01620866
18	of this report	0.01617484	0	0.01617484
19	in and in	0.01599854	0	0.01599854
20	million compared to	0.01592771	0	0.01592771
21	our business financial	0.01577839	0	0.01577839
22	this annual report	0.01575976	0	0.01575976
23	of the board	0.01547253	0	0.01547253
24	was primarily due	0.01546102	0	0.01546102
25	on form k	0.01532606	0	0.01532606
26	the increase was	0.01518592	0	0.01518592
27	the first half	0.01486857	0	0.01486857
28	to consolidated	0.02765893	0.01283139	0.01482754
29	was due to	0.01475306	0	0.01475306
30	increase of million	0.01469861	0	0.01469861
31	a decrease of	0.01463591	0	0.01463591
32	a increase in	0.01461792	0	0.01461792
33	half of the	0.01443845	0	0.01443845
34	the decline in	0.01439134	0	0.01439134
35	to million in	0.03235652	0.01807225	0.01428427
36	an increase of	0.02469892	0.01084511	0.01385381
37	primarily due to	0.02509335	0.01178719	0.01330616
38	from million in	0.02551922	0.01380522	0.011714
39	of million or	0.02236606	0.01232629	0.01003977
40	during the year	0.0313362	0.02172313	0.00961307
41	earnings per share	0.01987613	0.0108294	0.00904673
43	in the company	0.02795737	0.01892631	0.00903106
44	due primarily to	0.01929785	0.01099789	0.00829996
45	partially offset by	0.02250746	0.01431759	0.00818987
46	the fourth quarter	0.01959649	0.01190408	0.00769241
47	million in the	0.02013903	0.01325606	0.00688297
48	the company had	0.02116857	0.0145302	0.00663837
49	fourth quarter of	0.01778368	0.01119971	0.00658397
50	the united kingdom	0.01804855	0.01155169	0.00649686
51	compared to million	0.02335586	0.01713201	0.00622385
52	increased to million	0.01822774	0.01205432	0.00617342
53	fiscal year ended	0.01803165	0.01297604	0.00505561
54	note to the	0.02022641	0.01609239	0.00413402
55	research and	0.03334301	0.02958349	0.00375952
56	and million in	0.01670183	0.01317526	0.00352657
57	as a percentage	0.01606974	0.01254839	0.00352135
58	of the company's	0.03320327	0.02989858	0.00330469
59	revolving credit facility	0.01842736	0.01523948	0.00318788
60	million shares of	0.0143496	0.01122998	0.00311962

Table G.10: Most prominent trigrams in 'nf' reports with corresponding trigram in 'f' reports (MP).

Peer set (PS): Top 60 trigrams by greatest difference in tf-Idf scores (averaged) in fraud reports compared with trigrams in non-fraud reports

	Trigrams	Tf-Idf score 'f' reports (avg)	Tf-Idf score 'nf' reports (avg)	Difference
1	of the companys	0.001049713	0.000726807	0.000322905
2	the year ended	0.000629867	0.000357624	0.000272242
3	our common stock	0.000421015	0.000219748	0.000201268
4	during the period	0.000179811	0	0.000179811
5	management believes	0.000179494	0	0.000179494
6	million in cash	0.000170391	0	0.000170391
7	the company in	0.000164031	0	0.000164031
8	the acquisition of	0.000162386	0	0.000162386
9	shares of common	0.000154663	0	0.000154663
10	year ended december	0.000588672	0.000438658	0.000150014
11	of the acquisition	0.000149983	0	0.000149983
12	the market price	0.000147827	0	0.000147827
13	the financial	0.000280689	0.000135163	0.000145526
14	market price of	0.000144938	0	0.000144938
15	connection with the	0.000142918	0	0.000142918
16	in this report	0.000140537	0	0.000140537
17	the company the	0.000139951	0	0.000139951
18	shares of our	0.000139592	0	0.000139592
19	the result of	0.000139191	0	0.000139191
20	we entered into	0.000138669	0	0.000138669
21	of common stock	0.000136433	0	0.000136433
22	of our common	0.000317346	0.000184472	0.000132873
23	the fiscal year	0.000292612	0.000164276	0.000128337
24	the company has	0.000494314	0.00037951	0.000114804
25	of the company	0.000339337	0.000224813	0.000114524
26	for the year	0.000294037	0.000199814	9.42239E-05
27	we may be	0.000197086	0.000117688	7.93976E-05
28	the years ended	0.000214793	0.000140836	7.39564E-05
29	we may not	0.000200531	0.000131145	6.93856E-05
30	ended december and	0.000238829	0.000169788	6.90402E-05
31	for the years	0.000183322	0.000118337	6.49847E-05
32	if we are	0.000177169	0.000113856	6.33126E-05
33	our ability to	0.00032768	0.000269769	5.79104E-05
34	the company and	0.000185605	0.000128361	5.72438E-05
35	may not be	0.0001825	0.000126881	5.56188E-05
36	we believe that	0.000234961	0.000180015	5.49464E-05
37	selling general and	0.000233106	0.000178232	5.48742E-05
38	in the year	0.000222454	0.000168454	5.39992E-05
39	not be able	0.000167283	0.000115392	5.18916E-05
40	depreciation and	0.000167497	0.000116889	5.0608E-05
41	that the company	0.000165059	0.000115093	4.99667E-05
43	deferred tax assets	0.000163182	0.000113252	4.99292E-05
44	of our products	0.000281939	0.000232529	4.94099E-05
45	price of our	0.000159716	0.000115636	0.00004408
46	the company will	0.000187175	0.000143635	0.00004354
47	in north america	0.000186062	0.000144217	4.18455E-05
48	we do not	0.000198007	0.000156574	4.14336E-05
49	general and	0.000187316	0.000147085	4.02314E-05
50	to the company	0.000153334	0.000113627	3.97071E-05
51	be able to	0.000148226	0.000111671	3.65554E-05
52	effect on our	0.00021428	0.000178643	3.56375E-05
53	adverse effect on	0.000160192	0.000124638	3.55541E-05
54	adversely affect our	0.000224478	0.000190299	3.41787E-05
55	of approximately	0.000159068	0.000126208	3.286E-05
56	by the company	0.00024782	0.000215348	3.24719E-05
57	the consolidated	0.000167993	0.000139058	2.89357E-05
58	sales and marketing	0.000140362	0.000113703	2.66594E-05
59	charge of million	0.000139544	0.000116036	2.35081E-05
60	no assurance that	0.000150243	0.000127194	2.3049E-05

Table G.11: Most prominent trigrams in ‘f’ reports with corresponding trigram in ‘nf’ reports (PS).

Peer set: Top 60 trigrams by greatest difference in tf-Idf scores (averaged) in non-fraud reports compared with trigrams in fraud reports

	Trigrams	Tf-Idf score 'nf' reports (avg)	Tf-Idf score 'f' reports (avg)	Difference
1	in compared to	0.000332704	0	0.000332704
2	million in and	0.000227499	0	0.000227499
3	primarily due to	0.000215496	0	0.000215496
4	in and million	0.00019619	0	0.00019619
5	in and respectively	0.000188693	0	0.000188693
6	million at december	0.000185849	0	0.000185849
7	in the us	0.000174122	0	0.000174122
8	the notes to	0.000171841	0	0.000171841
9	the fourth quarter	0.000171757	0	0.000171757
10	due primarily to	0.000171568	0	0.000171568
11	compared to the	0.000156507	0	0.000156507
12	million in compared	0.000154333	0	0.000154333
13	of the company's	0.000437382	0.000283451	0.000153931
14	fourth quarter of	0.000152431	0	0.000152431
15	was primarily due	0.000151389	0	0.000151389
16	at the end	0.000144554	0	0.000144554
17	in and in	0.000143244	0	0.000143244
18	million to million	0.000140291	0	0.000140291
19	at december the	0.000135627	0	0.000135627
20	the second half	0.000135447	0	0.000135447
21	compared to in	0.000135377	0	0.000135377
22	to consolidated	0.000206684	0.000117502	8.91813E-05
23	as of december	0.000328356	0.000240367	8.79882E-05
24	notes to consolidated	0.000213821	0.000126605	8.7216E-05
25	from million in	0.000225231	0.000143753	8.14776E-05
26	earnings per share	0.000189341	0.000113488	7.58535E-05
27	of the year	0.000188544	0.00011478	7.37642E-05
28	an increase of	0.000193158	0.000125044	6.81142E-05
29	to million in	0.000226309	0.000175912	5.03971E-05
30	partially offset by	0.000190829	0.00014163	4.91987E-05
31	and million in	0.000176628	0.000130122	4.65058E-05
32	of million or	0.000163687	0.000119218	4.44682E-05
33	fair value of	0.000167955	0.000128769	3.91852E-05
34	increased to million	0.000165542	0.000126395	3.9147E-05
35	the company believes	0.000337044	0.000298656	3.83879E-05
36	during the year	0.000237328	0.0002016	3.57274E-05
37	intellectual property	0.000154487	0.00011897	3.55178E-05
38	research and	0.000348391	0.000313752	3.4639E-05
39	million and million	0.000201666	0.00016957	3.2096E-05
40	the united kingdom	0.000135327	0.000107373	2.79543E-05
41	million in the	0.000166446	0.000138799	2.7647E-05
43	the company had	0.000166048	0.000138707	2.73411E-05
44	our results of	0.000141528	0.000114537	2.69911E-05
45	december compared to	0.000135229	0.000109534	2.56954E-05
46	in the company	0.000194596	0.000173588	2.10081E-05
47	as a percentage	0.000157134	0.000136199	2.0935E-05
48	products and services	0.000159303	0.000141739	1.75638E-05
49	on our business	0.000142264	0.000124895	1.73693E-05
50	results of operations	0.000160067	0.000143108	1.69596E-05
51	of million in	0.00014912	0.000132223	1.68971E-05
52	a percentage of	0.000145537	0.00012924	1.62973E-05
53	consolidated financial	0.000151298	0.000135327	1.59711E-05
54	the company is	0.000268875	0.000253116	1.57593E-05
55	compared to million	0.000188755	0.000173903	1.48527E-05
56	cost of sales	0.000171295	0.000159813	1.14818E-05
57	and administrative	0.000177758	0.000166771	1.0987E-05
58	our operating results	0.000139725	0.000130281	9.44454E-06
59	for our products	0.000149897	0.000142745	7.15173E-06
60	revolving credit facility	0.000165131	0.000159708	5.42245E-06

Table G.12: Most prominent trigrams in 'nf' reports with corresponding trigram in 'f' reports (PS).

Matched Pair (MP): Top 60 concepts by greatest difference in concept scores (summed) in fraud reports compared with concepts in non-fraud reports

	Concept	'f' concept score	'nf' concept score	Difference
1	care.noun	56.62002	69.52835	-12.90833
2	charges.noun	62.39165	73.15756	-10.76591
3	production.noun	40.73049	51.43251	-10.70202
4	investments.noun	38.82079	46.00746	-7.18667
5	investment.noun	48.3486	52.68513	-4.33653
6	equity.noun	35.8955	38.96787	-3.07237
7	changes.noun	133.17102	135.61273	-2.44171
8	impact.noun	42.05837	43.99336	-1.93499
9	integration.noun	31.3315	33.01823	-1.68673
10	product.noun	48.09928	49.10763	-1.00835
11	businesses.noun	42.16642	42.91917	-0.75275
12	income.noun	36.15838	36.90094	-0.74256
13	customer.noun	42.40954	42.51524	-0.1057
14	process.noun	37.25314	36.91988	0.33326
15	include.verb	31.39645	31.05259	0.34386
16	delivery.noun	35.46582	35.05345	0.41237
17	products.noun	88.54478	87.94058	0.6042
18	service.noun	43.39577	42.63858	0.75719
19	including.verb	51.98321	51.01606	0.96715
20	channels.noun	33.14269	32.0235	1.11919
21	decision.noun	32.14113	30.88111	1.26002
22	facilities.noun	62.78174	61.35241	1.42933
23	operations.noun	149.2386	147.0435	2.1951
24	change.noun	75.02338	72.40208	2.6213
25	customers.noun	70.91416	67.63449	3.27967
26	completion.noun	33.63845	30.31584	3.32261
27	employees.noun	33.7995	30.21782	3.58168
28	existing.verb	33.09977	29.24284	3.85693
29	manufacturing.noun	35.49635	31.27382	4.22253
30	applications.noun	85.64642	81.15923	4.48719
31	interest.noun	30.69384	25.4702	5.22364
32	financing.noun	35.81703	30.49897	5.31806
33	expenses.noun	43.8688	38.4825	5.3863
34	order.noun	38.53378	33.07121	5.46257
35	condition.noun	47.32854	41.76725	5.56129
36	information.noun	32.17334	26.42933	5.74401
37	action.noun	59.03958	53.07357	5.96601
38	asset.noun	39.34334	33.24278	6.10056
39	connection.noun	34.67922	28.46918	6.21004
40	basis.noun	41.18056	34.45626	6.7243
41	managements.noun	30.77286	24.04404	6.72882
43	debt.noun	33.43023	26.55235	6.87788
44	equipment.noun	53.19711	46.30872	6.88839
45	disclosures.noun	31.43573	24.29145	7.14428
46	expense.noun	32.0662	24.83589	7.23031
47	decrease.noun	60.08361	52.81769	7.26592
48	marketing.noun	51.7166	44.15465	7.56195
49	group.noun	62.7155	54.99443	7.72107
50	case.noun	39.5609	31.73639	7.82451
51	loss.noun	36.78395	28.71988	8.06407
52	expansion.noun	35.85942	27.6179	8.24152
53	events.noun	48.05569	39.78431	8.27138
54	charge.noun	61.01798	52.51731	8.50067
55	markets.noun	89.87356	81.10624	8.76732
56	activity.noun	67.23534	58.40739	8.82795
57	amounts.noun	51.84655	42.99526	8.85129
58	cost.noun	64.75567	55.57471	9.18096
59	actions.noun	69.82077	60.31308	9.50769
60	obligations.noun	39.80622	29.93231	9.87391

Table G.13: Top 60 concepts in 'f' reports with the corresponding concept in 'nf' reports (MP).

Matched Pair: Top 60 concepts by greatest difference in concept scores (summed) in non-fraud reports compared with concepts in fraud reports

	Concept	'nf' concept score	'f' concept score	Difference
1	care.noun	69.52835	56.62002	12.90833
2	charges.noun	73.15756	62.39165	10.76591
3	production.noun	51.43251	40.73049	10.70202
4	contribution.noun	26.52175	19.26611	7.25564
5	investments.noun	46.00746	38.82079	7.18667
6	consumer.noun	32.74886	27.32463	5.42423
7	investment.noun	52.68513	48.3486	4.33653
8	channel.noun	26.89836	22.74063	4.15773
9	performance.noun	33.51527	30.44221	3.07306
10	equity.noun	38.96787	35.8955	3.07237
11	part.noun	29.63625	27.06914	2.56711
12	changes.noun	135.61273	133.17102	2.44171
13	approach.noun	27.53955	25.12024	2.41931
14	impact.noun	43.99336	42.05837	1.93499
15	integration.noun	33.01823	31.3315	1.68673
16	increase.noun	31.27565	29.98969	1.28596
17	product.noun	49.10763	48.09928	1.00835
18	businesses.noun	42.91917	42.16642	0.75275
19	income.noun	36.90094	36.15838	0.74256
20	adoption.noun	27.64928	27.42505	0.22423
21	customer.noun	42.51524	42.40954	0.1057
22	process.noun	36.91988	37.25314	-0.33326
23	include.verb	31.05259	31.39645	-0.34386
24	delivery.noun	35.05345	35.46582	-0.41237
25	products.noun	87.94058	88.54478	-0.6042
26	service.noun	42.63858	43.39577	-0.75719
27	including.verb	51.01606	51.98321	-0.96715
28	channels.noun	32.0235	33.14269	-1.11919
29	decision.noun	30.88111	32.14113	-1.26002
30	facilities.noun	61.35241	62.78174	-1.42933
31	operations.noun	147.0435	149.2386	-2.1951
32	change.noun	72.40208	75.02338	-2.6213
33	course.noun	27.27191	30.44862	-3.17671
34	customers.noun	67.63449	70.91416	-3.27967
35	completion.noun	30.31584	33.63845	-3.32261
36	employees.noun	30.21782	33.7995	-3.58168
37	factors.noun	26.94002	30.6503	-3.71028
38	existing.verb	29.24284	33.09977	-3.85693
39	manufacturing.noun	31.27382	35.49635	-4.22253
40	conduct.noun	25.92935	30.38297	-4.45362
41	applications.noun	81.15923	85.64642	-4.48719
43	financing.noun	30.49897	35.81703	-5.31806
44	expenses.noun	38.4825	43.8688	-5.3863
45	order.noun	33.07121	38.53378	-5.46257
46	condition.noun	41.76725	47.32854	-5.56129
47	information.noun	26.42933	32.17334	-5.74401
48	action.noun	53.07357	59.03958	-5.96601
49	asset.noun	33.24278	39.34334	-6.10056
50	connection.noun	28.46918	34.67922	-6.21004
51	basis.noun	34.45626	41.18056	-6.7243
52	debt.noun	26.55235	33.43023	-6.87788
53	equipment.noun	46.30872	53.19711	-6.88839
54	decrease.noun	52.81769	60.08361	-7.26592
55	marketing.noun	44.15465	51.7166	-7.56195
56	group.noun	54.99443	62.7155	-7.72107
57	case.noun	31.73639	39.5609	-7.82451
58	loss.noun	28.71988	36.78395	-8.06407
59	expansion.noun	27.6179	35.85942	-8.24152
60	events.noun	39.78431	48.05569	-8.27138

Table G.14: Top 60 concepts in 'nf' reports with the corresponding concept in 'f' reports (MP).

Peer set: Top 60 concepts by greatest difference in concept scores (averaged) in fraud reports compared with concepts in non-fraud reports

	Concept	'f' concept score	'nf' concept score	Difference
1	acquisition.noun	1.962218235	1.364031993	0.598186242
2	acquisitions.noun	1.399970196	1.157839706	0.24213049
3	company.noun	1.643904314	1.414173595	0.229730719
4	assets.noun	1.286926961	1.064202484	0.222724477
5	management.noun	1.273108137	1.070049935	0.203058203
6	ability.noun	0.799833824	0.623138856	0.176694967
7	ended.verb	0.615740098	0.444078399	0.171661699
8	event.noun	0.602680784	0.439515915	0.163164869
9	based.verb	0.740229412	0.581235131	0.158994281
10	capital.noun	1.019327745	0.861180686	0.158147059
11	market.noun	1.869063529	1.71289098	0.156172549
12	approval.noun	0.607439608	0.457756373	0.149683235
13	facility.noun	0.703885098	0.560475	0.143410098
14	design.noun	0.555798137	0.415721144	0.140076993
15	agreement.noun	0.435090098	0.297946503	0.137143595
16	services.noun	0.783681863	0.660375425	0.123306438
17	acquired.verb	0.549871569	0.428168464	0.121703105
18	cases.noun	0.494038627	0.379060196	0.114978431
19	insurance.noun	0.399782549	0.286523562	0.113258987
20	compliance.noun	0.666850098	0.554688203	0.112161895
21	borrowings.noun	0.611968529	0.50812549	0.103843039
22	development.noun	1.373215294	1.271049902	0.102165392
23	construction.noun	0.545948431	0.44421732	0.101731111
24	accordance.noun	0.548807255	0.452122124	0.096685131
25	asset.noun	0.38571902	0.297768301	0.087950719
26	approvals.noun	0.302026471	0.217290261	0.084736209
27	companies.noun	0.510732647	0.427251895	0.083480752
28	administration.noun	0.399710098	0.316335882	0.083374216
29	companys.noun	0.458083529	0.379323301	0.078760229
30	disclosures.noun	0.308193431	0.235582026	0.072611405
31	purchase.noun	0.359302255	0.288924248	0.070378007
32	application.noun	0.616443039	0.552183824	0.064259216
33	manufacturing.noun	0.348003431	0.285419346	0.062584085
34	amount.noun	0.670715588	0.614858824	0.055856765
35	activity.noun	0.65917	0.604245	0.054925
36	marketing.noun	0.50702549	0.455619412	0.051406078
37	action.noun	0.578819412	0.527723268	0.051096144
38	failure.noun	0.385188529	0.337239837	0.047948693
39	process.noun	0.365226863	0.318360523	0.04686634
40	business.noun	1.630249314	1.583906111	0.046343203
41	analysis.noun	0.660043529	0.61423902	0.04580451
43	credit.noun	0.365756667	0.320208497	0.04554817
44	order.noun	0.377782157	0.333051209	0.044730948
45	amounts.noun	0.50829951	0.465590359	0.04270915
46	completion.noun	0.329788725	0.287770033	0.042018693
47	group.noun	0.614857843	0.573043399	0.041814444
48	amortization.noun	0.521021176	0.484058922	0.036962255
49	efforts.noun	0.658023039	0.623405229	0.03461781
50	obligations.noun	0.390257059	0.360758268	0.029498791
51	loss.noun	0.360626961	0.332305458	0.028321503
52	expansion.noun	0.351562941	0.328854804	0.022708137
53	managements.noun	0.301694706	0.281208595	0.020486111
54	basis.noun	0.40373098	0.385206013	0.018524967
55	connection.noun	0.339992353	0.32183317	0.018159183
56	activities.noun	1.829572353	1.813122124	0.016450229
57	depreciation.noun	0.337546961	0.322736601	0.014810359
58	existing.verb	0.324507549	0.310631046	0.013876503
59	financing.noun	0.351147353	0.337849477	0.013297876
60	condition.noun	0.464005294	0.452715784	0.01128951

Table G.15: Top 60 concepts in 'f' reports with corresponding concept in 'nf' reports (PS).

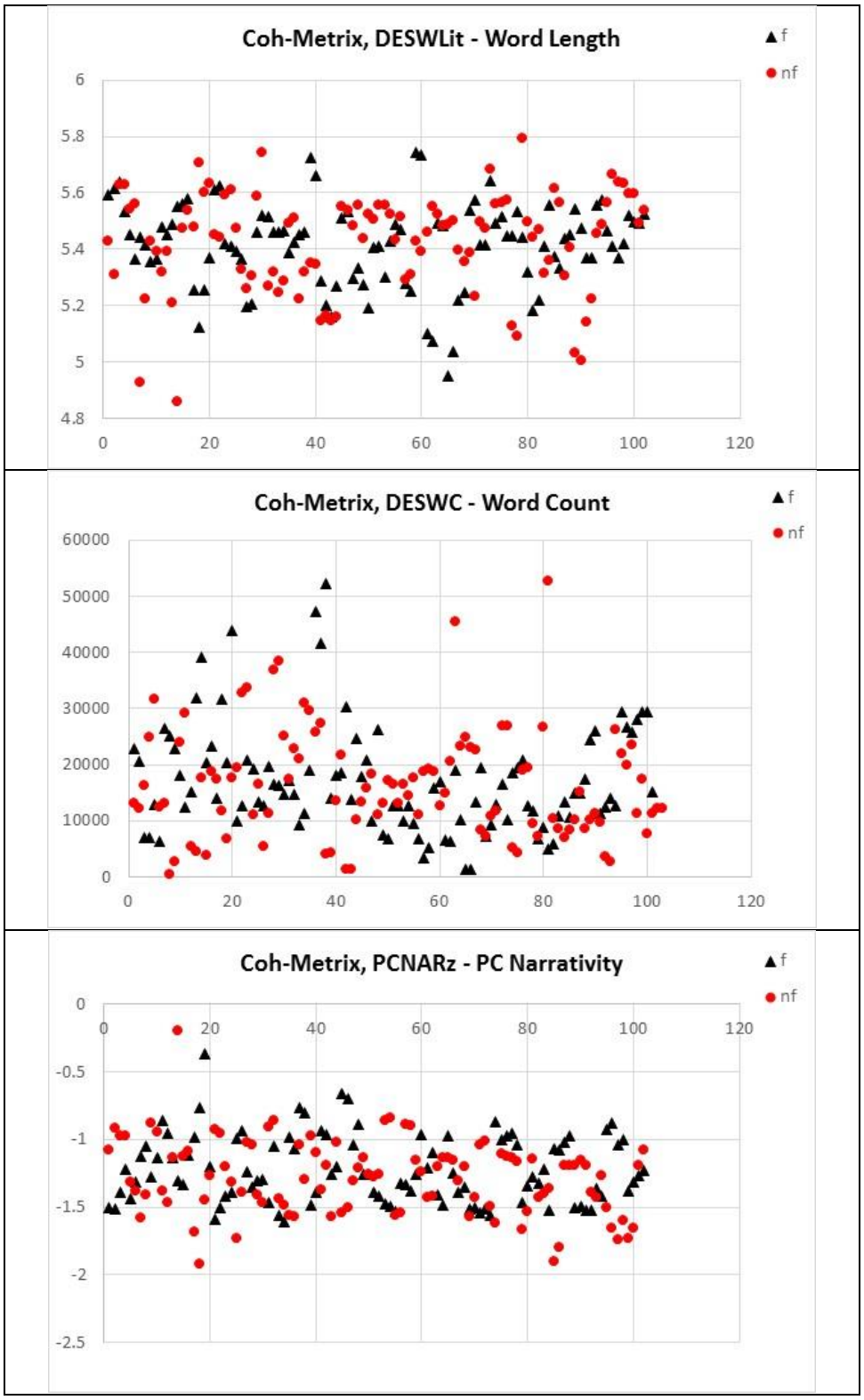
Peer set: Top 60 concepts by greatest difference in concept scores (averaged) in non-fraud reports compared with concepts in fraud reports

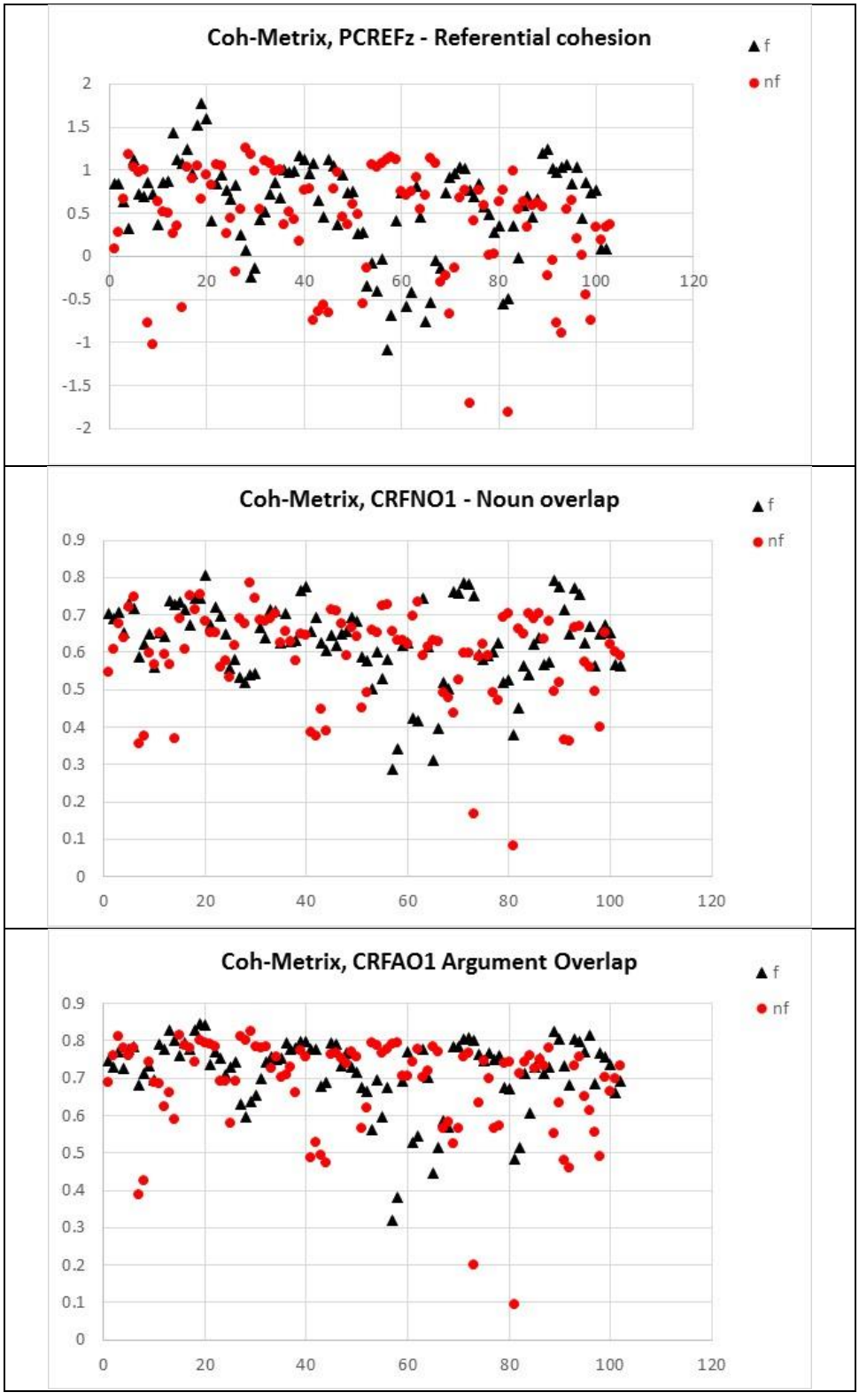
	Concept	'nf' concept score	'f' concept score	Difference
1	care.noun	0.739046569	0.555098235	0.183948333
2	charges.noun	0.774335131	0.611682843	0.162652288
3	changes.noun	1.424462549	1.305598235	0.118864314
4	markets.noun	0.988279804	0.881113333	0.107166471
5	channel.noun	0.326031503	0.222947353	0.10308415
6	impact.noun	0.50516732	0.412336961	0.092830359
7	production.noun	0.488896961	0.399318529	0.089578431
8	channels.noun	0.414360556	0.324928333	0.089432222
9	delivery.noun	0.431771307	0.347704118	0.08406719
10	operations.noun	1.54647634	1.463123529	0.08335281
11	equity.noun	0.434213889	0.351916667	0.082297222
12	investments.noun	0.459882614	0.38059598	0.079286634
13	investment.noun	0.546553693	0.474005882	0.07254781
14	customers.noun	0.765611993	0.695236863	0.070375131
15	consumer.noun	0.33804402	0.267888529	0.07015549
16	customer.noun	0.483453791	0.415779804	0.067673987
17	events.noun	0.537208268	0.471134216	0.066074052
18	decrease.noun	0.651548497	0.589055	0.062493497
19	businesses.noun	0.47504768	0.413396275	0.061651405
20	change.noun	0.785180425	0.735523333	0.049657092
21	including.verb	0.557365752	0.509639314	0.047726438
22	performance.noun	0.345267549	0.298453039	0.04681451
23	increase.noun	0.335535196	0.294016569	0.041518627
24	income.noun	0.389637451	0.354493922	0.035143529
25	debt.noun	0.362606732	0.327747353	0.034859379
26	facilities.noun	0.645801569	0.615507255	0.030294314
27	employees.noun	0.361195359	0.331367647	0.029827712
28	actions.noun	0.708122255	0.684517353	0.023604902
29	conduct.noun	0.320521732	0.297872255	0.022649477
30	charge.noun	0.619573203	0.59821549	0.021357712
31	equipment.noun	0.541112353	0.521540294	0.019572059
32	decision.noun	0.332293399	0.315109118	0.017184281
33	course.noun	0.31527281	0.298515882	0.016756928
34	product.noun	0.488024183	0.471561569	0.016462614
35	include.verb	0.320825556	0.307808333	0.013017222
36	factors.noun	0.312567059	0.300493137	0.012073922
37	competition.noun	0.295947026	0.285185686	0.01076134
38	information.noun	0.32293317	0.315424902	0.007508268
39	case.noun	0.394515327	0.387851961	0.006663366
40	cost.noun	0.637176601	0.63485951	0.002317092
41	competitors.noun	0.390528497	0.389876176	0.00065232
43	engineering.noun	0.299240425	0.298954902	0.000285523
44	expense.noun	0.314054183	0.31437451	-0.000320327
45	integration.noun	0.305719869	0.307171569	-0.001451699
46	applications.noun	0.836741046	0.839670784	-0.002929739
47	products.noun	0.863656144	0.868086078	-0.004429935
48	service.noun	0.416804739	0.425448725	-0.008643987
49	discussion.noun	0.525463464	0.535037843	-0.009574379
50	expenses.noun	0.419427386	0.430086275	-0.010658889
51	condition.noun	0.452715784	0.464005294	-0.01128951
52	financing.noun	0.337849477	0.351147353	-0.013297876
53	existing.verb	0.310631046	0.324507549	-0.013876503
54	depreciation.noun	0.322736601	0.337546961	-0.014810359
55	activities.noun	1.813122124	1.829572353	-0.016450229
56	connection.noun	0.32183317	0.339992353	-0.018159183
57	basis.noun	0.385206013	0.40373098	-0.018524967
58	expansion.noun	0.328854804	0.351562941	-0.022708137
59	loss.noun	0.332305458	0.360626961	-0.028321503
60	obligations.noun	0.360758268	0.390257059	-0.029498791

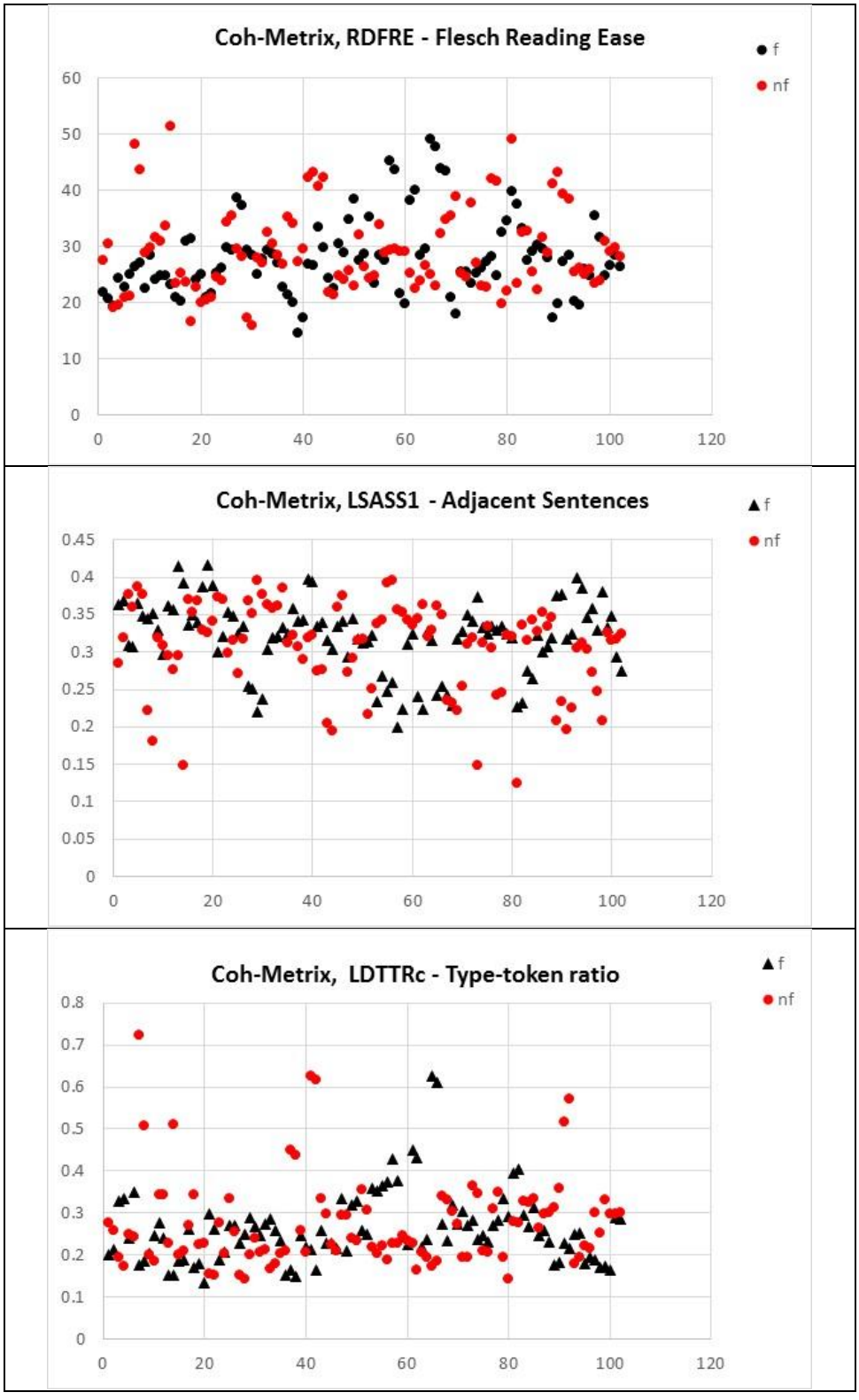
Table G.16: Top 60 concepts in 'nf' reports with corresponding concept in 'f' reports (PS)

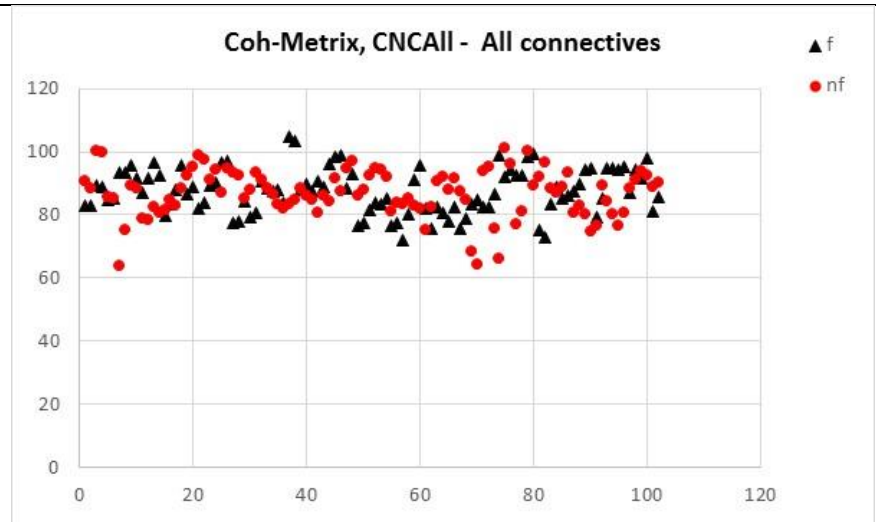
APPENDIX H

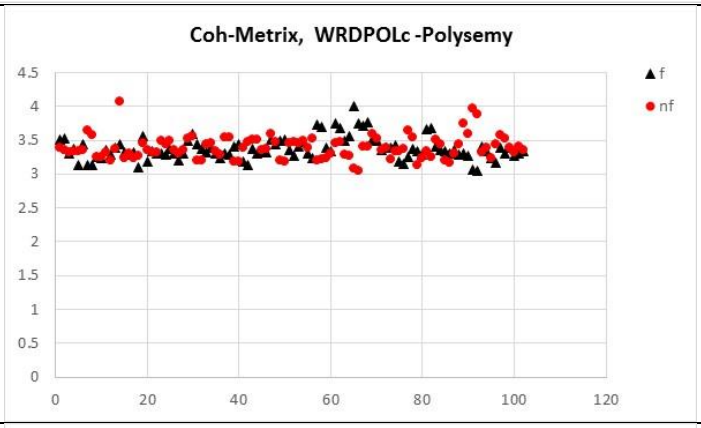
Sixteen Coh-Metrix Indices plotted for fraud and non-fraud reports

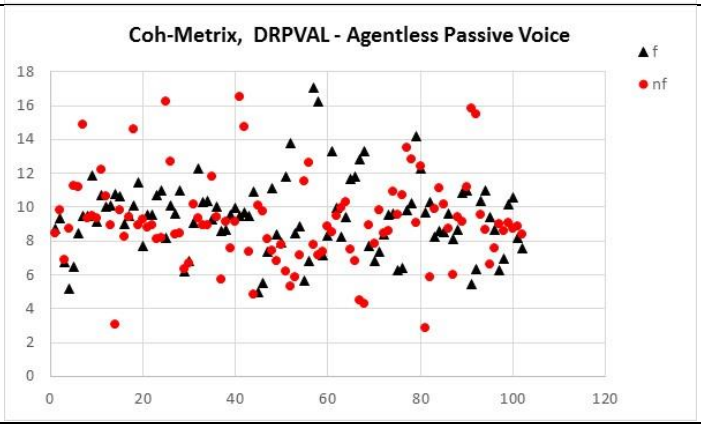


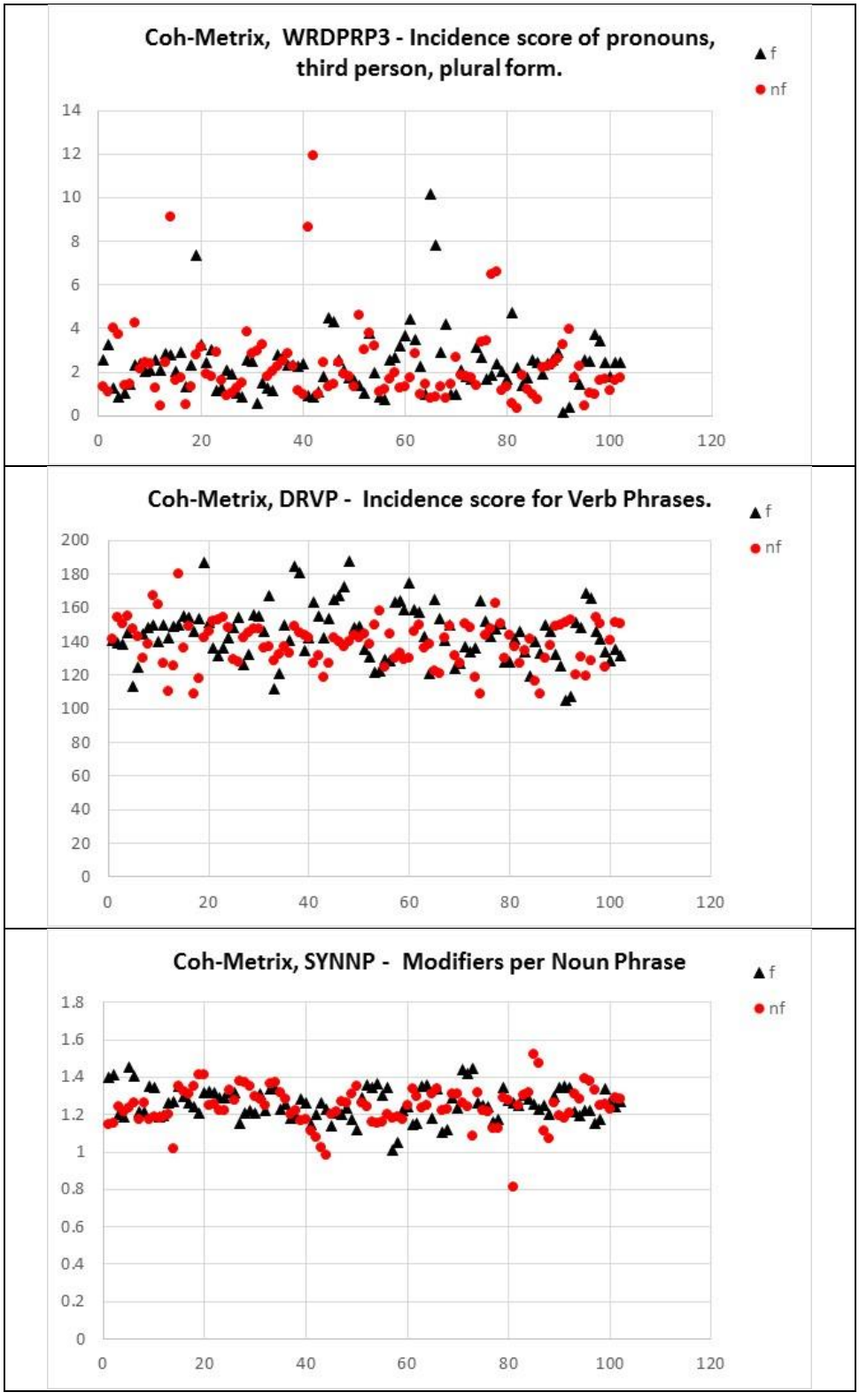

Coh-Metrix, PCREFz - Referential cohesion
f
nf
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
20
40
60
80
100
120
Coh-Metrix, CRFNO1 - Noun overlap
f
nf
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
Coh-Metrix, CRFAO1 Argument Overlap
f
nf
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120


Coh-Metrix, RDFRE - Flesch Reading Ease
f
nf
60
50
40
30
20
10
0
0
20
40
60
80
100
120
Coh-Metrix, LSASS1 - Adjacent Sentences
f
nf
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
20
40
60
80
100
120
Coh-Metrix, LDTTRc - Type-token ratio
f
nf
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120


Coh-Metrix, CNCAll - All connectives
f
nf
120
100
80
60
40
20
0
0
20
40
60
80
100
120


Coh-Metrix, WRDPOLc -Polysemy
f
nf
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
120


Coh-Metrix, DRPVAL - Agentless Passive Voice
f
nf
18
16
14
12
10
8
6
4
2
0
0
20
40
60
80
100
120


Coh-Metrix, WRDPRP3 - Incidence score of pronouns, third person, plural form.
f
nf
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120
Coh-Metrix, DRVP - Incidence score for Verb Phrases.
f
nf
0
20
40
60
80
100
120
140
160
180
200
0
20
40
60
80
100
120
Coh-Metrix, SYNNP - Modifiers per Noun Phrase
f
nf
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0
20
40
60
80
100
120

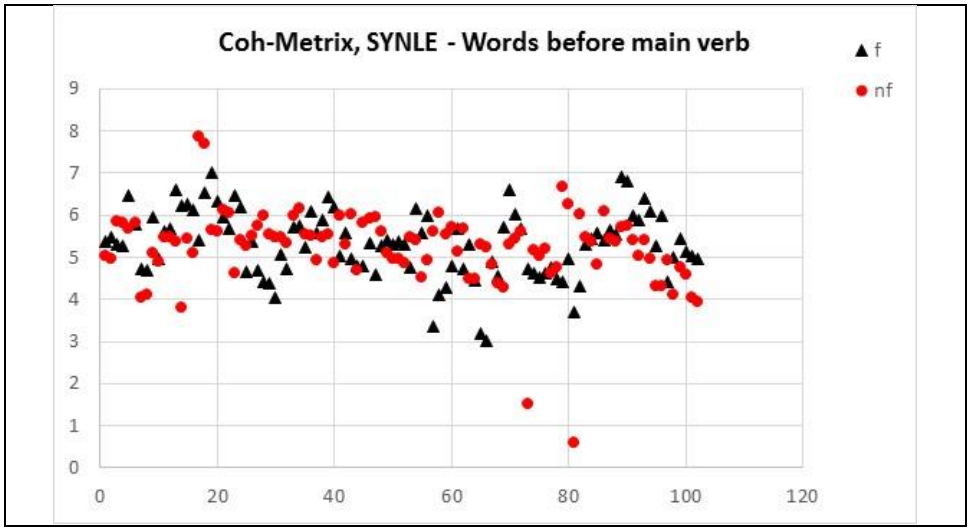


Table H.1: A few Coh-Metrix Indices plotted for fraud and non-fraud reports.

Extract produced when LIWC 2015 executed over a report

Filename	Segment	WC	Analytic	Clout	Authentic	Tone	WPS	Sixltr	Dic	function	pronoun	ppron
Verizon Comm Inc 2002	1	23166	97.41	70.82	32.94	62.68	26.57	35.37	73.63	37.68	5.24	3.07
			i	we	you	shehe	they	ipron	article	prep	auxverb	adverb
			0.04	2.78	0.06	0.03	0.17	2.17	5.80	17.11	2.65	1.64
			conj	negate	verb	adj	compare	interrog	number	quant	affect	posemo
			6.09	0.46	4.34	4.44	2.93	0.40	9.43	1.60	3.60	2.77
			negemo	anx	anger	sad	social	family	friend	female	male	cogproc
			0.82	0.10	0.04	0.57	5.29	0.01	0.04	0.00	0.04	7.08
			insight	cause	discrep	tentat	certain	differ	percept	see	hear	feel
			1.71	1.93	0.27	1.62	0.59	2.14	0.51	0.17	0.26	0.06
			bio	body	health	sexual	ingest	drives	affiliation	achieve	power	reward
			0.83	0.01	0.80	0.01	0.01	10.20	3.60	1.92	3.29	1.37
			risk	focuspast	focuspresen	focusfutur	relativ	motion	space	time	work	leisure
			0.97	2.01	2.88	0.76	14.77	2.09	8.47	4.27	8.31	0.20
			home	money	relig	death	informal	swear	netspeak	assent	nonflu	filler

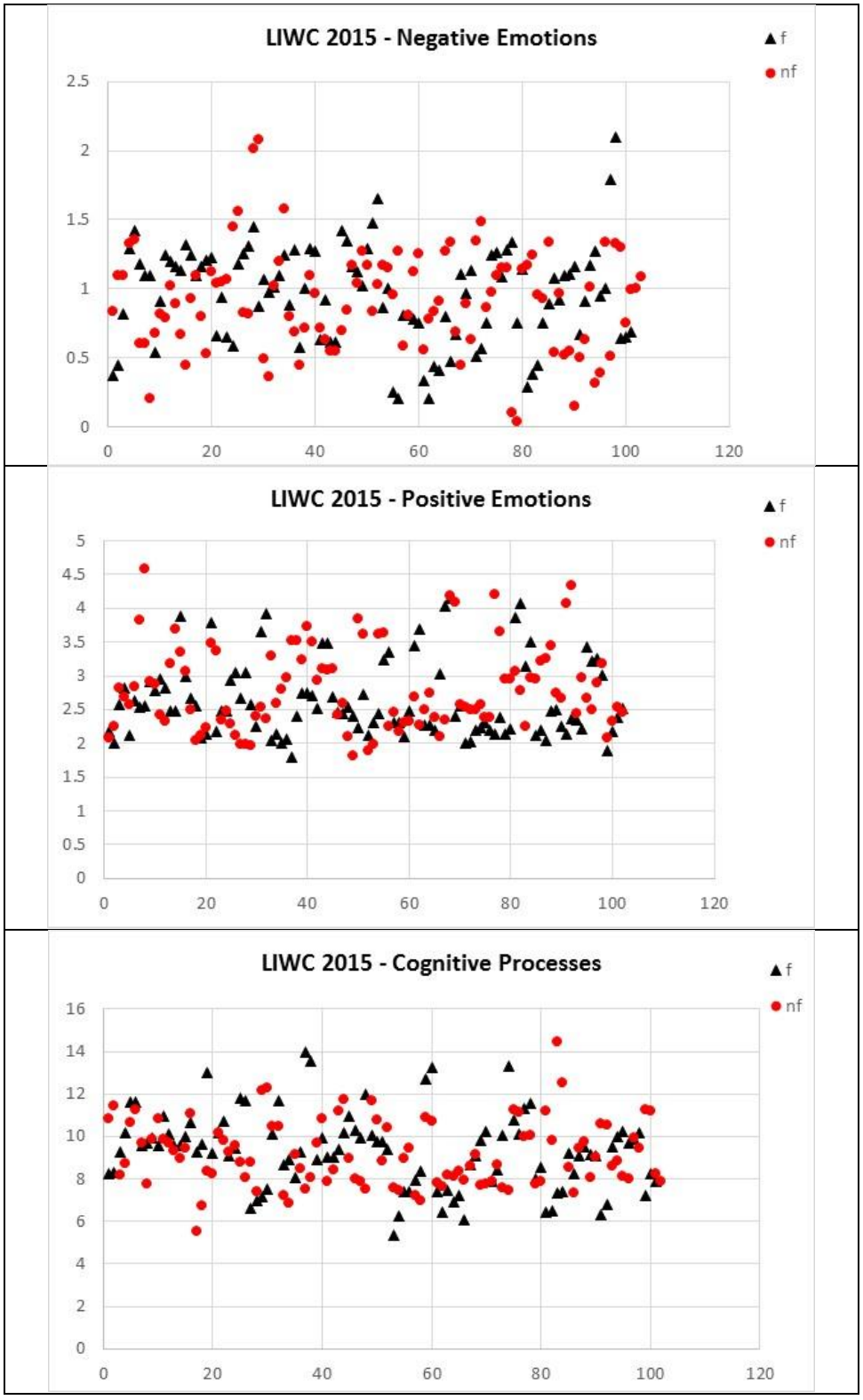
Figure H.1: LIWC output file produced when LIWC executed over a report.

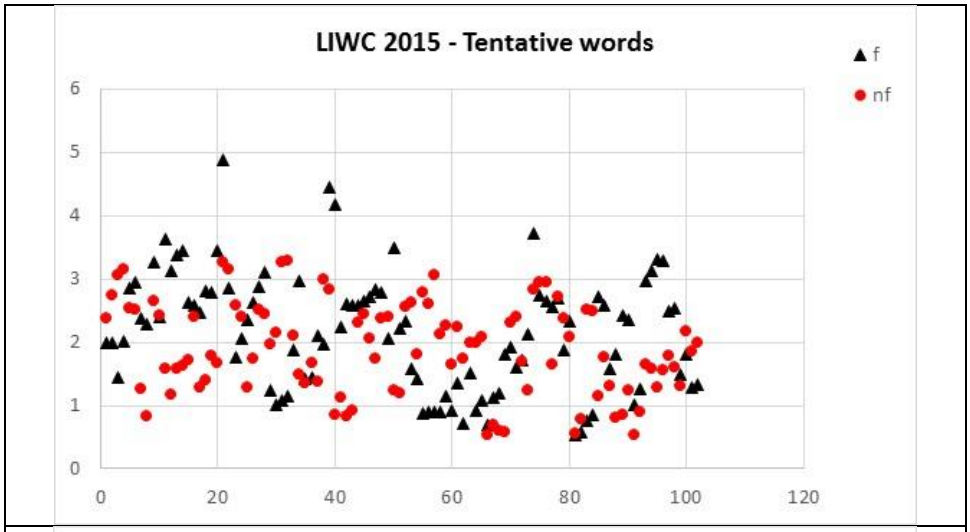
LIWC used to aid differentiation between fraud and non-fraud reports

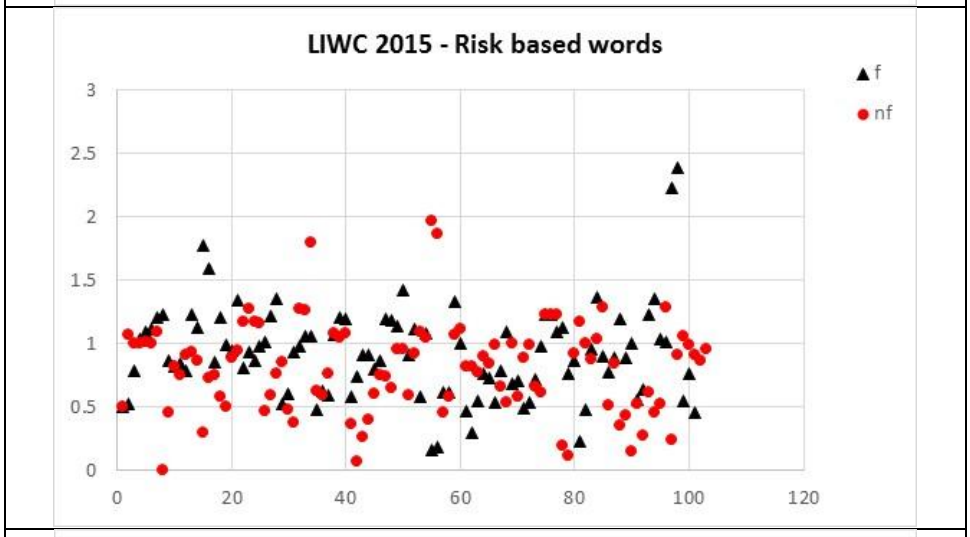
LIWC variables	Description
Words>6 letters	
Total Function Words	
Total Pronouns	I, them, itself
Personal Pronouns	I, them, her
1st person singular	I, me, mine
1st person plural	We, us, our
2nd person	You, your, thou
3rd pers singular	She, her, him
3rd pers plural	They, their, they'd
Impersonal pronouns	It, it's, those
Articles	A, an, the
Common verbs	Walk, went, see
Auxiliary verbs	Am, will, have
Past tense verbs	Went, ran, had
Present tense verbs	Is, does, hear
Future tense verbs	Will, gonna
Prepositions	To, with, above
Conjunction	And, but, whereas
Negations	No, not, never
Quantifiers	Few, many, much
Numbers	Second, thousand
Other Grammar	
Common verbs	
Common adjectives	
Comparisons	
Interrogatives	
Numbers	
Quantifiers	
Affective Processes	
Positive emotions	Love, nice, sweet
Negative emotions	Hurt, ugly, nasty
Cognitive mechanisms	cause, know, ought
Insight	think, know, consider
Causal	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain, stop
Inclusive	And, with, include
Exclusive	But, without, exclude
Relativity	Area, bend, exit, stop
Motion	Arrive, car, go
Space	Down, in, thin
Time	End, until, season
Drives	Drives
Affiliation	ally, friend, social
Achievement	Win, success, better
power	Superior, bully
reward	take, prize, benefit
risk	Danger, doubt
Time Orientations	
Past focus	Ago, did, talked
Present focus	Today, is, now
Future focus	May, will, soon

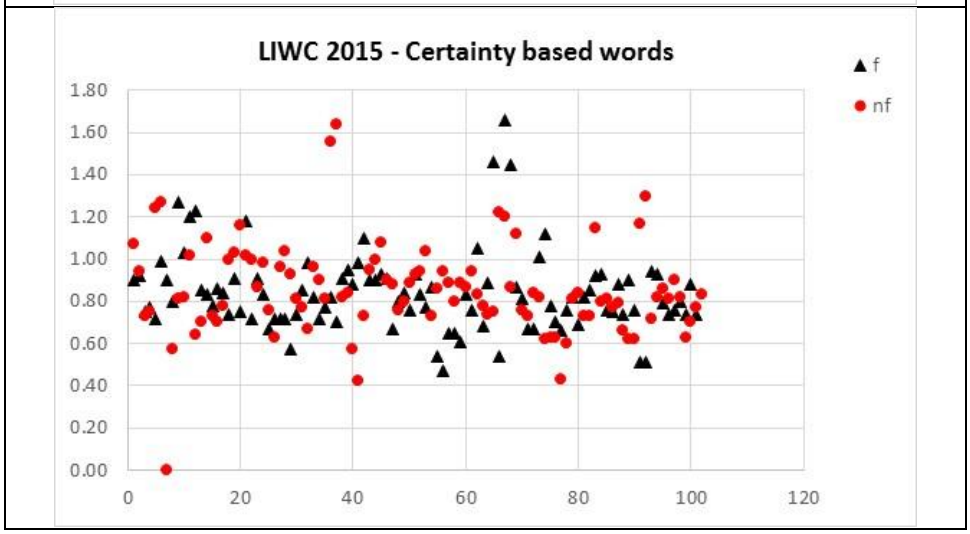
Table H.2: The LIWC variables used to form matrix.

Eight LIWC variables plotted for fraud and non-fraud reports




LIWC 2015 - Tentative words
f
nf
0
1
2
3
4
5
6
0
20
40
60
80
100
120


LIWC 2015 - Risk based words
f
nf
0
0.5
1
1.5
2
2.5
3
0
20
40
60
80
100
120


LIWC 2015 - Certainty based words
f
nf
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
0
20
40
60
80
100
120

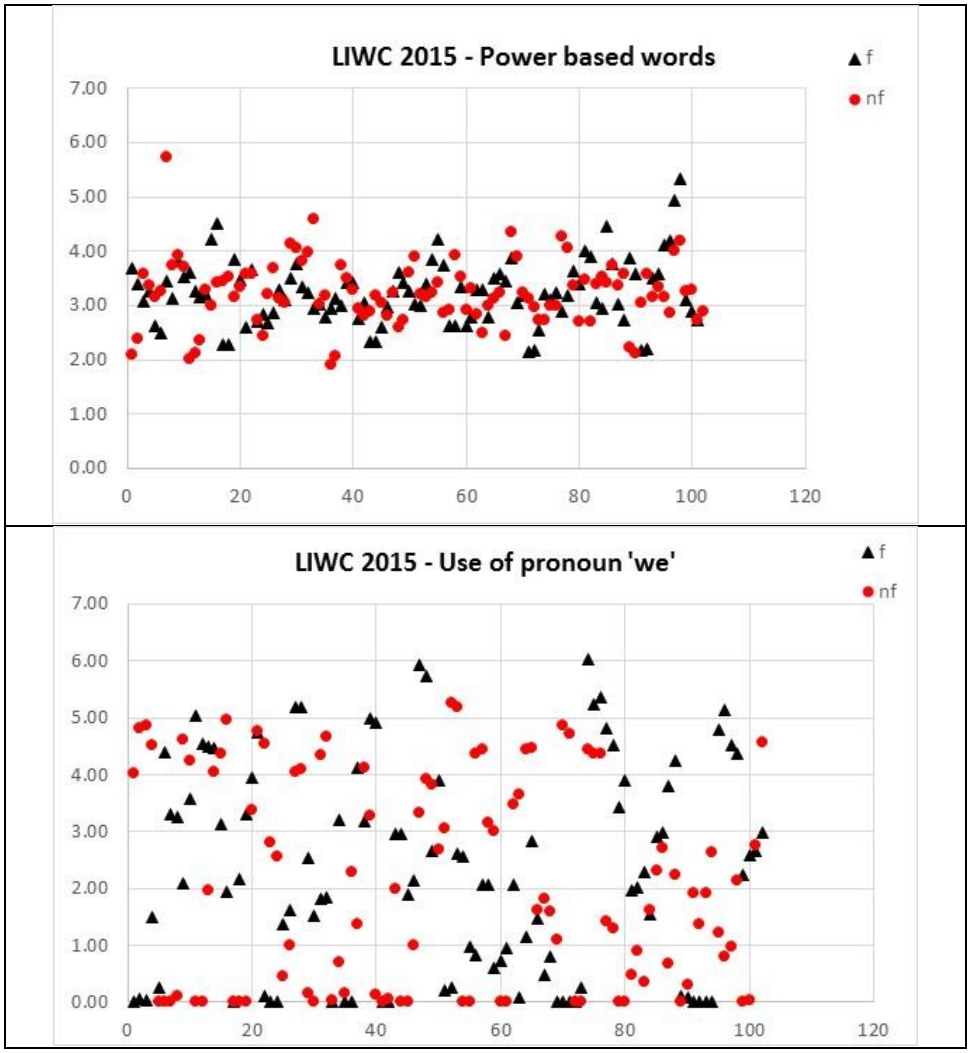
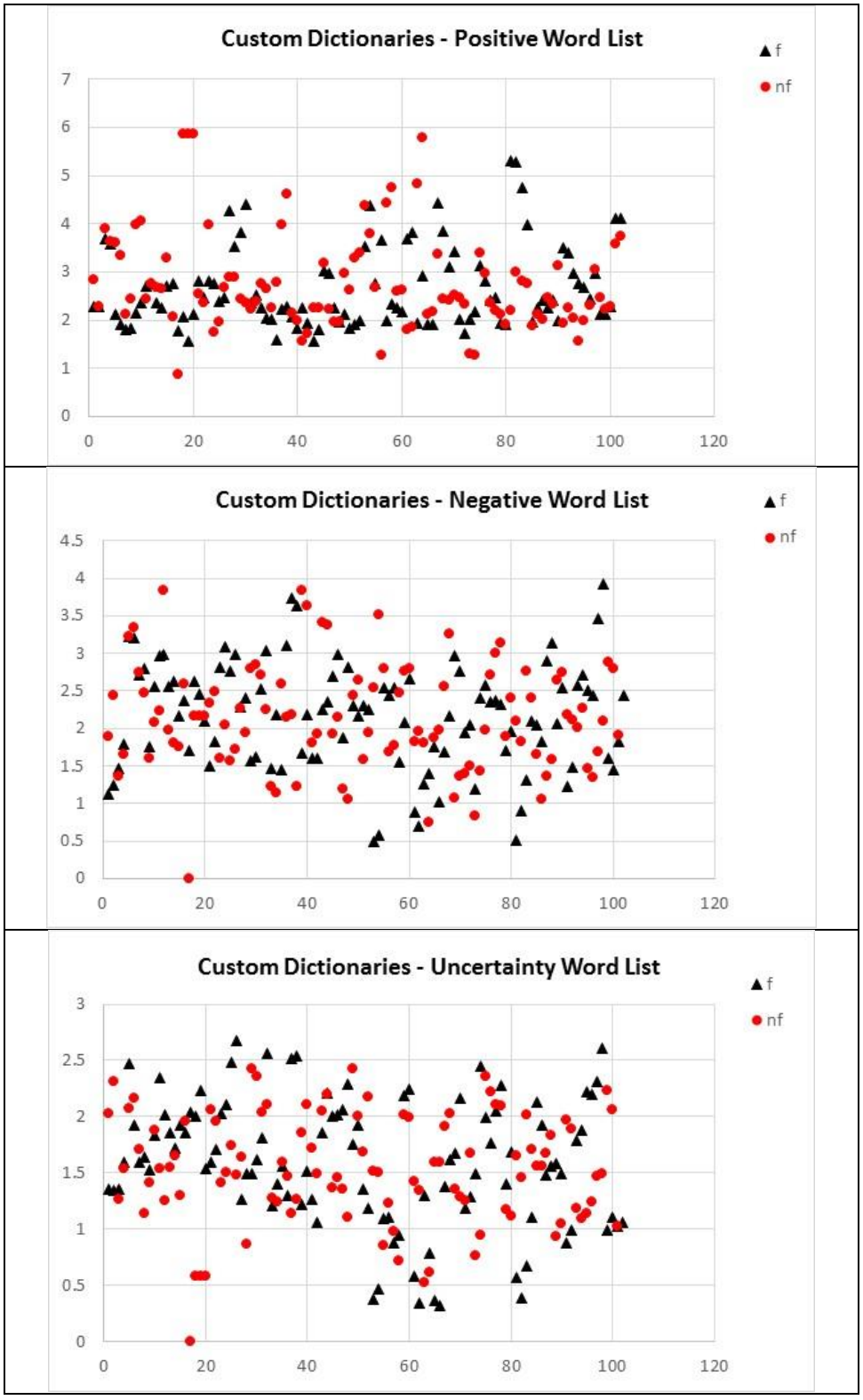
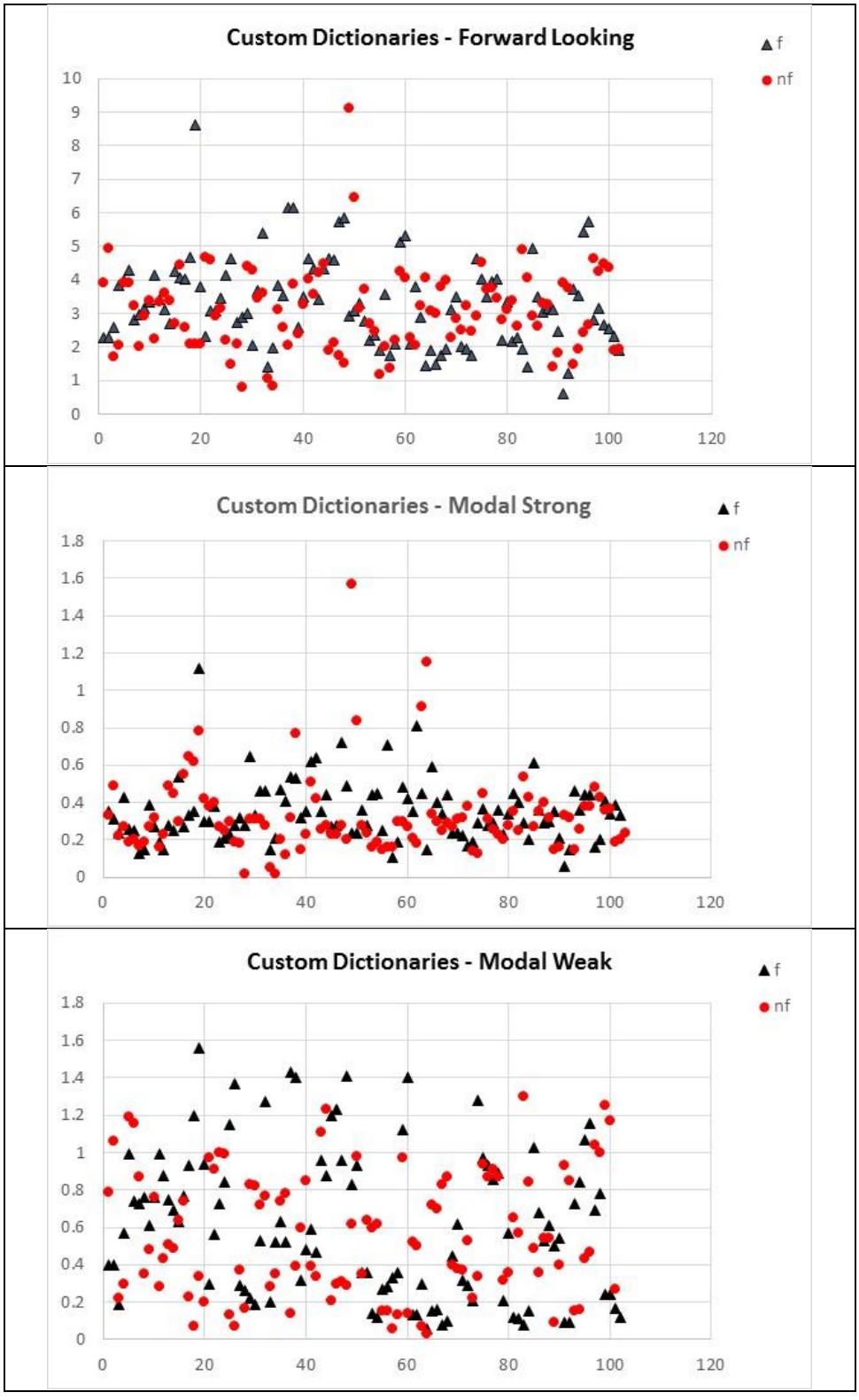


Table H.3: A few LIWC variables plotted for fraud and non-fraud reports.

Counts attained for words in 8 custom dictionaries plotted for fraud and non-fraud reports




Custom Dictionaries - Forward Looking
f
nf
10
9
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100
120
Custom Dictionaries - Modal Strong
f
nf
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120
Custom Dictionaries - Modal Weak
f
nf
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120

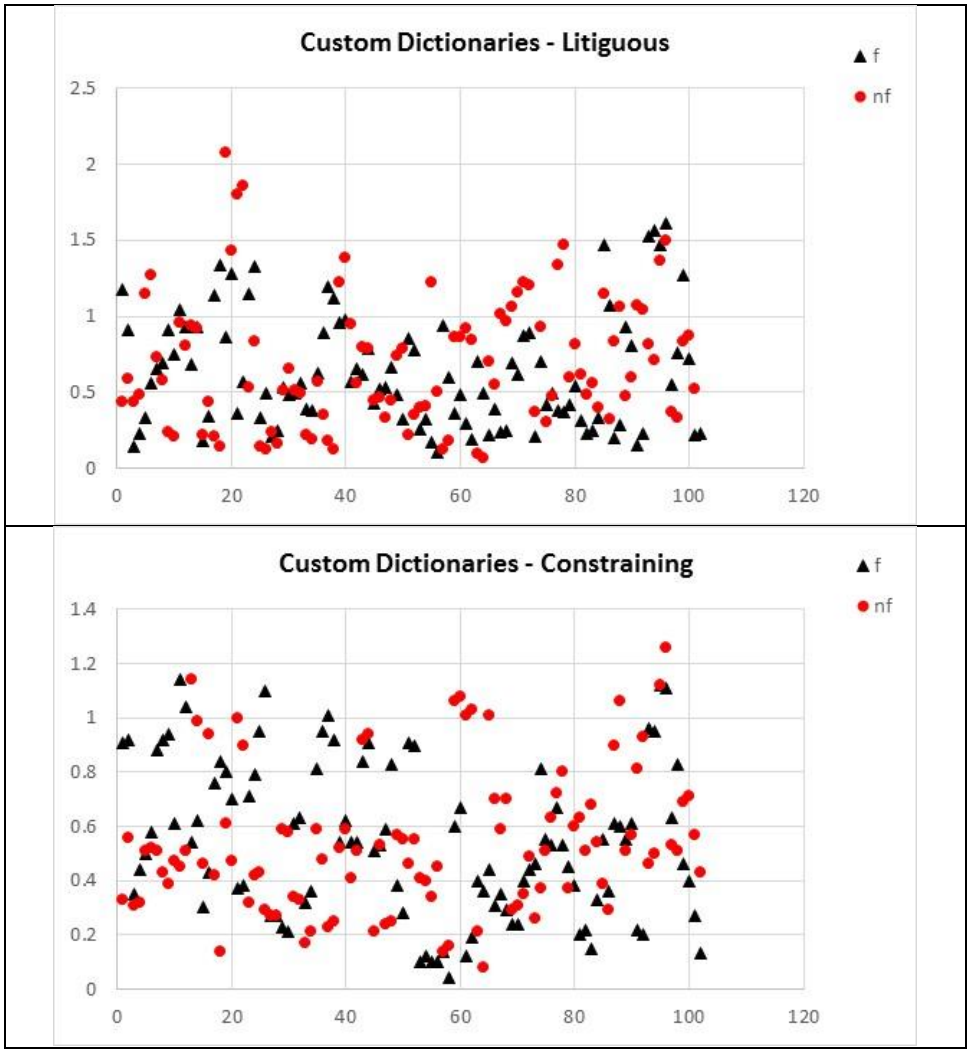
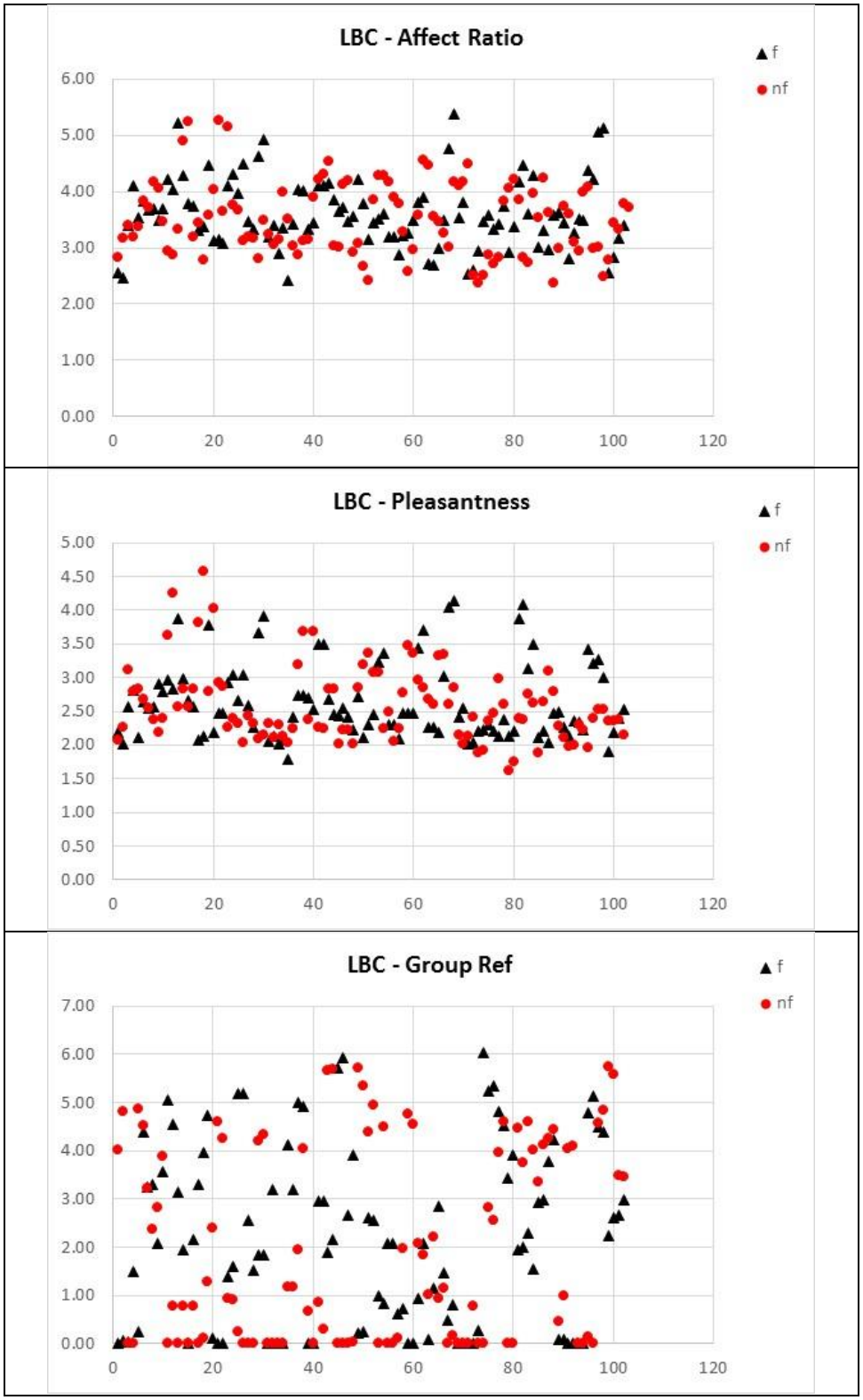
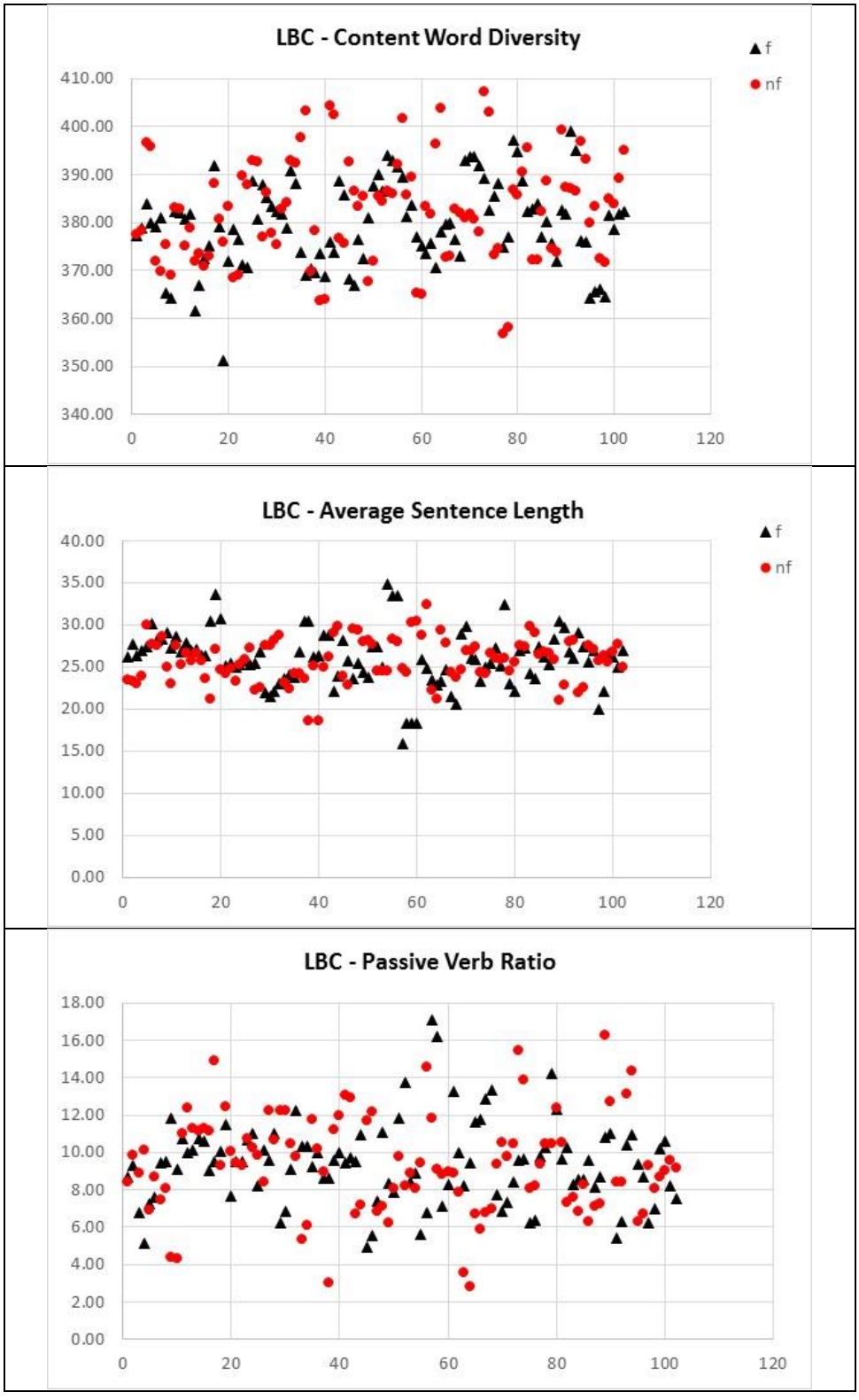


Table H.4: Counts of words in custom dictionaries plotted for fraud and non-fraud reports.

A sample (8) LBCs plotted derived from fraud and non-fraud reports




LBC - Content Word Diversity
f
nf
410.00
400.00
390.00
380.00
370.00
360.00
350.00
340.00
0
20
40
60
80
100
120
LBC - Average Sentence Length
f
nf
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
0
20
40
60
80
100
120
LBC - Passive Verb Ratio
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0
20
40
60
80
100
120

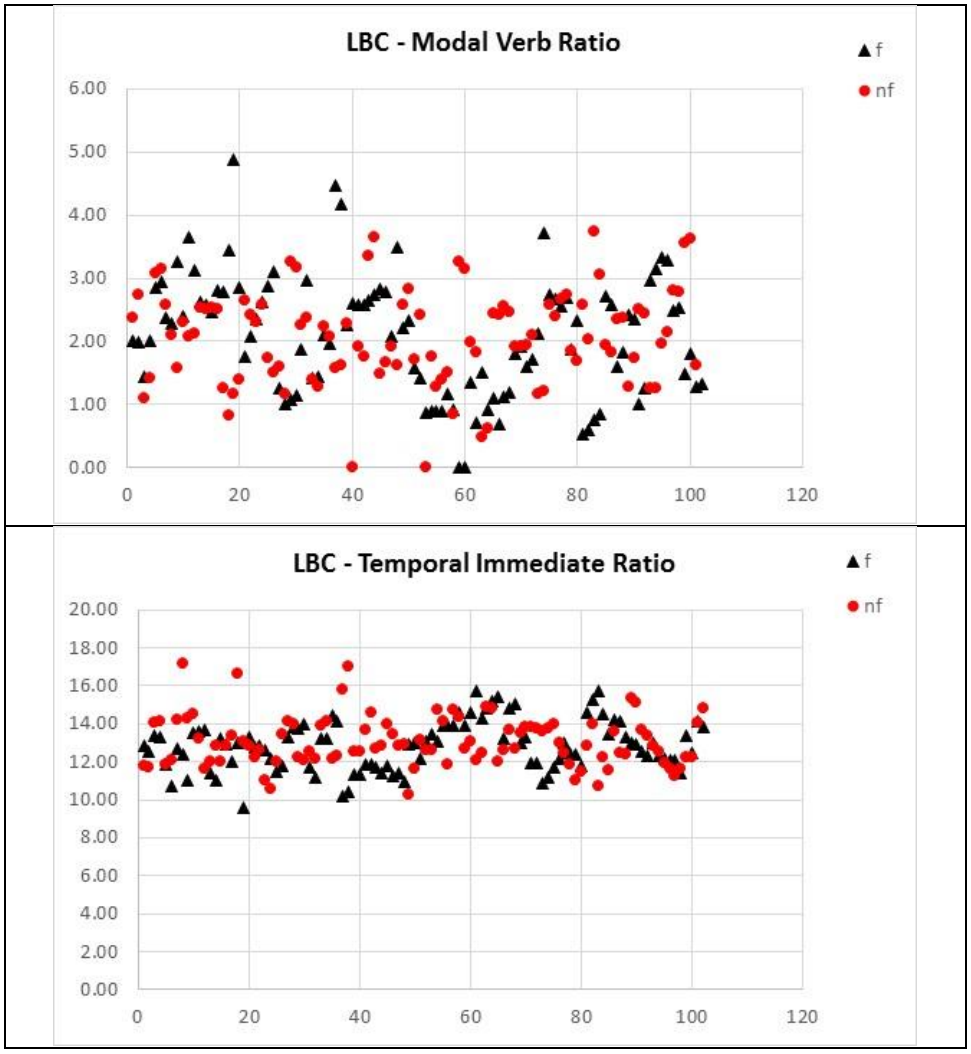
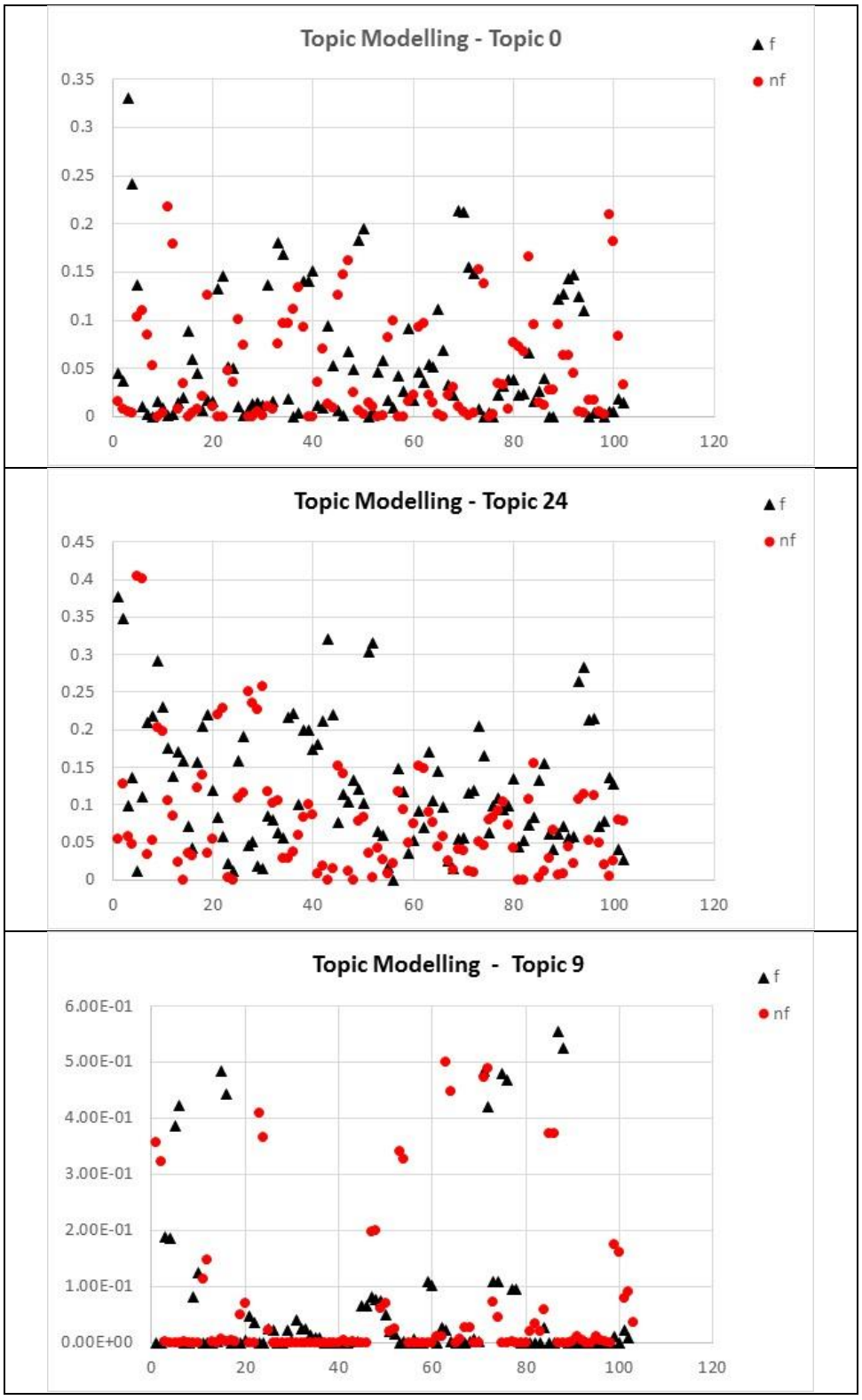
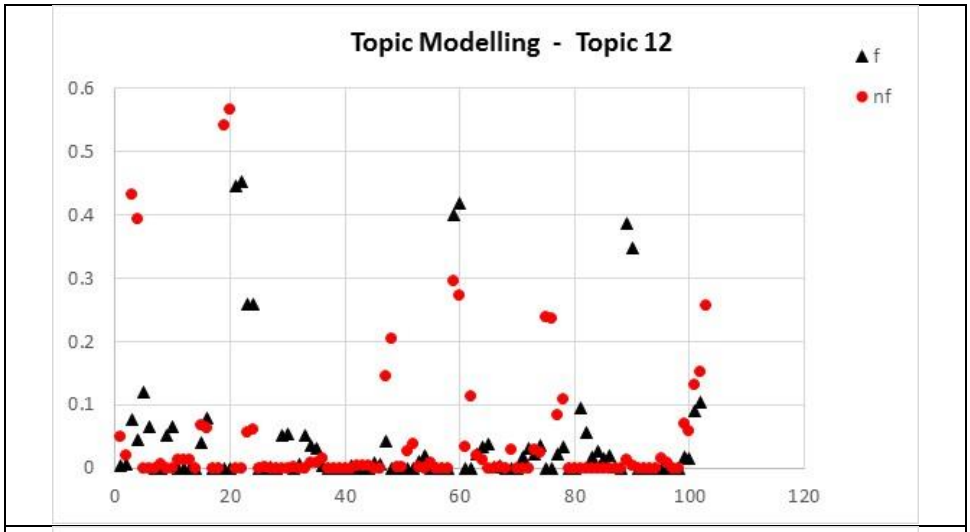
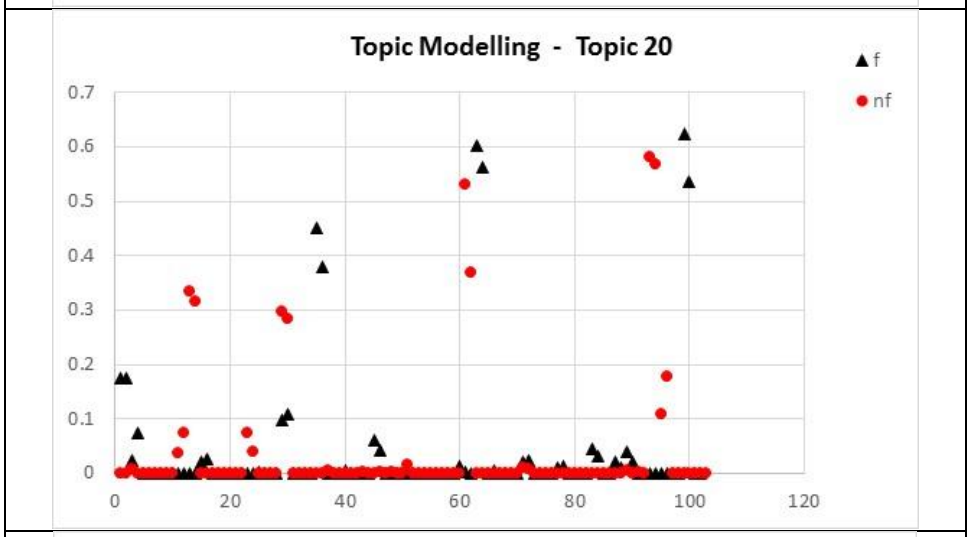


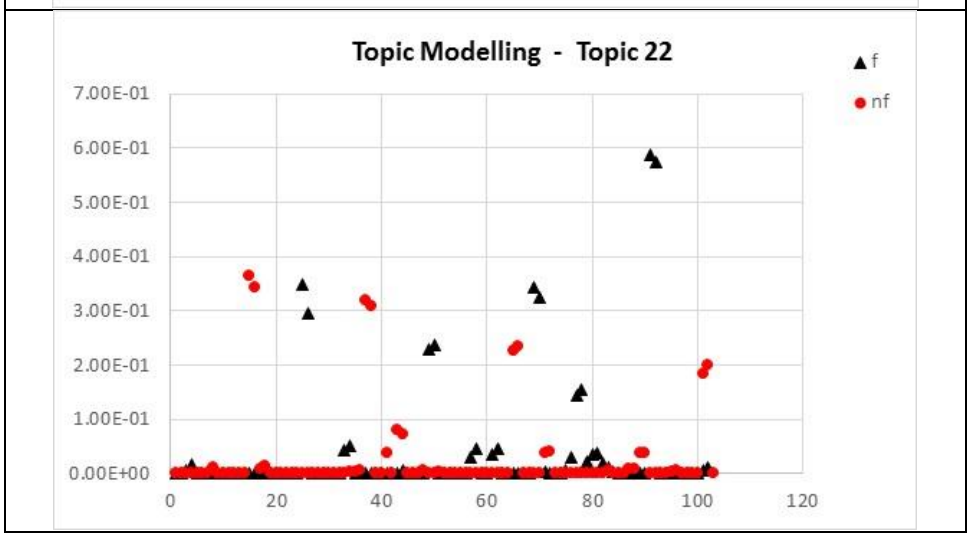
Table H.5: A few LBCs derived from reports.

A sample (8) Topics (weights derived from LDA) plotted for fraud and non-fraud reports




Topic Modelling - Topic 12
f
nf
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120


Topic Modelling - Topic 20
f
nf
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120


Topic Modelling - Topic 22
f
nf
7.00E-01
6.00E-01
5.00E-01
4.00E-01
3.00E-01
2.00E-01
1.00E-01
0.00E+00
0
20
40
60
80
100
120

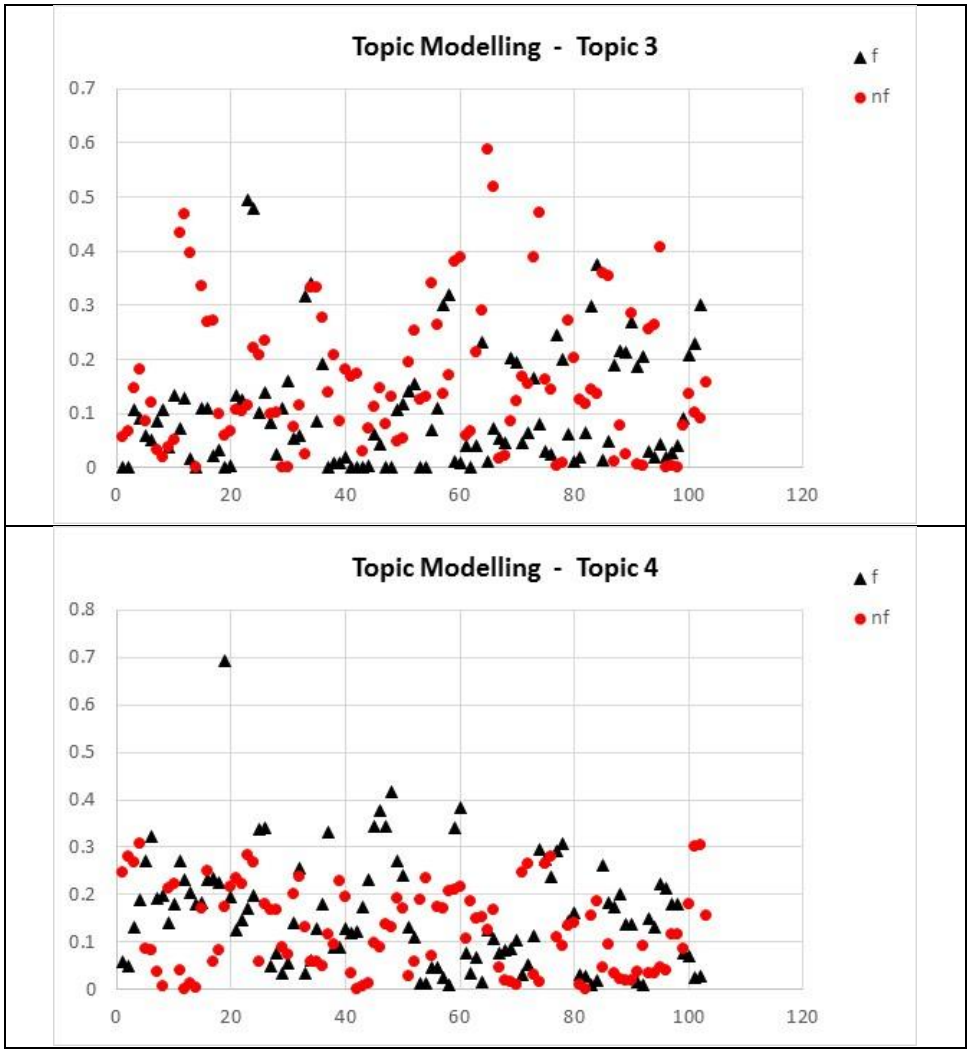
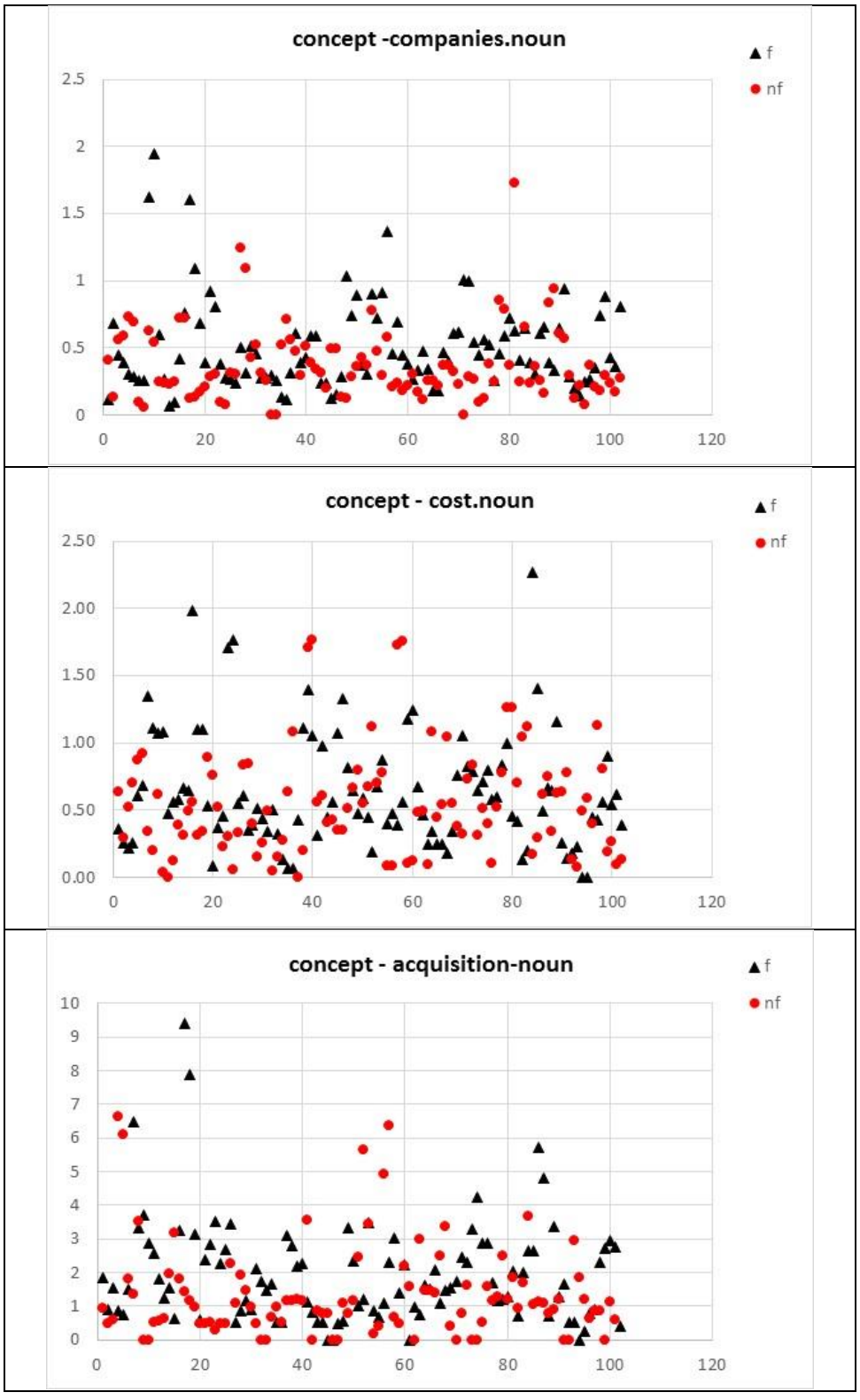
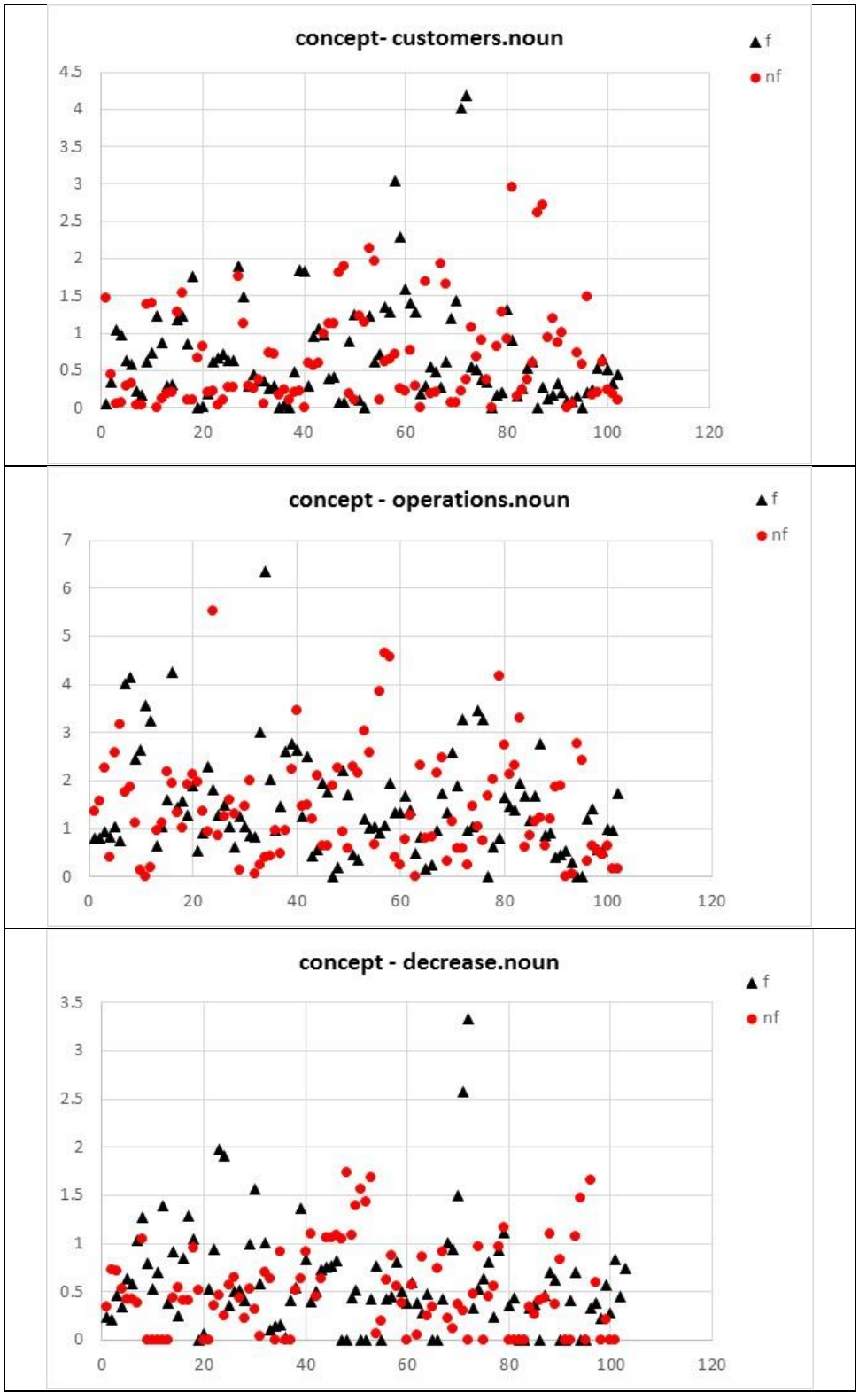


Table H.6: Topic weights plotted for fraud and non-fraud reports.

A sample (7) concepts (concept scores) plotted for fraud and non-fraud reports




concept- customers.noun
f
nf
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
120
concept - operations.noun
f
nf
7
6
5
4
3
2
1
0
0
20
40
60
80
100
120
concept - decrease.noun
f
nf
3.5
3
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
120

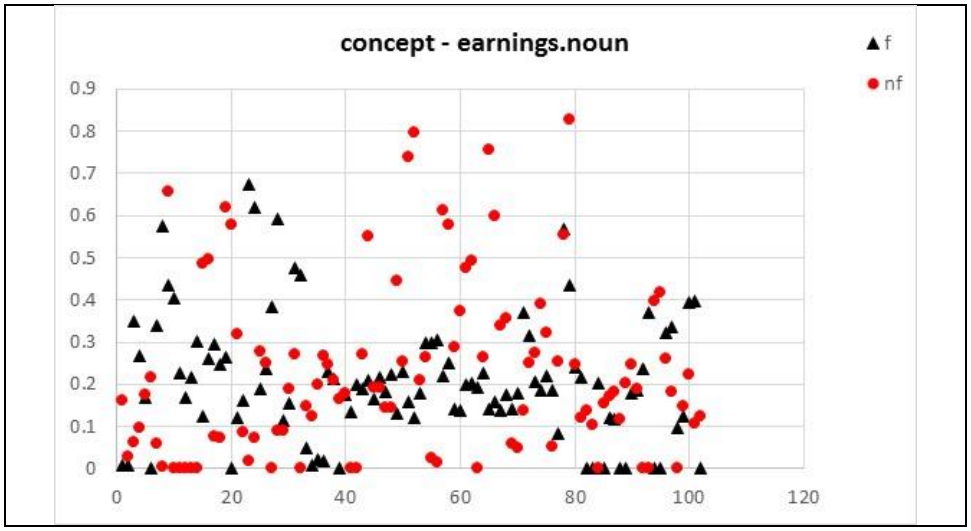


Table H.7: A selection of a few frequent concepts plotted.

Keywords identified from analysis conducted in chapter 3

ability	commission	export	inspection	net	procurement
according	commissions	failures	installation	networks	production
accounting	committed	fair	installed	notes	program
ability	common	favorable	insurance	officers	project
acquired	communications	filing	international	offset	purchase
acquisition	communities	flat	introduced	online	purpose
acquisitions	community	flexible	investment	open	quarter
across	companies	floating	investments	operating	rate
act	compared	flows	items	operations	read
adjustments	compensation	forward	joint	outlook	realized
advertising	component	fourth	july	overhead	receivable
agency	consisting	fuel	knowledge	partially	recorded
agreement	consumer	fund	launched	participation	reflecting
allocated	continued	gain	leases	parts	regulated
also	contract	gains	legislation	patent	related
approval	contributed	general	level	patents	remains
approvals	control	generation	licensing	pay	repurchase
areas	corporation	group	like	payable	required
assurance	currency	growth	limited	pension	reserves
average	customer	half	losses	percent	result
balances	date	health	lower	perform	results
bankruptcy	declined	hedges	management	performance	retained
believe	declines	higher	march	placement	secured
believed	decreased	impact	markets	platforms	see
billion	dollar	impacted	mass	portfolio	seek
brand	domestic	improved	medical	preferred	senior
brands	drive	income	members	pretax	services
buy	driven	increased	merger	prices	servicing
capital	due	incurred	million	primarily	settlement
carriers	earned	individuals	minority	private	share
certain	earnings	initial	mobile	problems	shipments
chain	education	initiatives	months	process	short
claims	employee	injury	necessary	processing	software
combination	euro	directors	volumes	stores	transfer
delivery	event	disruptions	warrants	store	traditional
demand	exclusive	division	website	top	users
derivative	expected	documents	without	storage	economic
source	sufficient	treasury	procedures	strong	affected
south	system	type	increased	stockholder	tools
spending	target	unable	solid	upon	time
staffing	tax	unfavorable	exceed	use	stock
state	telephone	unsecured	adverse		

Table H.8: Keywords used in the classification task.

Keywords identified based on Rutherford (2004) study

activity	products
asset	profit
borrowing	property
capital	rate
cash	reduce
company	result
completed	retail
continued	revenue
cost	risk
currency	sale
customer	services
debt	share
decrease	significant
development	store
division	strong
due	systems
end	tax
exchange	total
expenditure	trading
facility	turnover
financial	years
fixed	major
grow	make
growth	management
high	margin
higher	net
include	network
increase	new
increasingly	now
interest	number
investment	operating
item	operations
last	overall
level	performance
liability	previous
loss	lower

Table H.9: Keywords from Rutherford study (2004) used for Classification.

APPENDIX I

An extract of concepts found in fraud reports using LSA

Concept 0	Concept 1	Concept 2	Concept 3
million	able	brooke	million
may	adelphia	insurance	services
business	systems	loans	residences
company	television	gaming	credit
products	december	residences	loans
services	cable television	franchisees	medicaid
financial	adelphia business	credit	interest
year	business solutions	franchise	insurance
operations	adelphia business solutions	alc	group
sales	anicom	loan	sales
Concept 4	**Concept 5**	**Concept 6**	**Concept 7**
products	million	sales	natural
system	residences	products	natural gas
residences	development	customers	gas
sensei	companys	skandia	gaming
development	alc	fiscal	ended
sensei system	living	market	us
may	software	software	development
alc	units	development	cng
living	cme	gaming	ended december
procedures	net	lcos	december
Concept 8	**Concept 9**	**Concept 10**	**Concept 11**
company	business	may	company
development	cme	gaming	gas
software	brooke	cme	natural gas
companys	december	business	natural
assets	advertising	revenues	companys
systems	network	waste	december
tools	cmes	advertising	brooke
loans	may	cmes	income
brooke	skandia	net	increase
credit	stock	fiscal	capital
Concept 12	**Concept 13**	**Concept 14**	**Concept 15**
products	development	fiscal	million
network	services	business	management
sensei	assets	development	february
system	business	revenue	network
sales	results	may	ended february
skandia	cost	services	acquisition
sensei system	us	solutions	year ended february
procedures	costs	value	operating
global	enron	insurance	revenues
development	value	us	ended
Concept 16	**Concept 17**	**Concept 18**	**Concept 19**
services	fiscal	revenue	financial
network	rate	global	ended
company	companys	service	sales
companys	financial	services	based
operating	year ended december	wireless	loan
natural gas	skandia	result	business
capital	result	systems	material
qwest	billion	stock	cash
local	interest	ended	million
cme	average	customer	operating

Table I.1: Concepts from fraud reports uncovered by LSA.

An extract of concepts found in first 100 (1-100) non-fraud reports using LSA

Concept 0	Concept 1	Concept 2	Concept 3
million	products	december	year
company	ended	revenues	us
products	december	ended december	year ended
business	ended december	year ended december	ended
may	health	service	million
sales	year ended	credit	services
us	care	ended	china
december	software	television	communities
year	health care	programming	rmb
operations	year ended december		revenue
Concept 4	**Concept 5**	**Concept 6**	**Concept 7**
operations	company	occidental	business
year	care	occidentals	us
ended	communities	gas	health
may	companys	oil	costs
occidental	cable	care	due
per	living	fon	production
year ended	health	hat	per
cent	oil	red hat	year
per cent	million	ended february	impact
companys	programming	health	compared
Concept 8	**Concept 9**	**Concept 10**	**Concept 11**
us	business	us	medical waste
business	fon	financial	regulated medical
us million	pcs group	development	regulated medical waste
telephone	net	television	hotel
primarily	group	interest	customer
access	health	research	tax
products	cable	growth	waste
lecs	sprints	fon	polo
systems	pcs	amgen	travel
service	factors	rate	year ended
Concept 12	**Concept 13**	**Concept 14**	**Concept 15**
company	business	ervices	sales
million	interest	result	year
china	senior	interest	products
markets	company	occidental	hat
software	customer	cable	red hat
group	expenses	net	ended february
future	november	currency	year ended february
oil	net	due	foreign
result	increase	occidentals	financial
engine	include	approximately million	cable
Concept 16	**Concept 17**	**Concept 18**	**Concept 19**
products	product	oil	million
fiscal	increase	fiscal	net
year	could	production	income
management	costs	sales	operations
customers	year ended	operations	operating
credit	may	net	could
revenue	ended february	products	significant
fujifilm	year ended february	certain	market
operations	hat	exploration	expenses
communities	red hat	income	system

Table I.2: Concepts from 100 (1-100) non-fraud reports uncovered by LSA.

An extract of concepts found from 100 (101-200) non-fraud reports using LSA

Concept 0	Concept 1	Concept 2	Concept 3
million	aep	aep	business
company	products	regulatory	gas
business	generation	assets	insurance
sales	sales	generation	canon
products	electric	trading	chevron
companys	energy	revenues	natural gas
may	gas	costs	natural
operations	regulatory	million	management
services	product	electric	project
market	trading	risk	life
Concept 4	Concept 5	Concept 6	Concept 7
million	products	fiscal	paypal
billion	sales	company	services
market	product	equipment	ebay
cruise	fiscal	systems	segment
fiscal	development	gold	travel
communications	systems	asphalt	company
loans	technologies	software	senior
cruises	service	satellite	transaction
sales	network	segment	payment
products	communications	cash	merchant
Concept 8	Concept 9	Concept 10	Concept 11
services	fiscal	cable	may
customers	company	equipment	market
revenue	percent	canon	cruise
technology	gas	business	asphalt
business	development	asphalt	increase
staffing	production	used	cruises
fiscal	companys	paypal	loans
cdi	care	customers	financial
battery	revenue	group	could
percent	natural gas	also	equipment
Concept 12	Concept 13	Concept 14	Concept 15
million	paypal	company	game
company	company	sales	satellite
cable	ebay	companys	including
companys	companys	growth	interest
primarily	products	canon	echostar
million million	insurance	revenues	programming
applications	brand	cruise	credit
approximately	online	businesses	december
million	cruise	production	service
stock	cruises	cruises	certain
products			
Concept 16	Concept 17	Concept 18	Concept 19
paypal	products	gold	visteon
stock	costs	customers	customers
additional	revenues	satellite	ford
cruise	percent	echostar	visteons
battery	also	fcc	customer
natural	business	programming	revenues
technology	sales	insurance	total
ebay	senior living	customer	years
natural gas	retail	united states	xo
products	may	also	homebuilding

Table I.3: Concepts from 100 (101-200) non-fraud reports uncovered by LSA.

An extract of concepts found in 100 (201-300) non-fraud reports using LSA

Concept 0	Concept 1	Concept 2	Concept 3
million	accenture	products	systems
company	clients	sales	sales
services	services	customers	crude oil
business	fiscal	systems	crude
products	could	technology	natural gas
may	revenues	million	sla
operations	contracts	development	materials
financial	class	business	oil natural
new	waste	market	oil natural gas
could	us	product	technology
Concept 4	**Concept 5**	**Concept 6**	**Concept 7**
clients	fiscal	fiscal	crude oil
accenture	communities	communities	crude
us	operating	systems	level
natural	cash	revenues	disk
net	products	financial	products
fiscal	financial	company	natural
oil	living	income	natural gas
natural gas	software	products	suspension assemblies
revenues	scheringplough	living	suspension
crude	disney	disk	oil natural
Concept 8	**Concept 9**	**Concept 10**	**Concept 11**
million	company	company	company
cable	waste	million	systems
december	products	rmb	costs
level	scheringplough	operations	att
company	operations	waste	revenues
cox	companys	may	products
rate	states	revenues	mainframe
waste	merck	credit	license
due	scheringploughs	companies	distributed
bank	certain	suspension assemblies	services
Concept 12	**Concept 13**	**Concept 14**	**Concept 15**
services	products	cable	may
revenue	cable	systems	december
percent	cox	services	us
suspension	management	cox	communities
assemblies	compared	fcc	revenues
suspension	tax	communities	scheringplough
disney	may	operations	cash
disk	scheringplough	exchange	net
williams	could	level	living
customers	requirements	costs	subject
rate			
Concept 16	**Concept 17**	**Concept 18**	**Concept 19**
million	services	scheringplough	sales
systems	contracts	att	cable
companys	companys	united	industry
medicaid	company	rmb million	agreement
approximately	revenue	result	expenses
include	rmb billion	also	stores
management	systems	company	also
gas	health	software	us
us	risk	sales	scheringplough
medical	bone	credit	health

Table I.4: Concepts from 100 (201-300) non-fraud reports uncovered by LSA.

Terms from Concepts in Fraud Reports Repeating terms removed		
'ability'	'including'	'tools'
'accounting'	'income'	'units'
'acquired'	'increase'	'us'
'acquisition'	'increased'	'use'
'acquisitions'	'industry'	'value'
'activities'	'insurance'	'waste'
'addition'	'interest'	'wireless'
'additional'	'iron ore'	'would'
'advertising'	'living'	'year'
'also'	'loan'	'year ended'
'aluminium'	'loans'	'year ended december'
'amount'	'local'	'year ended february'
'approximately'	'loss'	'years'
'assets'	'management'	'gaming'
'average'	'market'	'gas'
'based'	'marketing'	'general'
'billion'	'markets'	'global'
'business'	'material'	'group'
'business combination'	'may'	'stock'
'business solutions'	'medicaid'	'system'
'cable'	'million'	'systems'
'cable television'	'natural'	'tax'
'capital'	'natural gas'	'technology'
'cash'	'net'	'growth'
'certain'	'network'	'homes'
'changes'	'new'	'television'
'companies'	'number'	'time'
'company'	'operating'	
'companys'	'operations'	
'contract'	'per'	
'control'	'price'	
'cost'	'primarily'	
'costs'	'procedures'	
'could'	'product'	
'credit'	'products'	
'customer'	'purchase'	
'customers'	'rate'	
'debt'	'rates'	
'december'	'related'	
'development'	'required'	
'due'	'requirements'	
'effect'	'residences'	
'ended'	'respectively'	
'ended december'	'result'	
'ended february'	'results'	
'enron'	'results operations'	
'expense'	'revenue'	
'expenses'	'revenues'	
'facility'	'sales'	
'february'	'securities'	
'financial'	'service'	
'financial statements'	'services'	
'fiscal'	'shares'	
'fiscal year'	'significant'	
'franchise'	'software'	
'franchisees'	'solutions'	
'future'	'state'	

Table I.5: Unique terms that constitute the concepts uncovered by LSA for fraud reports.

Terms from Concepts in Non-Fraud Reports **Repeating terms removed**			
‘ability',	'acquisition',	'operations',	'per cent',
'access',	'addition',	'pcs',	'percent',
'activities',	'adverse',	'per',	'polo',
'additional',	'agreement',	'period',	'primarily',
'adversely',	'approximately',	'price',	'production',
'also',	'based',	'product',	'program',
'approximately million',	'billion',	'products',	'property',
'assets',	'business',	'programming',	'purchase',
'bank',	'capital',	'provide',	'rates',
'basis',	'cash',	'rate',	'regulated medical waste',
'bone',	'certain',	'red hat'	'related',
'cable',	'china',	'regulated medical',	'research',
'care',	'clients',	'requirements',	'results operations',
'cent',	'communities',	'result',	'revenue',
'changes',	'company',	'results',	'risk',
'class',	'compared',	'revenues',	'sales',
'common',	'contracts',	'segment',	'senior',
'companies',	'corporation',	'service',	'services',
'companys',	'costs',	'share',	'shares',
'consolidated',	'cox',	'significant',	'sla',
'cooper',	'crude oil',	'software',	'sprints',
'cost',	'currency',	'state',	'statements',
'could',	'customer',	'states',	'stock',
'credit',	'data',	'stores',	'subject',
'crude',	'decrease',	'suspension assemblies',	'suspension',
'currently',	'disk',	'system',	'systems',
'customers',	'distributed',	'tax',	'technology',
'december',	'ended december',	'telephone',	'television',
'development',	'ended',	'time',	'total',
'disney',	'exchange',	'travel',	'united',
'due',	'expense',	'us million',	'us',
'ended february',	'exploration',	'use',	'value',
'engine',	'factors',	'waste',	'well',
'expected',	'financial statements',	'year ended december',	'would',
'expenses',	'fiscal',	'year ended',	'year ended february',
'facility',	'foreign',	'years',	'year',
'fcc',	'future',	'occidental',	'mainframe',
'financial',	'general',	'occidentals',	'market',
'fon',	'group',	'merck',	'material',
'fujifilm',	'hat',	'million',	
'gas',	'health',	'natural',	
'generally',	'homebuilding',	'new',	
'growth',	'impact',	'million million',	
'health care',	'include',	'natural gas',	
'higher'}	'income',	'net',	
'hotel',	'increased',	'november',	
'inc',	'information',	'medical waste',	
'including',	'interest',	'medical',	
'increase',	'lecs',	may	
'industry',	'license',		
'insurance',	'loan',		
'investment',	'local',		
'level',	'lower',		
'living',	'management',		
'loans',	'markets',		
'loss',	'materials',		

Table I.6: Unique terms that constitute the concepts uncovered by LSA for non-fraud reports.

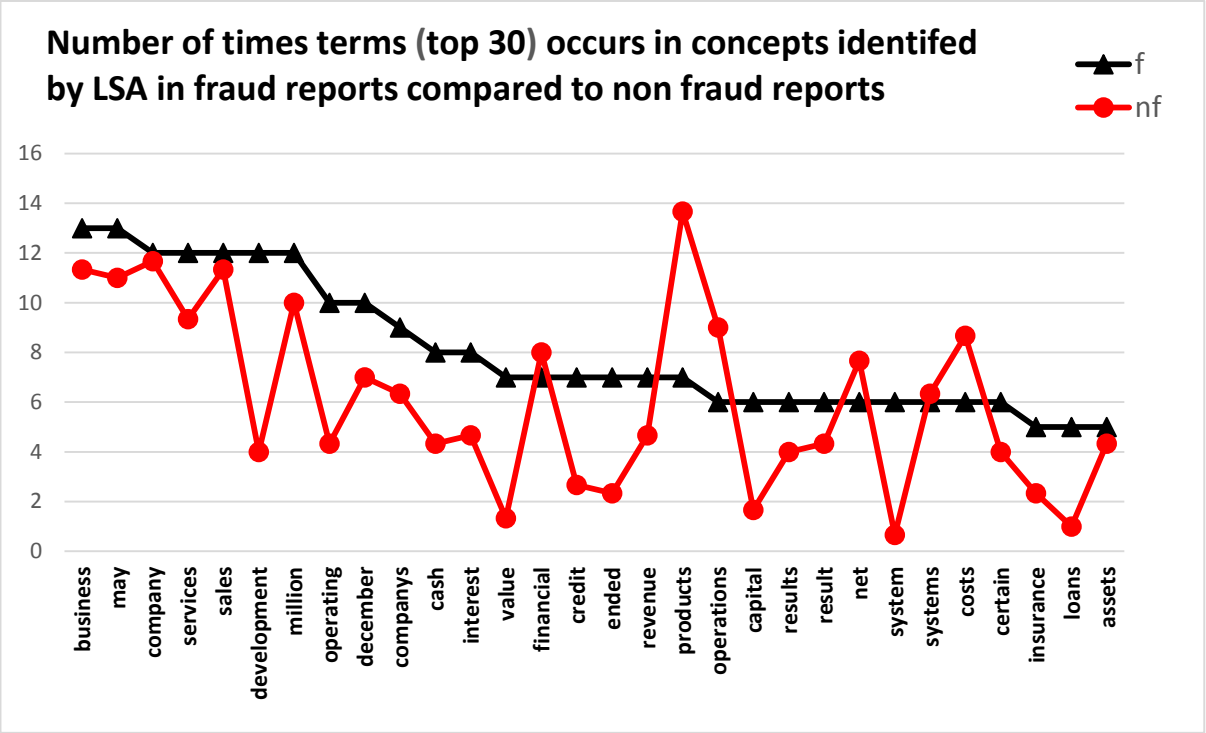


Figure I.1: Terms from concepts more prominent in fraud reports compared with non-fraud reports.

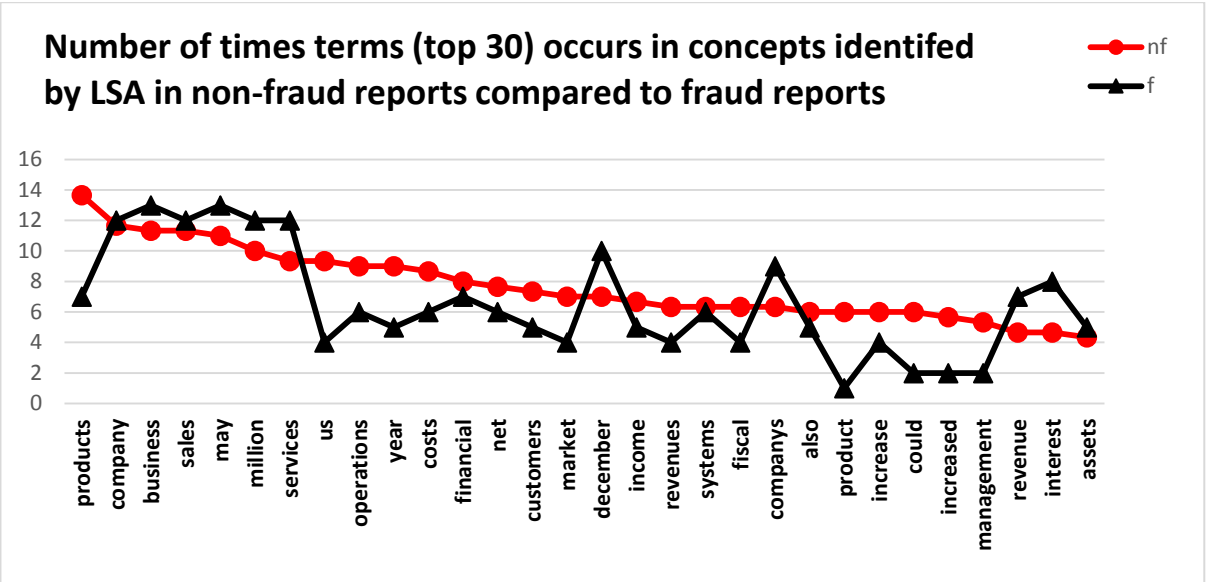
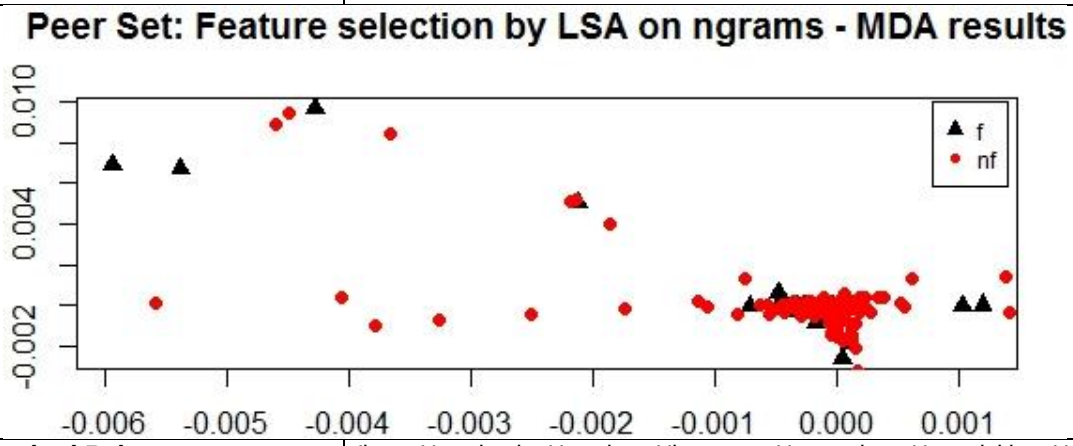


Figure I.2: Terms from concepts more prominent in non-fraud reports compared with fraud reports.

APPENDIX J

Table J.1: Fraud and non-fraud reports compared using MDA based on LSA terms that constitute concepts.

Document Representation	Chosen Features
Peer Set: ngrams	'loans','production','products','insurance','procedures','acquisitions','operations','management','debt','decrease','result','revenue','securities','certain','cash','interest','may','required','amount','net','communities','decrease','lower','compared','consolidated','total','operating','respectively','accounting','effect'



Matched Pair: ngrams	'loans','production','products','insurance','procedures','acquisitions','operations','management','debt','decrease','result','revenue','securities','certain','cash','interest','may','required','amount','net','communities','decrease','lower','compared','consolidated','total','operating','respectively','accounting','effect'

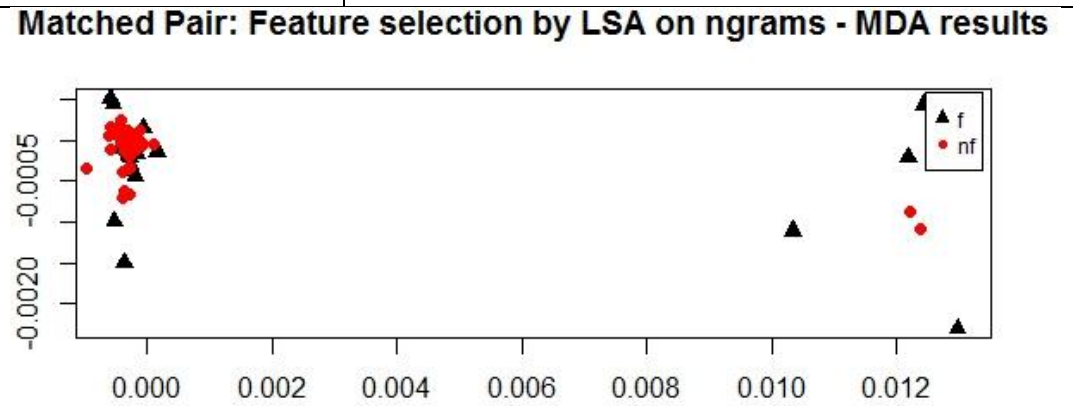


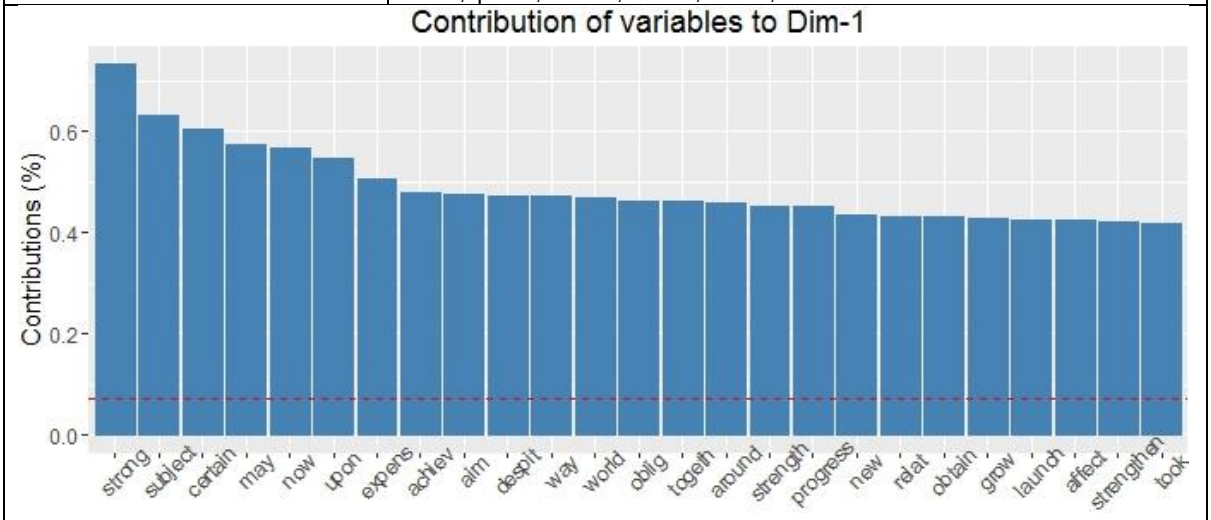
Table J.1: Fraud and non-fraud reports compared using MDA based on LSA terms that constitute concepts.

APPENDIX K

APPENDIX K

Feature Selection using PCA, graphed based on MDA computation

Document Representation	Chosen Features
Peer Set: Unigrams	'strong', 'subject', 'certain', 'may', 'now', 'upon', 'expens', 'achiev', 'aim', 'despit', 'way', 'world', 'oblig', 'togeth', 'around', 'strength', 'progress', 'new', 'relat', 'obtain', 'grow', 'launch', 'affect', 'strengthen', 'took', 'great', 'requir', 'addit', 'portion', 'enjoy', 'incur', 'import', 'avail', 'group', 'determin', 'move', 'toward', 'restrict', 'bring', 'start', 'primarili', 'peopl', 'top', 'advers', 'alreadi', 'liabil', 'profit', 'effect', 'excel', 'clear',



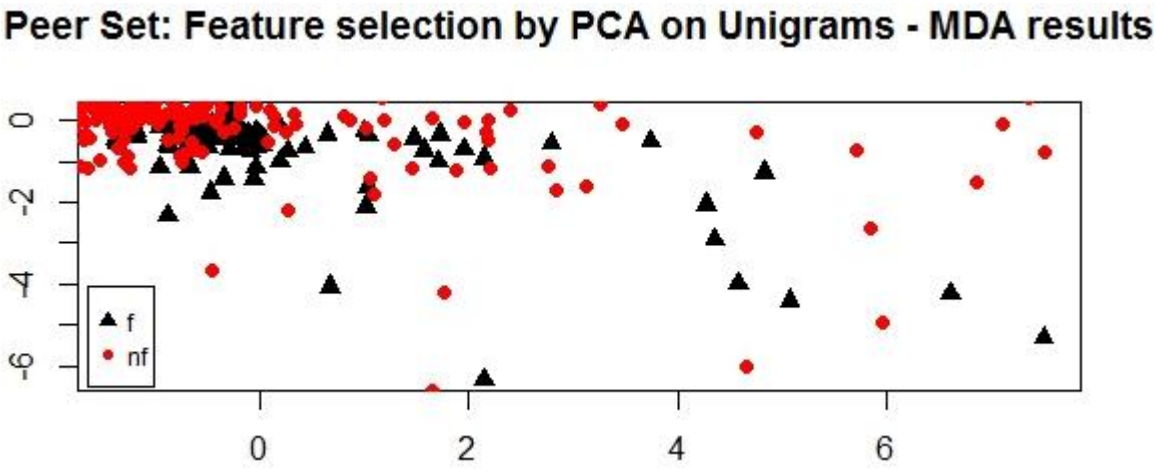
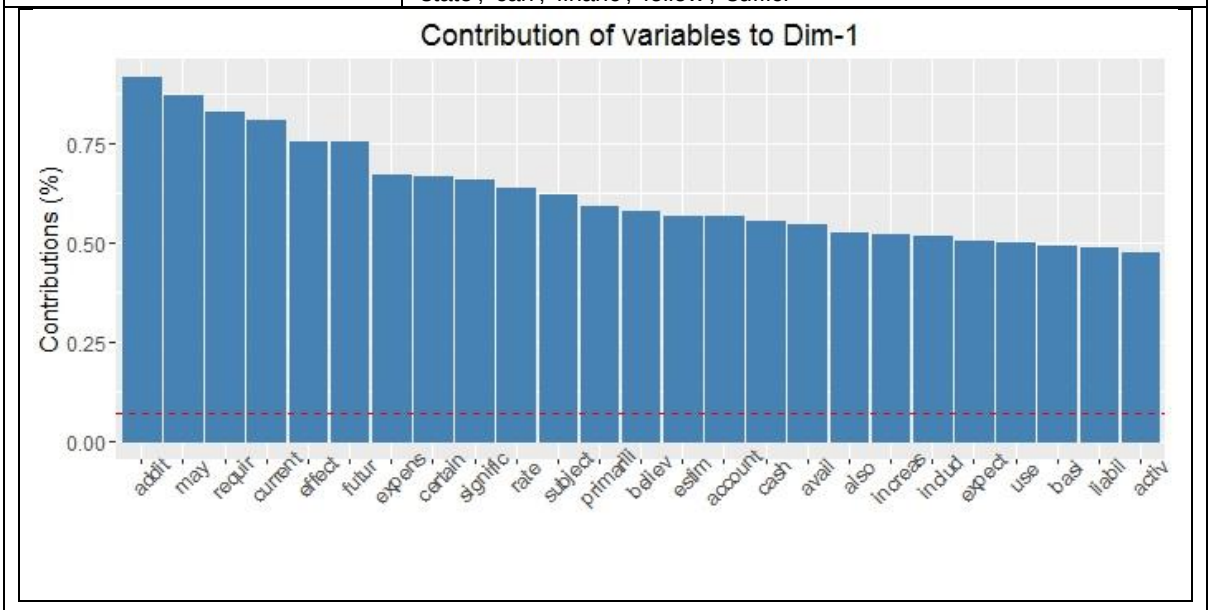


Table K.1: Unigrams chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Unigrams	'addit', 'may', 'requir', 'current', 'effect', 'futur', 'expens', 'certain', 'signific', 'rate', 'subject', 'primarili', 'believ', 'estim', 'account', 'cash', 'avail', 'also', 'increas', 'includ', 'expect', 'use', 'basi', 'liabil', 'activ', 'base', 'revenu', 'respect', 'sell', 'oblig', 'obtain', 'upon', 'loss', 'period', 'purpos', 'need', 'tax', 'approxim', 'actual', 'credit', 'general', 'decreas', 'due', 'portion', 'third', 'state', 'can', 'financ', 'follow', 'suffici'



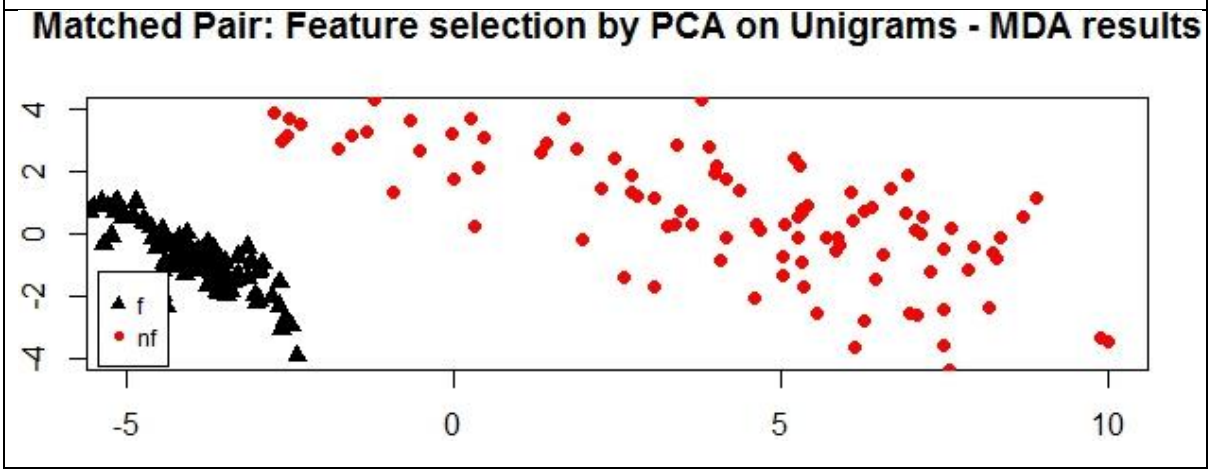
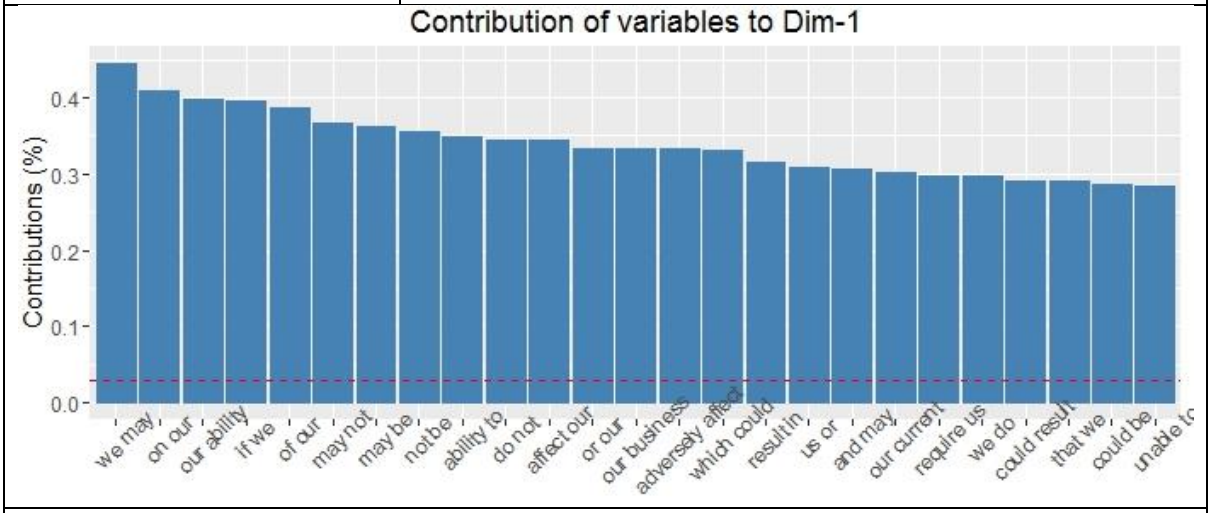


Table K.2: Unigrams chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Bigrams	'we may', 'on our', 'our ability', 'if we', 'of our', 'may not', 'may be', 'not be', 'ability to', 'do not', 'affect our', 'or our', 'our business', 'adversely affect', 'which could', 'result in', 'us or', 'and may', 'our current', 'require us', 'we do', 'could result', 'that we', 'could be', 'unable to', 'we believe', 'or other', 'and our', 'although we', 'that our', 'our revenues', 'fail to', 'be able', 'our operating', 'could adversely', 'are unable', 'and could', 'our financial', 'our future', 'subject to', 'our stock', 'we could', 'longterm debt ', 'could have', 'to obtain', 'our competitors', 'we cannot', 'our cash', 'our products', 'liquidity and



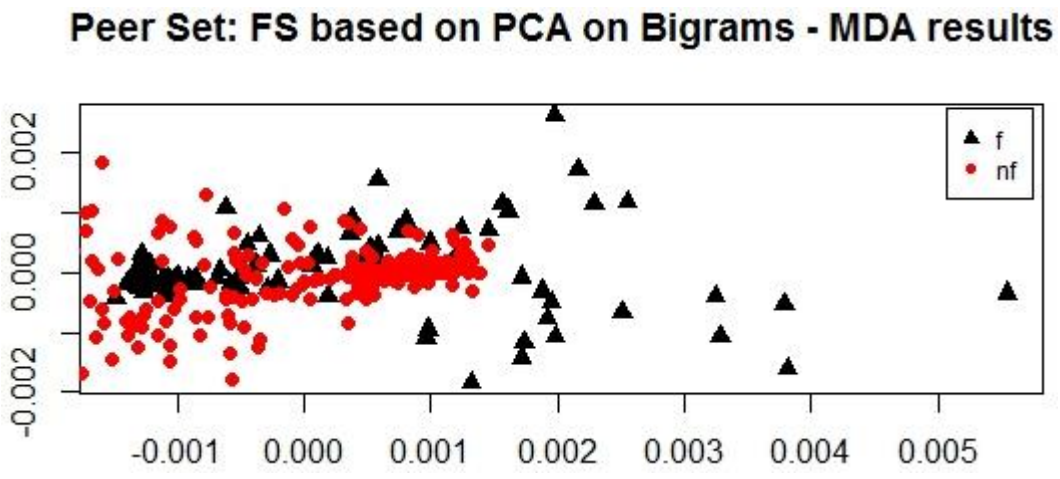
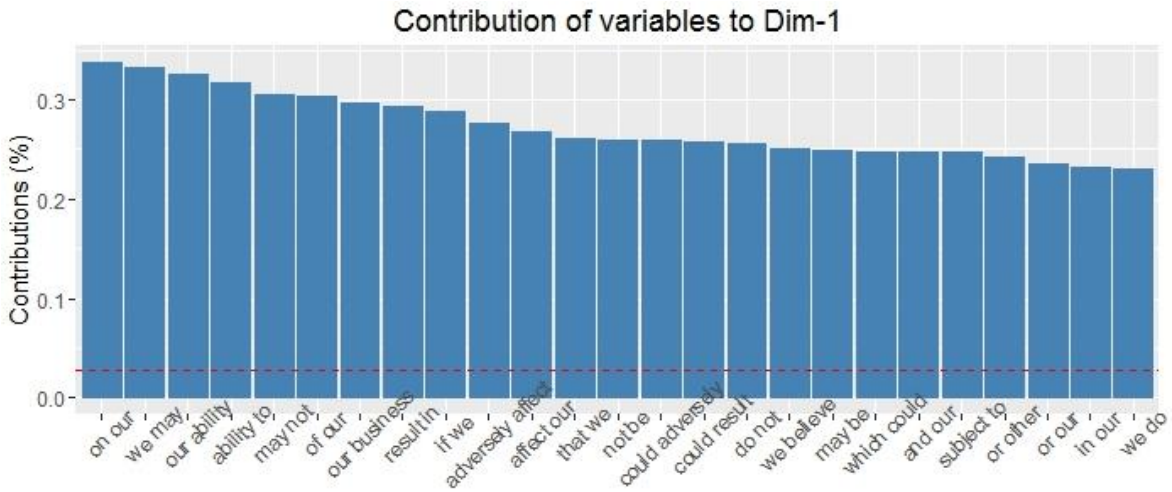


Table K.3: Bigrams chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Bigrams	'on our', 'we may', 'our ability', 'ability to', 'may not', 'of our', 'our business', 'result in', 'if we', 'adversely affect', 'affect our', 'that we', 'believes that', 'could adversely', 'could result', 'do not', 'we believe', 'may be', 'which could', 'and our', 'subject to', 'operating activities', 'or our', 'in our', 'we do', 'and may', 'are subject', 'could have', 'our current', 'that our', 'the outstanding', 'could cause', 'our financial', 'be able', 'our operating', 'could be', 'and could', 'may result', 'upon our', 'of operations', 'for our', 'interest rate', 'cash flow', 'require us', 'although we', 'we currently', 'if our', 'to obtain', 'results of', 'reduce our',



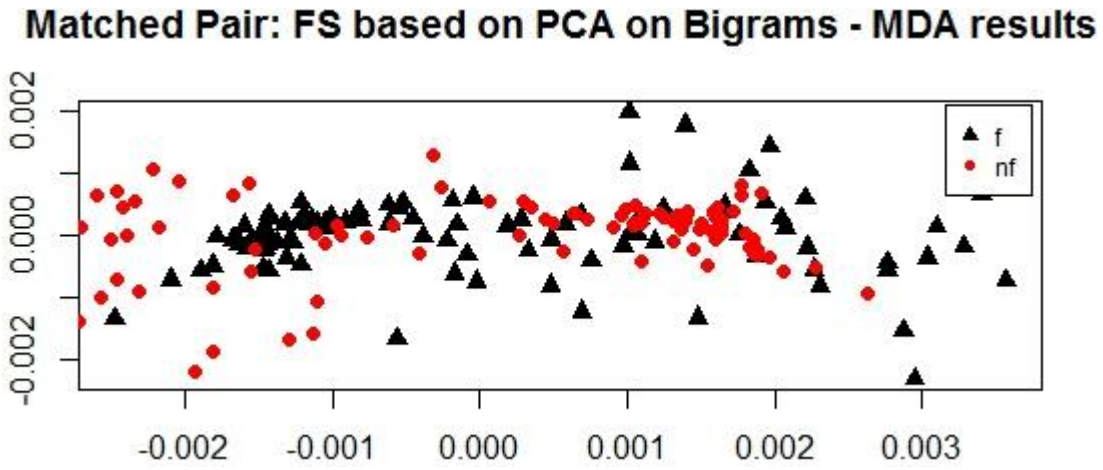
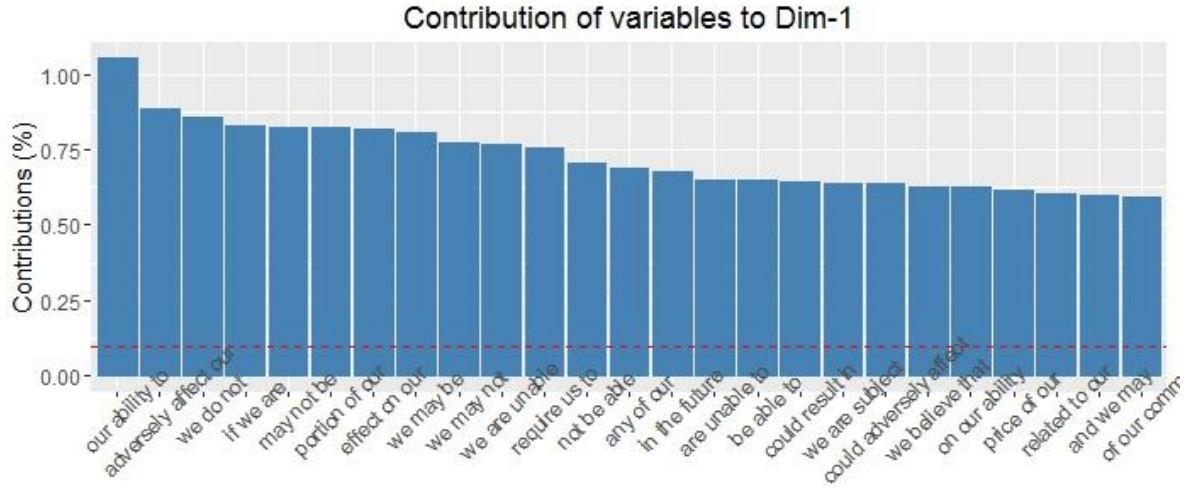


Table K.4: Bigrams chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Trigrams	'our ability to', 'adversely affect our', 'we do not', 'if we are', 'may not be', 'portion of our', 'effect on our', 'we may be', 'we may not', 'we are unable', 'require us to', 'not be able', 'any of our', 'in the future', 'are unable to', 'be able to', 'could result in', 'we are subject', 'could adversely affect', 'we believe that', 'on our ability', 'price of our', 'related to our', 'and we may', 'of our common', 'affect our business', 'in addition we', 'we are not', 'all of our', 'affect our ability', 'we fail to', 'our results of', 'our operating results', 'on our business', 'some of our', 'of our products', 'we are required', 'if we do', 'in addition our', 'our common stock', 'we could be', 'beyond our control', 'if any of', 'if we fail', 'of our competitors', 'the market price', 'be adversely affected', 'cause us to', 'that we will', 'be unable to',



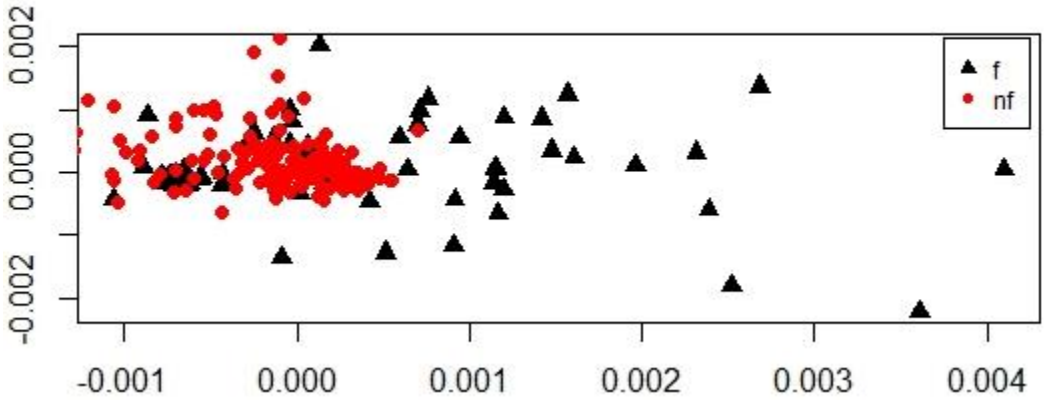
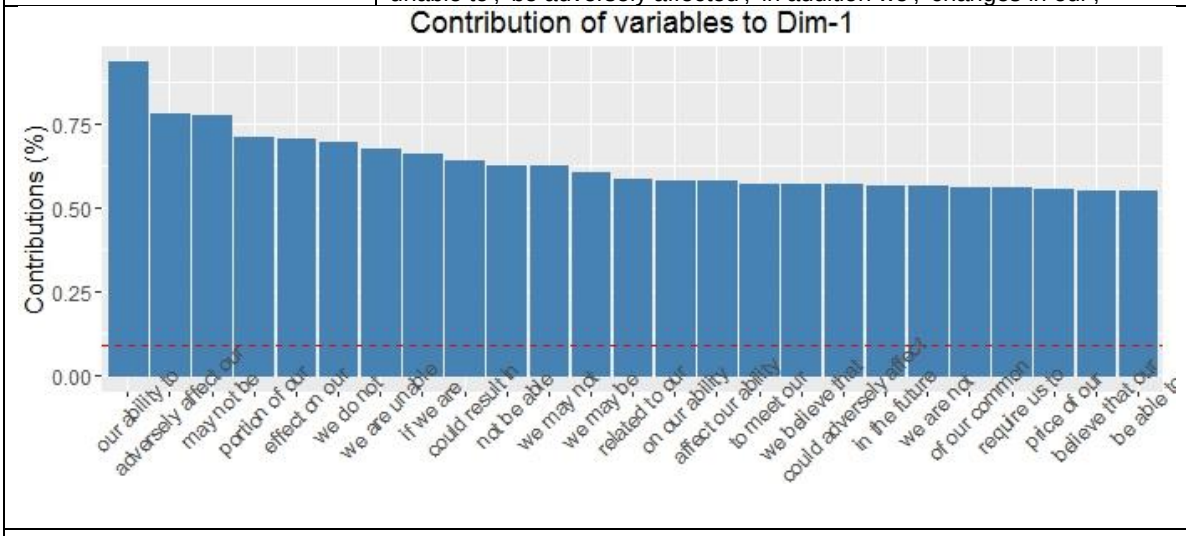


Table K.5: Trigrams chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Trigrams	'our ability to', 'adversely affect our', 'may not be', 'portion of our', 'effect on our', 'we do not', 'we are unable', 'if we are', 'could result in', 'not be able', 'we may not', 'we may be', 'related to our', 'on our ability', 'affect our ability', 'to meet our', 'we believe that', 'could adversely affect', 'in the future', 'we are not', 'of our common', 'require us to', 'price of our', 'believe that our', 'be able to', 'we are subject', 'that we will', 'are subject to', 'any of our', 'we fail to', 'our operating results', 'limit our ability', 'all of our', 'which could adversely', 'if we fail', 'we are required', 'results of operations', 'to our business', 'affect our business', 'our results of', 'our business and', 'cause us to', 'of our competitors', 'our common stock', 'may result in', 'which could have', 'are unable to', 'be adversely affected', 'in addition we', 'changes in our',



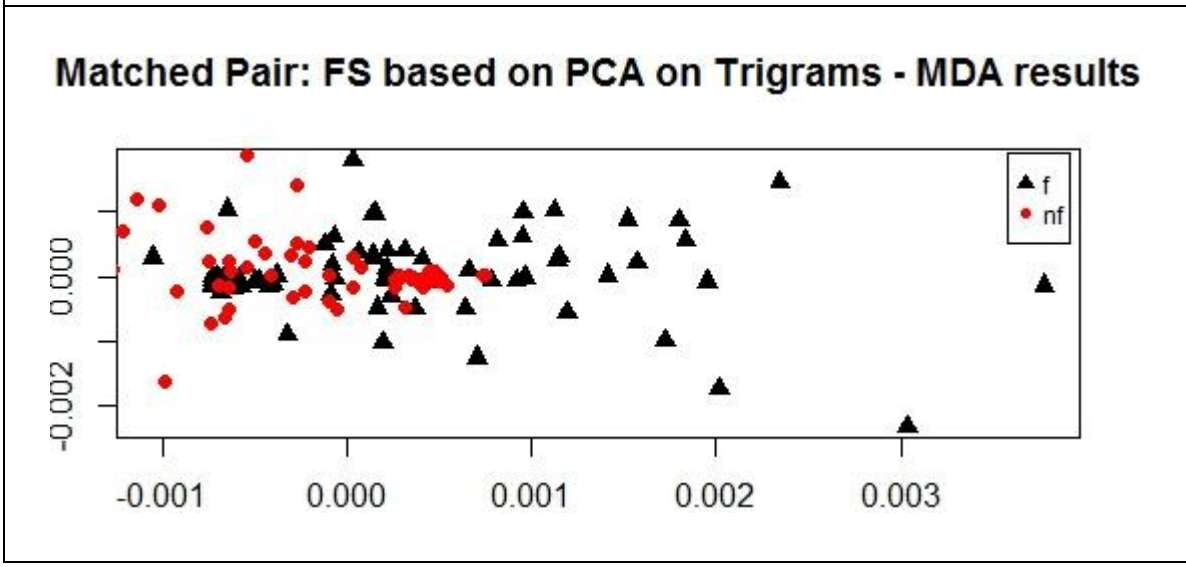


Table K.6: Trigrams chosen by PCA for matched pair data set up and results of MDA computation.

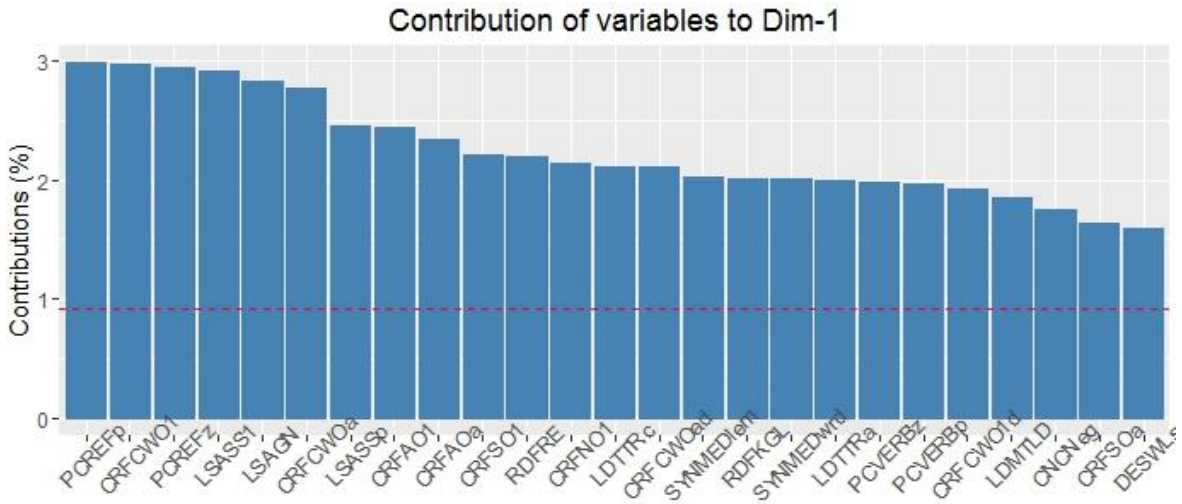
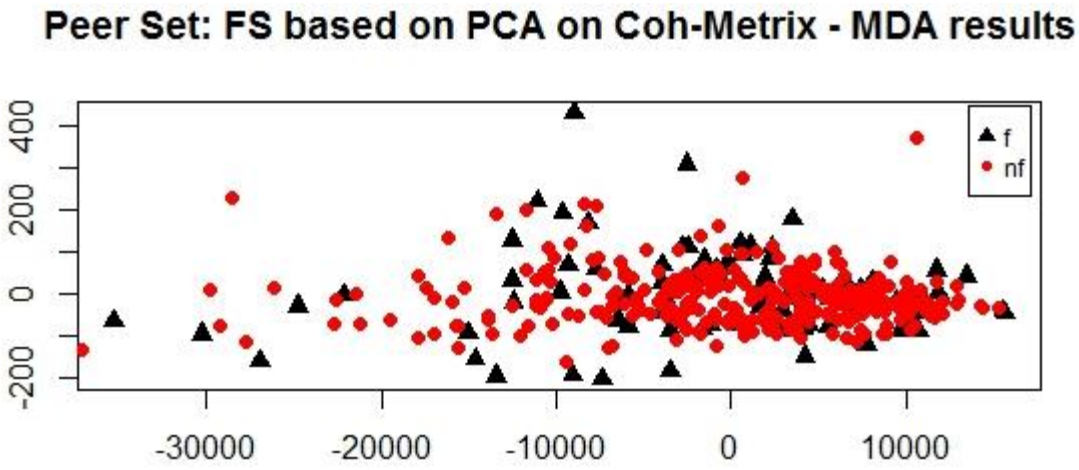
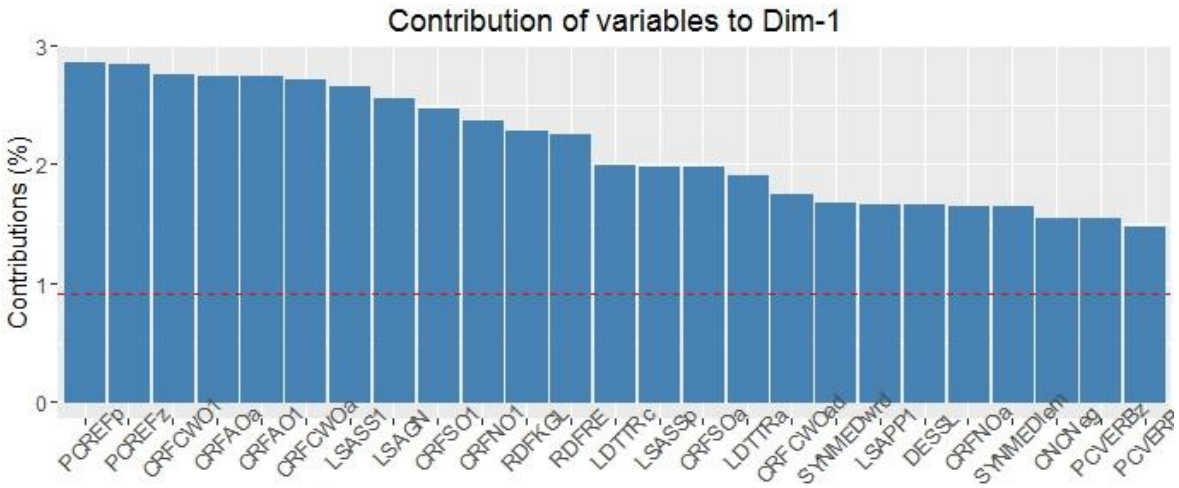
Document Representation	Chosen Features
Peer Set: Coh-Metrix	'PCREFp', 'CRFCWO1', 'PCREFz', 'LSASS1', 'LSAGN', 'CRFCWOa', 'LSASSp', 'CRFAO1', 'CRFAOa', 'CRFSO1', 'RDFRE', 'CRFNO1', 'LDTTRc', 'CRFCWOad', 'SYNMEDlem', 'RDFKGL', 'SYNMEDwrd', 'LDTTRa', 'PCVERBz', 'PCVERBp', 'CRFCWO1d', 'LDMTLD', 'CNCNeg', 'CRFSOa', 'DESWLsy', 'WRDFRQc', 'CNCADC', 'WRDAOAc', 'CRFNOa', 'LSASSpd', 'WRDMEAc', 'DESWC', 'WRDPOLc', 'WRDFAMc', 'CNCAll', 'CNCLogic', 'PCCONNz', 'PCCONNp', 'DESWLsyd', 'DRNEG', 'CNCAdd', 'DESWLltd', 'LSASS1d', 'WRDIMGc', 'DESSL', 'DESWLlt', 'WRDFRQa', 'WRDHYPn', 'DESSC', 'SMCAUSv'
	
	

Table K.7: Coh-Metrix Indices chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Coh-Metrix	'PCREFp', 'CRFCWO1', 'PCREFz', 'LSASS1', 'LSAGN', 'CRFCWOa', 'LSASSp', 'CRFAO1', 'CRFAOa', 'CRFSO1', 'RDFRE', 'CRFNO1', 'LDTTRc', 'CRFCWOad', 'SYNMEDlem', 'RDFKGL', 'SYNMEDwrd', 'LDTTRa', 'PCVERBz', 'PCVERBp', 'CRFCWO1d', 'LDMTLD', 'CNCNeg', 'CRFSOa', 'DESWLsy', 'WRDFRQc', 'CNCADC', 'WRDAOAc', 'CRFNOa', 'LSASSpd', 'WRDMEAc', 'DESWC', 'WRDPOLc', 'WRDFAMc', 'CNCAll', 'CNCLogic', 'PCCONNz', 'PCCONNp', 'DESWLsyd', 'DRNEG', 'CNCAdd', 'DESWLltd', 'LSASS1d', 'WRDIMGc', 'DESSL', 'DESWLlt', 'WRDFRQa', 'WRDHYPn', 'DESSC', 'SMCAUSv'



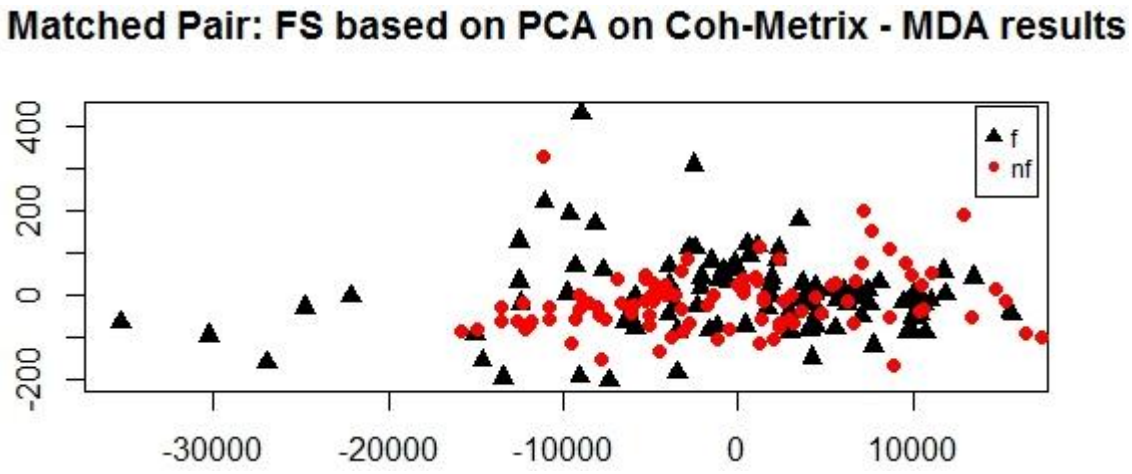
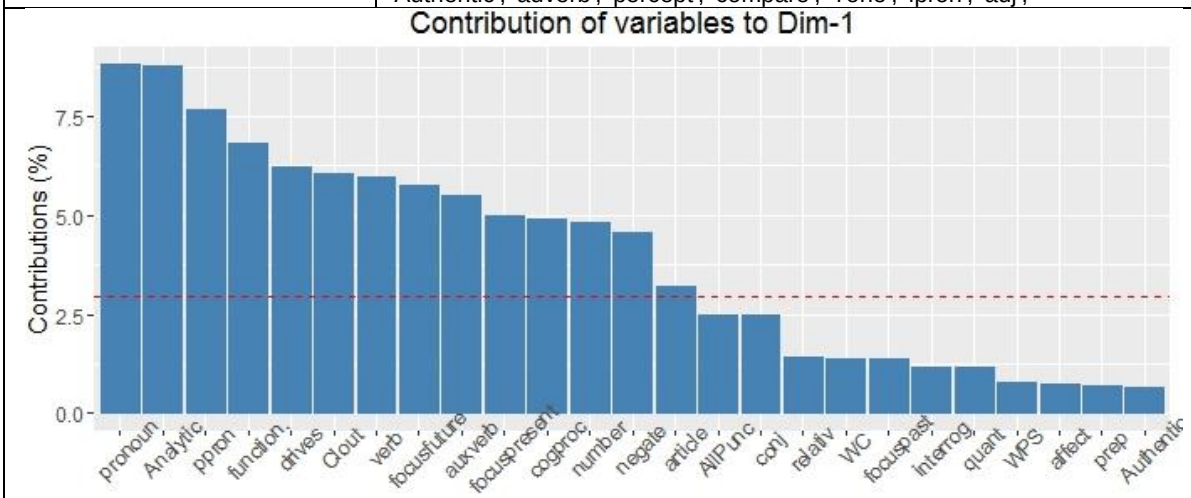


Table K.8: Coh-Metrix Indices chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: LIWC 2015	'pronoun', 'Analytic', 'ppron', 'function.', 'drives', 'Clout', 'verb', 'focusfuture', 'auxverb', 'focuspresent', 'cogproc', 'number', 'negate', 'article', 'AllPunc', 'conj', 'relativ', 'WC', 'focuspast', 'interrog', 'quant', 'WPS', 'affect', 'prep', 'Authentic', 'adverb', 'percept', 'compare', 'Tone', 'ipron', 'adj',



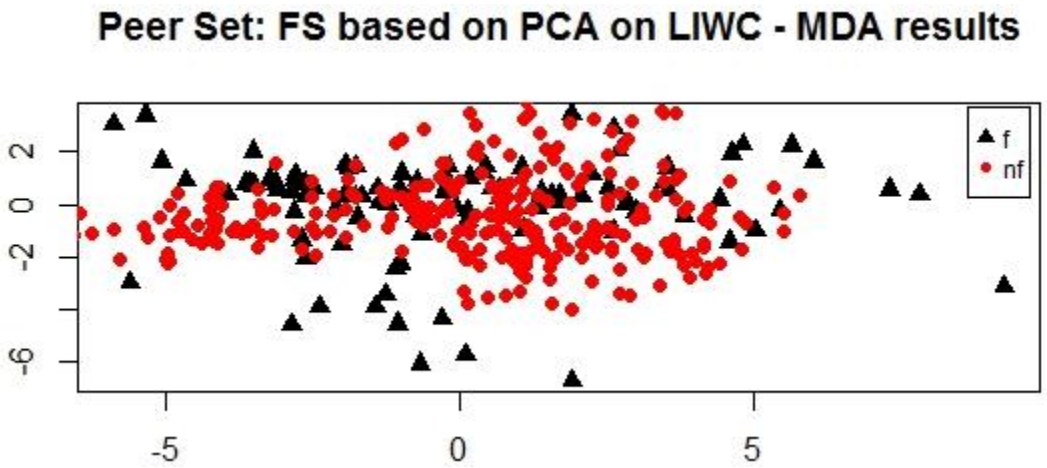
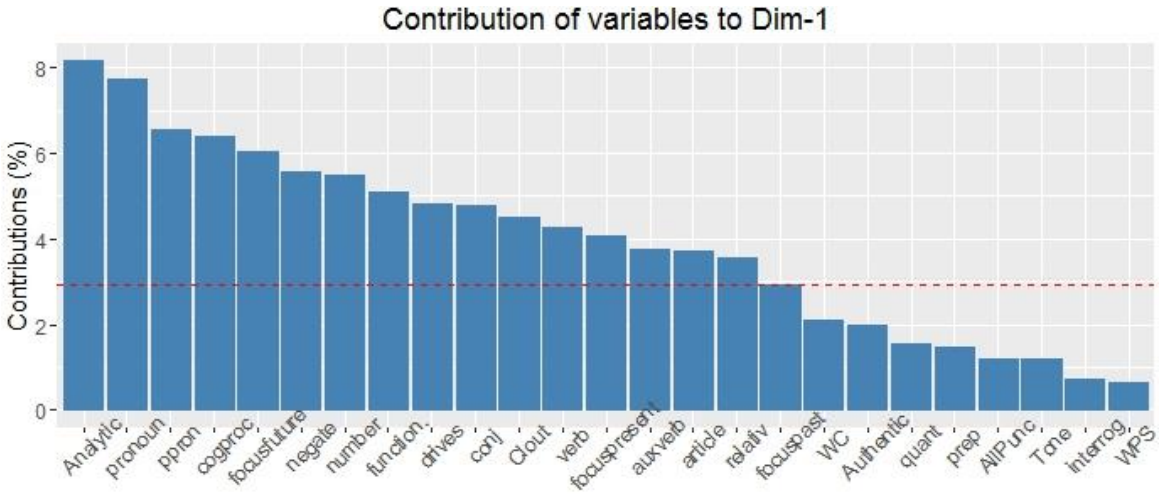


Table K.9: LIWC variables chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: LIWC	'Analytic', 'pronoun', 'ppron', 'cogproc', 'focusfuture', 'negate', 'number', 'function.', 'drives', 'conj', 'Clout', 'verb', 'focuspresent', 'auxverb', 'article', 'relativ', 'focuspast', 'WC', 'Authentic', 'quant', 'prep', 'AllPunc', 'Tone', 'interrog', 'WPS', 'percept', 'compare', 'ipron', 'Tone', 'ipron', 'adj',



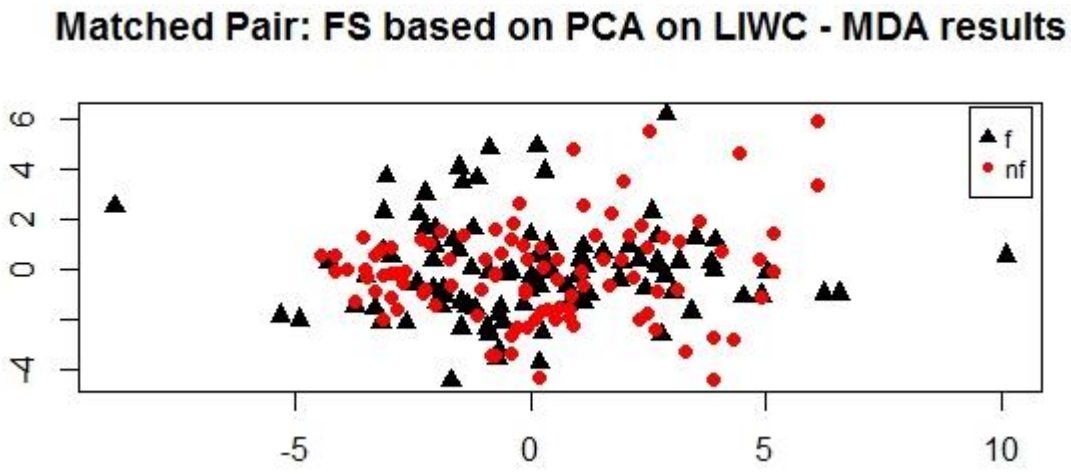
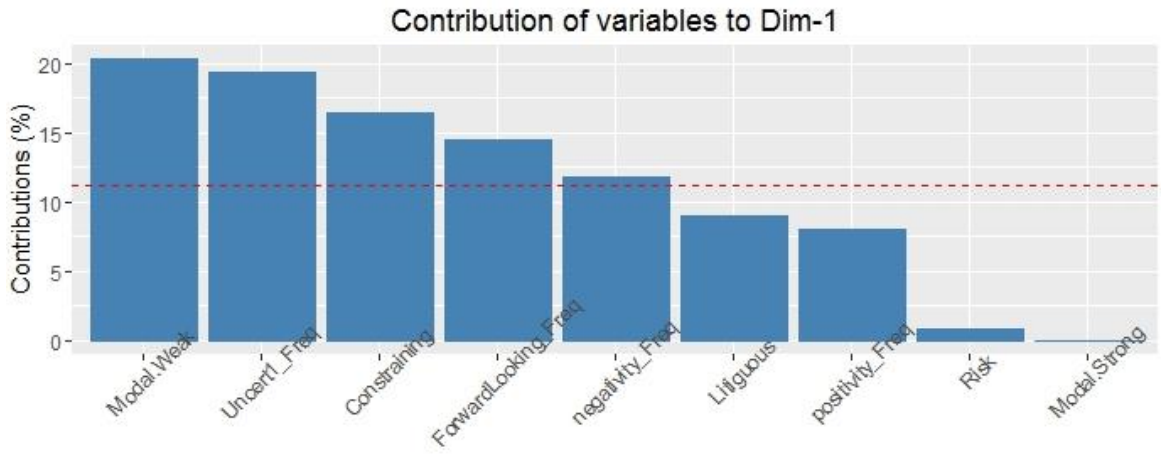


Table K.10: LIWC variables chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Custom Dictionaries	'Modal.Weak', 'Uncert1_Freq', 'Constraining', 'ForwardLooking_Freq', 'negativity_Freq', 'Litiguous', 'positivity_Freq', 'Risk',



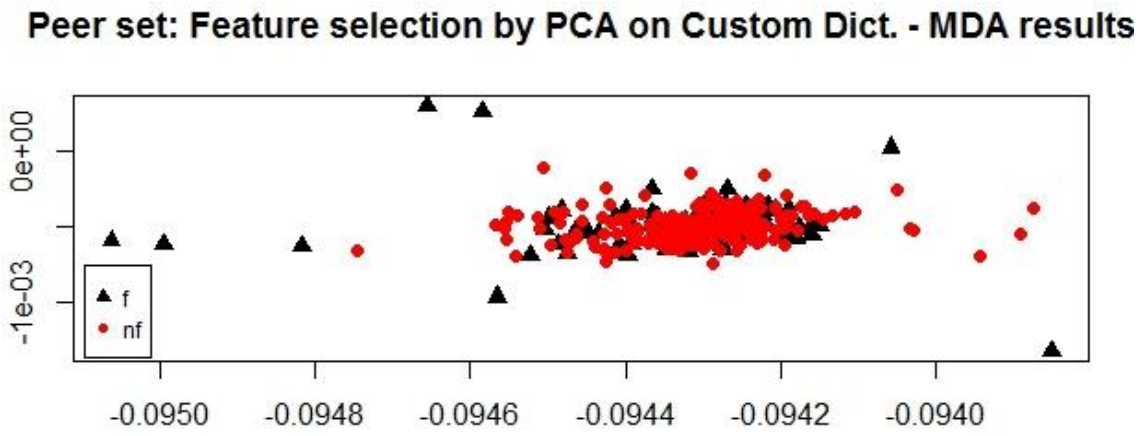
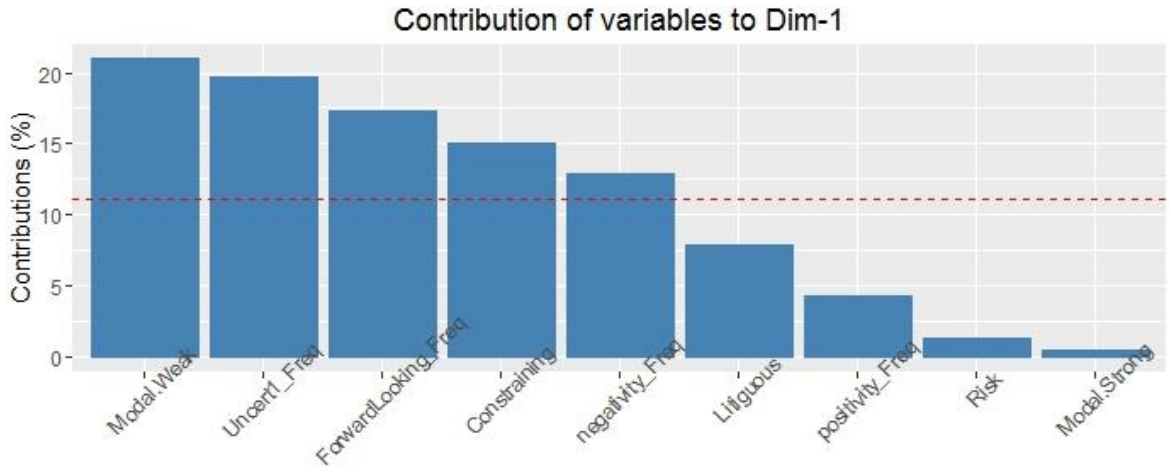


Table K.11: Custom word lists chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair : Custom Dictionaries	'Modal.Weak', 'Uncert1_Freq', 'Constraining', 'ForwardLooking_Freq', 'negativity_Freq', 'Litiguous', 'positivity_Freq', 'Risk',



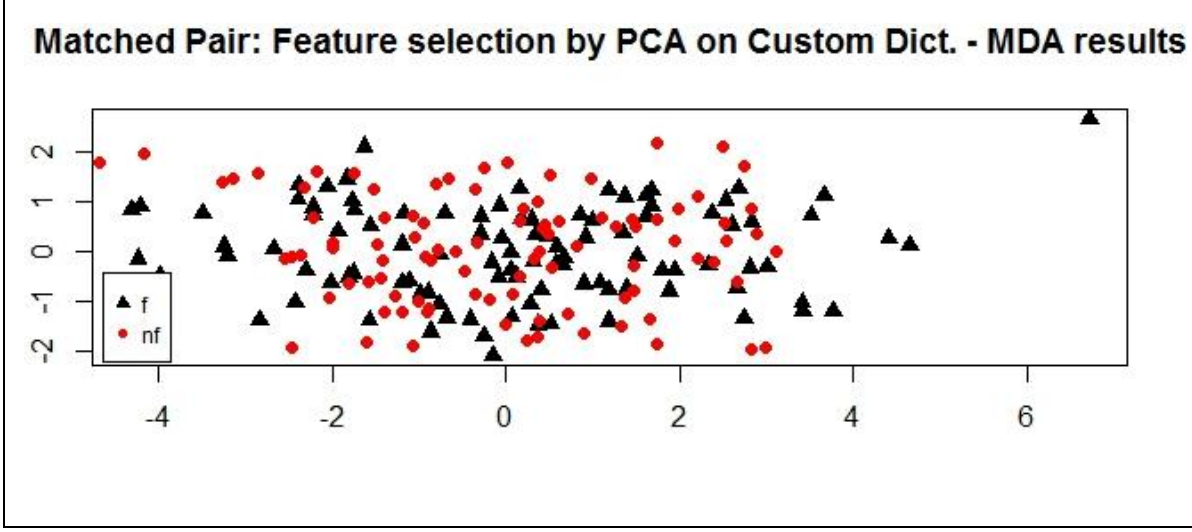
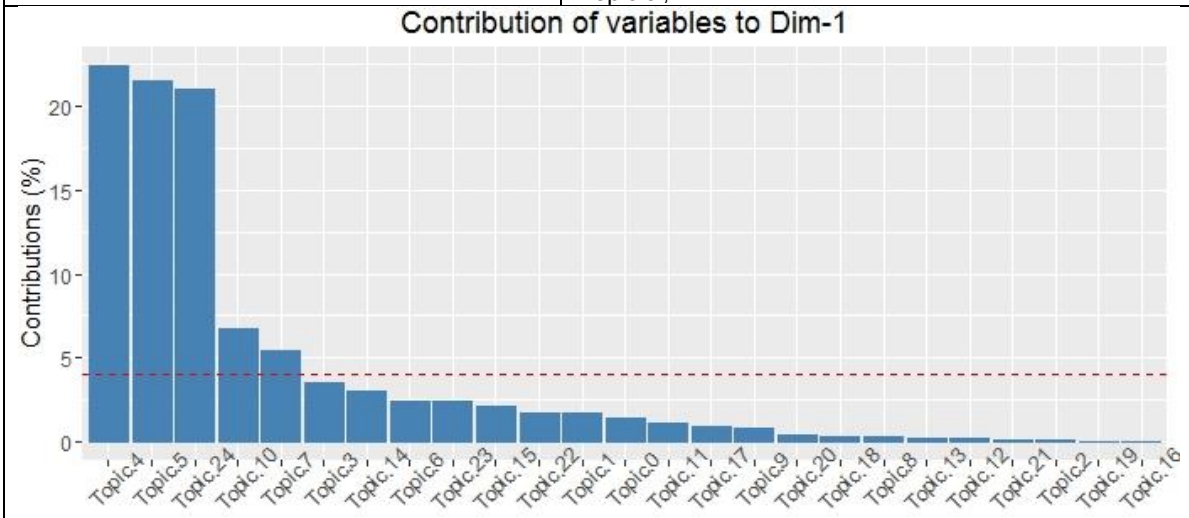


Table K.12: Custom word lists chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Topic Modelling	'Topic.4', 'Topic.5', 'Topic.24', 'Topic.10', 'Topic.7', 'Topic.3', 'Topic.14', 'Topic.6', 'Topic.23', 'Topic.15', 'Topic.22', 'Topic.1', 'Topic.0', 'Topic.11', 'Topic.17', 'Topic.9',



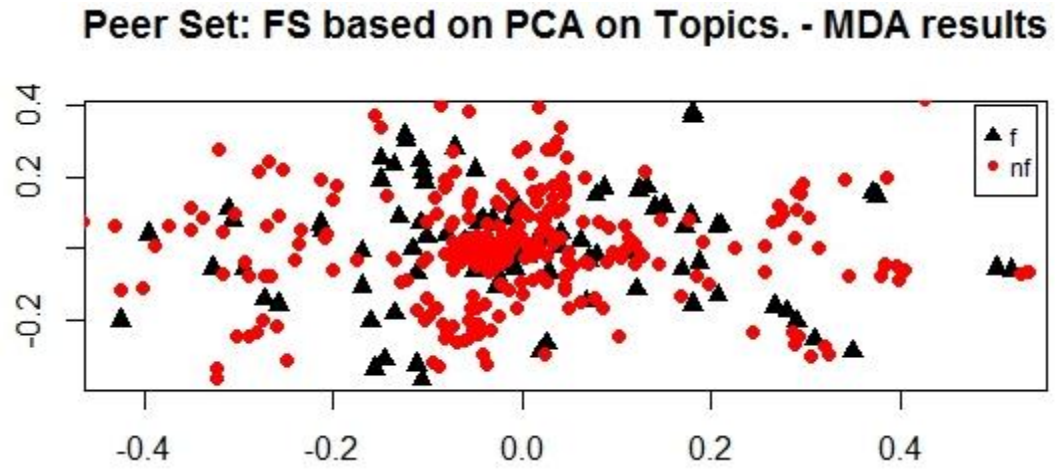
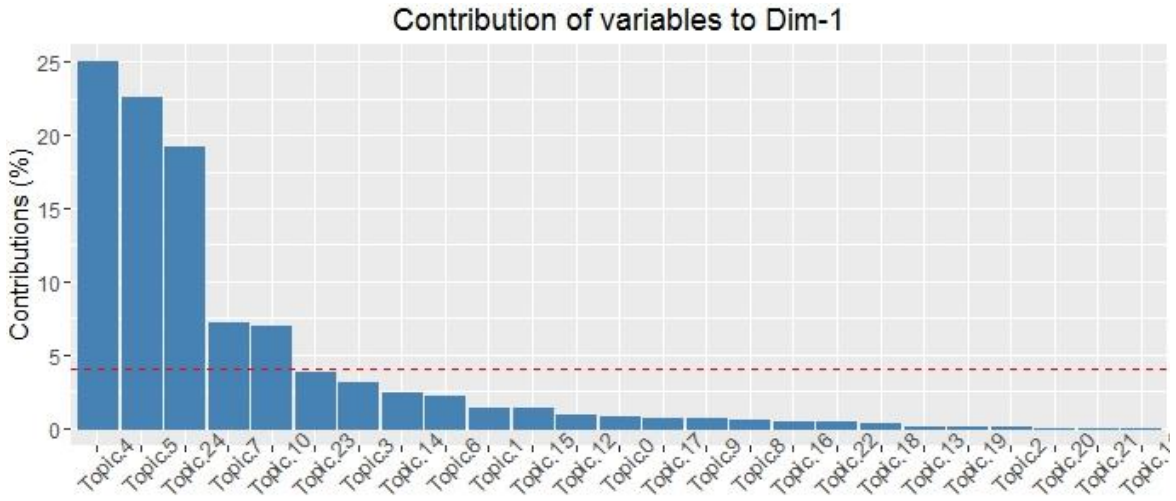


Table K.13: Topics chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Topic Modelling	'Topic.4', 'Topic.5', 'Topic.24', 'Topic.7', 'Topic.10', 'Topic.23', 'Topic.3', 'Topic.14', 'Topic.6', 'Topic.1', 'Topic.15', 'Topic.12', 'Topic.0', 'Topic.11', 'Topic.17', 'Topic.9',



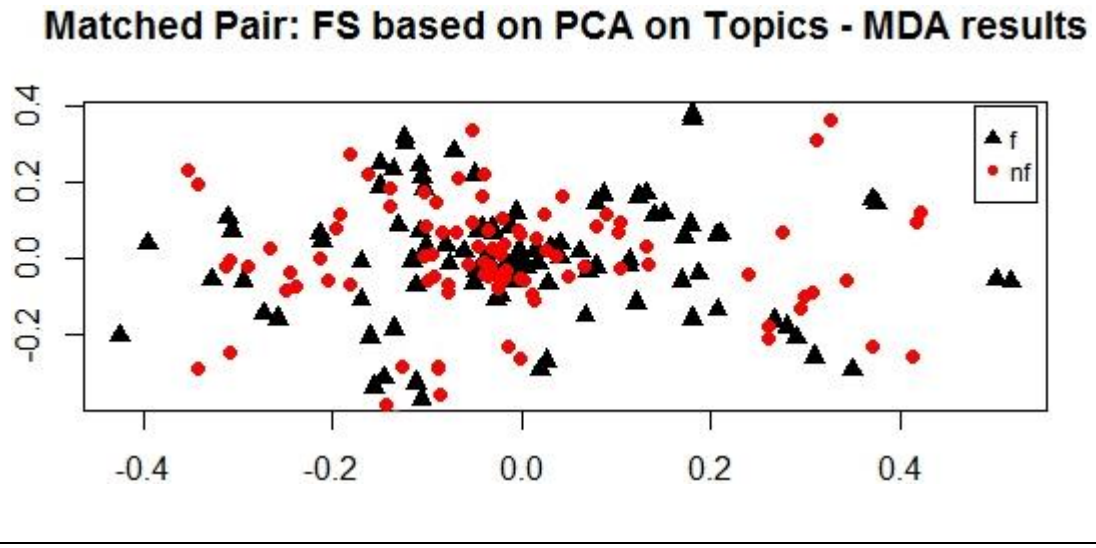
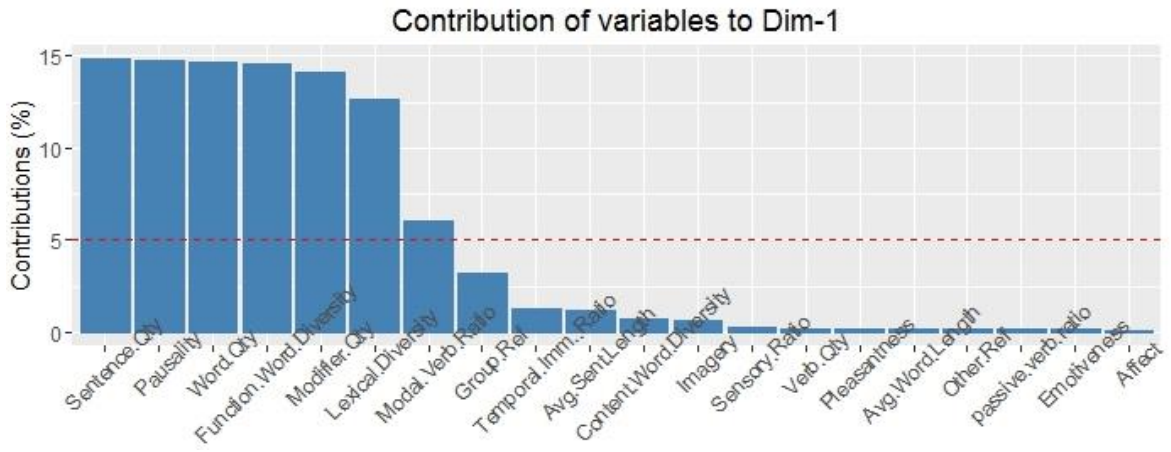


Table K.14: Topics chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Linguistics Based Cues	'Sentence.Qty', 'Pausality', 'Word.Qty', 'Function.Word.Diversity', 'Modifier.Qty', 'Lexical.Diversity', 'Modal.Verb.Ratio', 'Group.Ref', 'Temporal.Imm..Ratio', 'Avg.Sent.Length', 'Content.Word.Diversity', 'Imagery', 'Sensory.Ratio', 'Verb.Qty', 'Pleasantness', 'Avg.Word.Length',



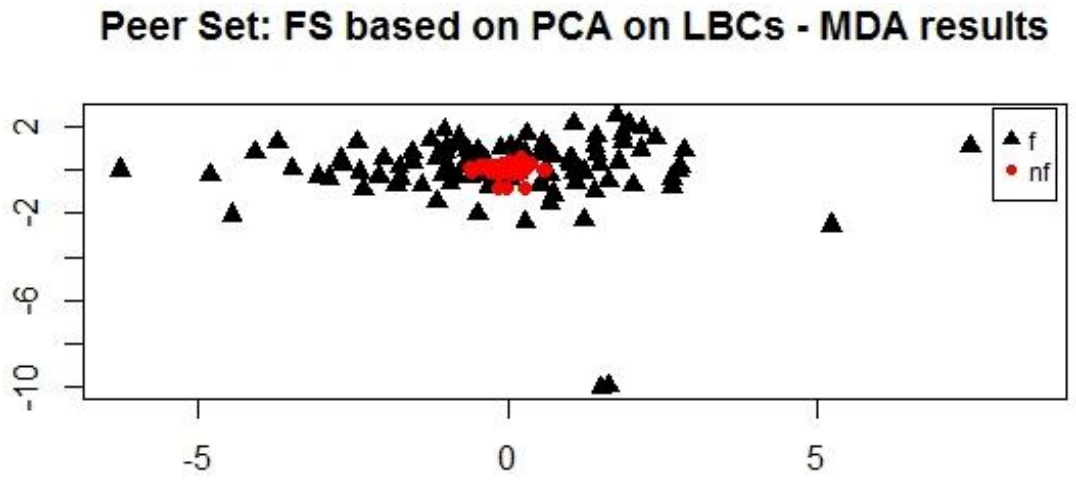
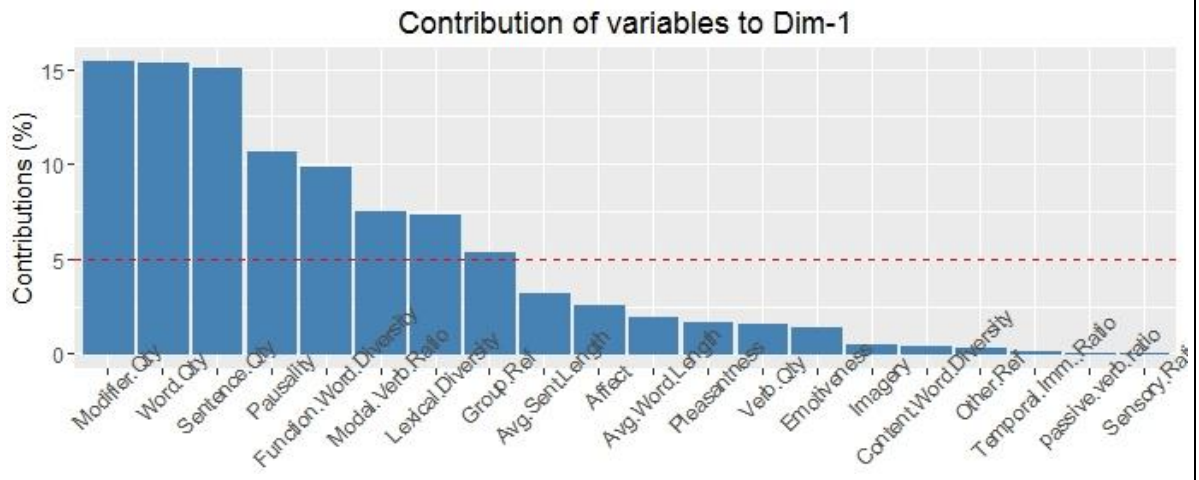


Table K.15: LBCs chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Linguistics Based Cues	'Modifier.Qty', 'Word.Qty', 'Sentence.Qty', 'Pausality', 'Function.Word.Diversity', 'Modal.Verb.Ratio', 'Lexical.Diversity', 'Group.Ref', 'Avg.Sent.Length', 'Affect', 'Avg.Word.Length', 'Pleasantness', 'Verb.Qty','Emotiveness','Imagery', 'Content.Word.Diversity',



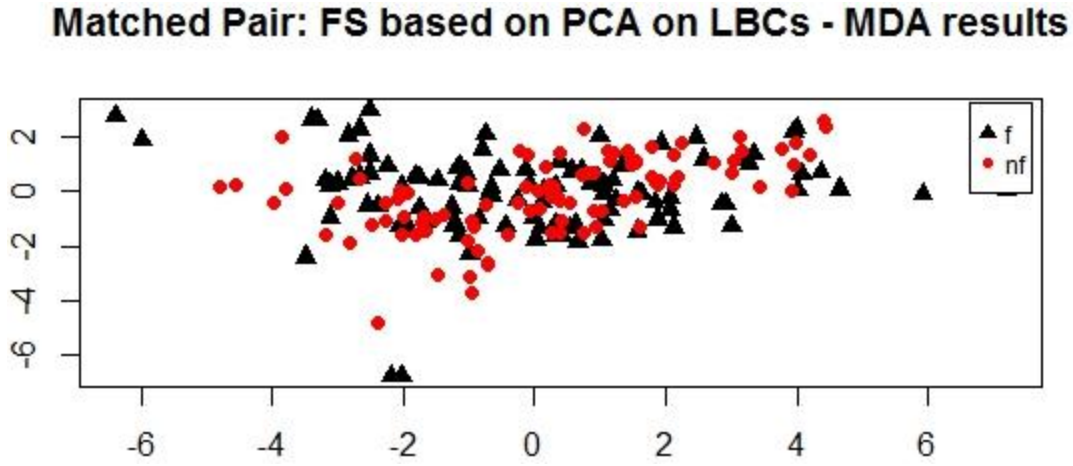


Table K.16: LBCs chosen by PCA for matched pair data set up and results of MDA computation.

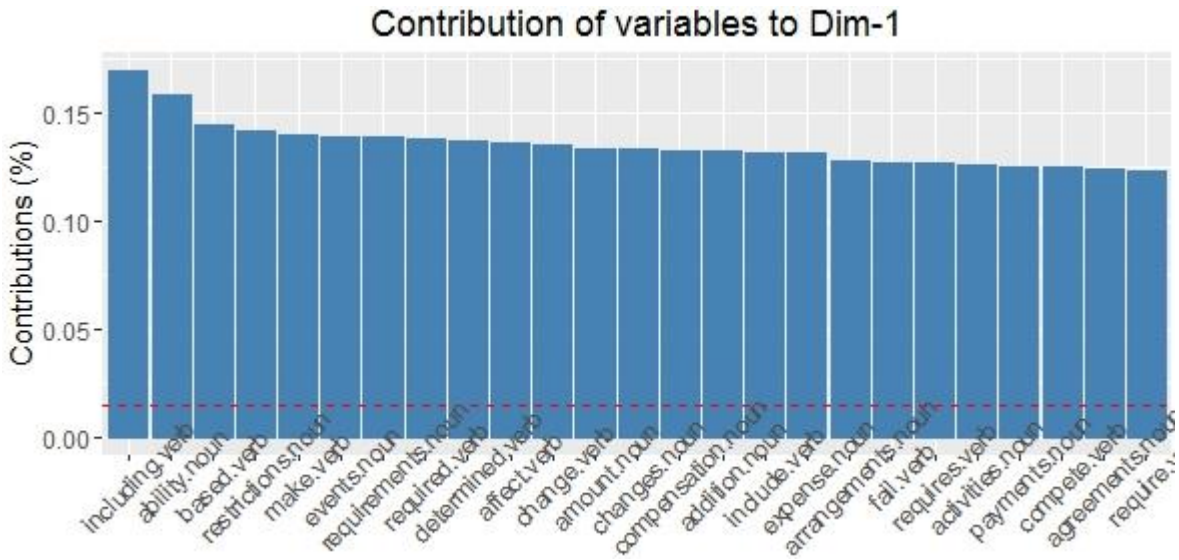
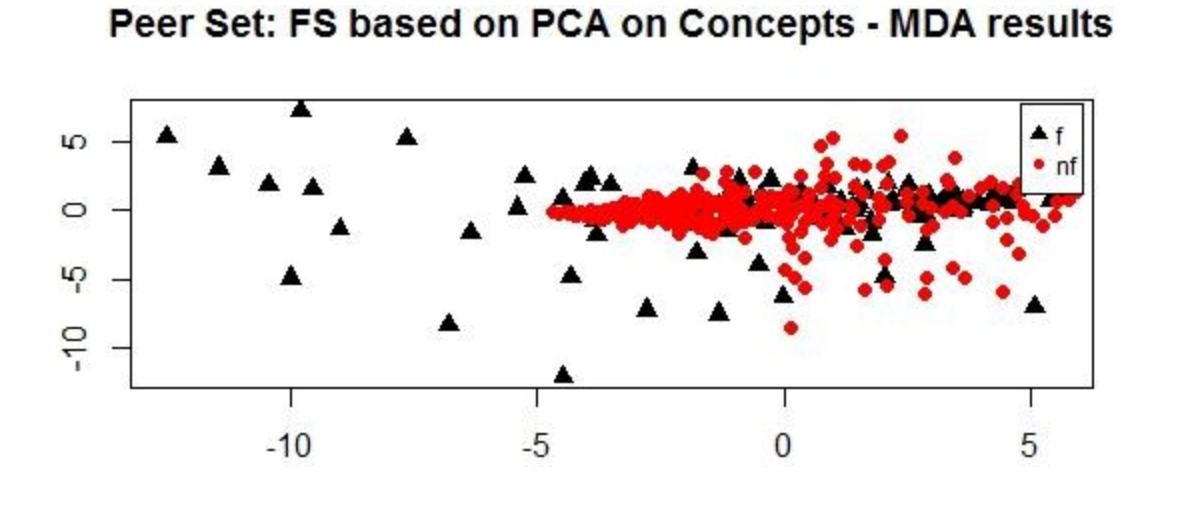
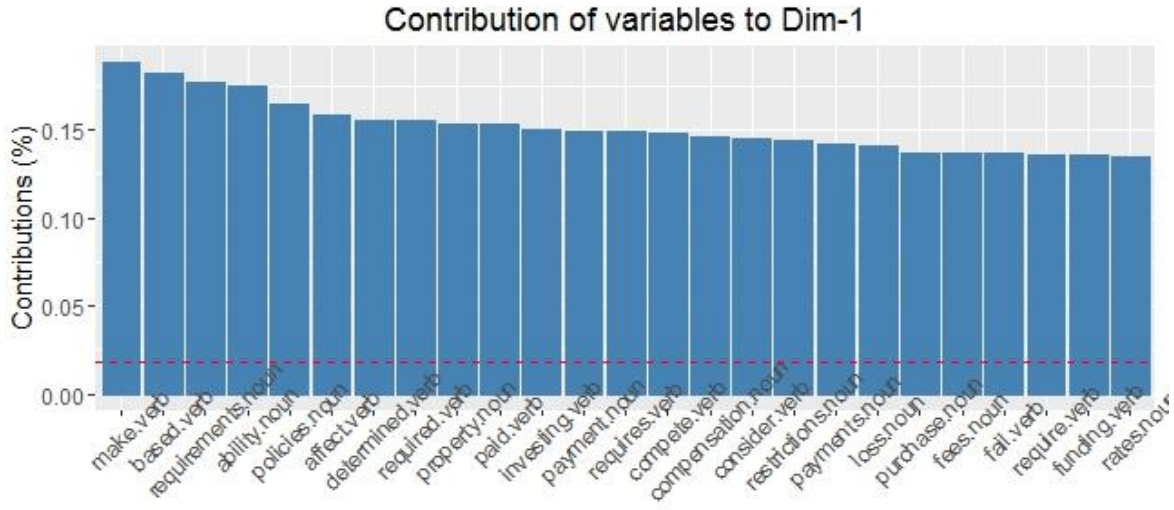
Document Representation	Chosen Features
Peer Set: Concept Mining	'including.verb', 'ability.noun', 'based.verb', 'restrictions.noun', 'make.verb', 'events.noun', 'requirements.noun', 'required.verb', 'determined.verb', 'affect.verb', 'change.verb', 'amount.noun', 'changes.noun', 'compensation.noun', 'addition.noun', 'include.verb', 'expense.noun', 'arrangements.noun', 'fail.verb', 'requires.verb', 'activities.noun', 'payments.noun', 'compete.verb', 'agreements.noun', 'require.verb', 'purposes.noun', 'credit.noun', 'circumstances.noun', 'event.noun', 'obligations.noun', 'requiring.verb', 'parties.noun', 'loss.noun', 'purchase.noun', 'existing.verb', 'assets.noun', 'disclosure.noun', 'number.noun', 'portion.noun', 'apply.verb', 'failure.noun', 'regulations.noun', 'issued.verb', 'accounting.noun', 'action.noun', 'laws.noun', 'rate.noun', 'provision.noun', 'equity.noun', 'issuance.noun',
	
	

Table K.17: Concepts chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Concept Mining	'make.verb', 'based.verb', 'requirements.noun', 'ability.noun', 'policies.noun', 'affect.verb', 'determined.verb', 'required.verb', 'property.noun', 'paid.verb', 'investing.verb', 'payment.noun', 'requires.verb', 'compete.verb', 'compensation.noun', 'consider.verb', 'restrictions.noun', 'payments.noun', 'loss.noun', 'purchase.noun', 'fees.noun', 'fail.verb', 'require.verb', 'funding.verb', 'rates.noun', 'limitations.noun', 'purposes.noun', 'provide.verb', 'agreements.noun', 'meet.verb', 'acquiring.verb', 'obligations.noun', 'sale.noun', 'determine.verb', 'liability.noun', 'expense.noun', 'liabilities.noun', 'portion.noun', 'including.verb', 'maintain.verb', 'types.noun', 'losses.noun', 'action.noun', 'activities.noun', 'fee.noun', 'agencies.noun', 'occur.verb', 'evaluate.verb', 'transaction.noun', 'include.verb'



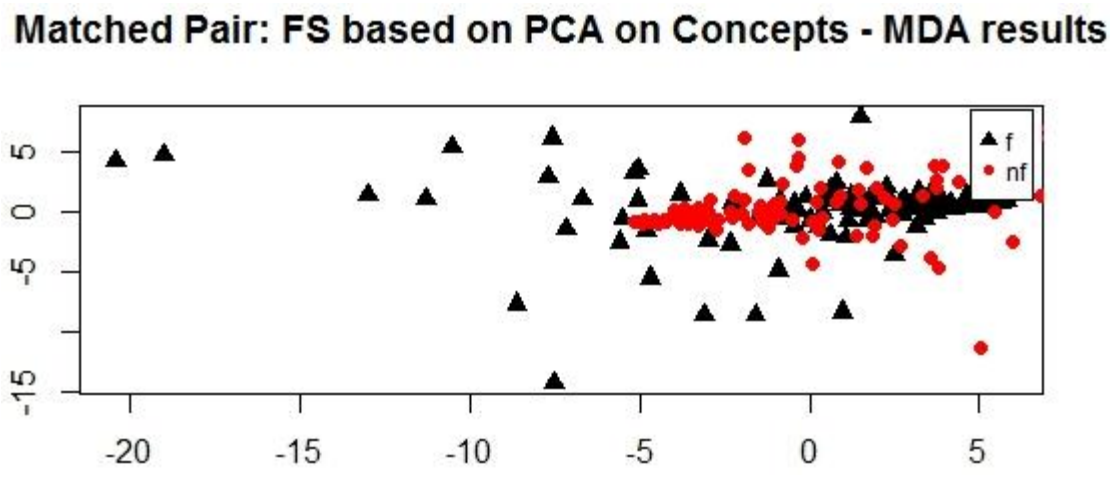
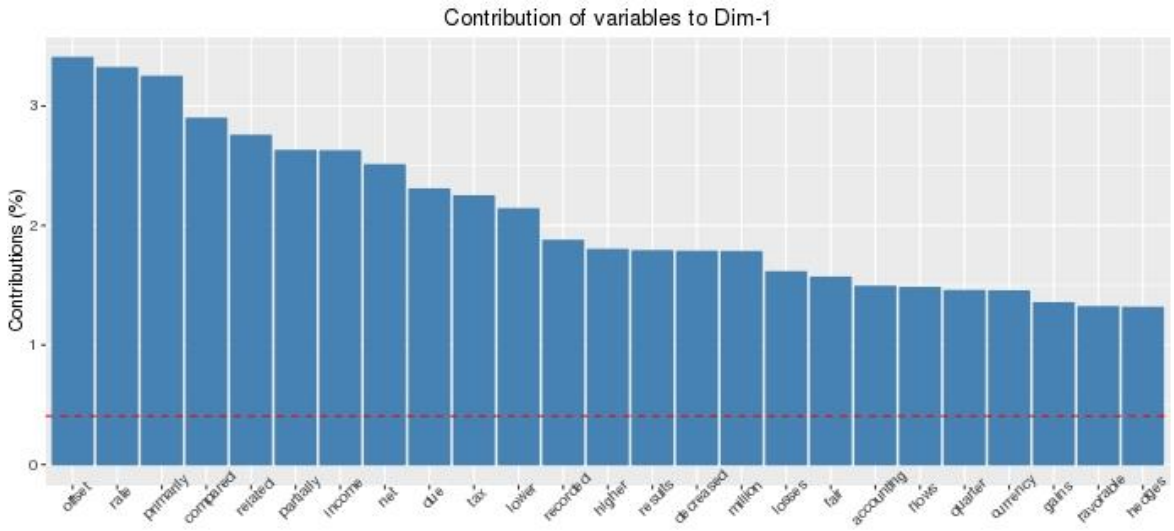


Table K.18: Concepts chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Keywords	'offset', 'rate', 'primarily', 'compared', 'related', 'partially', 'income', 'net', 'due', 'tax', 'lower', 'recorded', 'higher', 'results', 'decreased', 'million', 'losses', 'fair', 'accounting', 'flows', 'quarter', 'currency', 'gains', 'favorable', 'hedges', 'capital', 'fourth', 'unfavorable', 'earnings', 'investments', 'unsecured', 'impact', 'declines', 'pretax', 'pension', 'certain', 'impacted', 'driven', 'gain', 'result', 'notes', 'management', 'adjustments', 'also', 'realized', 'compensation', 'system', 'dollar', 'failures', 'stockholder'



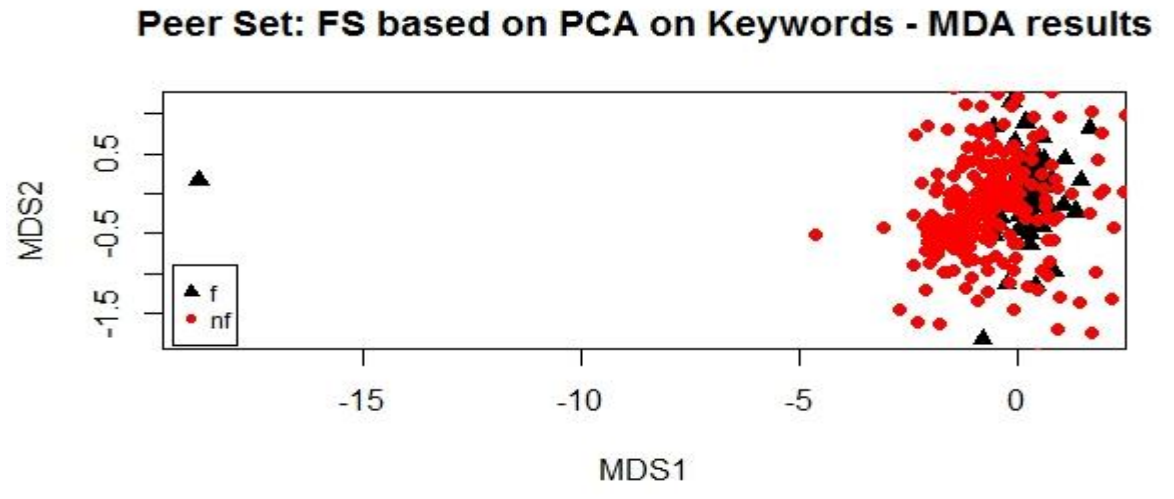
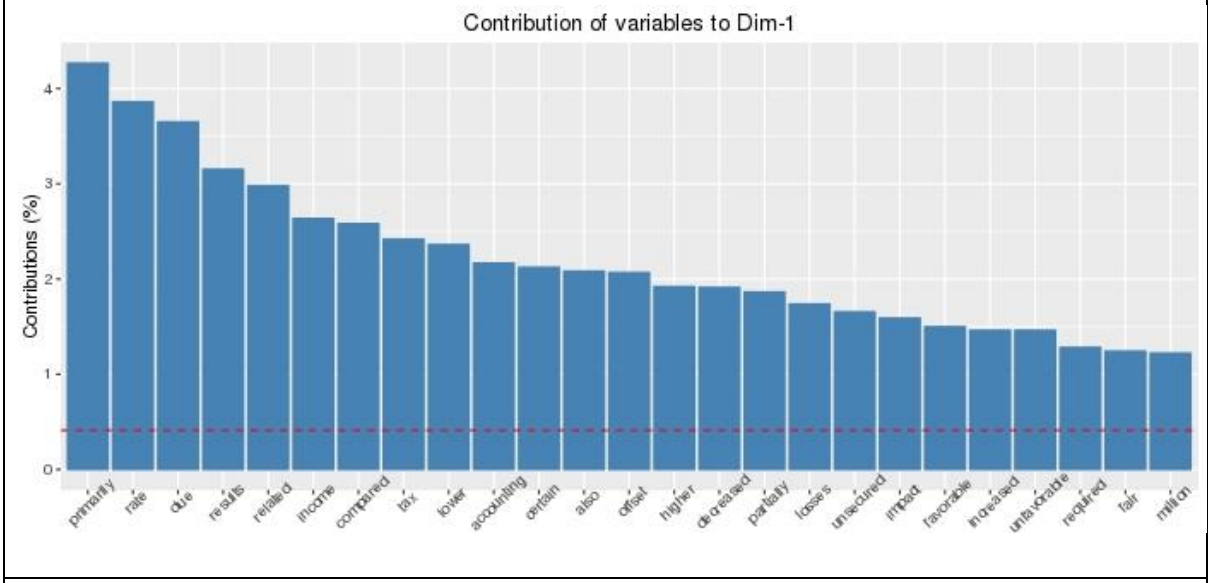


Table K.19: Keywords chosen by PCA for peer set data set up and results of MDA computation.

Document Representation	Chosen Features
Matched Pair: Keywords	'primarily', 'rate', 'due', 'results', 'related', 'income', 'compared', 'tax', 'lower', 'accounting', 'certain', 'also', 'offset', 'higher', 'decreased', 'partially', 'losses', 'unsecured', 'impact', 'favorable', 'increased', 'unfavorable', 'required', 'fair', 'million', 'investments', 'flows', 'net', 'fourth', 'adjustments', 'ability', 'result', 'compensation', 'growth', 'realized', 'approvals', 'earnings', 'currency', 'date', 'quarter', 'flat', 'gains', 'believe', 'acquisition', 'general', 'average', 'chain', 'upon', 'operating', 'shipments'



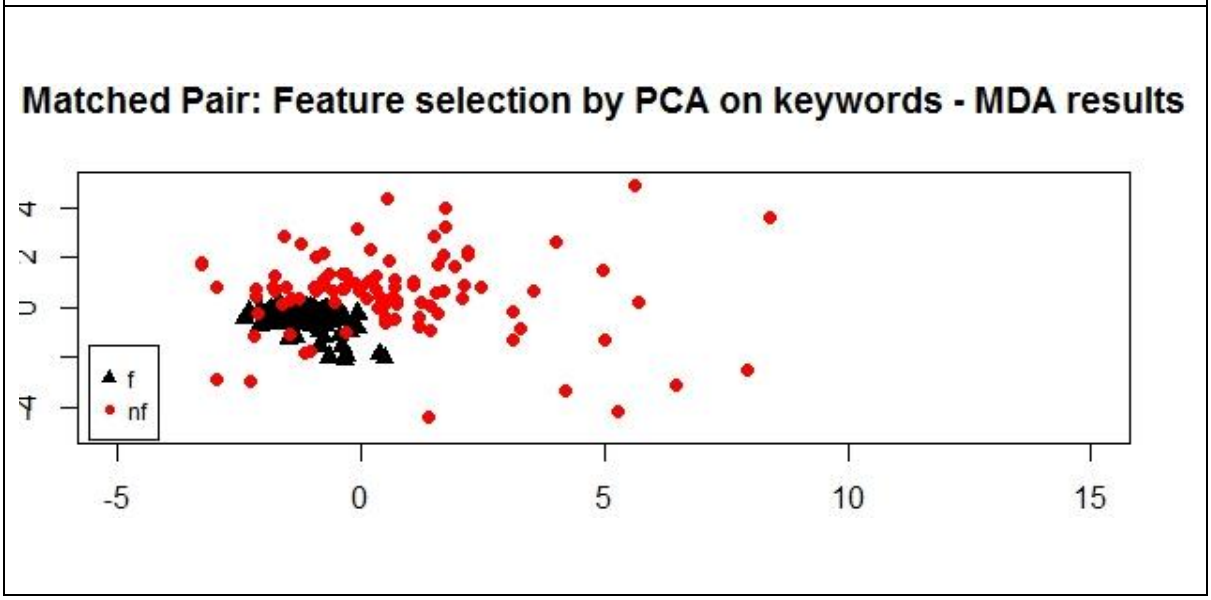
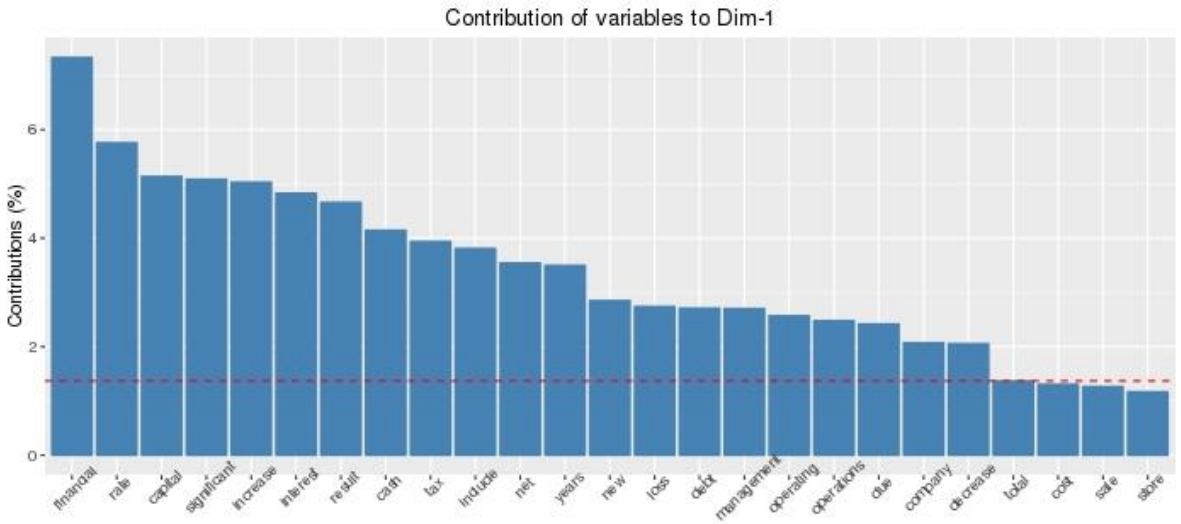


Table K.20: Keywords chosen by PCA for matched pair data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Rutherford	'financial', 'rate', 'capital', 'significant', 'increase', 'interest', 'result', 'cash', 'tax', 'include', 'net', 'years', 'new', 'loss', 'debt', 'management', 'operating', 'operations', 'due', 'company', 'decrease', 'total', 'cost', 'sale', 'store', 'growth', 'lower', 'make', 'revenue', 'higher', 'currency', 'item', 'activity', 'exchange', 'reduce', 'products', 'turnover', 'liability', 'division', 'margin', 'now', 'strong', 'borrowing', 'investment', 'risk', 'services', 'previous', 'property', 'asset', 'trading'



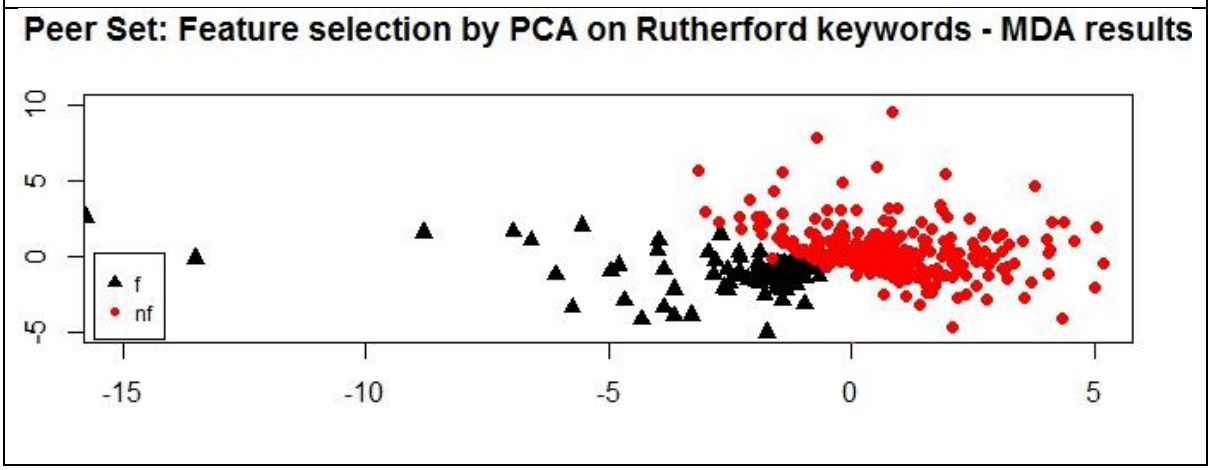


Table K.21: Keywords chosen by PCA for peer set data set up and results of MDA computation.

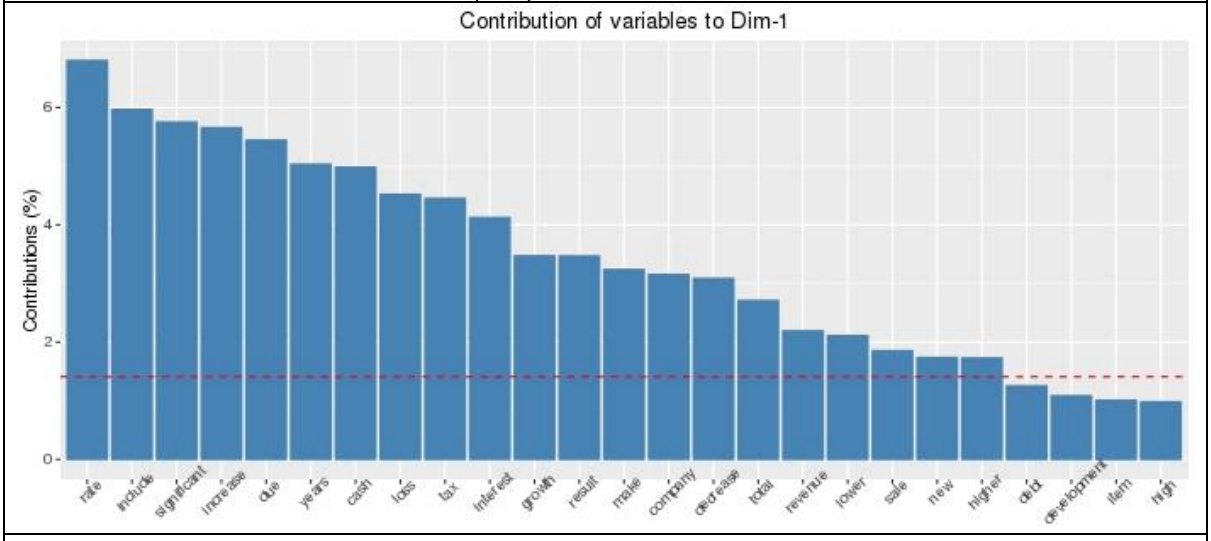
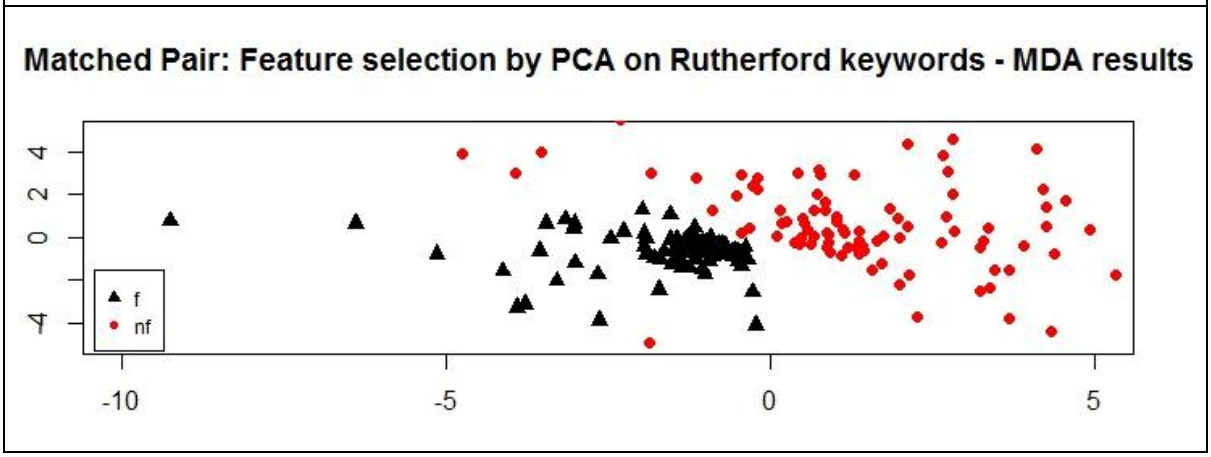
Document Representation	Chosen Features
Matched Pair: Rutherford	'rate', 'include', 'significant', 'increase', 'due', 'years', 'cash', 'loss', 'tax', 'interest', 'growth', 'result', 'make', 'company', 'decrease', 'total', 'revenue', 'lower', 'sale', 'new', 'higher', 'debt', 'development', 'item', 'high', 'store', 'operating', 'activity', 'risk', 'number', 'services', 'currency', 'investment', 'turnover', 'share', 'completed', 'liability', 'strong', 'net', 'exchange', 'division', 'property', 'overall', 'continued', 'retail', 'asset', 'operating', 'shipments'
	
	

Table K.22: Keywords chosen by PCA for matched pair data set up and results of MDA computation.

APPENDIX L

Feature Selection using Boruta graphed, based on MDA computation

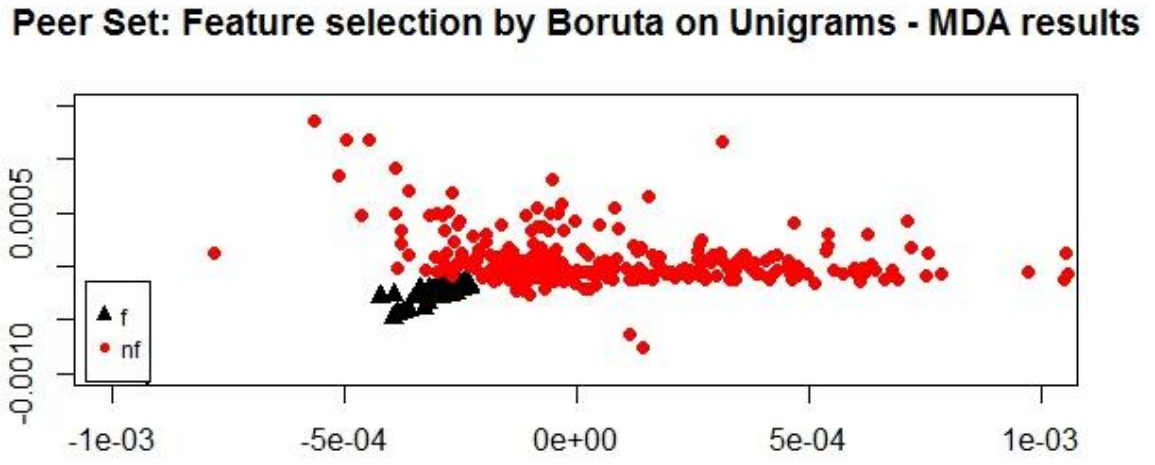
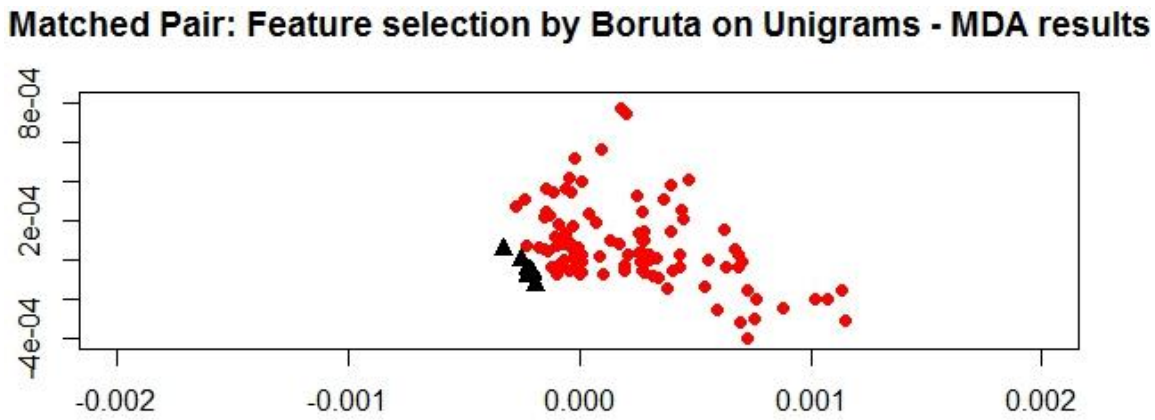
Document Representation	Chosen Features
Peer Set: Unigrams	'interest', 'invest', 'expect', 'made', 'manag', 'meet', 'increas', 'order', 'princip', 'provid', 'outlook', 'recent', 'statement', 'tax', 'revenu', 'unit', 'account', 'addit', 'signific', 'asset', 'capit', 'current', 'loss', 'financi', 'futur', 'growth', 'make', 'requir', 'share', 'satisfactori'
	
Matched Pair: Unigrams	'growth', 'interest', 'asset', 'increas', 'signific', 'plan', 'revenu', 'maintain', 'statement', 'loss'
	

Table L.1: Unigrams chosen by Boruta for PS and MP data set up and results of MDA computation.

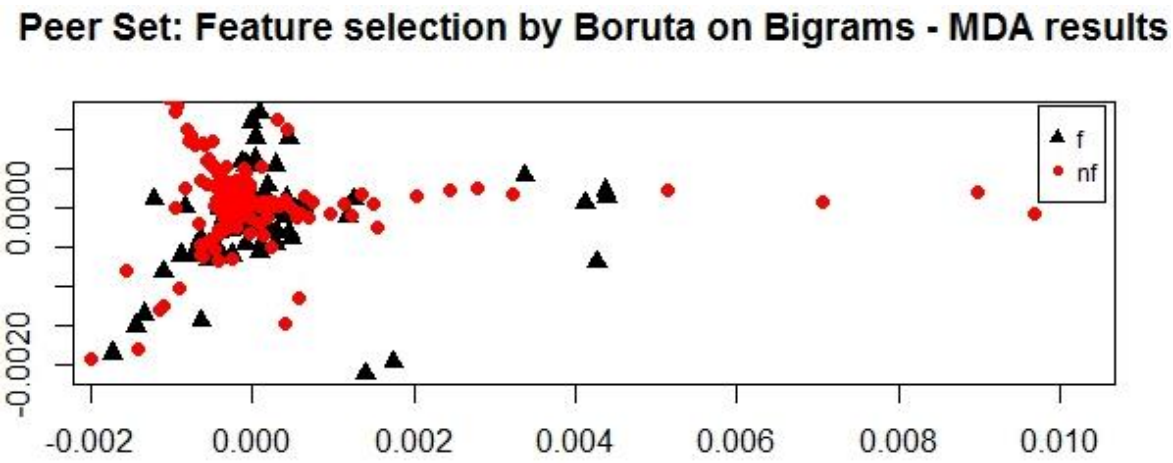
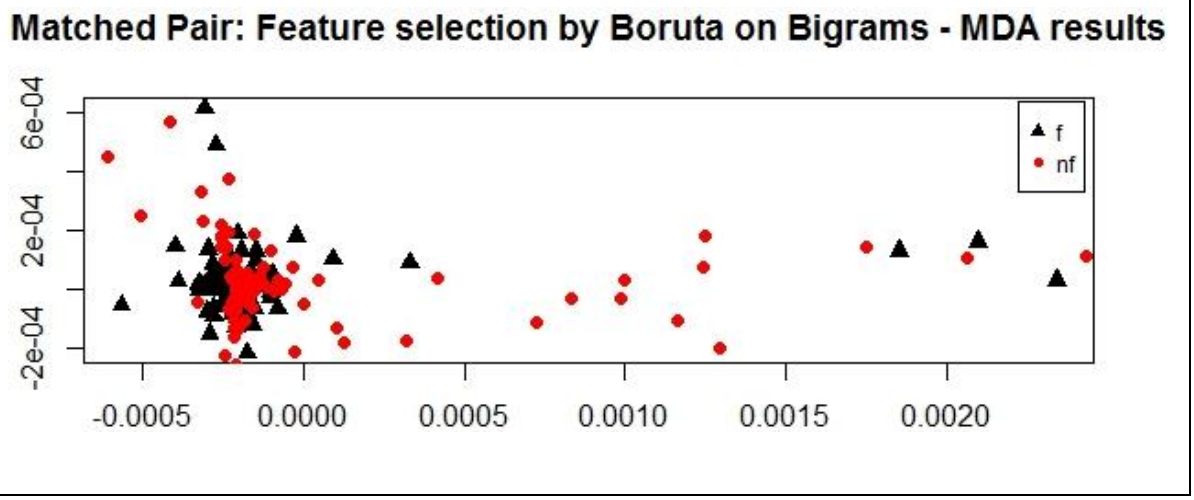
Document Representation	Chosen Features
Peer Set: Bigrams	'accounted for', 'acquisition of', 'and sale', 'annual report', 'be required', 'company in', 'continued to', 'designed for', 'due to', 'event that', 'experience in', 'for fiscal', 'group of', 'in and', 'in compared', 'into a', 'legal and', 'market our', 'necessary to', 'of approximately', 'our management', 'our own', 'purchase price', 'the acquisition', 'the fiscal', 'to conduct', 'year ended', 'total revenues', 'borrowings under', 'to obtain'
	
Matched Pair: Bigrams	'acquisition of', 'be limited', 'continued to', 'operating income', 'commitment to', 'contributed to', 'greater financial', 'primarily due', 'have greater', 'income increased'
	

Table L.2: Bigrams chosen by Boruta for PS and MP data set up and results of MDA computation.

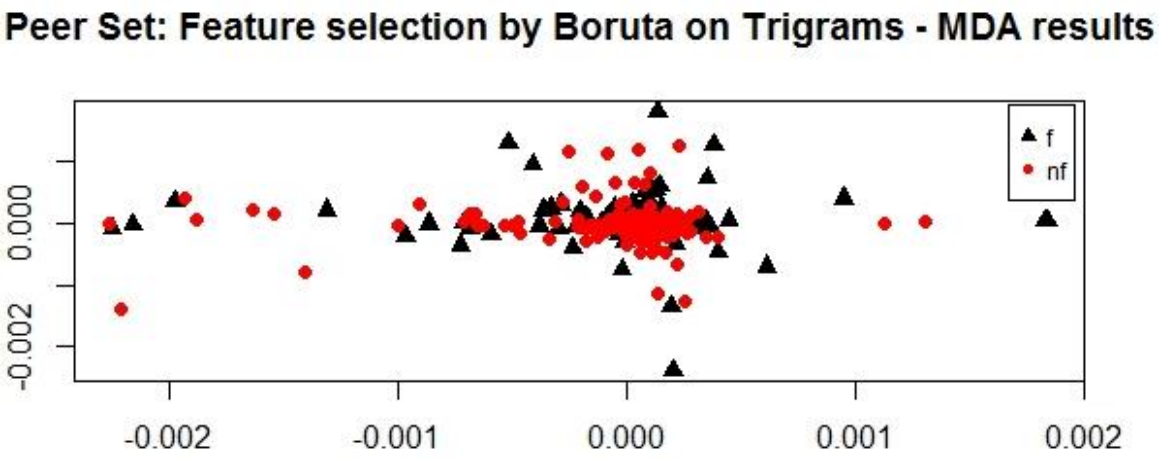
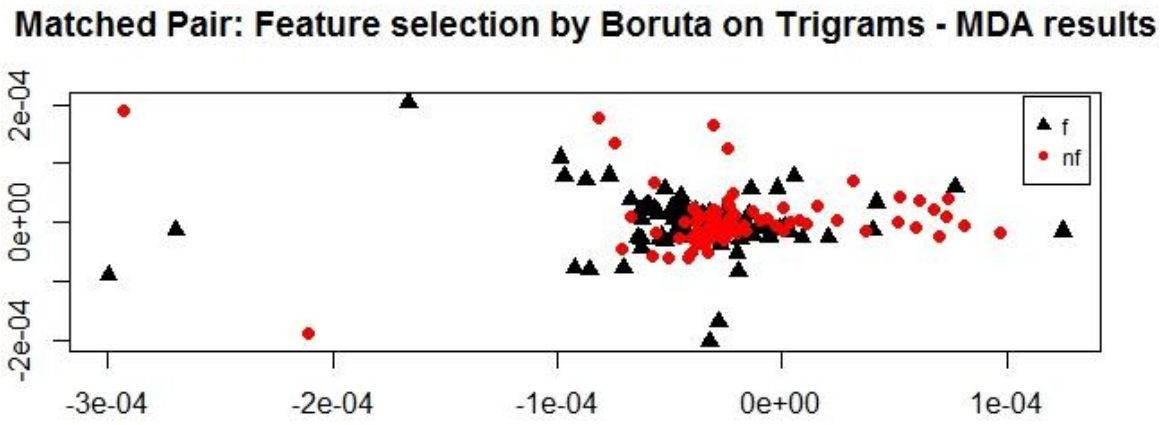
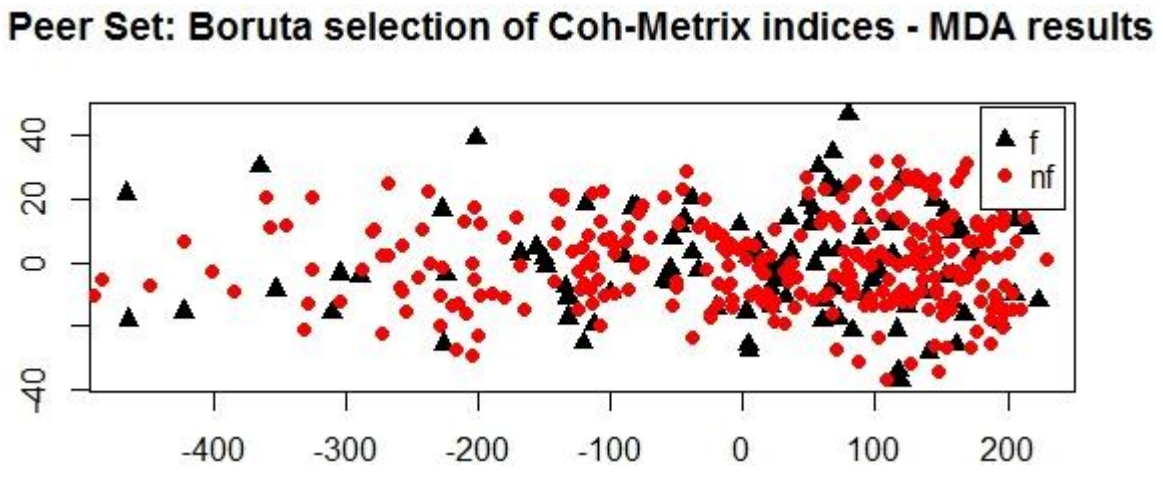
Document Representation	Chosen Features
Peer Set: Trigrams	'an adverse effect', 'and sale of', 'at the time', 'during the period', 'entered into a', 'for the year', 'in the event', 'may be required', 'million at december', 'million in cash', 'million of cash', 'not believe that', 'of our common', 'our common stock', 'primarily as a', 'primarily due to', 'provided by financing', 'pursuant to the', 'shares of common', 'the acquisition of', 'the company in', 'the fiscal year', 'the impact of', 'the results of', 'the year ended', 'use of the'
	
Matched Pair: Trigrams	'comply with the', 'the acquisition of', 'the event that', 'we may not', 'which may be', 'primarily due to', 'in compared to', 'and will be', 'in addition the', 'the end of',
	

Table L.3: Trigrams chosen by Boruta for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Coh-Metrix	'CNCADC', 'CNCAdd', 'CNCNeg', 'CNCTempx', 'CRFANP1', 'CRFNO1', 'CRFNOa', 'CRFSOa', 'DRGERUND', 'DRINF', 'DRPVAL', 'DRVP', 'LSASS1d', 'PCCNCz', 'PCCONNz', 'PCNARz', 'PCVERBz', 'RDFKGL', 'SMCAUSlsa', 'SMCAUSwn', 'SYNLE', 'SYNSTRUTa', 'SYNSTRUTt', 'WRDADJ', 'WRDAOAc', 'WRDFRQa', 'WRDIMGc', 'WRDMEAc', 'WRDVERB'



Matched Pair: Coh-Metrix	'LSAPP1d', 'WRDPOSin', 'SYNSTRUTt', 'DRVP', 'DRPP', 'WRDADJ', 'WRDFRQa', 'WRDHYPnv'

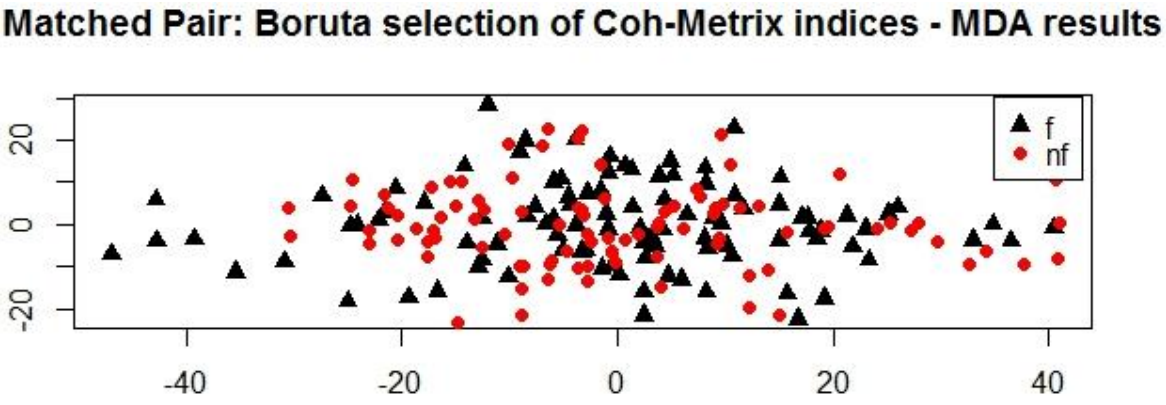


Table L.4: Coh-Metrix chosen by Boruta for PS and MP data set up and results of MDA computation.

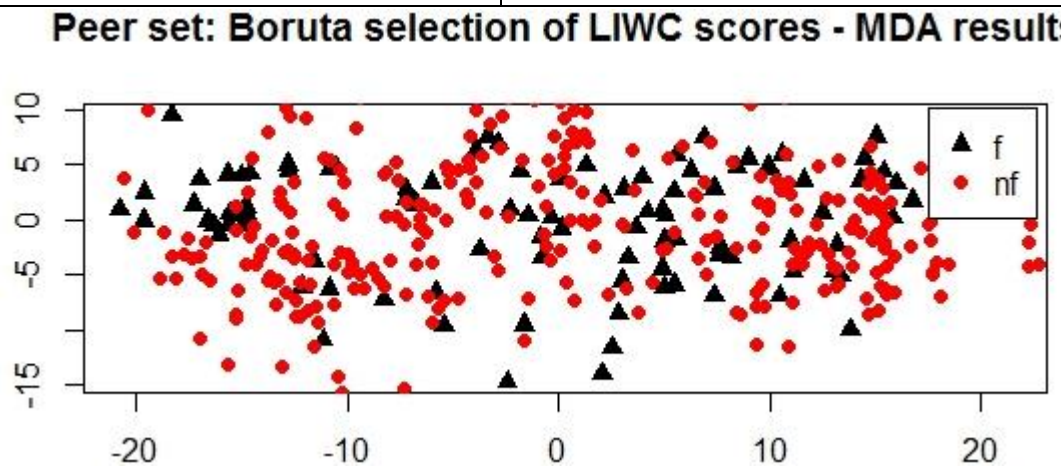
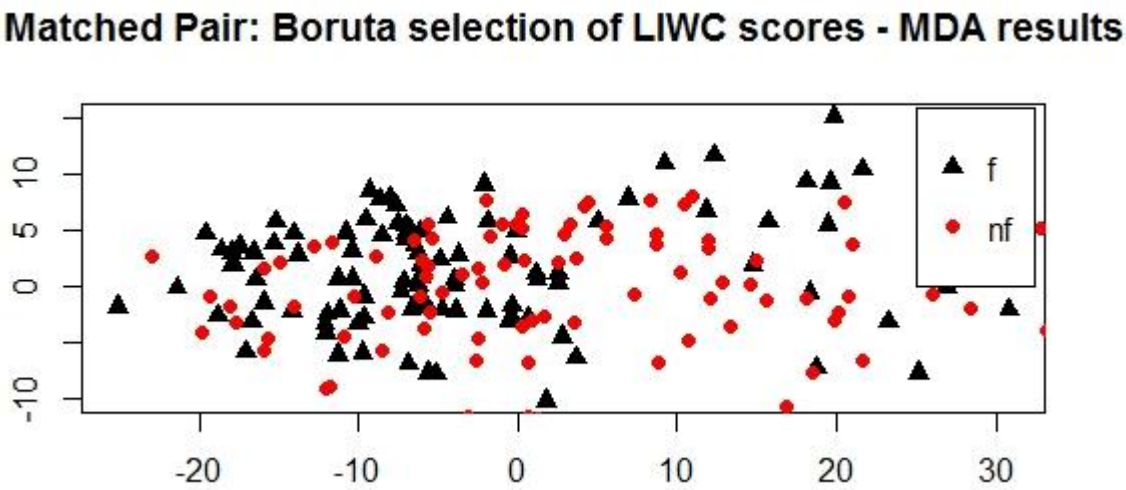
Document Representation	Chosen Features
Peer Set: LIWC	'Clout', 'auxverb', 'focuspresent', 'Authentic', 'verb', 'focusfuture', 'function.', 'adj', 'relativ', 'pronoun', 'compare', 'ppron', 'interrog', 'article', 'cogproc',
	
Matched Pair: LIWC	'Analytic', 'interrog', 'Authentic', 'relativ', 'Tone', 'ipron', 'prep', 'auxverb', 'adj'
	

Table L.5: LIWC variables chosen by Boruta for PS and MP data set up and results of MDA computation.

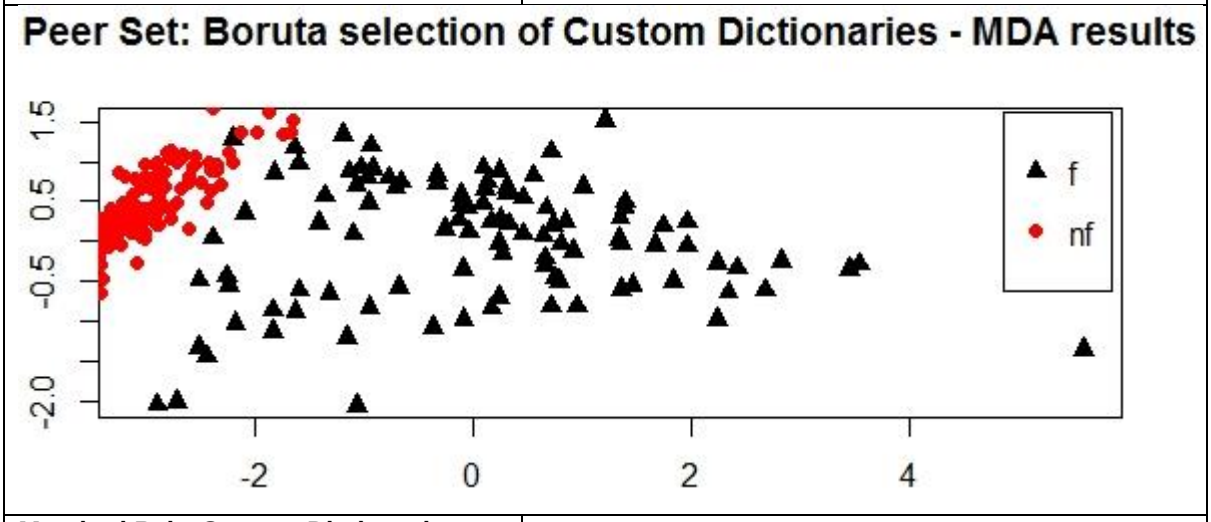
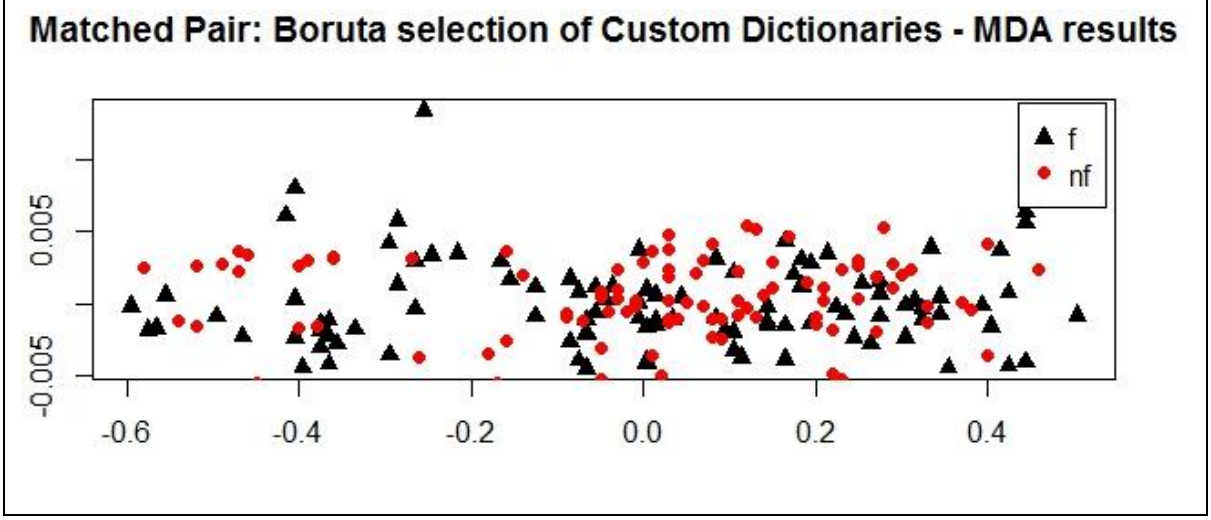
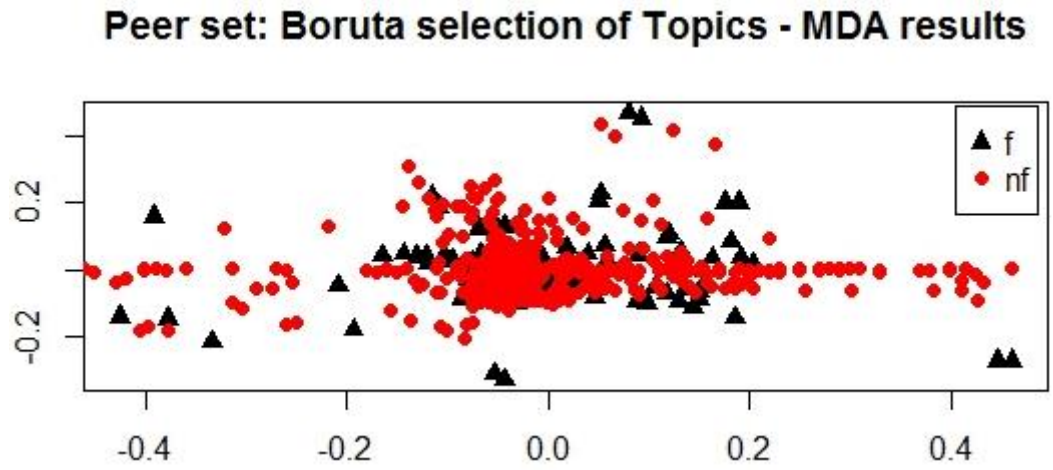
Document Representation	Chosen Features
Peer Set: Custom Dictionaries	'positivity_Freq', 'ForwardLooking_Freq', 'Constraining', 'Uncert1_Freq', 'Litiguous', 'negativity_Freq', 'Modal.Weak',
	
Matched Pair: Custom Dictionaries	'All
	

Table L.6: Custom Dictionaries chosen by Boruta for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set: Topics	'Topic.3', 'Topic.24', 'Topic.4', 'Topic.5', 'Topic.12', 'Topic.18', 'Topic.22',



Matched Pair: Topics	'Topic.3', 'Topic.4', 'Topic.21', 'Topic.24',

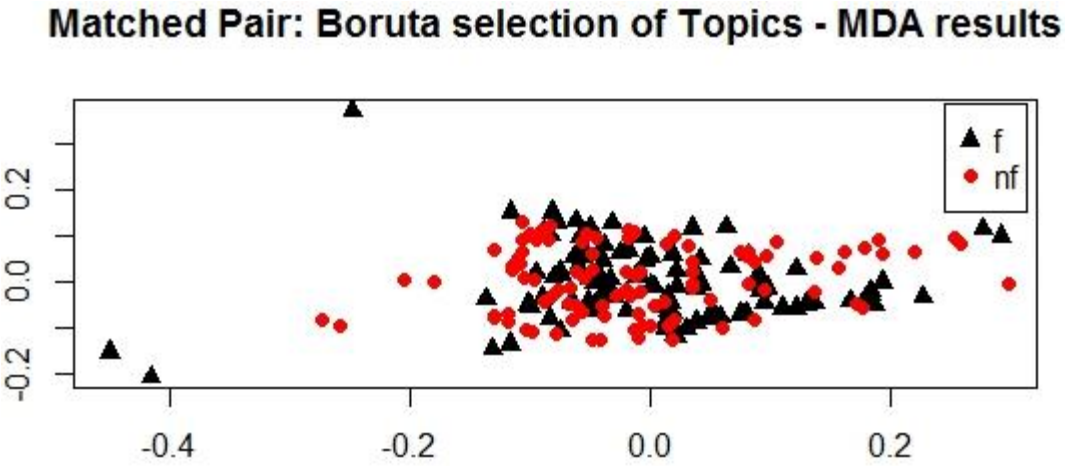


Table L.7: Topics chosen by Boruta for PS and MP data set up and results of MDA computation.

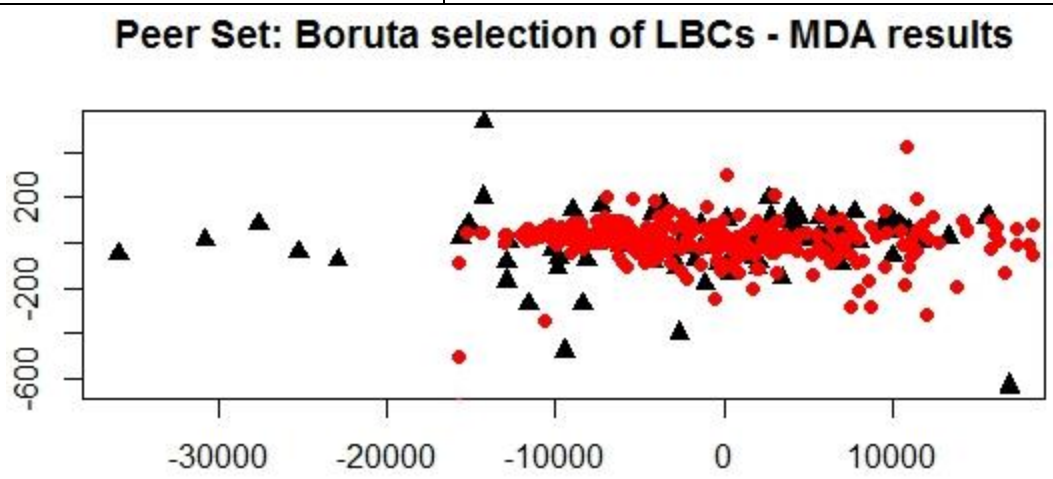
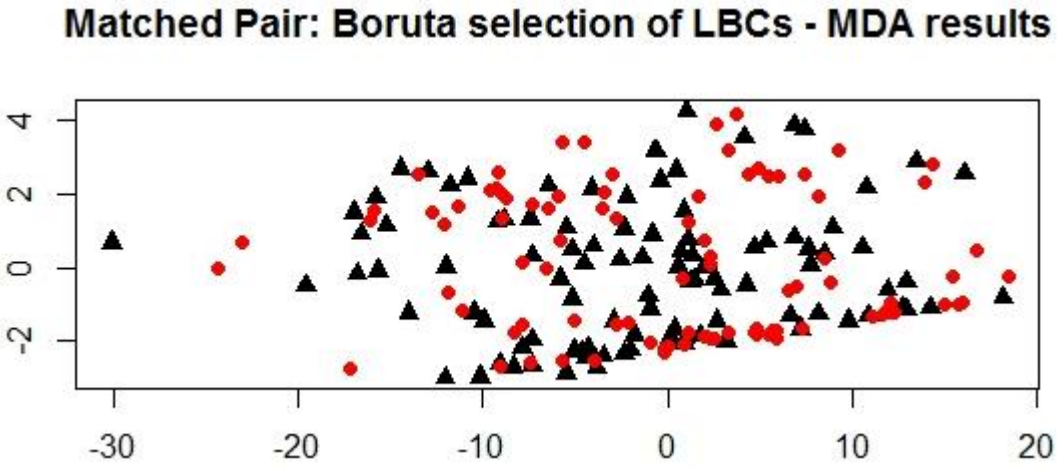
Document Representation	Chosen Features
Peer Set: LBCs	'passive.verb.ratio', 'Sentence.Qty', 'Word.Qty', 'Modal.Verb.Ratio', 'Pausality', 'Function.Word.Diversity', 'Group.Ref', 'Modifier.Qty', 'Verb.Qty', 'Temporal.Imm..Ratio',
	
Matched Pair: LBCs	'Content.Word.Diversity', 'Other.Ref', 'Group.Ref', 'Modal.Verb .Ratio',
	

Table L.8: LBCs chosen by Boruta for PS and MP data set up and results of MDA computation.

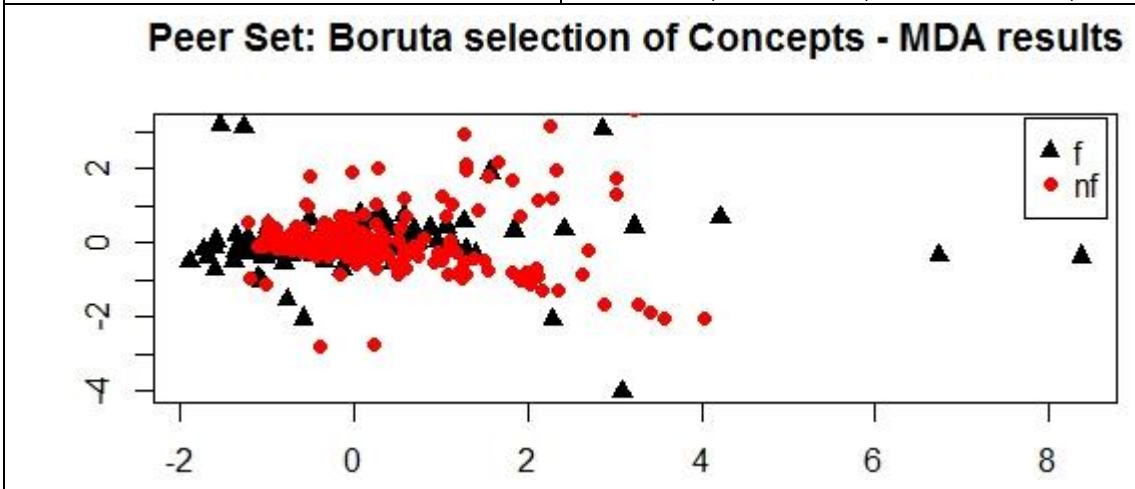
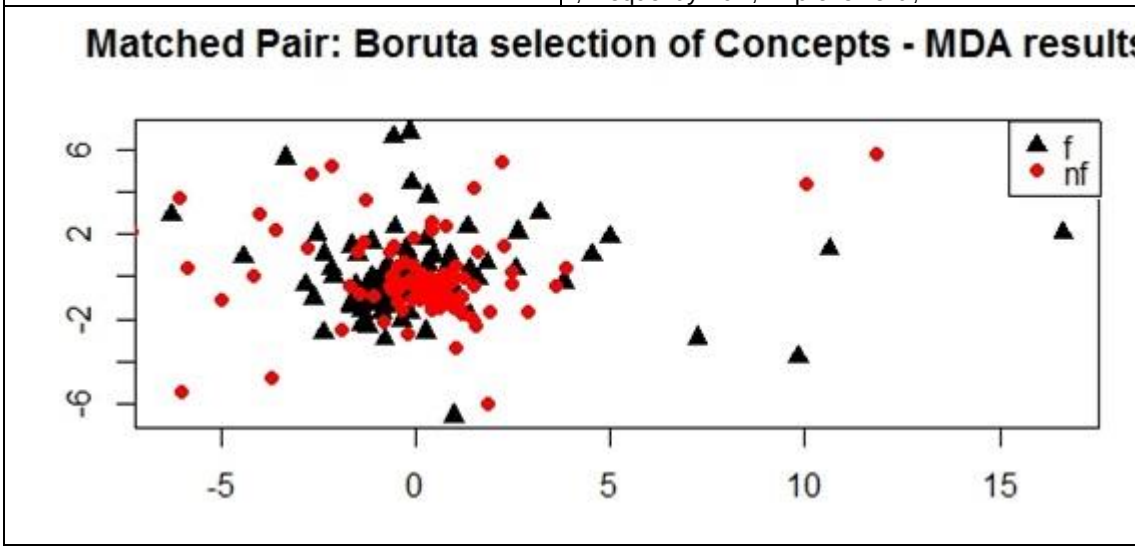
Document Representation	Chosen Features
Peer Set: Concepts	'accounts.noun', 'acquired.verb', 'acquisition.noun', 'amounting.verb', 'arranged.verb', 'assigned.verb', 'clients.noun', 'collect.verb', 'companies.noun', 'consisting.verb', 'continued.verb', 'credit.noun', 'discourage.verb', 'employee.noun', 'ended.verb', 'entered.verb', 'event.noun', 'expenses.noun', 'frequently.Adv', 'improve.verb', 'losses.noun', 'manufactures.noun', 'obtained.verb', 'performed.verb', 'procedure.noun', 'purchase.noun', 'put.verb', 'relates.verb', 'results.noun', 'stockholder.noun',
	
Matched Pair: Concepts	'accounts.noun', 'acquired.verb', 'acquisition.noun', 'amounting.verb', 'arranged.verb', 'assigned.verb', 'clients.noun', 'collect.verb', 'companies.noun', 'consisting.verb', 'continued.verb', 'credit.noun', 'discourage.verb', 'employee.noun', 'ended.verb', 'entered.verb', 'event.noun', 'expenses.noun', 'frequently.Adv', 'improve.verb',
	

Table L.9: Concepts chosen by Boruta for PS and MP data set up and results of MDA computation.

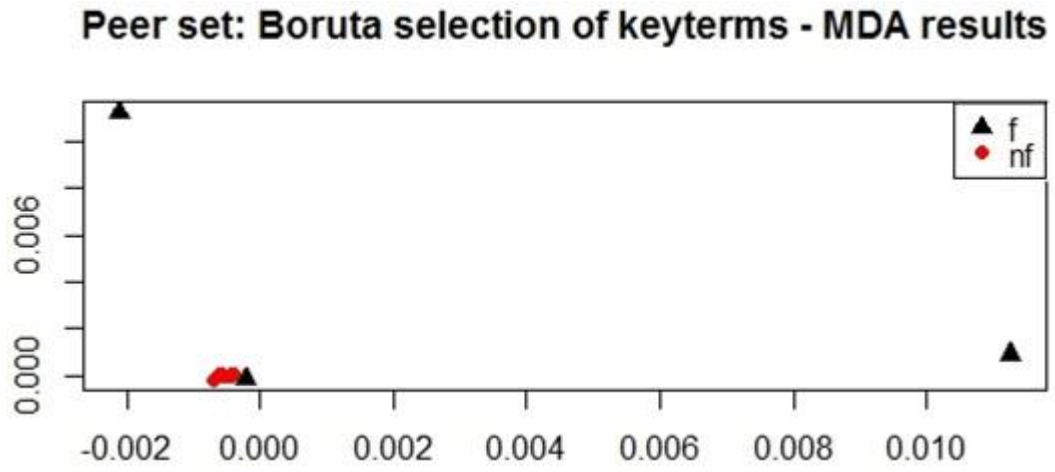
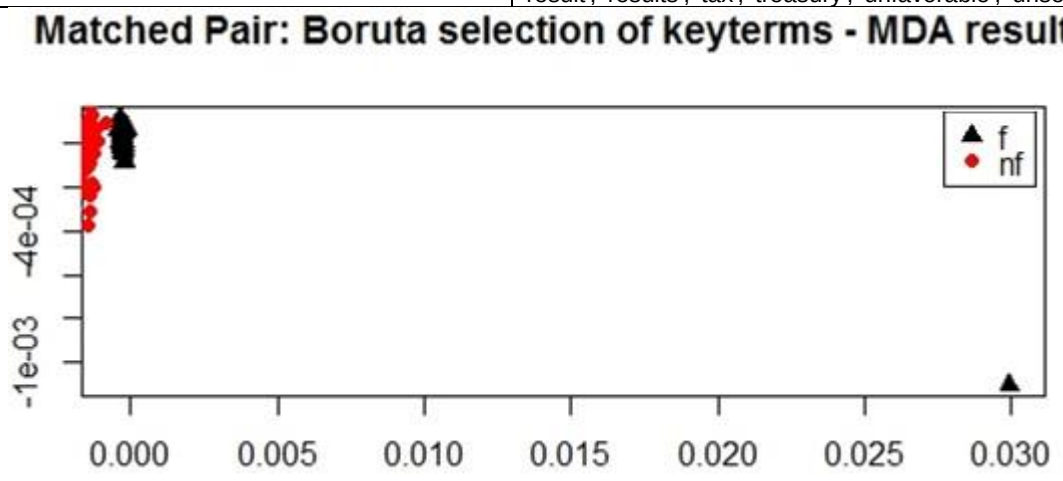
Document Representation	Chosen Features
Peer Set: Keywords	'accounting', 'communities', 'management', 'operations', 'purchase', 'tax', 'acquisition', 'capital', 'companies', 'growth', 'outlook', 'staffing', 'compared', 'expected', 'hedges', 'required', 'stockholder', 'bankruptcy', 'certain', 'failures', 'unfavorable', 'believed', 'operating', 'procurement', 'results', 'unsecured',
	
Matched Pair: Keywords	'accounting', 'acquisition', 'bankruptcy', 'believed', 'certain ' communities', 'companies', 'compared', 'demand', 'econo mic', 'flexible', 'growth', 'losses', 'primarily', 'procurement', ' result', 'results', 'tax', 'treasury', 'unfavorable', 'unsecured',
	

Table L.10: Keywords chosen by Boruta for PS and MP data set up and results of MDA computation.

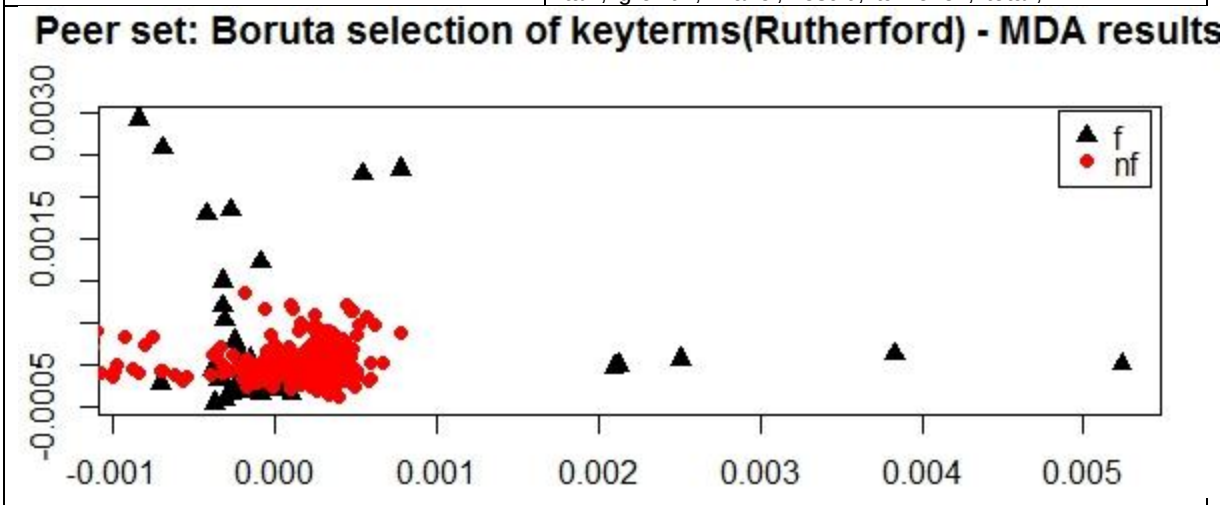
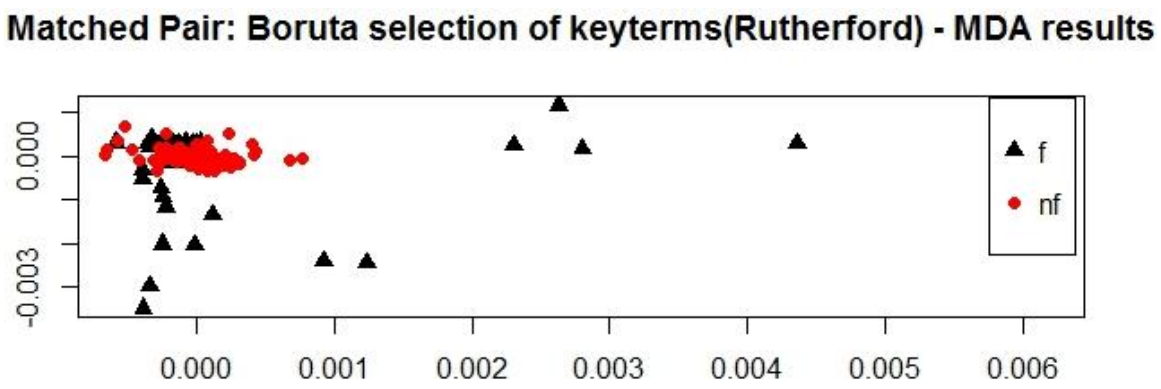
Document Representation	Chosen Features
Peer Set: Rutherford keywords	'activity', 'high', 'management', 'revenue', 'years', 'capital', 'include', 'net', 'significant', 'loss', 'company', 'increase', 'operating', 'store', 'rate', 'financial', 'interest', 'operations', 'tax', 'growth', 'make', 'result', 'turnover', 'total',
	
Matched Pair: Rutherford keywords	'cash', 'include', 'loss', 'result', 'total', 'company', 'increase', 'make', 'sale', 'turnover', 'due', 'interest', 'new', 'significant', 'years', 'growth', 'item', 'number', 'store', 'high', 'last', 'rate', 'tax',
	

Table L.11: Keywords (Rutherford) chosen by Boruta for PS and MP data set up and results of MDA computation.

APPENDIX M

Feature selection using Information Gain (IG) graphed based on MDA computation

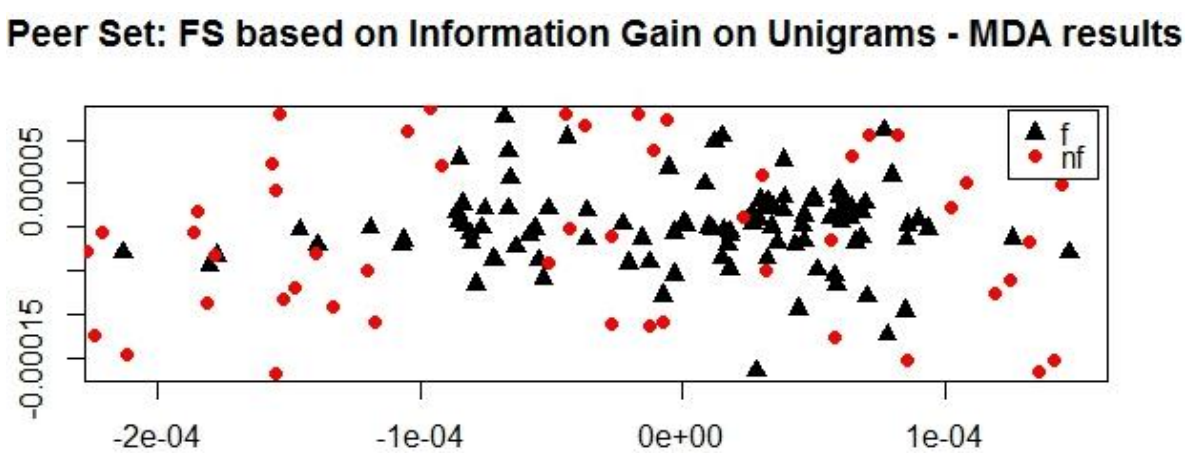
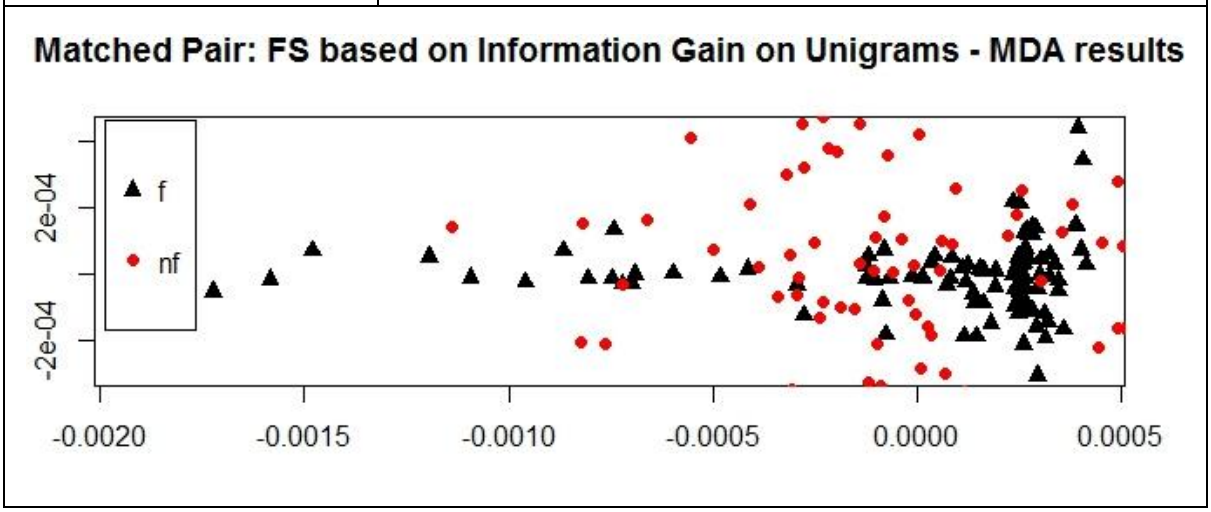
Document Representation	Chosen Features
Peer Set: Unigrams	'increas', 'will', 'expect', 'signific', 'addit', 'financi', 'share', 'tax', 'futur', 'sale', 'capit', 'current', 'servic', 'intern', 'period', 'requir', 'invest', 'account', 'respect', 'growth', 'basi', 'make', 'loss', 'purpos', 'asset', 'revenu', 'includ', 'annual', 'cost', 'debt'
	
Matched Pair: Unigrams	'activ', 'end', 'includ', 'expect', 'increas', 'part', 'signific', 'manag', 'account', 'indebted', 'follow', 'intern', 'respect', 'sale', 'servic', 'interest', 'valu', 'addit', 'futur', 'growth', 'period', 'expens', 'asset', 'current', 'make', 'revenu', 'primarili', 'loss', 'requir', 'purpos'
	

Table M.1: Unigrams chosen by IG for PS and MP data set up and results of MDA computation.

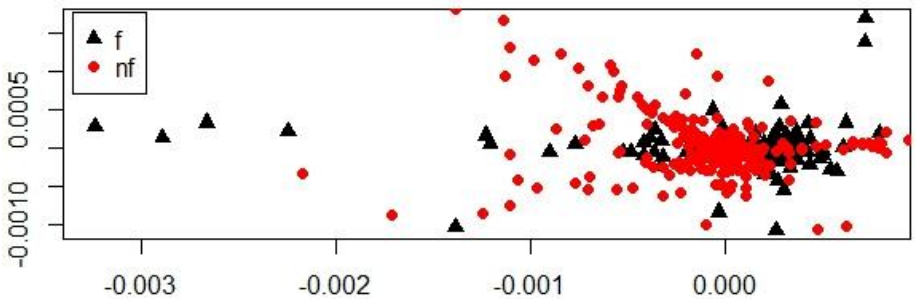
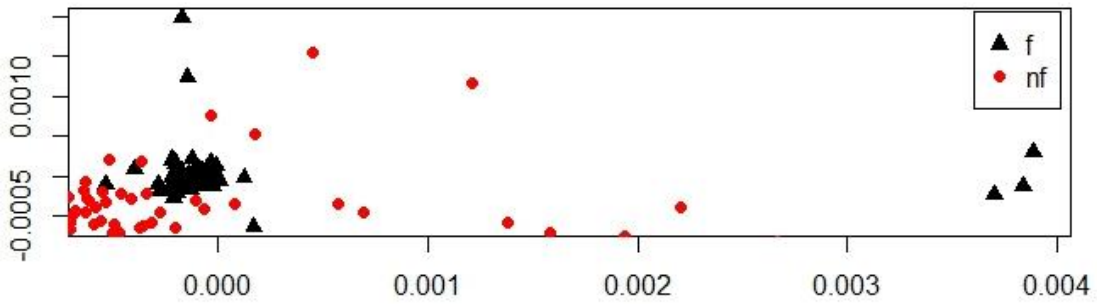
Document	Chosen Features
Peer Set: Bigrams	'to date', 'to higher', 'in and', 'income increased', 'efficiency and', 'market our', 'of approximately', 'primarily due', 'continued to', 'improvement in', 'commitment to', 'acquisition of', 'income in', 'annual report', 'into a', 'of any', 'purchase price', 'year ended', 'in increased', 'volume of', 'customers with', 'them to', 'event that', 'markets and', 'provider of', 'necessary to', 'service and', 'failure to', 'accounting and', 'presented in',
Peer Set: FS based on Information Gain on Bigrams - MDA results 	
Matched Pair: Bigrams	'greater financial', 'failure to', 'have greater', 'event that', 'in and', 'obtain a', 'presented in', 'continued to', 'contributed to', 'service and', 'income increased', 'and lower', 'resources than', 'purchase price', 'also include', 'efficiency and', 'capital expenditures', 'primarily due', 'operating income', 'fail to', 'market our', 'the first', 'a decline', 'a increase', 'in increased', 'in compared', 'may be', 'to date', 'approximately and', 'approved by')
Matched Pair: FS based on Information Gain on Bigrams - MDA results 	

Table M.2: Bigrams chosen by IG for PS and MP data set up and results of MDA computation.

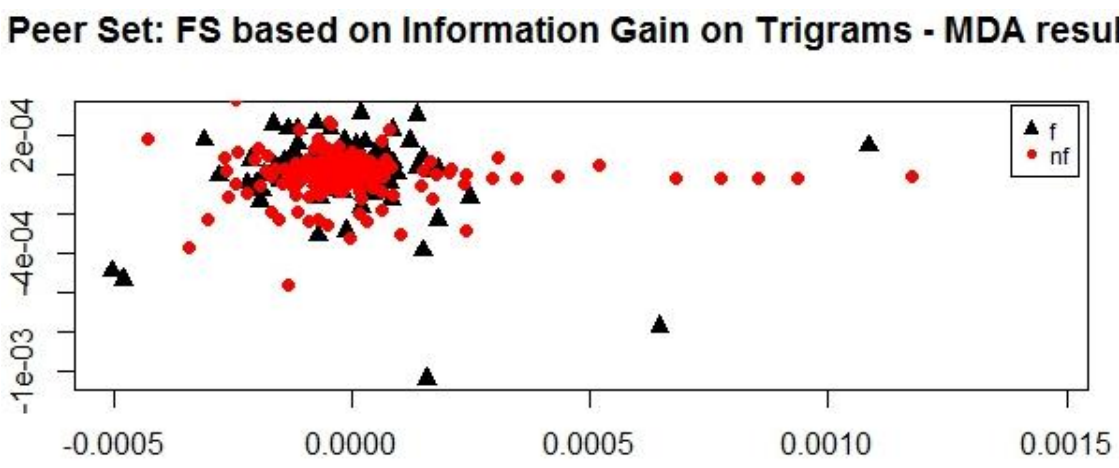
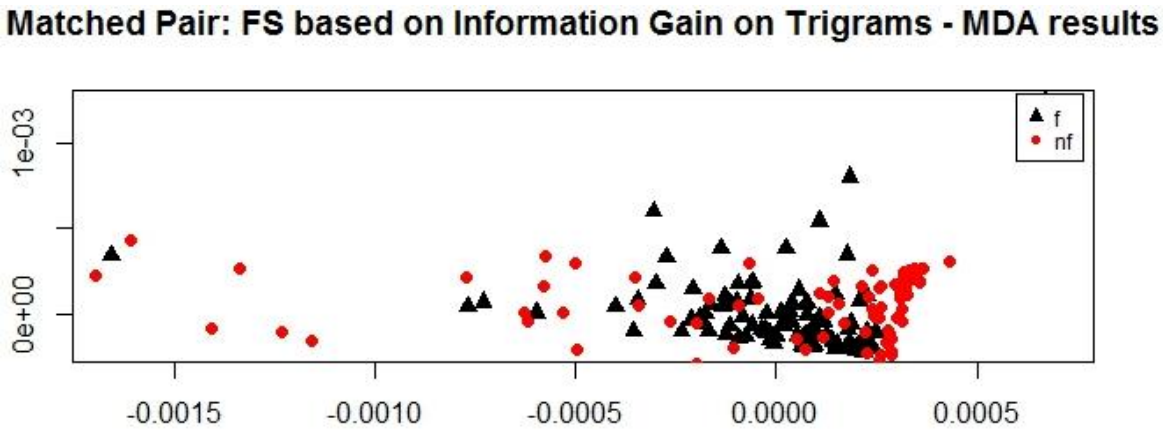
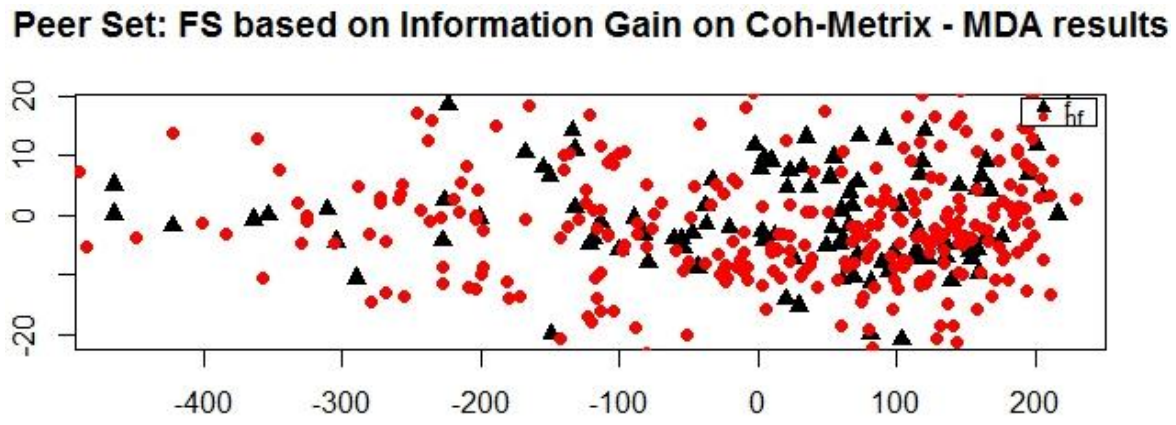
Document Representation	Chosen Features
Peer Set: Trigrams	'are expected to', 'the event that', 'of shares of', 'entered into an', 'the acquisition of', 'a range of', 'part of our', 'in the fair', 'should be read', 'a broad range', 'a change in', 'a combination of', 'a decline in', 'a decrease in', 'a increase in', 'a loss of'
	
Matched Pair: Trigrams	'the event that', 'ability to provide', 'in the event', 'primarily due to', 'based on the', 'one or more', 'entered into a', 'in the fair', 'comply with the', 'to do so', 'we are a', 'the end of', 'the event of', 'pursuant to the', 'of shares of', 'due to lower', 'in compared to',
	

Table M.3: Trigrams chosen by IG for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set – Coh-Metrix	'DESPL', 'PCNARz', 'SMCAUSwn', 'DESPC', 'SYNSTRUTt', 'PCNARp', 'LSAPP1d', 'CNCTempx', 'WRDAOAc', 'SYNSTRUTa', 'WRDHYPnv',



Matched Pair – Coh-Metrix	'LSAPP1d', 'DRPP', 'SYNSTRUTt', 'PCNARz', 'PCNARp', 'DRINF', 'WRDFRQa', 'WRDMEAc',

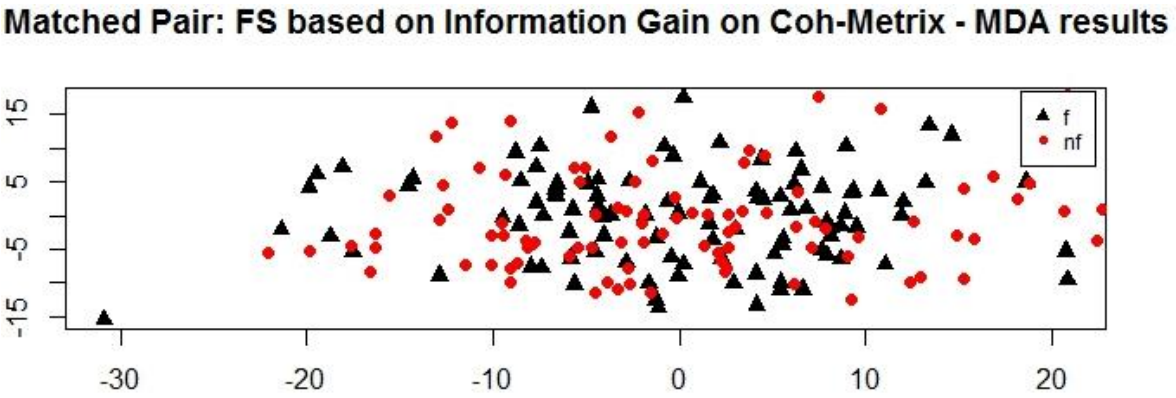


Table M.4: Coh-Metrix chosen by IG for PS and MP data set up and results of MDA computation.

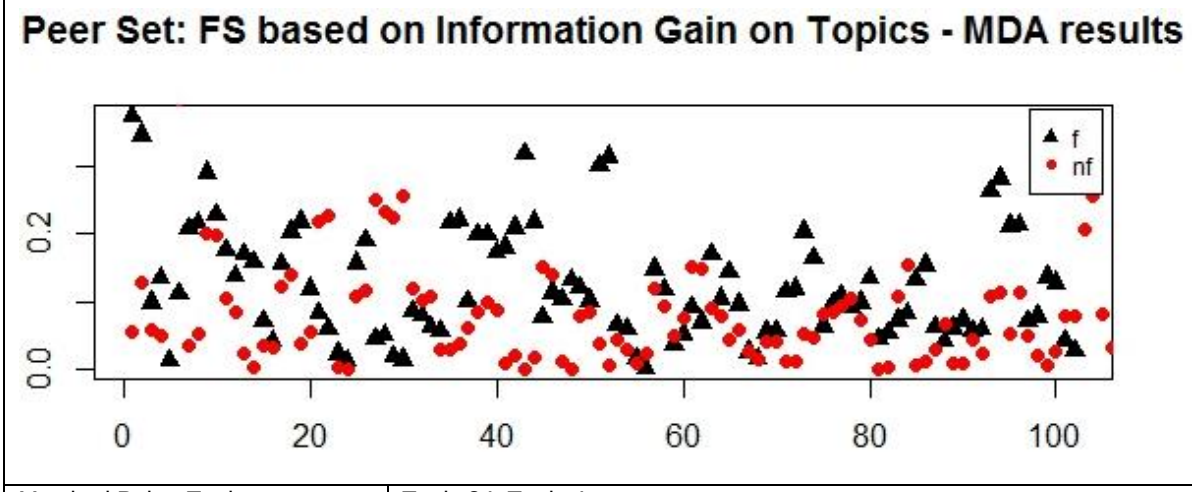
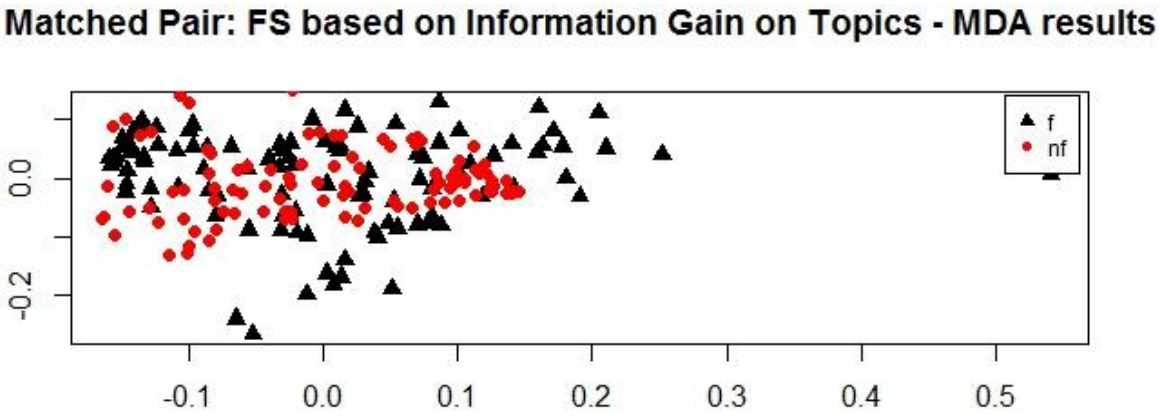
Document Representation	Chosen Features
Peer Set – Topics	Topic.24
	
Matched Pair – Topics	Topic.24, Topic 4
	

Table M.5: Topics chosen by IG for PS and MP data set up and results of MDA computation.

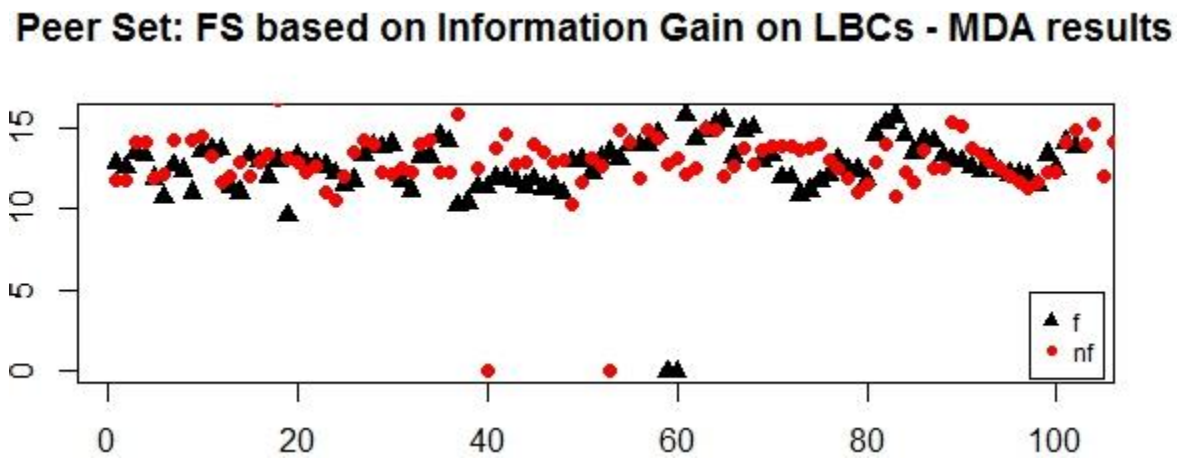
Document Representation	Chosen Features
Peer Set – LBC's	Temporal.Imm..Ratio
	
Matched Pair – LBCs	No features selected

Table M.6: LBCs chosen by IG for PS and MP data set up and results of MDA computation.

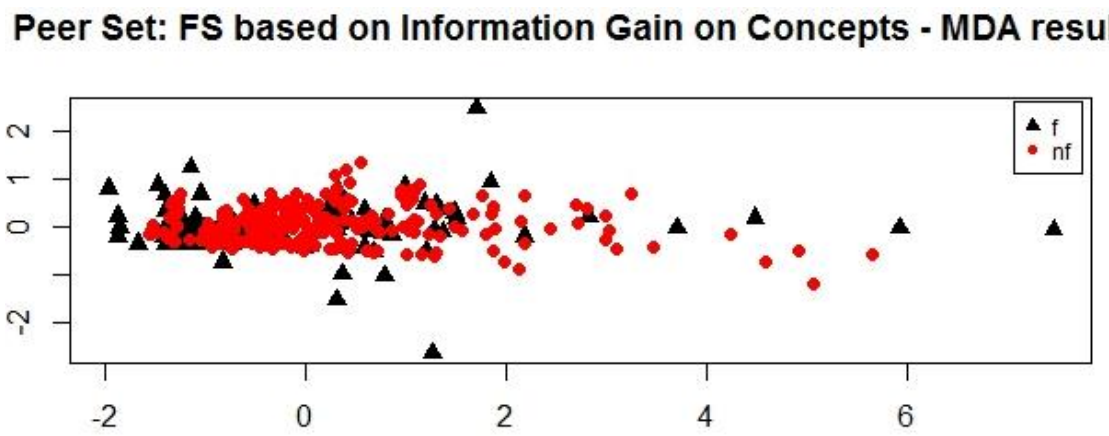
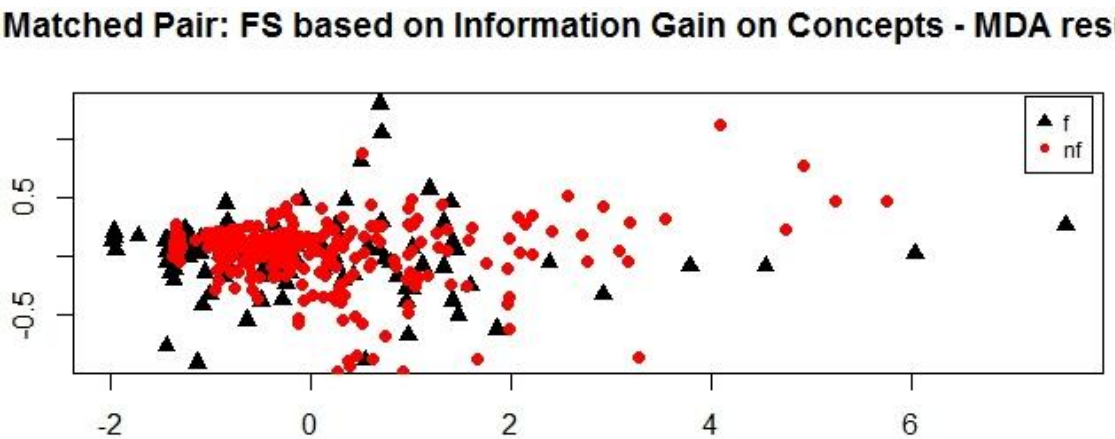
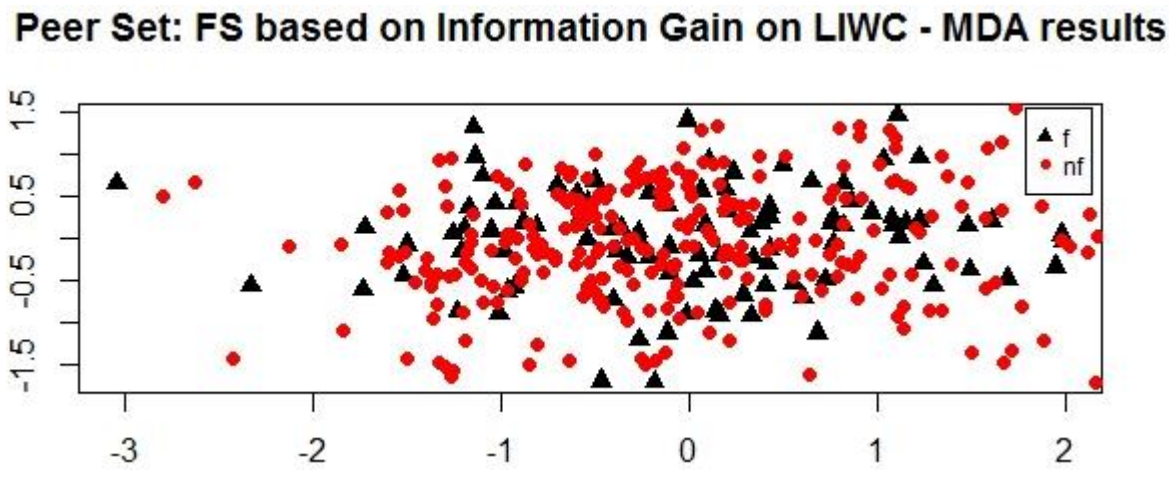
Document Representation	Chosen Features
Peer Set – Concepts	'stockholder.noun', 'procedure.noun', 'obtaining.verb', 'obtain.verb', 'performed.verb', 'refurbishment.noun', 'obtained.verb', 'offset.verb', 'continued.verb', 'identify.verb', 'acquisition.noun', 'event.noun', 'posted.verb', 'flat.Adj', 'failures.noun', 'impacted.verb', 'platforms.noun', 'mainly.Adv', 'depth.noun', 'device.noun', 'fragments.noun', 'mapped.verb', 'patent.verb', 'seizure.noun', 'signatures.noun', 'stockholm.noun', 'road.noun', 'forward.noun', 'improving.verb'
	
Matched Pair – Concepts	'improving.verb', 'distribute.verb', 'precedent.noun', 'obtained.verb', 'obtaining.verb', 'improve.verb', 'restoration.noun', 'acquisition.noun', 'cease.verb', 'merchandisers.noun', 'switches.noun', 'acquired.verb', 'ceased.verb', 'standards.noun', 'manufacturer.noun', 'division.noun', 'results.noun', 'association.noun', 'incurrence.noun', 'avoiding.verb', 'bankers.noun', 'caution.noun', 'discounts.noun', 'predictions.noun',
	

Table M.7: Concepts chosen by IG for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set – LIWC	"relative", "prep"



Matched Pair – LIWC	"relative", "prep"

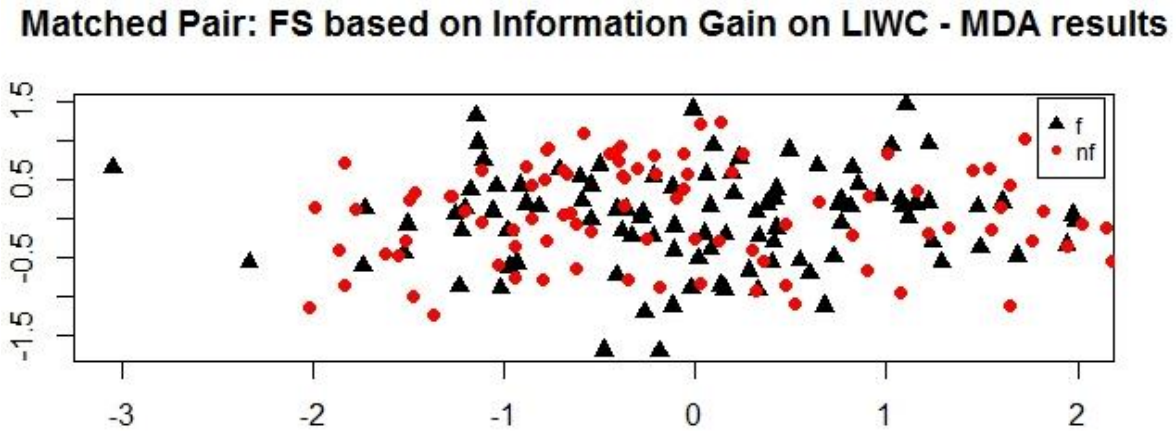
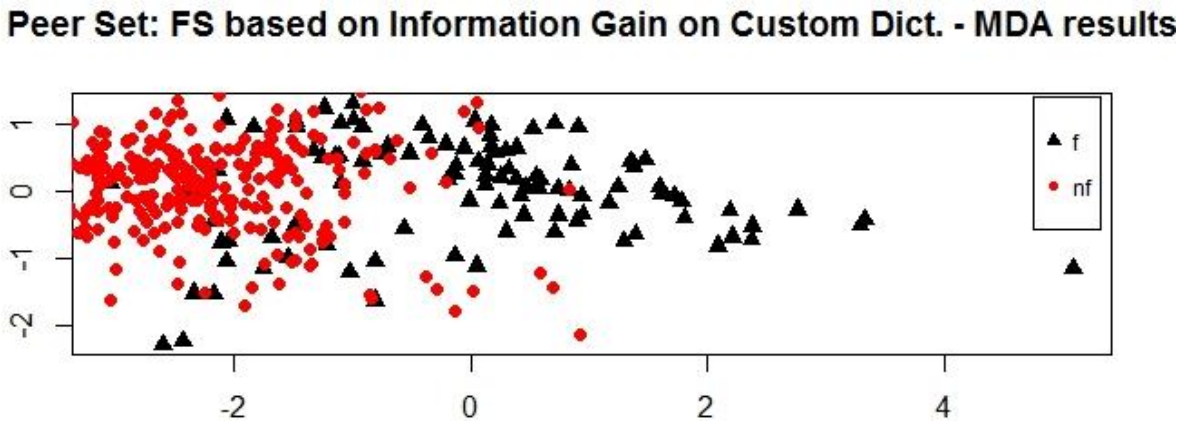


Table M.8: LIWC variables chosen by IG for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set – Custom Dictionaries	'ForwardLooking_Freq', 'negativity_Freq', 'positivity_Freq' , 'Uncert1_Freq'



Matched Pair – Custom Dictionaries	'ForwardLooking_Freq', 'negativity_Freq', 'positivity_Freq' , 'Uncert1_Freq'

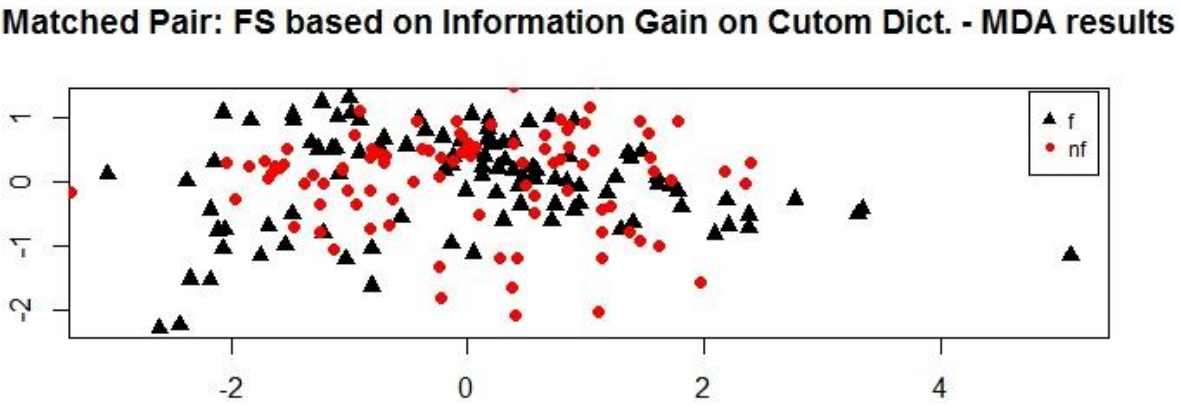
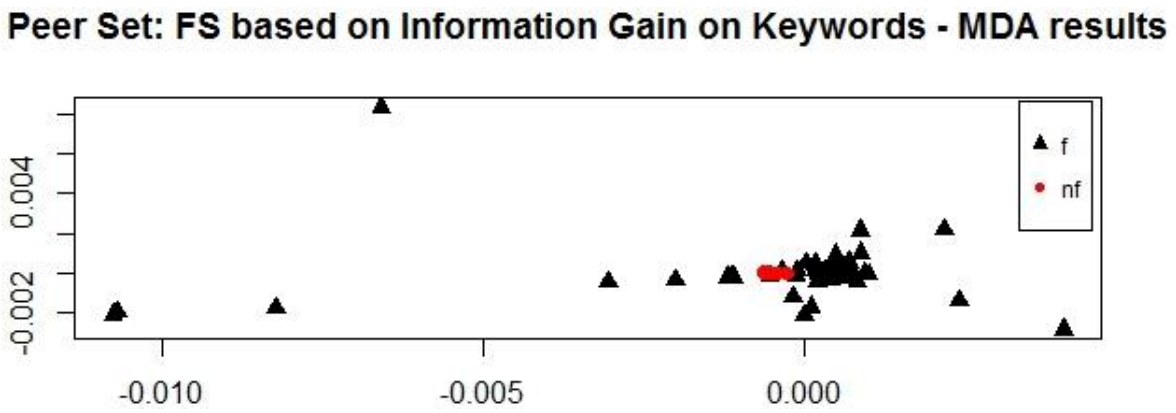


Table M.9: Custom Dictionaries chosen by IG for PS and MP data set up and results of MDA computation.

Document Representation	Chosen Features
Peer Set – Keywords	'also', 'capital', 'companies', 'growth', 'result', 'operations', 'tax', 'accounting', 'bankruptcy', 'operating', 'communities', 'management', 'expected', 'stockholder', 'event', 'placement', 'servicing', 'acquisition', 'net', 'documents', 'certain', 'believed', 'failures', 'losses', 'carriers', 'procurement', 'unsecured', 'increased', 'injury', 'inspection',



Matched Pair – Keywords	'also', 'companies', 'growth', 'result', 'required', 'accounting', 'results', 'primarily', 'certain', 'losses', 'general', 'compared', 'approvals', 'demand', 'flexible', 'retained', 'due', 'unfavorable', 'unsecured', 'stockholder', 'tax', 'education', 'treasury', 'international', 'outlook', 'placement', 'servicing', 'solid', 'bankruptcy', 'decreased',

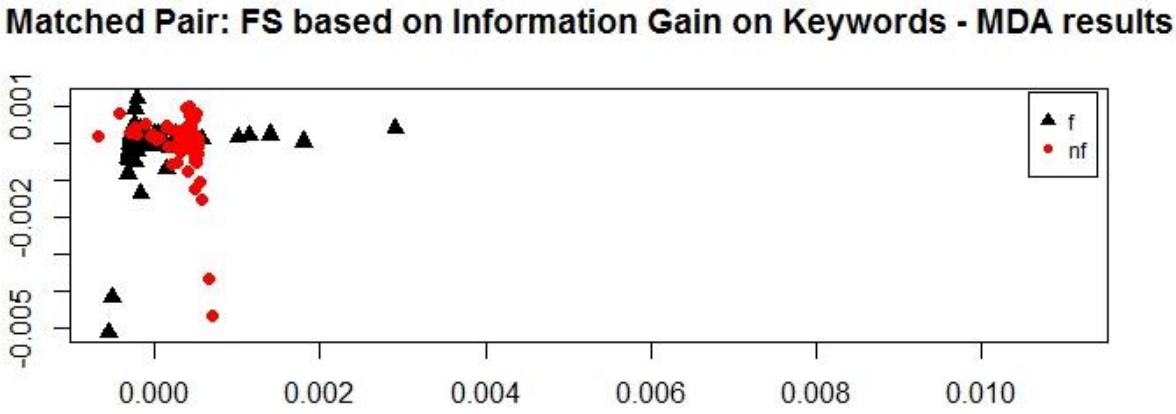


Table M.10: Keywords chosen by IG for PS and MP data set up and results of MDA computation.

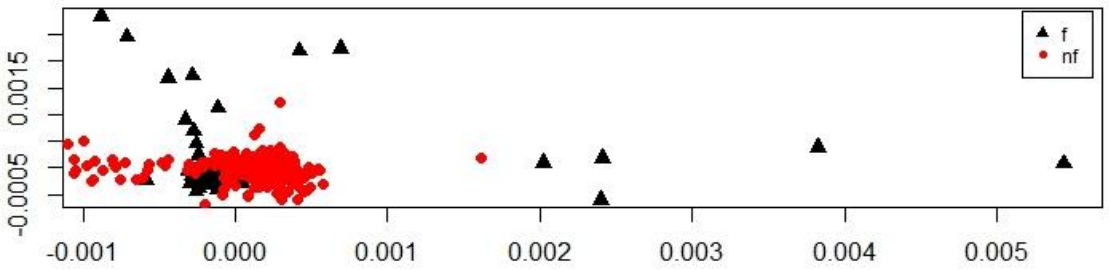
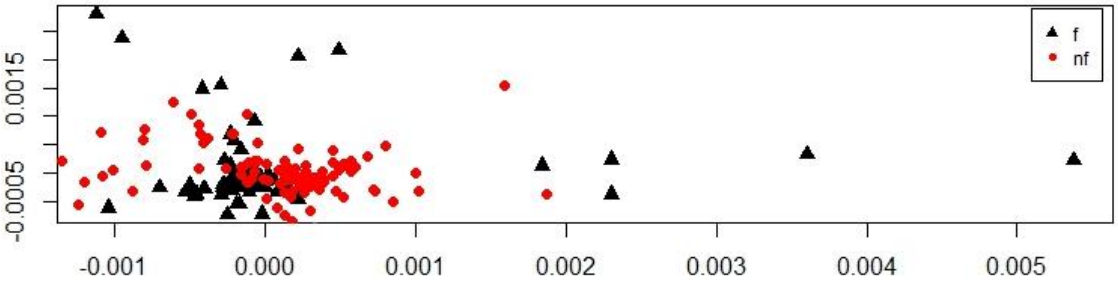
Document Representation	Chosen Features
Peer Set – Rutherford keywords	'years', 'company', 'financial', 'increase', 'significant', 'capital', 'growth', 'include', 'result', 'operations', 'loss', 'tax', 'operating', 'make', 'management', 'store', 'net', 'turnover', 'rate', 'item', 'cash', 'cost', 'interest', 'high', 'total', 'revenue', 'new', 'debt', 'sale', 'reduce',
Peer Set: FS based on Information Gain on Rutherford Keywords - MDA results 	
Matched Pair – Rutherford keywords	'years', 'company', 'significant', 'growth', 'increase', 'include', 'result', 'loss', 'rate', 'make', 'item', 'new', 'cash', 'due', 'sale', 'tax', 'store', 'number', 'total', 'turnover', 'interest', 'high', 'revenue', 'decrease', 'risk', 'major', 'lower', 'development', 'services', 'last',
Matched Pair: FS based on Information Gain on Rutherford Keywords - MDA results 	

Table M.11: Keywords (Rutherford) chosen by IG for PS and MP data set up and results of MDA computation.

APPENDIX N

Classification results for PCA selected Unigrams

PCA feature selection on Unigrams - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.96,1	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.96,1	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.96,1	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.96,1	0.75	3.355e-13	1	1	1
SVM	0.91	0.92	0.98	0.97	0.9,0.9	0.75	2.185e-09	0.95	0.97	0.95

Table N.1: Classification results (Unigrams-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
addit	100	addit	100	addit	100	addit	100	addit	100
requir	98.75	requir	19	requir	46	requir	46	requir	97.9
may	61.67	expens	1	may	3	may	13	may	57.64
achiev	56.61	primarili	1	expens	3	expens	3	achiev	57.28
certain	56.38	new	0.1	achiev	2.31	achiev	2.31	certain	53.71
expens	47.8	excel	0.1	certain	1.95	certain	1.95	expens	47.19
new	42.07	subject	0.06	excel	0.96	excel	0.98	primarili	40.08
upon	39.91	import	0.06	group	0.78	group	0.78	new	39.89
primarili	39.23	strong	0	primarili	0.49	primarili	0.45	liabil	36.56
excel	37.77	start	0	liabil	0.41	liabil	0.41	relat	35.95

Table N.2: Unigrams chosen by classifier for (Unigrams-PCA-PS).

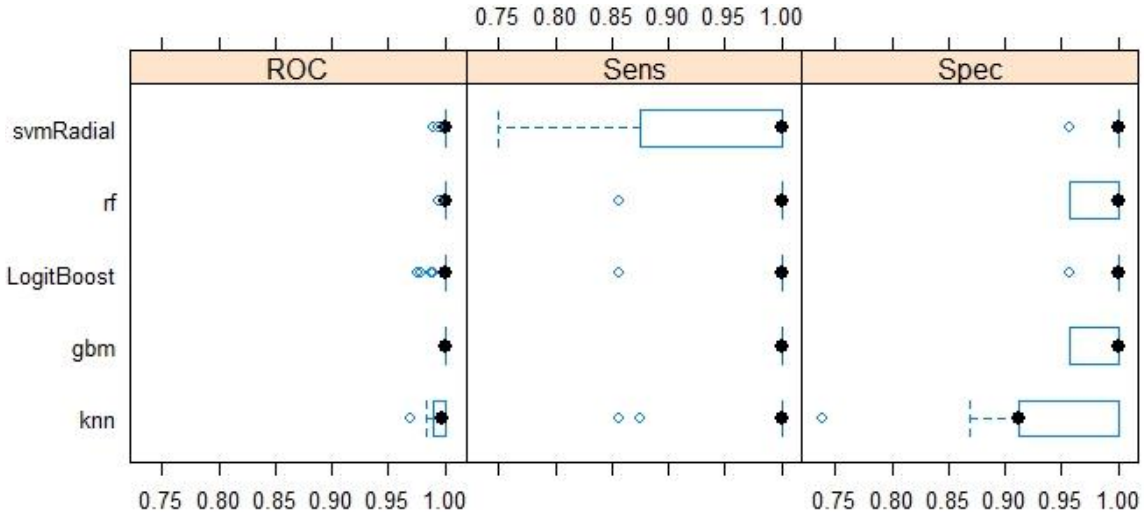


Figure N.1: ROC, Sensitivity and Specificity for classifiers (Unigrams-PCA-PS).

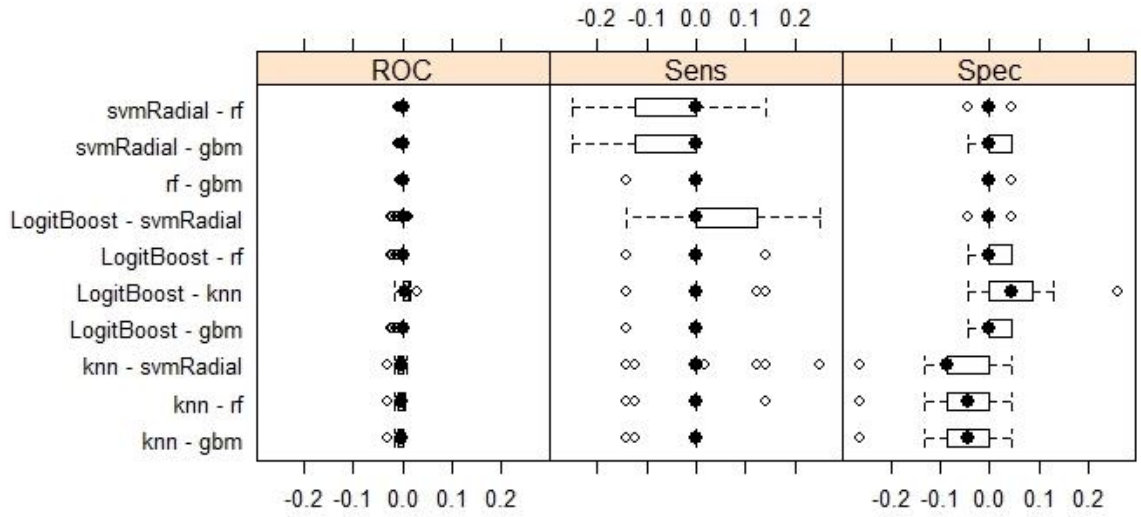


Figure N.2: Performance compared between classifiers (Unigrams-PCA-PS).

PCA feature selection on Unigrams – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.96	1	0.96	0.98	0.89, 0.99	0.5	4.53e-14	0.96	1	0.98
SGB	0.96	1	1	0..98	0.89, 0.99	0.5	4.53e-14	0.96	1	0.98
RF	1	1	1	1	0.92, 1	0.5	8.882e-16	1	1	1
kNN	1	1	1	1	0.92,1	0.5	8.882e-16	1	1	1
SVM	**1**	**1**	**1**	**1**	0.92,1	0.5	8.882e-16	1	1	1

Table N.3: Table N.1: Classification results (Unigrams-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
signific	100.00	signific	100	also	100	signific	100	increas	100
active	98.62	addit	90.772	increas	92.91	activ	98.62	signific	100
expect	98.62	includ	42.180	respect	89.68	expect	98.62	respect	100
addit	98.62	activ	0.23379	addit	86.7	addit	98.62	activ	100
includ	98.62	follow	0.0742	follow	82.44	respect	98.62	also	100
respect	98.62	use	0.05493	basi	81.38	increas	98.62	follow	100
increas	98.62	basi	0.04932	expect	77.61	includ	98.62	expect	100
account	97.25	respect	0.02401	signific	76.14	follow	97.25	includ	100
follow	97.25	account	0	futur	73.89	basi	97.25	futur	98.53
period	97.25	liabil	0	requir	73.28	also	97.25	account	98.53

Table N.4: Unigrams chosen by classifier as significant (Unigrams-PCA-MP).

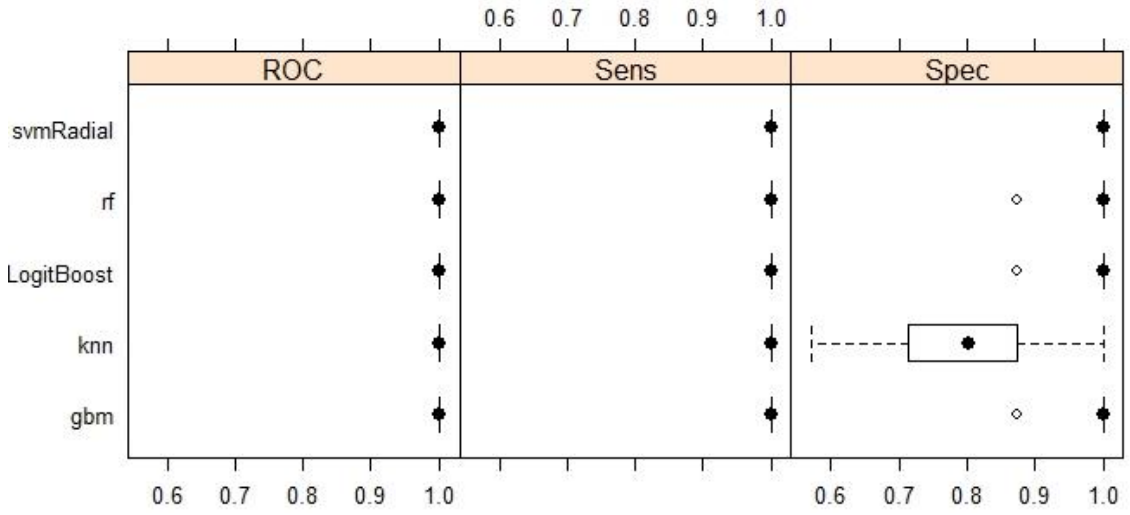


Figure N.3: ROC, Sensitivity and Specificity for classifiers (Unigrams-PCA-MP).

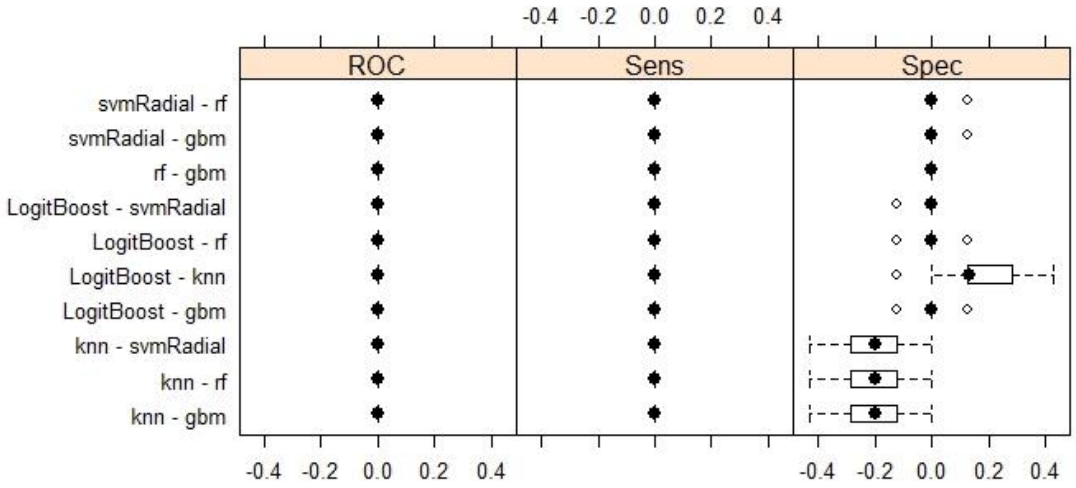


Figure N.4: Performance compared between classifiers (Unigrams-PCA-MP).

Classification results for Boruta selected Unigrams

Boruta feature selection on Unigrams - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.96, 1	0.75	3.35e-13	1	1	1
SGB	1	1	1	1	0.96, 1	0.75	3.35e-13	1	1	1
RF	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1
kNN	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1
SVM	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1

Table N.5: Classification results (Unigrams-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
increas	100	increas	100	expect	100	increas	100	increas	100
share	99.68	signific	50.592	increas	98.88	expect	99.68	signific	100
expect	99.68	share	45.125	futur	89.17	signific	99.68	expect	99.7
signific	99.68	financi	29.611	share	85.15	share	99.68	financi	99.7
addit	99.36	capit	17.035	addit	84.85	futur	99.36	addit	99.41
financi	99.36	expect	14.078	requir	81.19	financi	99.36	capit	99.41
futur	99.36	order	0.02344	capit	78.07	addit	99.36	share	99.41
capit	99.04	tax	0.01264	signific	77.8	capit	99.04	current	99.41
current	98.72	interest	0	financi	72.03	current	98.72	futur	99.41
requir	98.4	made	0	current	65.52	requir	98.4	growth	99.11
account	97.76	outlook	0	account	54.85	account	97.76	account	99.11

Table N.6: Unigrams chosen by classifier as significant (Unigrams-Boruta-PS).

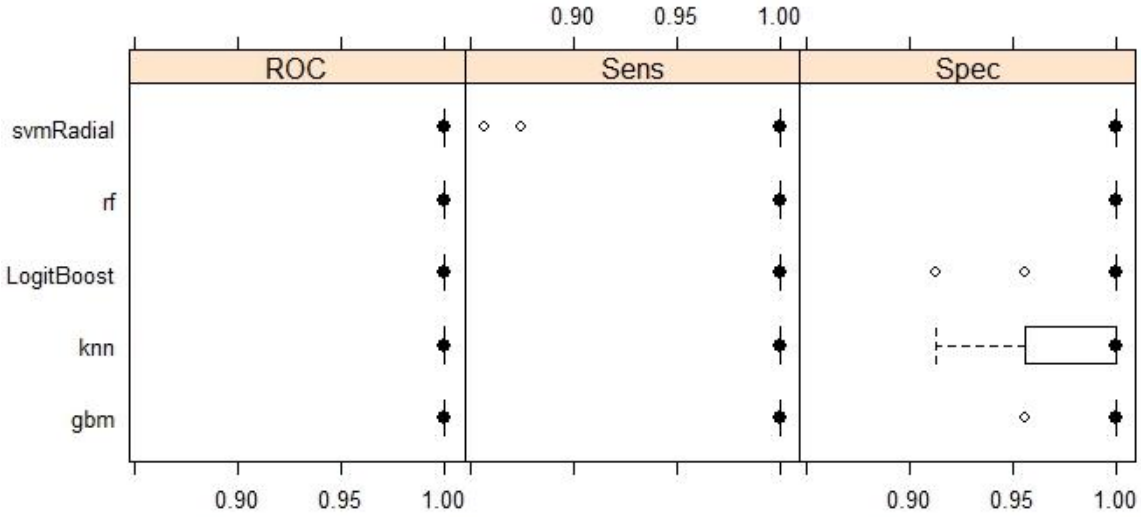


Figure N.5: ROC, Sensitivity and Specificity for classifiers (Unigrams-Boruta-PS).

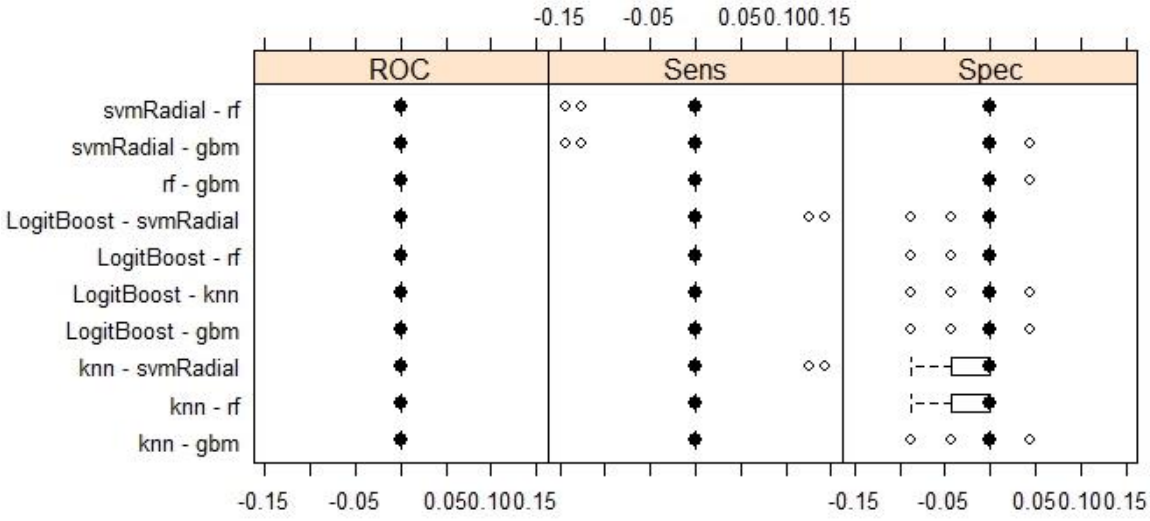


Figure N.6 Performance compared between classifiers (Unigrams-Boruta-PS).

Boruta feature selection on Unigrams - Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.96	1	0.96	0.98	0.89,0.99	0.5	4.53e-14	0.96	1	0.98
SGB	1	1	1	1	0.92,1	0.5	8.88e-16	1	1	1
RF	1	1	1	1	0.92,1	0.5	8.88e-16	1	1	1
kNN	0.88	1	0.88	0.94	0.83,0.98	0.5	0.8	0.89	1	0.94
SVM	0.96	0.92	1	0.96	0.86,0.99	0.5	1.13e-12	1	0.92	0.96

Table N.7: Classification results (Unigrams-Boruta-MP).

Matched Pair– Variable Importance									
LR		***SVM***		***RF***		***kNN***		***SVM***	
signific	100	increas	100	asset	100	signific	100	increas	100
asset	95.3	signific	58.81	signific	96.958	increas	100	signific	100
increas	95.39	asset	26.39	increas	88.959	asset	95.813	growth	95.092
growth	86.06	growth	4.93	loss	70.566	growth	91.626	revenu	85.277
loss	86.07	interest	0.341	growth	53.803	revenu	87.439	asset	80.37
revenu	76.79	statement	0.10	revenu	49.682	loss	83.252	loss	80.37
interest	29.04	maintain	0	plan	6.326	plan	18.651	plan	16.061
statement	14.88	revenu	0	interest	5.64	maintain	11.691	interest	11.026
plan	1.808	loss	0	stateme	4.375	interest	8.646	maintain	9.114
maintain	0	plan	0	maintain	0	stateme	0	stateme	0

Table N.8: Unigrams chosen by classifier as significant (Unigrams-Boruta-MP)

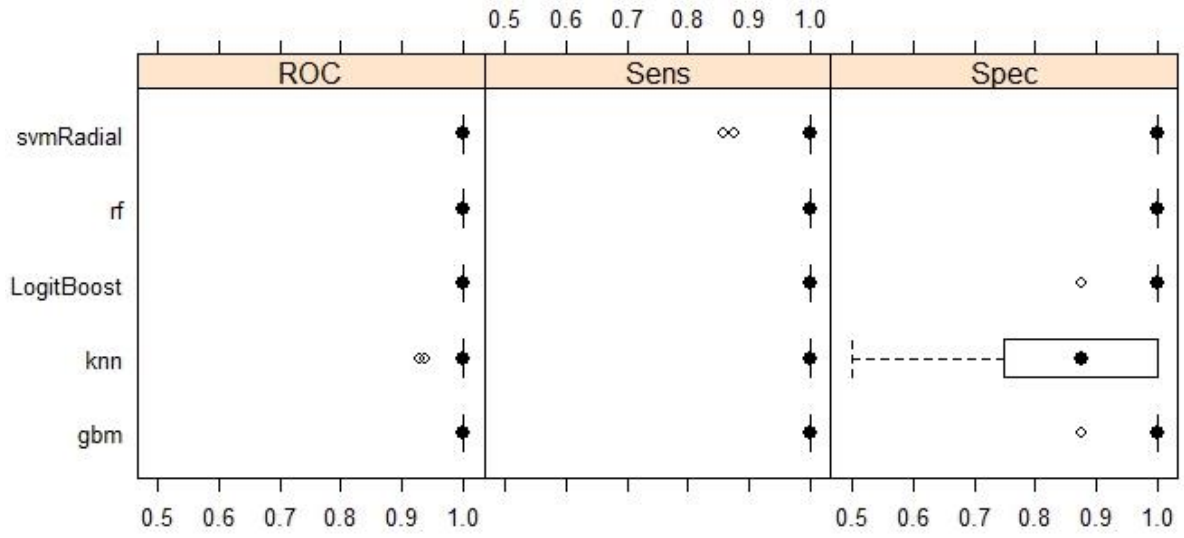


Figure N.7: ROC, Sensitivity and Specificity for classifiers (Unigrams-Boruta-MP).

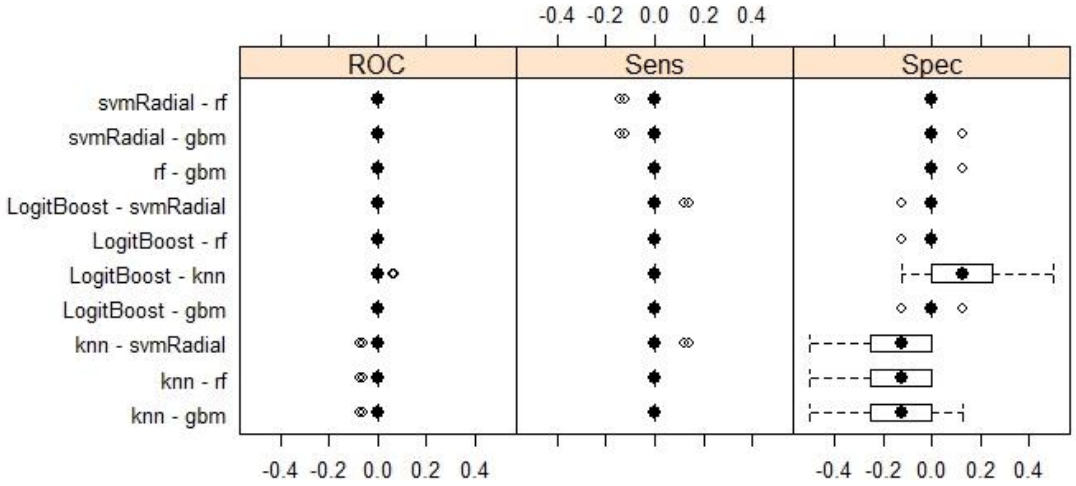


Figure N.8: Performance compared between classifiers (Unigrams-Boruta-MP).

Classification results for Information Gain (IG) selected Unigrams

IG feature selection on Unigrams – Peer set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.96, 1	0.75	3.35e-13	1	1	1
SGB	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1
RF	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1
kNN	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1
SVM	1	1	1	1	0.96,1	0.75	1	1	1	1

Table N.9: Classification results (Unigrams-IG-PS).

Peer set– Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
increas	100	increas	1.00E+02	increa	100	increas	100	increas	100
will	100	share	3.84E+01	will	94.25	expect	99.23	share	100
expect	99.29	will	3.81E+01	signific	86.39	will	99.23	will	100
signific	99.29	signific	2.86E+01	share	82.49	signific	98.46	signific	99.22
share	99.29	capit	1.49E+01	futur	64	servic	98.46	expect	99.22
financi	98.57	financi	5.54E+00	intern	59.67	sale	98.46	addit	98.44
addit	98.57	annual	1.15E-01	period	54.26	addit	98.46	financi	98.44
futur	98.57	period	1.10E-01	expect	53.5	share	97.69	futur	98.44
sale	97.86	cost	7.68E-02	addit	53.39	futur	97.69	current	97.67
capit	97.86	tax	3.77E-02	servic	53.12	financi	97.69	servic	97.67

Table N.10: Unigrams chosen by classifier as significant (Unigrams-IG-PS).

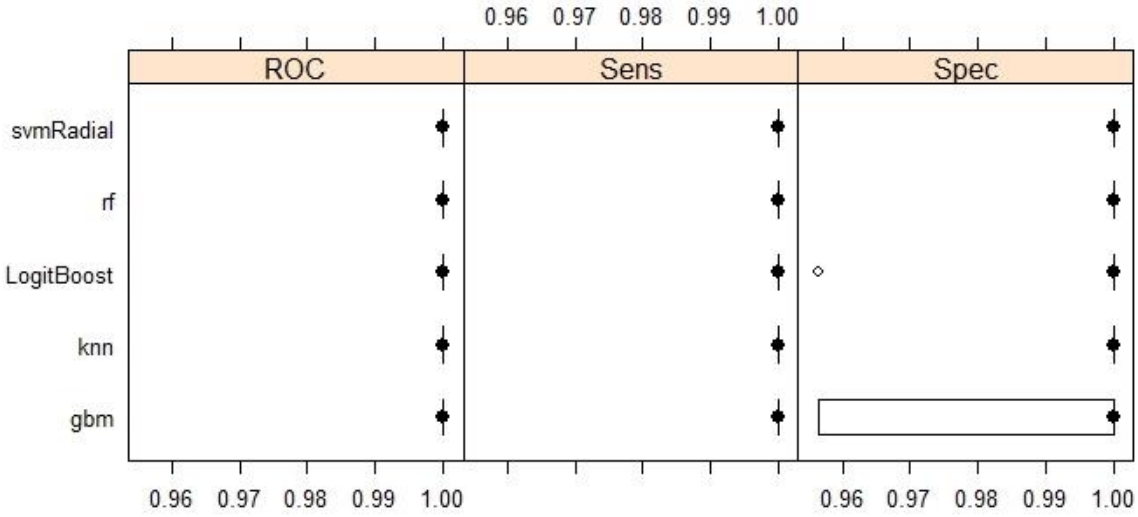


Figure N.9: Performance compared between classifiers (Unigrams-IG-PS).

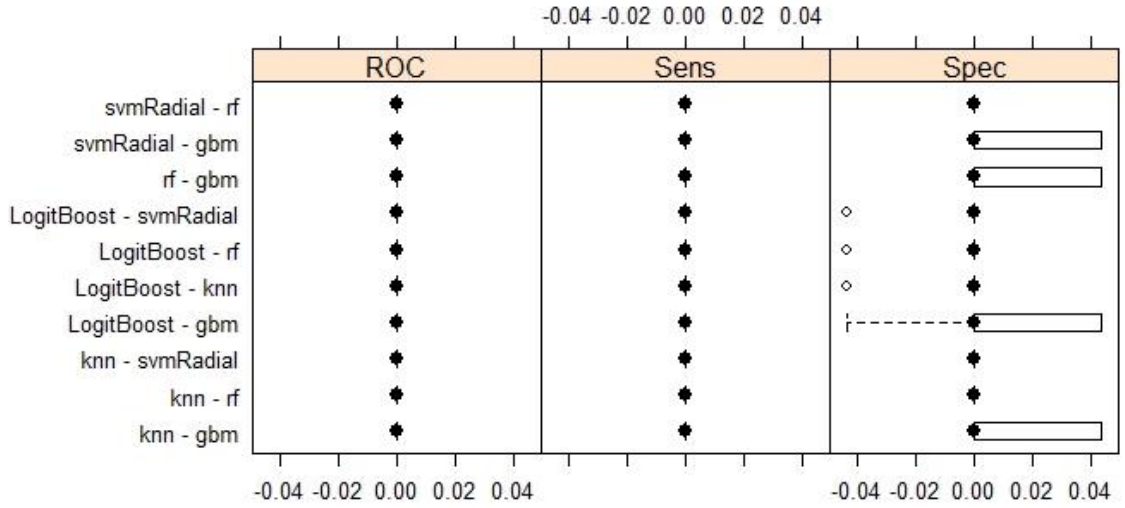


Figure N.10: Performance compared between classifiers (Unigrams-Boruta-MP).

IG feature selection on Unigrams – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.96	1	0.96	0.98	0.89, 0.99	0.5	4.53e-14	0.96	0.96	0.98
SGB	1	1	1	1	0.92, 1	0.5	8.88e-16	1	1	1
RF	1	1	1	1	0.92, 1	0.5	8.88e-16	1	1	1
kNN	1	1	1	1	0.92, 1	0.5	8.88e-16	1	1	1
SVM	1	1	1	1	0.92, 1	0.5	8.88e-16	1	1	1

Table N.11: Classification results (Unigrams-IG-MP).

Peer set– Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
signific	100	end	1.00E+02	asset	100	signific	100	signific	100
increas	99.09	includ	8.77E+01	signific	95.56	part	99.09	includ	100
intern	99.09	expect	8.37E+01	increas	91.62	respect	99.09	part	100
expect	99.09	activ	5.79E+01	activ	79.36	expect	99.09	valu	100
activ	99.09	part	2.76E+01	intern	74.31	addit	99.09	manag	100
asset	99.09	signific	1.48E+01	account	68.6	includ	99.09	end	97.37
includ	99.09	expens	1.52E-01	part	65.98	end	99.09	activ	97.37
part	99.09	interest	4.66E-02	end	64.9	sale	99.09	expect	97.37
addit	99.09	current	4.52E-02	addit	63.63	activ	99.09	increas	97.37
respect	99.09	primarili	3.25E-02	sale	55.54	increas	99.09	follow	97.37
end	99.09	respect	2.18E-02	manag	54.6	asset	99.09	intern	97.37
sale	99.09	growth	1.85E-02	includ	54.29	intern	99.09	addit	94.75

Table N.12: Unigrams chosen by classifier as significant (Unigrams-IG-MP).

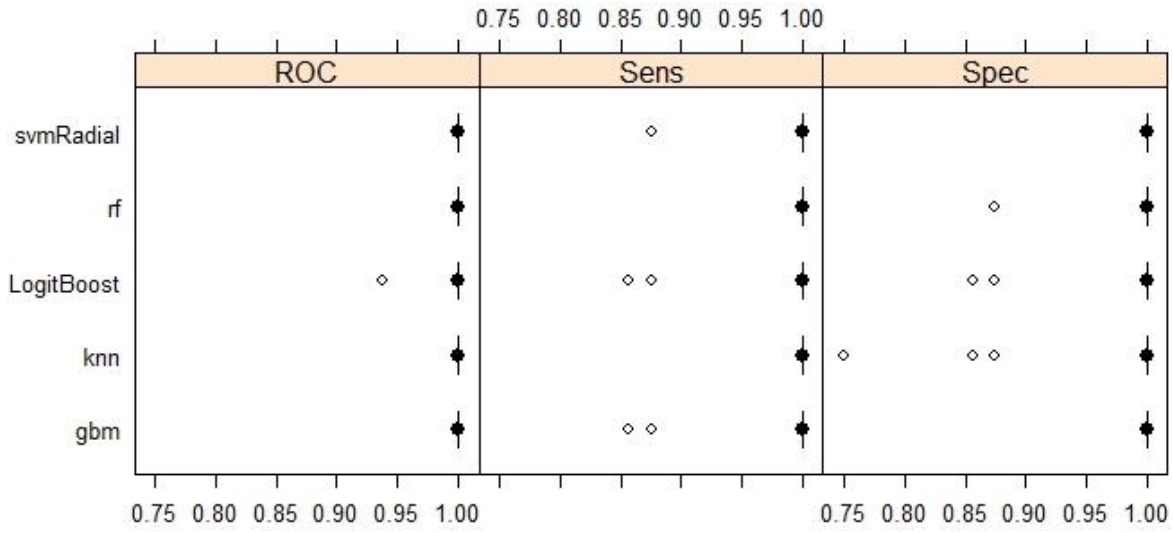


Figure N.11: ROC, Sensitivity and Specificity for classifiers (Unigrams-IG-MP).

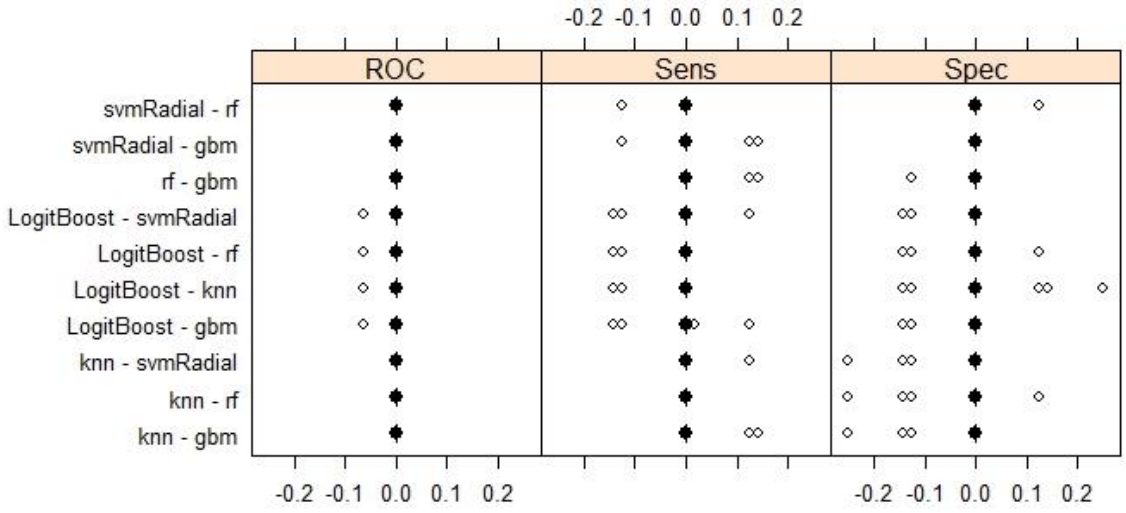


Figure N.12: Performance compared between classifiers (Unigrams-IG-MP).

APPENDIX O

Classification results for PCA selected Bigrams

PCA feature selection on Bigrams - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.2	0.40	0.84	0.73	0.63, 0.81	0.75	0.72	0.45	0.81	0.62
SGB	0.3	0.32	0.93	0.78	0.68, 0.85	0.75	0.28	0.61	0.80	0.62
RF	0.059	0.04	1	0.76	0.66, 0.84	0.75	0.46	1	0.76	0.52
kNN	0	0	1	0.75	0.65, 0.83	0.75	0.55	0	0.75	0.50
SVM	0.05	0.12	0.92	0.72	0.62, 0.80	0.75	0.79	0.33	0.76	0.52

Table O.1: Classification results (Bigrams-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
to obtain	100	result in	100	ability to	100	to obtain	100	ability to	100
ability to	80.38	ability to	80.38	may be	91.17	ability to	80.38	to obtain	99.55
be able	78.09	not be	78.09	to obtain	90.69	be able	78.09	be able	90.35
do not	77.66	subject to	77.66	subject to	84.28	do not	77.66	adversely affect	84.13
may be	75.47	or other	75.47	do not	84.05	may be	75.47	may be	82.94
our revenues	74.59	do not	74.59	result in	81.51	our revenues	74.59	which could	81.27
not be	70.35	could be	70.35	not be	70.47	not be	70.35	may not	77
we believe	68.35	we do	68.35	or other	69.55	we believe	68.35	and could	70.16
we may	68.25	that our	68.25	be able	66.33	we may	68.25	not be	68.71
may not	67.77	may be	67.77	could be	65.75	may not	67.77	could adversely	67.8
unable to	67.45	may not	67.45	adversely affect	65.73	unable to	67.45	unable to	66.53

Table O.2: Bigrams chosen by classifier for (Bigrams-PCA-PS).

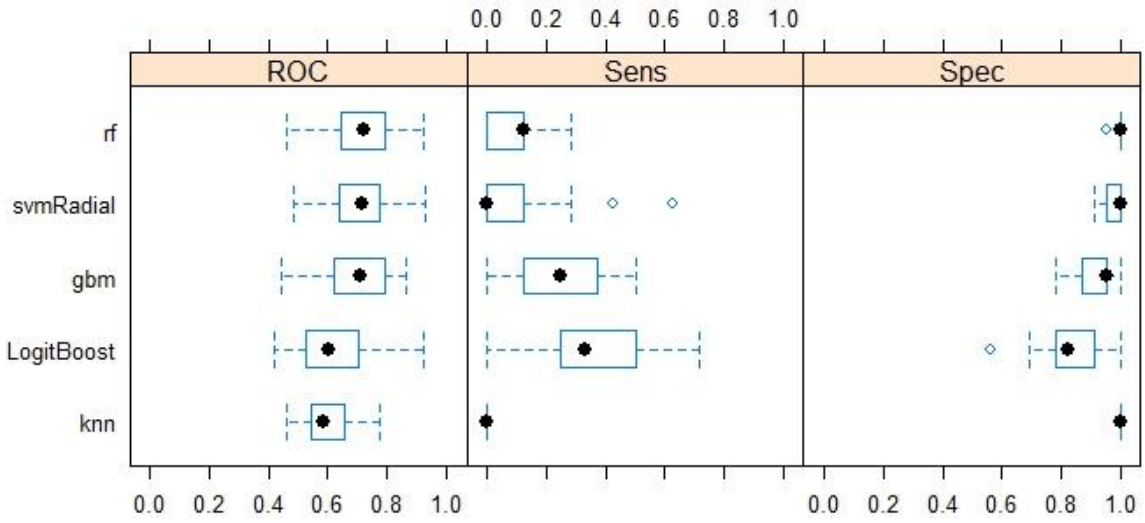


Figure O.1: ROC, Sensitivity and Specificity for classifiers (Bigrams-PCA-PS).

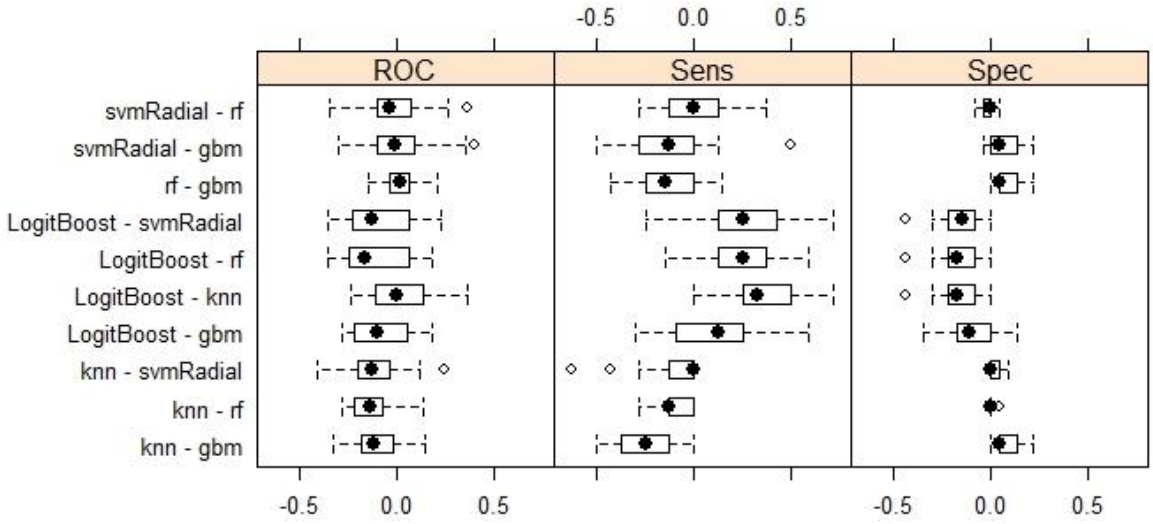


Figure O.2: Performance compared between classifiers (Bigrams-PCA-PS).

PCA feature selection on Bigrams – Matched Pair (MP)										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.64	0.60	0.62	0.47, 0.75	0.5	0.05	0.61	0.62	0.62
SGB	0.52	0.80	0.72	0.76	0.61, 0.86	0.5	0.0001	0.74	0.78	0.76
RF	0.36	0.64	0.72	0.68	0.53, 0.80	0.5	0.0076	0.69	0.66	0.68
kNN	0	0.48	0.52	0.50	0.35, 0.64	0.5	0.55	0.50	0.50	0.50
SVM	0.28	0.60	0.68		0.64	0.5	0.03	0.65	0.62	0.64

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
to obtain	100	may be	100	be able	100	to obtain	100	may be	100
be able	99	be able	99	subject to	99	be able	99	be able	99
may be	93.2	ability to	93.2	to obtain	93.2	may be	93.2	ability to	93.2
ability to	87.4	interest rate	87.4	may be	87.4	ability to	87.4	to obtain	87.4
that we	82.2	results of	82.2	interest rate	82.2	that we	82.2	may not	82.2
we may	81.27	believes that	81.2	ability to	81.27	we may	81.27	which could	81.27
may not	76.73	that we	76.7	that we	76.73	may not	76.73	could be	76.73
could be	75.93	to obtain	75.9	of operations	75.93	could be	75.93	subject to	75.93
we believe	74.87	of operations	74.8	do not	74.87	we believe	74.87	adversely affect	74.87
subject to	74.8	and may	74.8	results of	74.8	subject to	74.8	could result	74.8

Table O.4: Bigrams chosen by classifier as significant (Bigrams-PCA-MP).

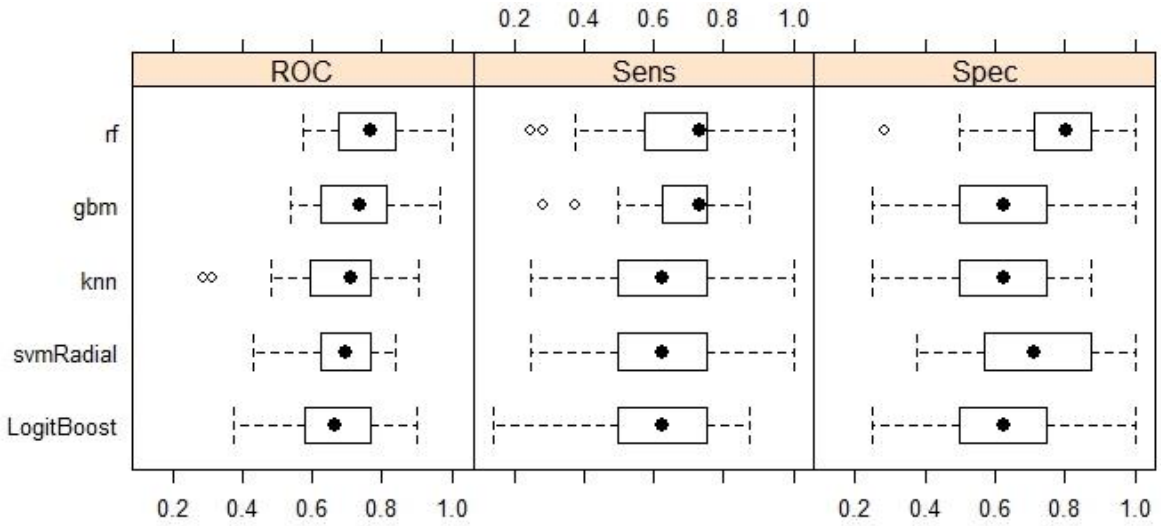


Figure O.3: ROC, Sensitivity and Specificity for classifiers (Bigrams-PCA-MP).

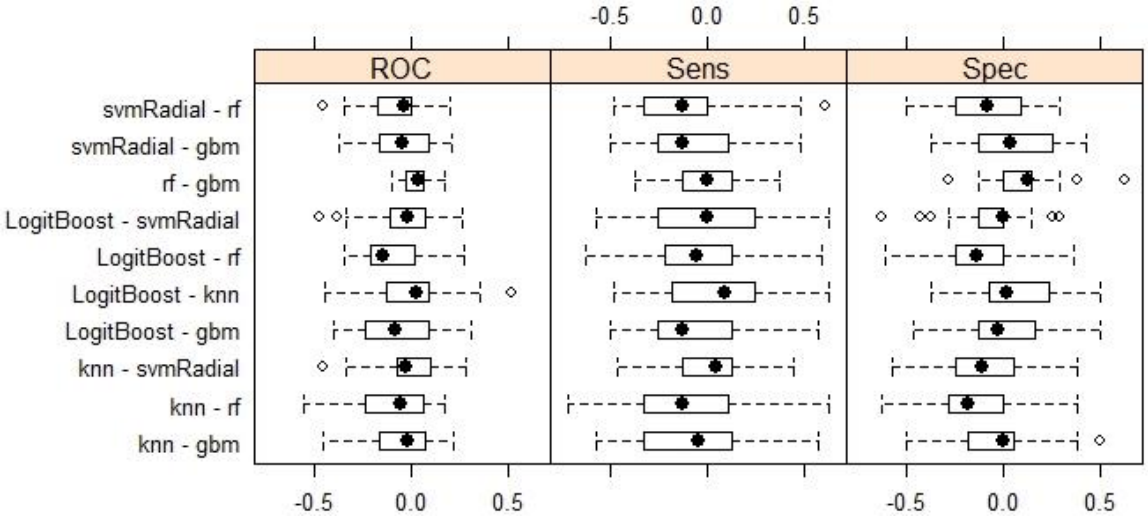


Figure O.4: Performance compared between classifiers (Bigrams-PCA-MP).

Classification results for Boruta selected Bigrams

Boruta feature selection on Bigrams – Peer Set (PS)										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.52	0.52	0.94	0.84	0.75, 0.90	0.75	0.02	0.76	0.85	0.73
SGB	0.76	0.72	0.98	0.92	0.84, 0.96	0.75	1.236e-05	0.94	0.91	0.85
RF	0.61	0.52	1.0	0.88	0.80, 0.93	0.75	0.0014	1	0.86	0.76
kNN	0.07	0.08	0.97	0.75	0.65, 0.83	0.75	0.55	0.50	0.76	0.52
SVM	0.59	0.60	0.94	0.86	0.77, 0.92	0.75	0.0054	0.78	0.87	0.77

Table O.5: Classification results (Bigrams-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
into a	100	to date	100	into a	100	into a	100	into a	100
purchase price	91	into a	82	to date	88	to date	96	acquisition of	87
necessary to	86	state of	77	in and	77	them to	90	continued to	87
in and	85	the fiscal	72	of approximate	72	in and	84	necessary to	86
to obtain	84	of approximate	72	the event	62	to obtain	83	in and	84
of approximat	83.	volume of	54	them to	61	of approximat	82	year ended	84
borrowings under	82	in and	41	acquisition of	56	borrowings under	81	the acquisition	80
continued to	82	acquisition of	39	borrowings under	53	acquisition of	79	due to	80
year ended	81	the event	37	volume of	47	the event	76.7	purchase price	79
be required	81	borrowings under	37	accounted for	46	the acquisition	73	to obtain	79
acquisition of	80	them to	36.	the acquisition	44	to enter	72	of approximat	77
annual report	79	event that	35	to obtain	41	the state	68	annual report	75
the acquisition	75	subsidiaries are	29	state of	41	state of	65.7	borrowings under	74

Table O.6: Bigrams chosen by classifier as significant (Bigrams-Boruta-PS).

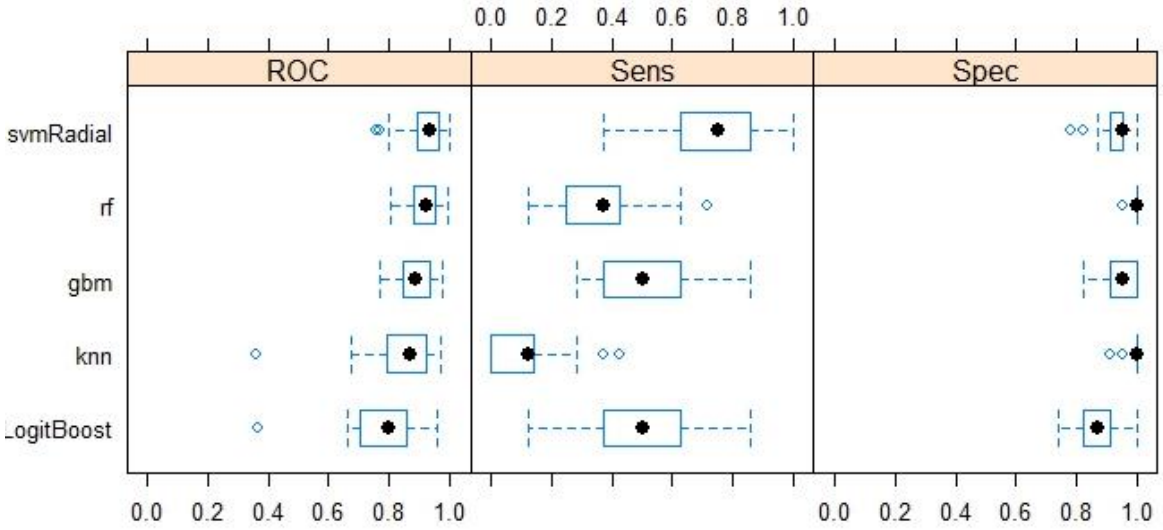


Figure O.5: ROC, Sensitivity and Specificity for classifiers (Bigrams-Boruta-PS).

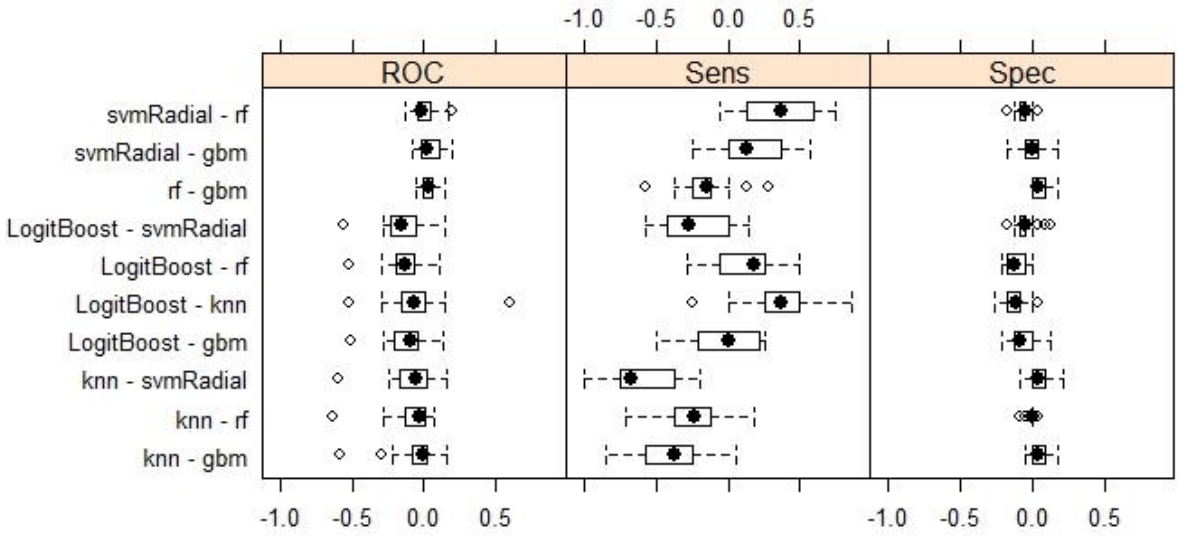


Figure O.6: Performance compared between classifiers (Bigrams-Boruta-PS).

Boruta feature selection on Bigrams – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.6	0.88	0.72	0.8	0.66, 0.89	0.5	1.19e-05	0.75	0.85	0.80
SGB	0.6	0.96	0.64	0.8	0.66, 0.89	0.5	1.193e-05	0.72	0.94	0.80
RF	0.64	0.92	0.72	0.82	0.68, 0.91	0.5	2.807e-06	0.76	0.90	0.82
kNN	0.4	0.72	0.68	0.7	0.55, 0.82	0.5	0.0033	0.69	0.70	0.70
SVM	0.72	0.84	0.88	0.86	0.73, 0.94	0.5	1.049e-07	0.87	0.84	0.86

Table O.7: Classification results (Bigrams-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
acquisition of	100	acquisition of	100	acquisition of	100	acquisition of	100	continued to	100
continued to	97	primarily due	95	primarily due	92	continued to	97	operating income	93
operating income	95	contributed to	88	contributed to	81	operating income	95	acquisition of	79
primarily due	89	continued to	78	continued to	68	primarily due	89	commitment to	74
contributed to	86	operating income	67	greater financial	55	contributed to	86	contributed to	73
income increased	83	greater financial	62	operating income	51	income increased	83	income increased	72
commitment to	82	be limited	24	income increased	27	commitment to	82	primarily due	69
have greater	25	income increased	10	commitment to	25	have greater	25.8	have greater	11
be limited	2	commitment to	1	be limited	13	be limited	2	be limited	3
greater financial	0	have greater	0	have greater	0	greater financial	0	greater financial	0

Table O.8: Bigrams chosen by classifier as significant (Bigrams-Boruta-MP).

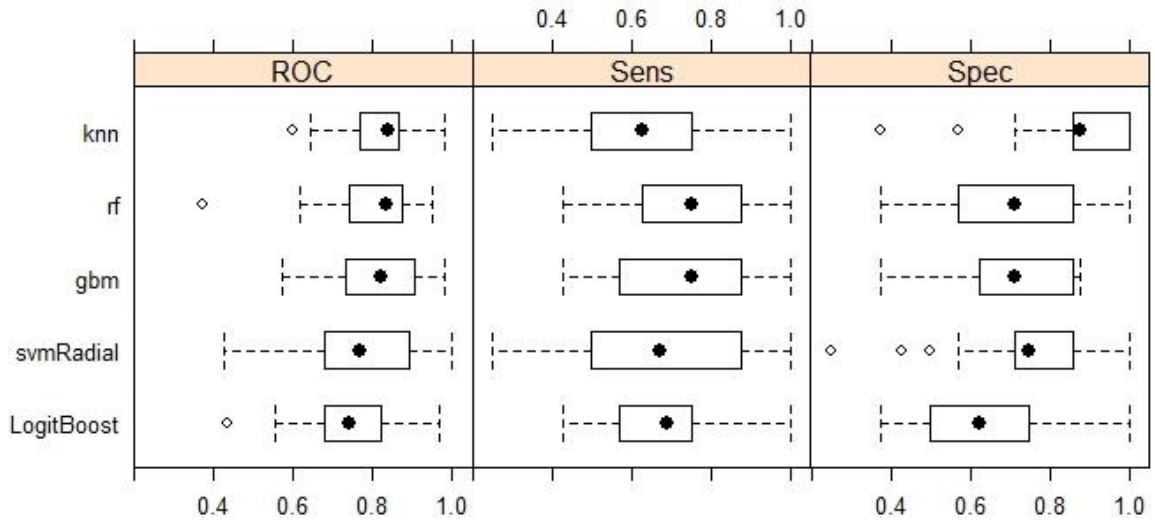


Figure O.7: ROC, Sensitivity and Specificity for classifiers (Bigrams-Boruta-MP).

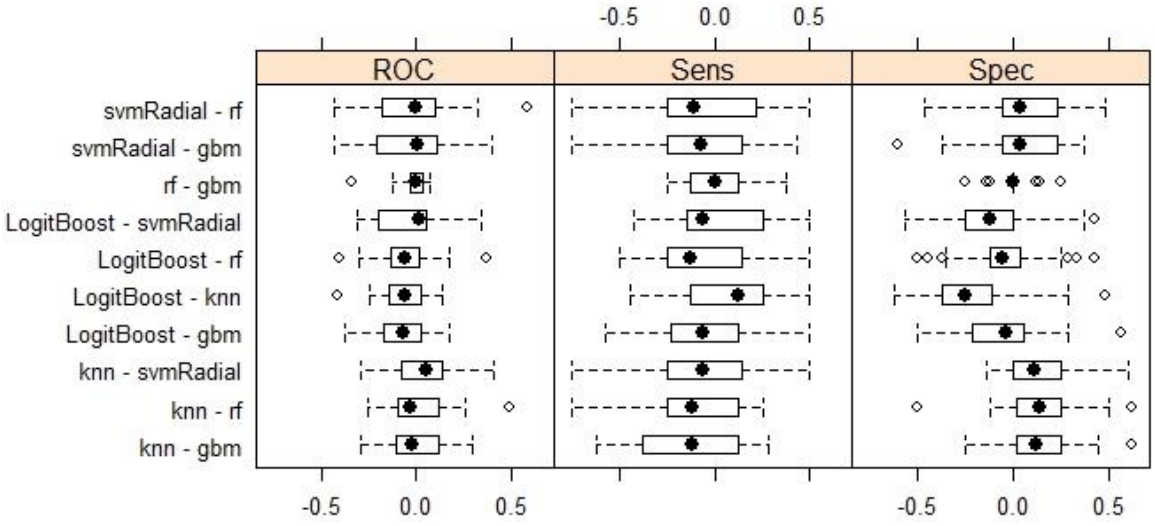


Figure O.8: Performance compared between classifiers (Bigrams-Boruta-MP).

Classification results for Information Gain (IG) selected Bigrams

IG feature selection on Bigrams – Peer Set (PS)										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.50	0.60	0.89	0.82	0.73, 0.89	0.75	0.06	0.65	0.87	0.74
SGB	0.61	0.60	0.96	0.87	0.79, 0.92	0.75	0.002	0.83	0.87	0.78
RF	0.50	0.40	1	0.85	0.76, 0.91	0.75	0.01	1	0.83	0.70
kNN	0.36	0.36	0.94	0.80	0.71, 0.87	0.75	0.14	0.69	0.81	0.65
SVM	0.4	0.60	0.86	0.80	0.71 0.87	0.75	0.14	0.60	0.86	0.73

Table O.9: Classification results (Bigrams-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
to date	100	acquisition of	100	to date	100	to date	100	to date	100
purchase price	94	purchase price	59	acquisition of	90	purchase price	94	acquisition of	93
accounting and	91	to date	57	of approximately	87	accounting and	91	income in	91
customers with	89	accounting and	48	customers with	86	customers with	89	of approximately	89
necessary to	88	customers with	45	year ended	80	necessary to	88	primarily due	89
to higher	86	event that	44	between the	78	to higher	86	commitment to	88
of approximatel	85	year ended	43	necessary to	77	of approximately	85	annual report	88
income in	84	of approximately	38	accounting and	77	income in	84	continued to	87
markets and	84	between the	38	purchase price	77	markets and	84	markets and	87
continued to	84	necessary to	37	accounted for	76	continued to	84	necessary to	85
year ended	83	accounted for	35	markets and	68	year ended	83	purchase price	84

Table O.10: Bigrams chosen by classifier as significant (Bigrams-IG-PS).

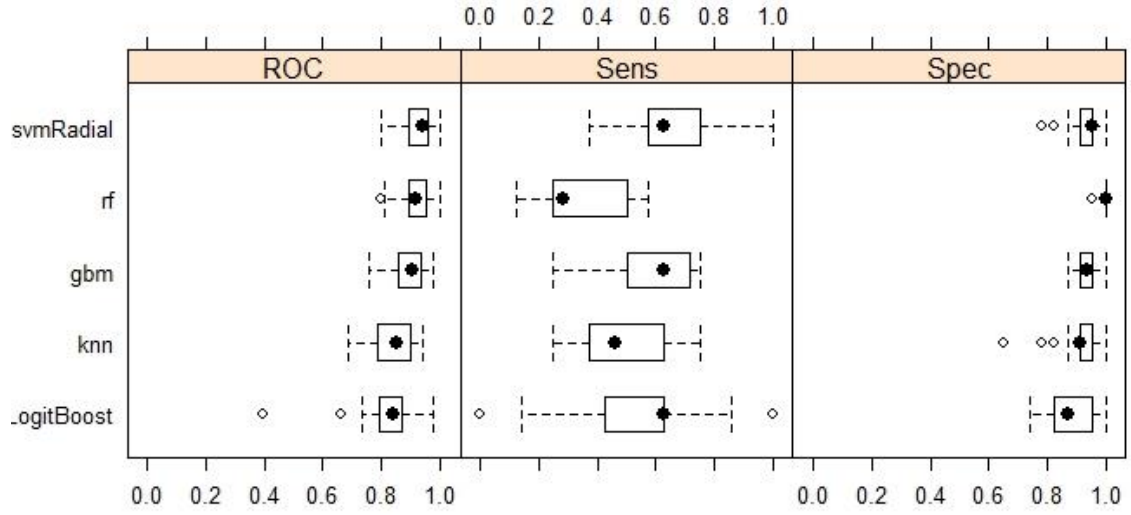


Figure O.9: ROC, Sensitivity and Specificity for classifiers (Bigrams-IG-PS).

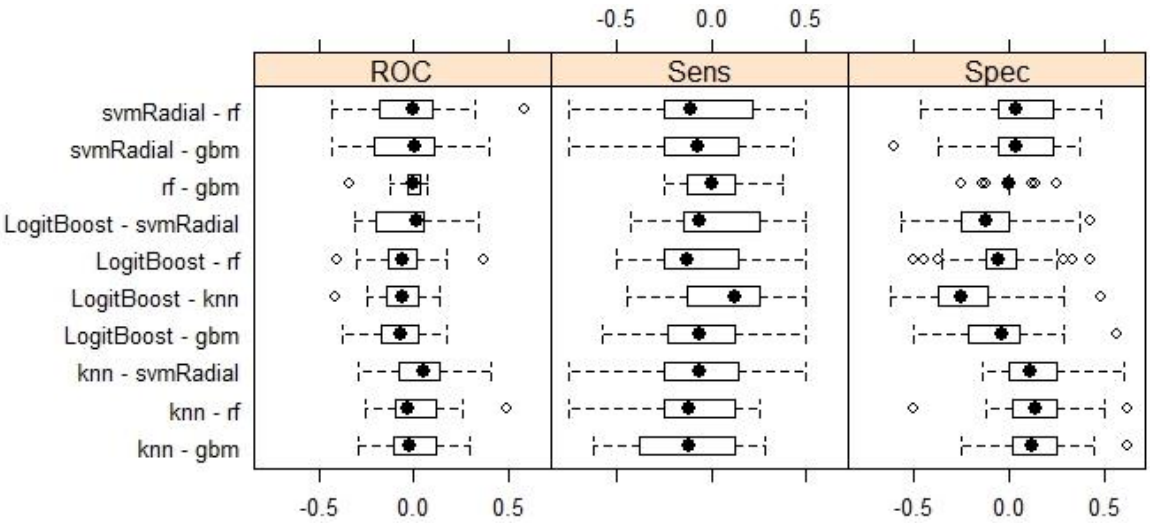


Figure O.10: Performance compared between classifiers (Bigrams-IG-PS).

IG feature selection on Bigrams – Matched Pair (MP)										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.6	0.8	0.8	0.8	0.66, 0.89	0.5	1.19e-05	0.8	0.8	0.8
SGB	0.6	0.84	0.76	0.8	0.66 0.89	0.5	1.193e-05	0.77	0.82	0.80
RF	0.76	1	0.76	0.88	0.75 0.95	0.5	1.622e-08	0.80	1	0.88
kNN	0.8	0.88	0.72	0.8	0.66 0.89	0.5	1.19e-05	0.75	0.85	0.80
SVM	0.6	0.76	0.84	0.8	0.66 0.89	0.5	1.193e-05	0.82	0.77	0.80

Table O.11: Classification results (Bigrams-IG-MP)..

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
purchase price	100	purchase price	100	purchase price	100	purchase price	100	purchase price	100
in compared	95	in compared	62	in and	95	in compared	95	may be	95
in and	93	in and	61	in compared	66	in and	93	continued to	94
failure to	86	event that	57	may be	65	failure to	86	in and	92
may be	86	contributed to	54	contributed to	63.8	may be	86	failure to	88
to date	85	primarily due	49	to date	60	to date	85	operating income	88
continued to	83	to date	46	service and	60	continued to	83	in compared	87
the first	83	approximately and	39	the first	59	the first	83	the first	77
operating income	82	service and	39	primarily due	57	operating income	82	and lower	74 2
approved by	78	operating income	37	failure to	54	approved by	78	a decline	70
primarily due	76	obtain a	37	continued to	51	primarily due	76	contributed to	70

Table O.12: Bigrams chosen by classifier as significant (Bigrams-IG-MP).

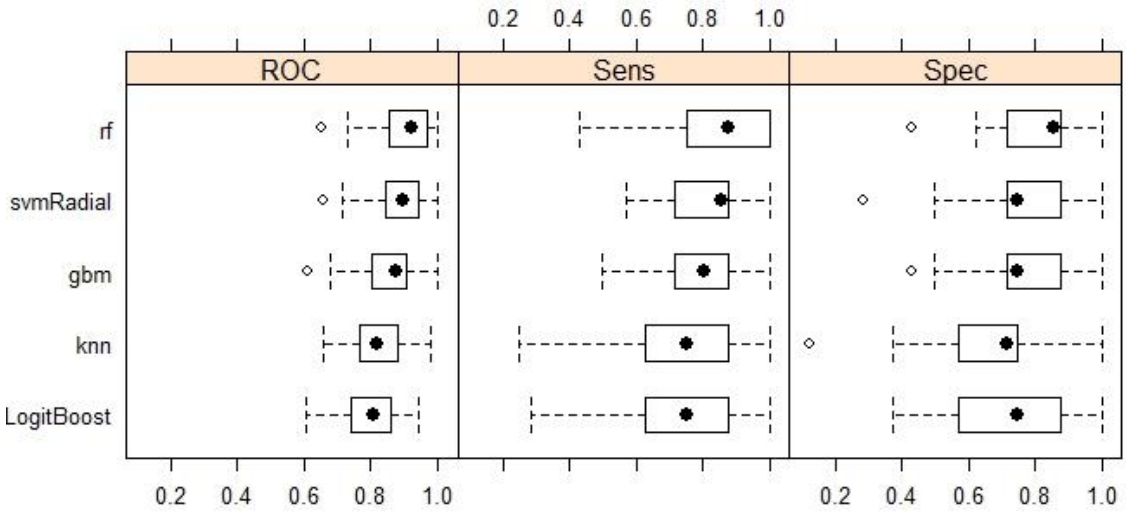


Figure O.11: ROC, Sensitivity and Specificity for classifiers (Bigrams-IG-MP).

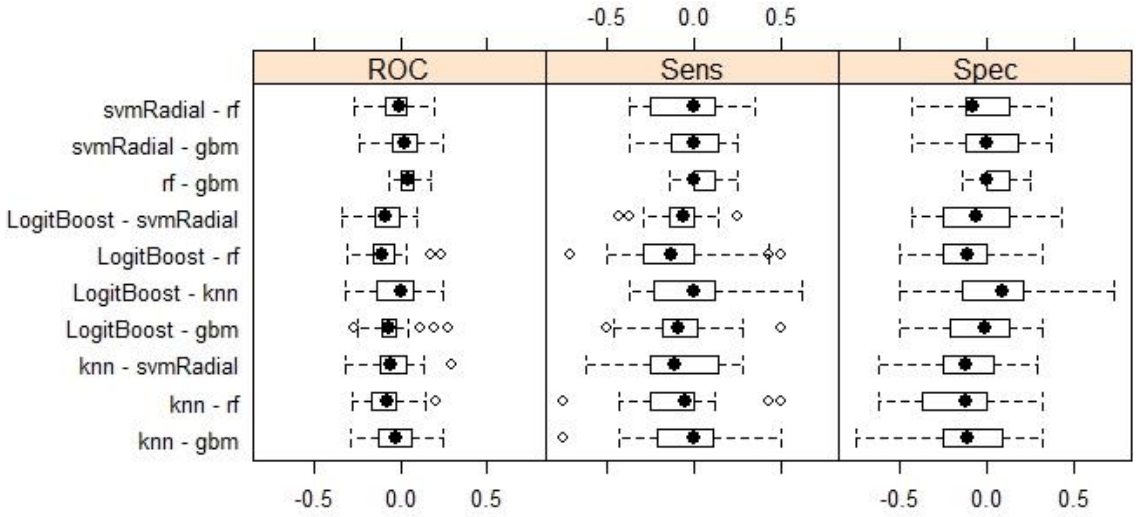


Figure O.12: Performance compared between classifiers (Bigrams-IG-MP).

APPENDIX P

Classification results for PCA selected Trigrams

PCA feature selection on Trigrams - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.04	0.28	0.76	0.64	0.54 0.73	0.75	0.99	0.28	0.76	0.52
SGB	0.01	0.20	0.81	0.66	0.56 0.75	0.75	0.98	0.26	0.75	0.50
RF	0.21	0.24	0.93	0.76	0.66 0.84	0.75	0.46	0.54	0.78	0.58
kNN	0.03	0.04	0.98	0.75	0.65 0.83	0.75	0.55	0.50	0.75	0.51
SVM	0.05	0.04	1	0.76	0.66, 0.84	0.75	0.46	1	0.76	0.52

Table P.1: Classification results (Trigrams-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
be able to	100	be able to	100	in the future	100	be able to	100	be able to	100
we believe that	97	may not be	96	be able to	88	we believe that	97	we may be	92
not be able	96	in the future	92.4	may not be	85	not be able	96	not be able	85
may not be	95	we believe that	82	could result in	64	may not be	95	may not be	83
we may be	92	could result in	64	we believe that	59.7	we may be	92	in the future	82
we do not	88	could adversely affect	54.7	we do not	50.3	we do not	88	we believe that	82
effect on our	79	we do not	53	we may be	42	effect on our	79	we do not	78
could adversely affect	75	be adversely affected	50	could adversely affect	42	could adversely affect	75	could adversely affect	78
in addition we	75	the market price	48	the market price	41	in addition we	75	could result in	75
our ability to	75	some of our	46	in addition we	40.6	our ability to	75	effect on our	73.8
all of our	74	our operating results	42.7	be adversely affected	39	all of our	74	our ability to	72

Table P.2: Trigrams chosen by classifier as significant (Trigrams-PCA-PS).

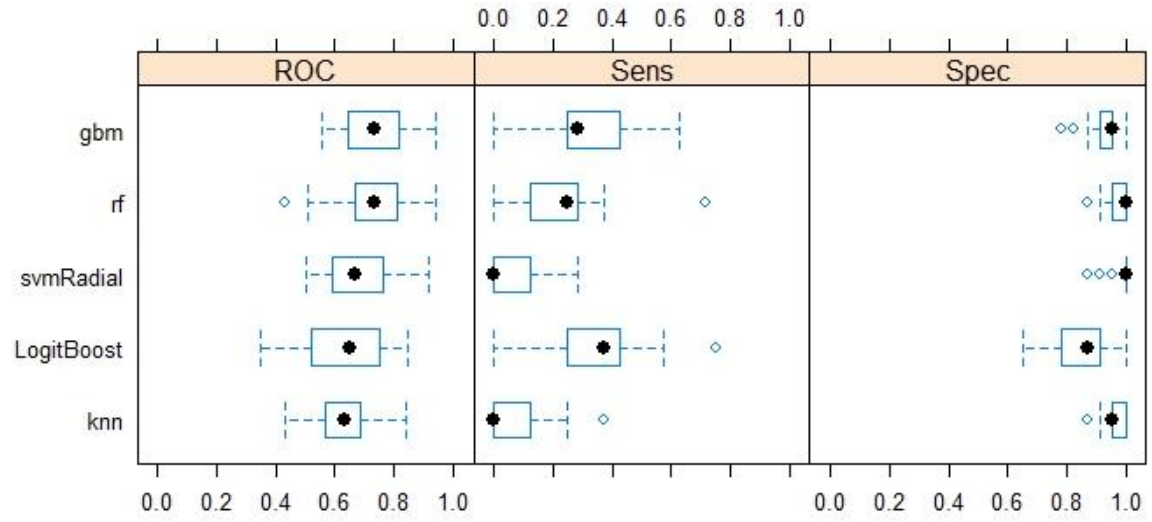


Figure P.1: ROC, Sensitivity and Specificity for classifiers (Trigrams-PCA-PS).

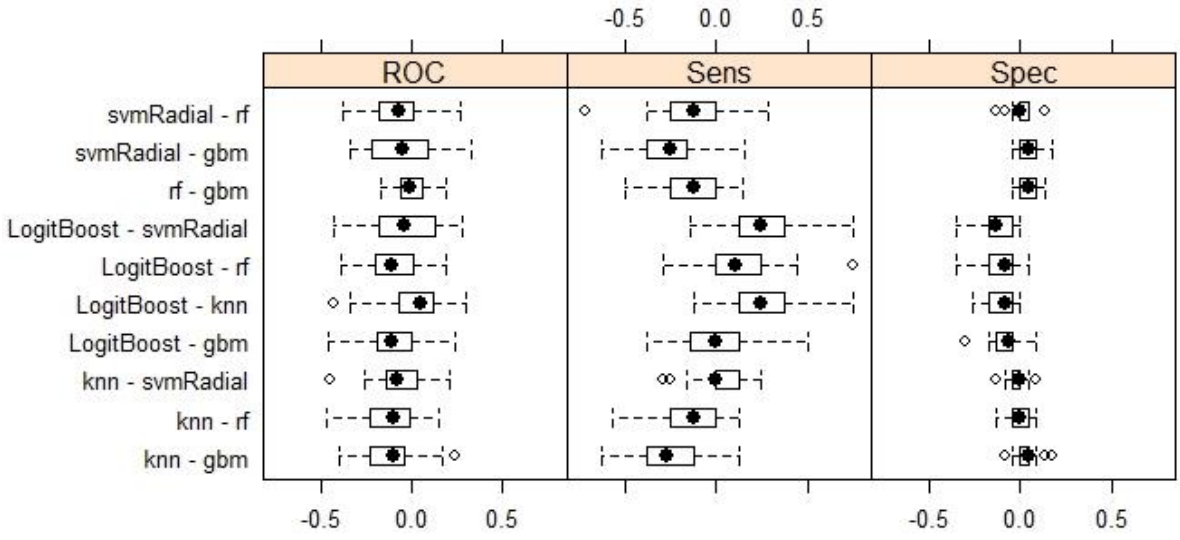


Figure P.2: Performance compared between classifiers (Trigrams-PCA-PS).

PCA feature selection on Trigrams – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.60	0.64	0.62	0.47 0.75	0.5	0.05	0.62	0.61	0.62
SGB	0.4	0.64	0.76	0.7	0.55 0.82	0.5	0.003	0.72	0.67	0.70
RF	0.4	0.56	0.84	0.7	0.55 0.82	0.5	0.0033	0.77	0.65	0.70
kNN	0.2	0.72	0.48	0.6	0.45 0.73	0.5	0.101	0.58	0.63	0.60
SVM	0.32	0.56	0.76	0.66	0.51 0.78	0.5	0.01	0.70	0.63	0.66

Table P.3: Classification results (Trigrams-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
be able to	100	results of operations	100	results of operations	100	be able to	100	be able to	100
results of operations	87	be able to	64	in the future	67	we may be	87	our operating results	74
in addition we	86	in the future	44	be able to	64	we believe that	85	limit our ability	68
we believe that	84	we may be	43	could result in	53	not be able	84	could result in	67
not be able	83	are subject to	29	be adversely affected	49	may not be	81	not be able	66
may not be	81	could result in	29	may not be	36	in addition we	78	be adversely affected	66
in the future	75	be adversely affected	22	we may be	30	in the future	76	portion of our	66
our ability to	73	we believe that	21	are subject to	28	our ability to	74	may not be	65
be adversely affected	68	portion of our	20	that we will	24	could adversely affect	70	affect our ability	65
could result in	63	may result in	20	we do not	23	we do not	69	believe that our	62
we do not	62	we do not	19	not be able	22	be adversely affected	68	could adversely affect	60

Table P.4: Trigrams chosen by classifier as significant (Trigrams-PCA-MP).

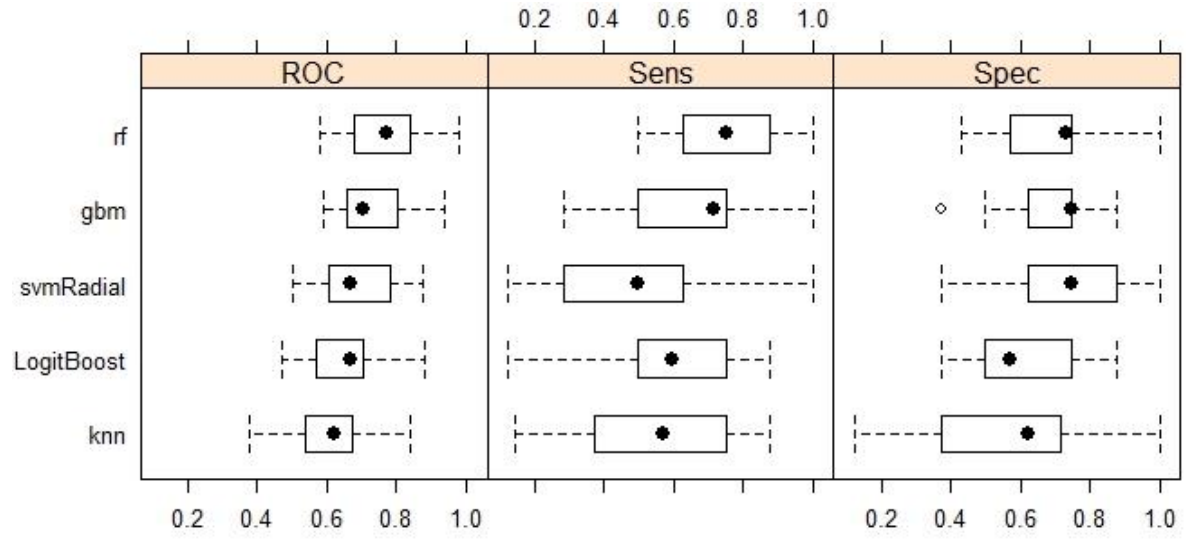


Figure P.3: ROC, Sensitivity and Specificity for classifiers (Trigrams-PCA-MP).

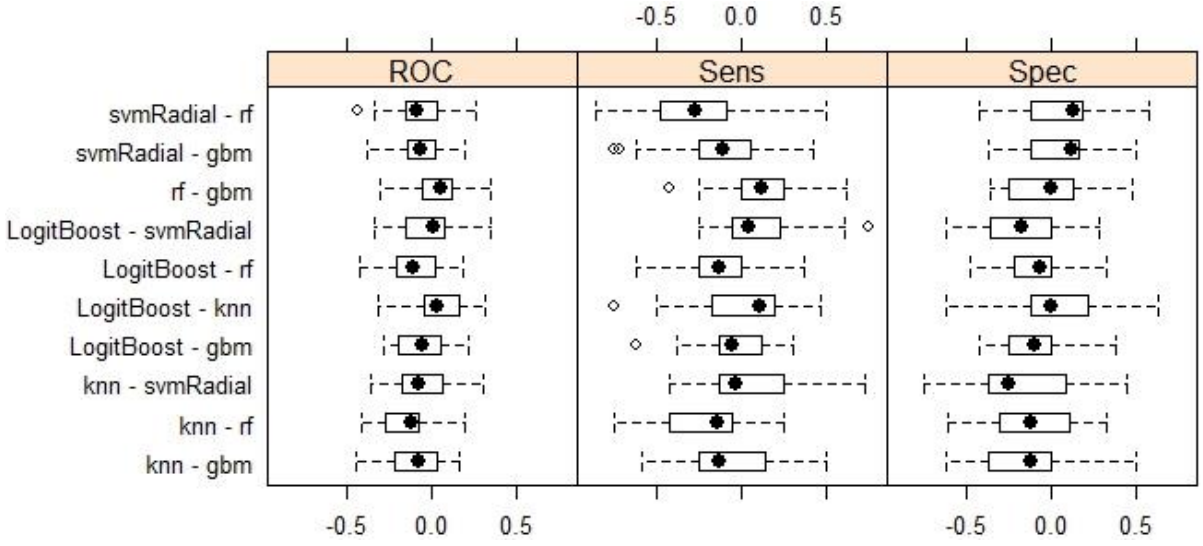


Figure P.4: Performance compared between classifiers (Trigrams-PCA-MP).

Classification results for Boruta selected Trigrams

Boruta feature selection on Trigrams - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.50	0.48	0.96	0.84	0.75 0.90	0.75	0.02	0.80	0.84	0.72
SGB	0.64	0.60	0.97	0.88	0.80 0.93	0.75	0.001	0.88	0.88	0.78
RF	0.50	0.40	1	0.85	0.76 0.91	0.75	0.01	1	0.83	0.70
kNN	0	0	1	0.75	0.65 0.83	0.75	0.55	0	0.75	0.50
SVM	0.43	0.52	0.89	0.80	0.71 0.87	0.75	0.75	0.61	0.85	0.70

Table P.5: Classification results (Trigrams-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
the acquisition of	100	the results of	100	the acquisition of	100	the year ended	100	the year ended	100
the year ended	93	the impact of	82	in the event	82.9	in the event	96	the acquisition of	94
primarily due to	90	the acquisitio n of	77	the year ended	77.9	the acquisition of	95	primarily due to	85
in the event	87	provided by financing	70	for the year	66	the results of	93	the impact of	83.4
the impact of	82	at the time	67	the results of	66	for the year	93	of our common	82
the results of	82	use of the	60	at the time	64	may be required	92	for the year	82
may be required	77	in the event	59	pursuant to the	47.7	the impact of	88	may be required	82
for the year	77	may be required	58	entered into a	47	million at december	84	in the event	81
at the time	77	for the year	57	the impact of	46.7	entered into a	83.9	at the time	78
million at december	75	million in cash	56.7	primarily due to	45.4	at the time	80	the results of	74
entered into a	69	primarily due to	55	not believe that	40	primarily due to	74	entered into a	71
primarily as a	33	and sale of	52.4	may be required	39	primarily as a	48	million at december	68
million of cash	29.4	the fiscal year	51	during the period	36	during the period	42	primarily as a	32
million in cash	20	the year ended	48.5	primarily as a	33	the company in	39	the company in	28

Table P.6: Trigrams chosen by classifier as significant (Trigrams-Boruta-PS).

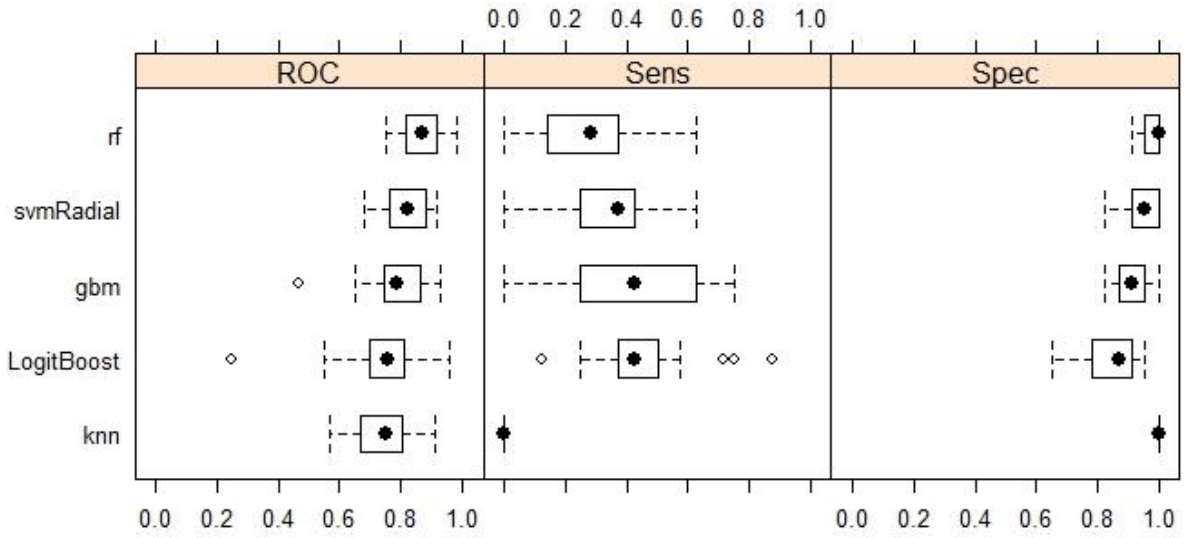


Figure P.5: ROC, Sensitivity and Specificity for classifiers (Trigrams-Boruta-PS).

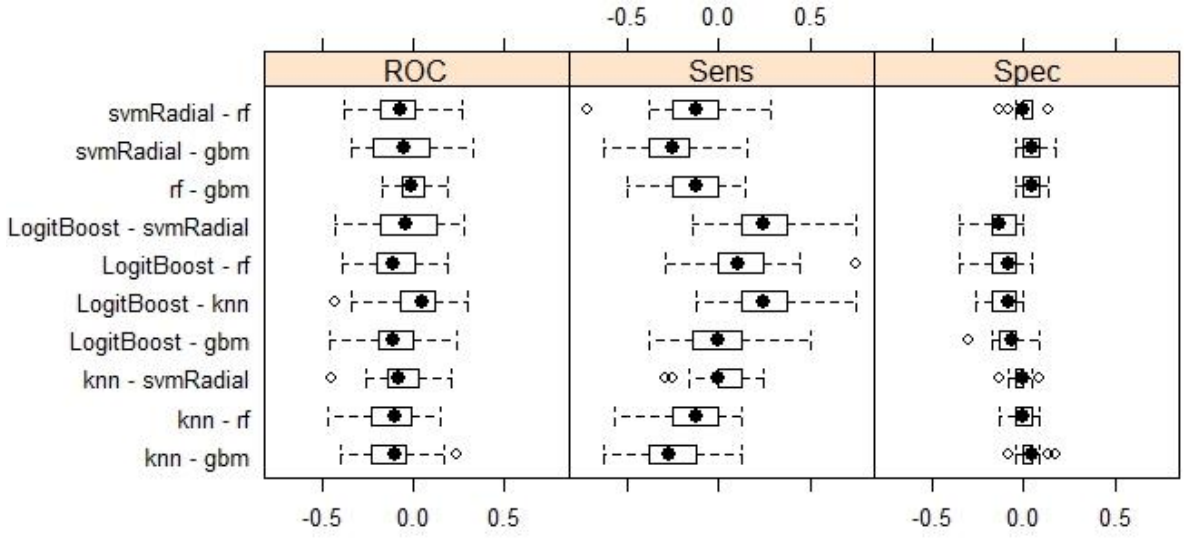


Figure P.6: Performance compared between classifiers (Trigrams-Boruta-PS).

Classification results for Boruta selected Trigrams

Boruta feature selection on Trigrams –Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.4	0.72	0.68	0.70	0.55 0.82	0.5	0.0033	0.69	0.70	0.70
SGB	0.52	0.80	0.72	0.76	0.61 0.86	0.5	0.0001	0.74	0.78	0.76
RF	0.56	0.72	0.84	0.78	0.64 0.88	0.5	4.511e-05	0.81	0.75	0.78
kNN	0.32	0.64	0.68	0.66	0.51 0.78	0.5	0.01	0.66	0.65	0.66
SVM	0.64	0.84	0.80	0.82	0.68 0.91	0.5	2.807e-06	0.80	0.83	0.82

Table P.7: Classification results (Trigrams-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
the end of	100	the end of	100	the end of	100	the acquisition of	100	the end of	100
in compared to	99	primarily due to	80	primarily due to	80	the end of	96	primarily due to	93
primarily due to	99	in addition the	69	in addition the	69	in compared to	93	the acquisition of	87
the acquisition of	90	in compared to	64	in compared to	64	which may be	93	in addition the	83
in addition the	82	the acquisition of	58	the acquisition of	58	primarily due to	87	in compared to	76
we may not	18	and will be	22	and will be	22	in addition the	77	and will be	15
which may be	12	we may not	19	we may not	19	the event that	18	we may not	14
and will be	7	the event that	17	the event that	17	we may not	3	the event that	11
comply with the	6	comply with the	13	comply with the	13	comply with the	1	comply with the	6
the event that	0	which may be	0			and will be	0	which may be	0

Table P.8: Trigrams chosen by classifier as significant (Trigrams-Boruta-MP).

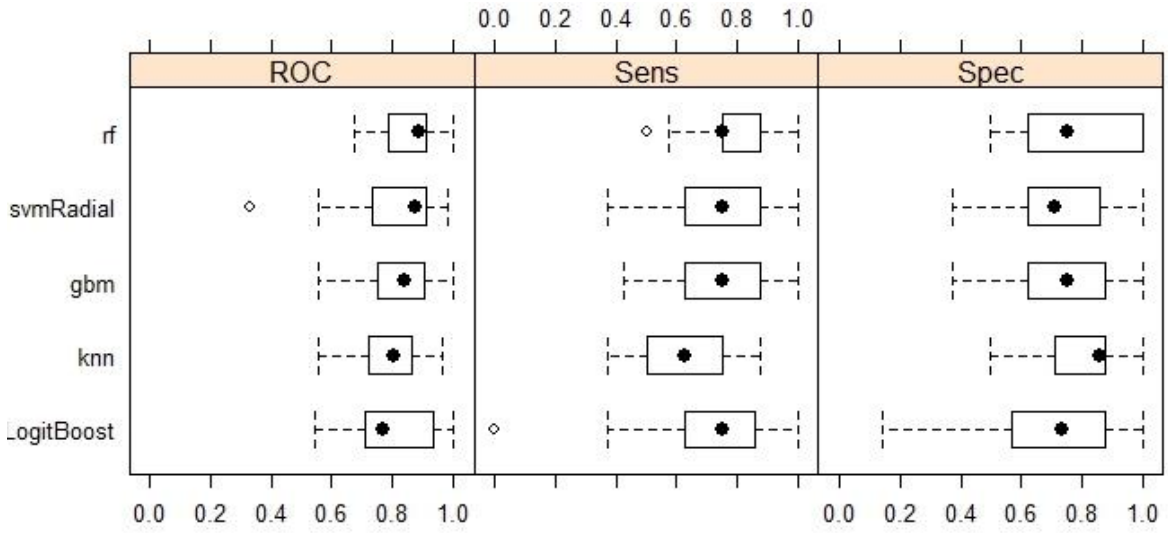


Figure P.7: ROC, Sensitivity and Specificity for classifiers (Trigrams-Boruta-MP).

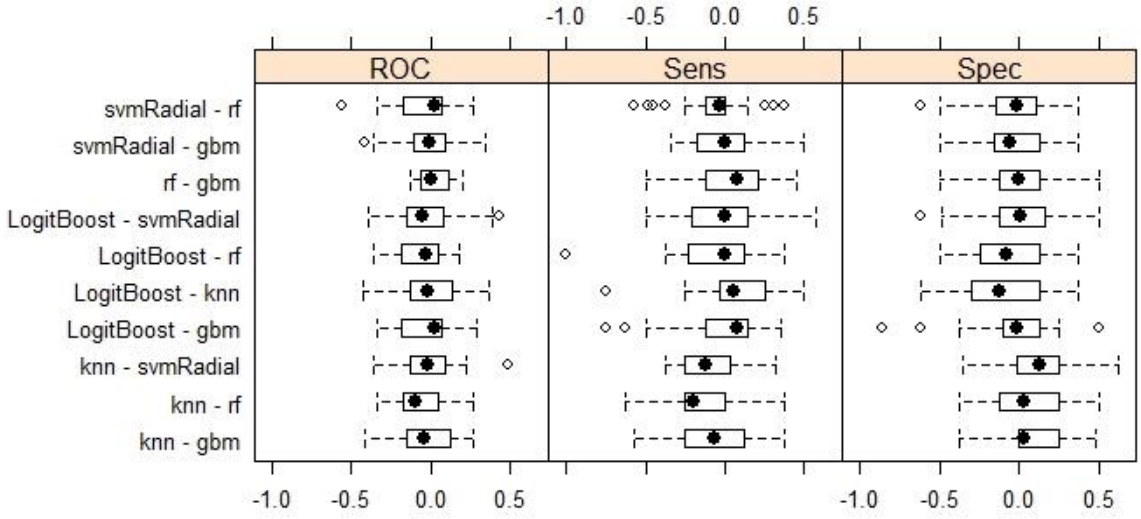


Figure P.8: Performance compared between classifiers (Trigrams-Boruta-MP).

Classification results for IG selected Trigrams

IG feature selection on Trigrams – Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.25	0.28	0.93	0.77	0.67 0.84	0.75	0.37	0.58	0.79	0.60
SGB	0.40	0.40	0.94	0.81	0.72 0.88	0.75	0.09	0.71	0.82	0.67
RF	0.11	0.08	1	0.77	0.67 0.84	0.75	0.37	1	0.76	0.54
kNN	0.39	0.52	0.86	0.78	0.68 0.85	0.75	0.28	0.56	0.84	0.69
SVM	0.39	0.32	0.98	0.82	0.73 0.89	0.75	0.06	0.88	0.81	0.65

Table P.9: Classification results (Trigrams-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
the acquisition of	100	are expected to	100	the acquisition of	100	the acquisition of	100	entered into an	100
a decline in	77	the acquisition of	99	the event that	72	are expected to	95	the acquisition of	96
are expected to	72	entered into an	56	are expected to	70	in the fair	82	a decline in	83
in the fair	69	the event that	42	entered into an	57	a decline in	74	are expected to	72
a range of	58	a range of	29	a change in	40	a loss of	60	in the fair	67
a decrease in	56	a decline in	29	a decline in	39	a change in	59	a change in	60
a change in	56	a broad range	27	a decrease in	36	a range of	57	a decrease in	58
a combination of	55	part of our	17	of shares of	33	part of our	54	a increase in	57
should be read	55	a decrease in	17	a combination of	28	a increase in	52	a range of	56
a increase in	50	should be read	16	a broad range	27	should be read	50	part of our	56
part of our	49	a combination of	14	a range of	24	a broad range	47	a loss of	54
a loss of	47	a change in	13	part of our	24	a decrease in	46	should be read	49

Table P.10: Trigrams chosen by classifier as significant (Trigrams-IG-PS).

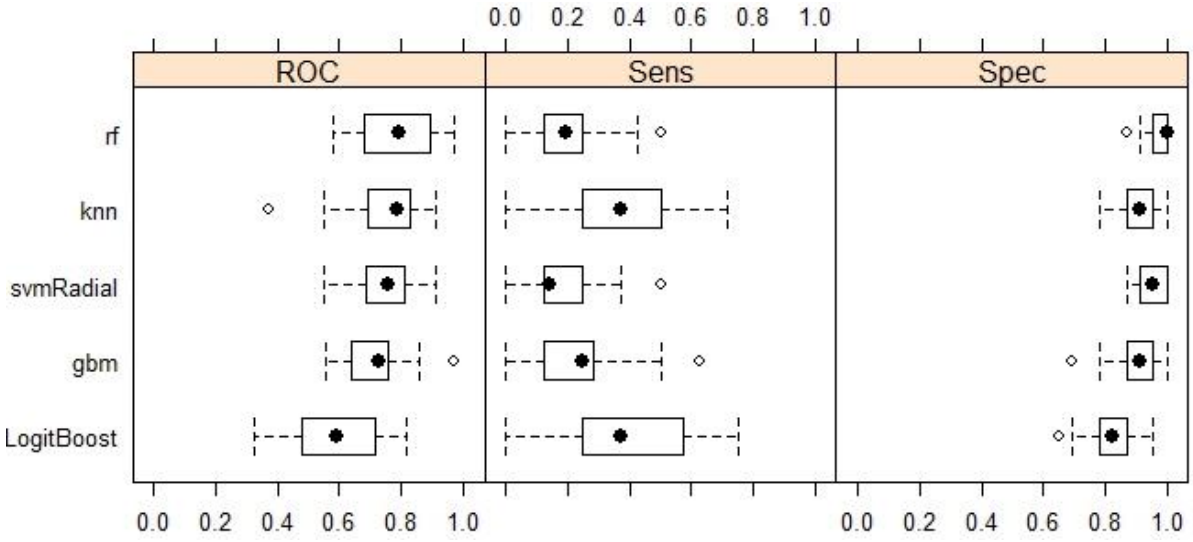


Figure P.9: ROC, Sensitivity and Specificity for classifiers (Trigrams-IG-PS).

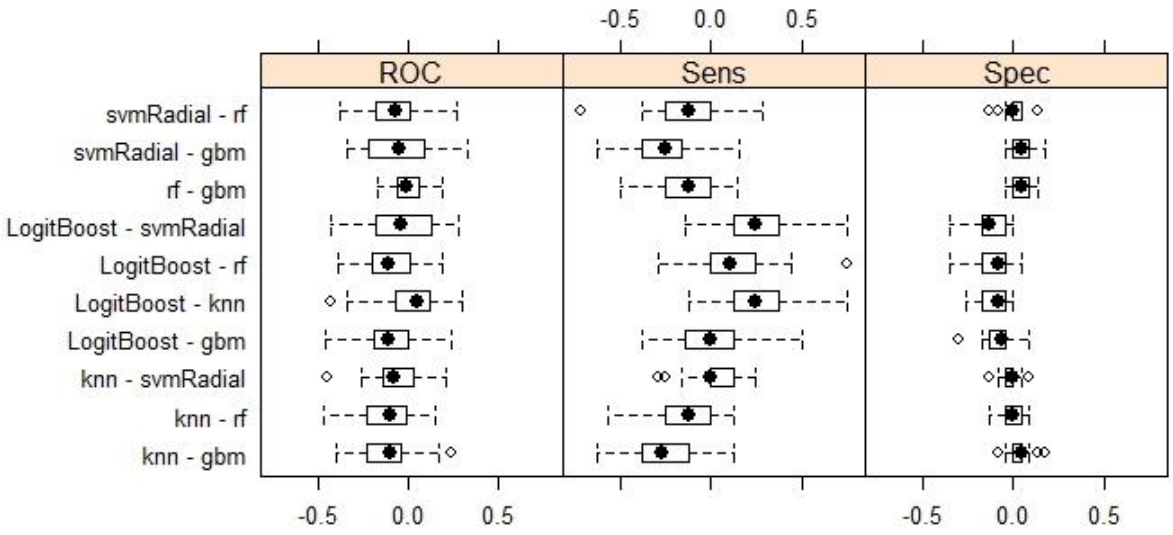


Figure P.10: Performance compared between classifiers (Trigrams-IG-PS).

IG feature selection on Trigrams – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.60	0.64	0.62	0.47 0.75	0.5	0.05	0.62	0.61	
SGB	0.56	0.84	0.72	0.78	0.64 0.88	0.5	4.511e-05	0.75	0.81	
RF	0.68	0.84	0.84	0.84	0.70 0.92	0.5	5.818e-07	0.84	0.84	
kNN	0.36	0.68	0.68	0.68	0.53 0.80	0.5	0.0076	0.68	0.68	
SVM	0.4	0.72	0.68	0.7	0.55 0.81	0.5	0.0033	0.69	0.70	

Table P.11: Classification results (Trigrams-IG-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
the end of	100	in compared to	100	based on the	100	based on the	100	in the event	100
in the event	92	based on the	85	the end of	772	in the event	965	the end of	986
based on the	86	primarily due to	72	primarily due to	576	one or more	925	based on the	977
in the fair	83	one or more	70	in the event	558	the event of	90	the event of	884
the event of	80	the end of	52	one or more	404	the end of	906	entered into a	858
one or more	77	entered into a	51	entered into a	288	in compared to	86	primarily due to	828
entered into a	71	comply with the	41	the event of	214	in the fair	866	in compared to	81
in compared to	70	we are a	41	in compared to	206	entered into a	85	one or more	787
primarily due to	68	in the event	33	comply with the	187	primarily due to	815	in the fair	713
due to lower	59	in the fair	29	of shares of	124	due to lower	771	due to lower	70
pursuant to the	28	of shares of	15	ability to provide	124	pursuant to the	22	pursuant to the	295

Table P.12: Trigrams chosen by classifier as significant (Trigrams-IG-MP).

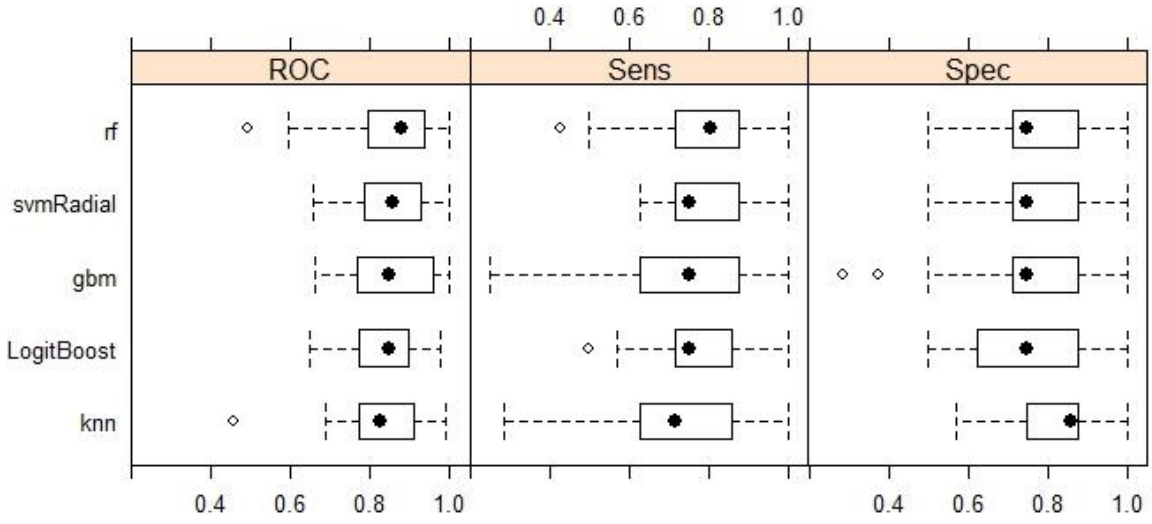


Figure P.11: ROC, Sensitivity and Specificity for classifiers (Trigrams-IG-MP).

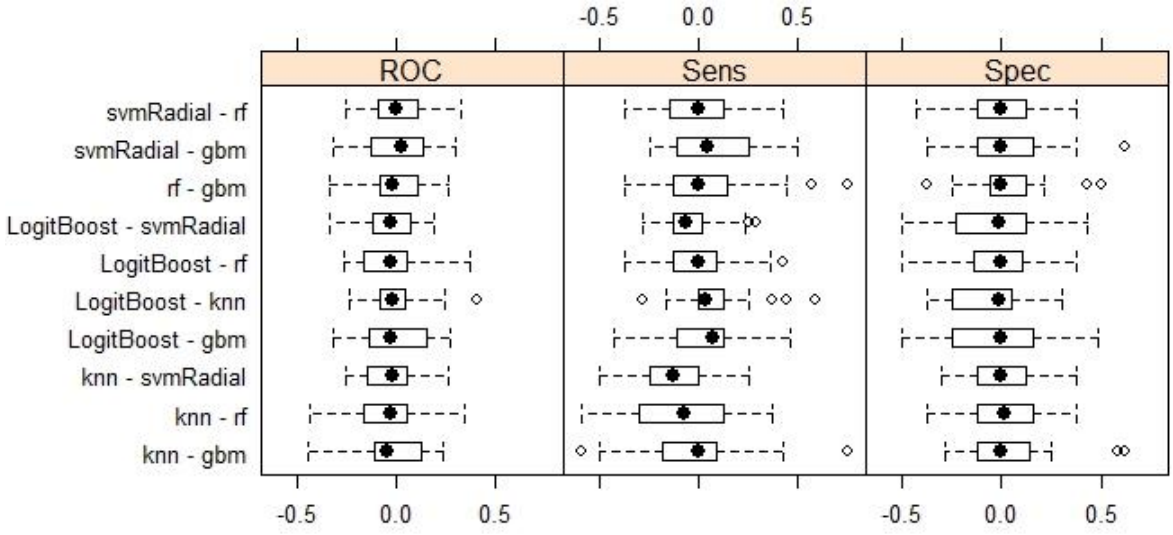


Figure P.12: Performance compared between classifiers (Trigrams-IG-MP).

APPENDIX Q

Classification results for PCA selected Coh-Metrix Indices

PCA feature selection on Coh-Metrix Indices - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.36	0.86	0.74	0.64 0.82	0.75	0.64	0.47	0.80	0.61
SGB	0.16	0.28	0.86	0.72	0.62 0.80	0.75	0.79	0.41	0.78	0.57
RF	0.14	0.12	0.98	0.77	0.67 0.84	0.75	0.37	0.75	0.77	0.55
kNN	0.16	0.28	0.86	0.72	0.62 0.80	0.75	0.79	0.41	0.78	0.57
SVM	0.45	0.40	0.97	0.83	0.74 0.89	0.75	0.037	0.83	0.83	0.68

Table Q.1: Classification results (CohMetrix-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
WRDIMGc	100	WRDAOAc	100	CNCLogic	100	WRDIMGc	100	WRDFRQa	100
WRDFRQa	80	LSASS1d	70.2	CNCNeg	93	DRNEG	80	WRDIMGc	92
WRDAOAc	67	DESSL	66	CNCADC	88.7	CNCLogic	74	DRNEG	72
DRNEG	64	DRNEG	55	RDFKGL	87	WRDFRQa	74	WRDAOAc	70
LDMTLD	63	WRDIMGc	50	DESSL	83	LSASS1d	68	WRDMEAc	70
PCVERBp	62	PCVERBz	44	WRDFRQa	82	WRDAOAc	64	LSASS1d	63
PCVERBz	62	WRDFAMc	42	WRDAOAc	74.5	LDMTLD	62	LDMTLD	59
PCCONNz	61	WRDFRQa	40	DRNEG	73.5	PCVERBp	60	DESWLlt	57
PCCONNp	61	DESWLltd	39	CRFNO1	71.6	PCVERBz	59.8	CNCLogic	55
CNCLogic	59	WRDMEAc	38	WRDIMGc	71.6	LSASSp	59.8	PCCONNz	52
PCREFz	57	CNCNeg	36.5	PCREFz	65	DESSC	56	PCVERBz	48.7

Table Q.2: CohMetrix Indices chosen by classifier for (CohMetrix-PCA-PS).

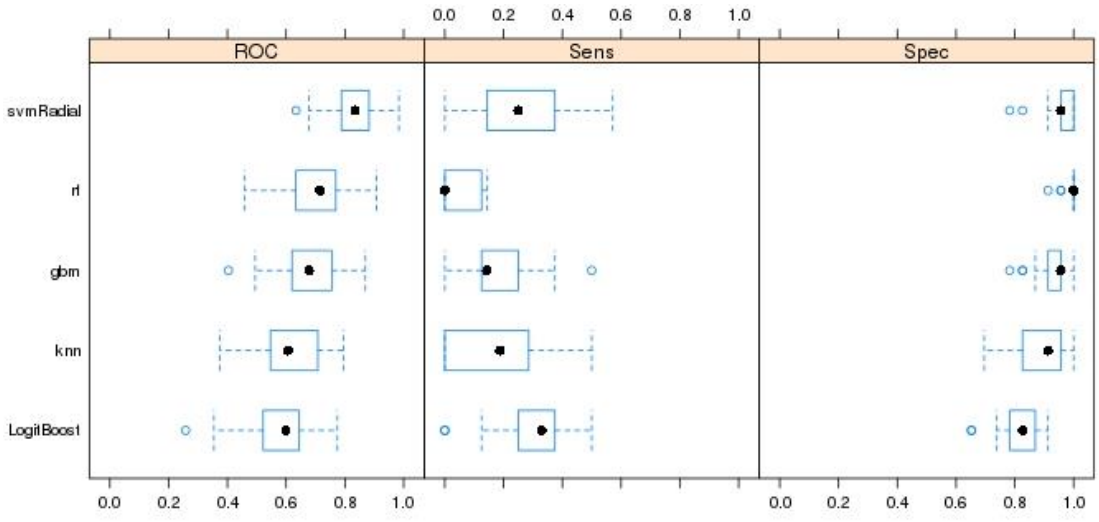


Figure Q.1: ROC, Sensitivity and Specificity for classifiers (CohMetrix-PCA-PS).

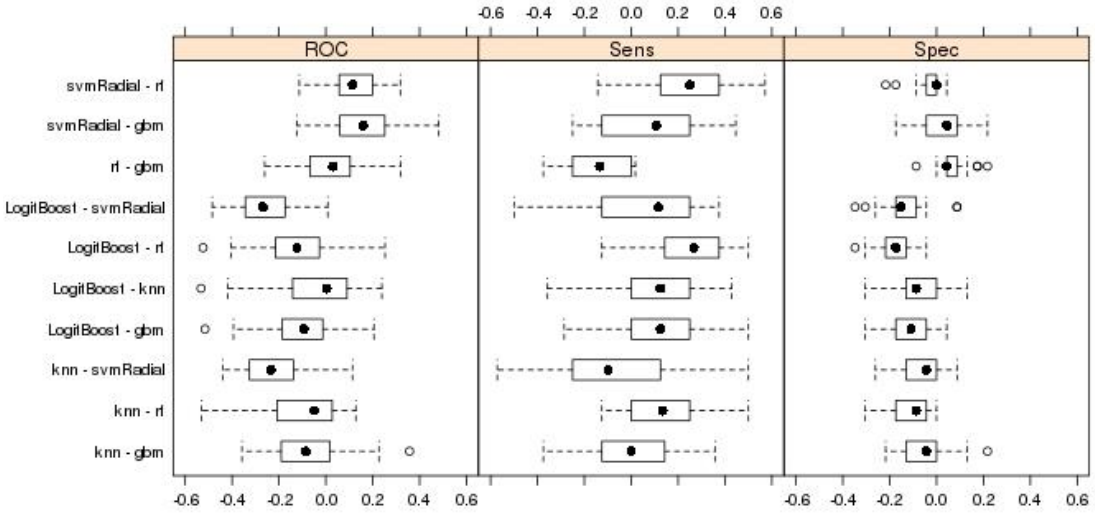


Figure Q.2: Performance compared between classifiers (CohMetrix-PCA-PS).

PCA feature selection on Coh-Metrix Indices – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.04	0.56	0.48	0.52	0.37 0.66	0.5	0.44	0.51	0.52	0.52
SGB	0.4	0.68	0.76	0.72	0.55 0.82	0.5	0.001	0.73	0.70	0.72
RF	0.16	0.44	0.72	0.58	0.43 0.71	0.5	0.16	0.61	0.56	0.58
kNN	-0.24	0.52	0.24	0.38	0.24 0.52	0.5	0.96	0.40	0.33	0.38
SVM	0.48	0.68	0.80	0.74	0.59 0.85	0.5	0.0004	0.77	0.71	0.74

Table Q.3: Classification results (CohMetrix-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
LDMTLD	100	DRNEG	100	CNCADC	100	CRFCWOa	100	SMCAUSv	100
WRDMEAc	94	CNCADC	97	WRDIMGc	59	WRDIMGc	99	CRFCWOa	99
WRDIMGc	94	CRFCWOa	87	DRNEG	58	CRFAOa	93	WRDIMGc	99
CRFCWOa	93	WRDIMGc	86	WRDMEAc	39	WRDMEAc	93	LDMTLD	98
SMCAUSv	93	RDFRE	70	WRDFRQa	28	PCREFp	86	WRDMEAc	89
PCREFp	92	CNCNeg	63	SMCAUSv	26	PCREFz	86	CRFAOa	88
PCREFz	92	PCCONNz	63	CNCNeg	25	CRFNOa	85	WRDAOAc	84
CRFAOa	88	WRDPOLc	62	LDMTLD	24	LDMTLD	82	PCREFp	82
LSAGN	85	WRDAOAc	54	WRDPOLc	23.8	PCCONNp	81	PCREFz	82
CRFNOa	80.5	DESWLlt	53.4	CRFNOa	23	PCCONNz	80	PCCONNp	82.4
CRFSOa	80	WRDFRQa	52.8	CNCAll	20	CRFSOa	80	CRFCWOa d	82

Table Q.4: CohMetrix Indices chosen by classifier as significant (CohMetrix-PCA-MP).

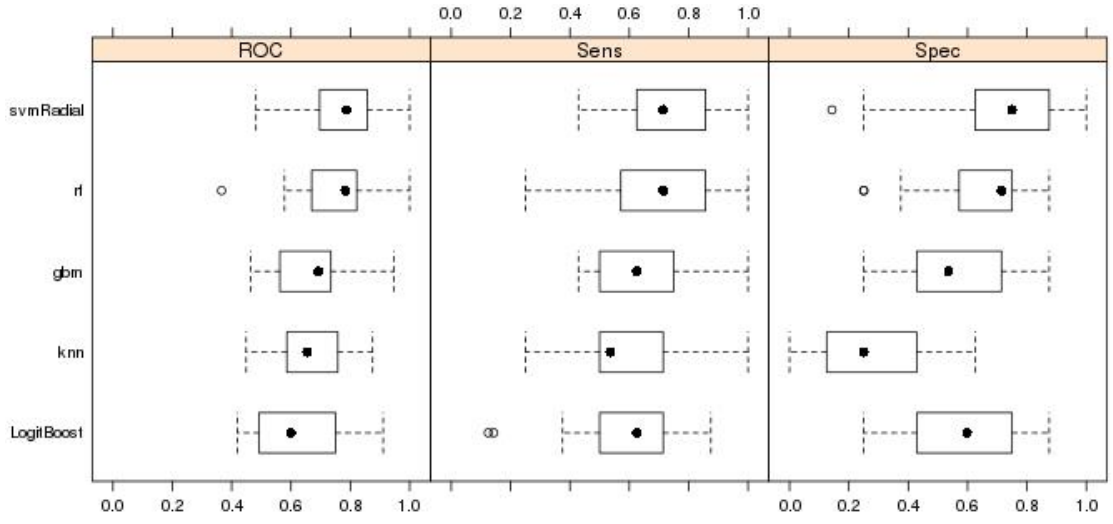


Figure Q.3: ROC, Sensitivity and Specificity for classifiers (CohMetrix-PCA-MP).

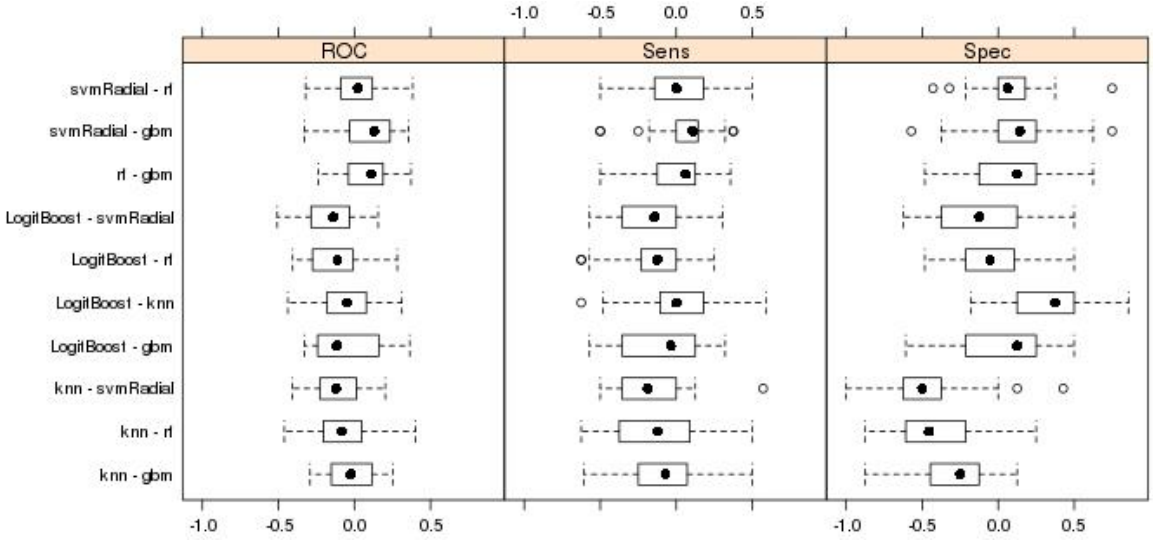


Figure Q.4: Performance compared between classifiers (CohMetrix-PCA-MP).

Classification results for Boruta selected Coh-Metrix Indices

Boruta feature selection on Coh-Metrix Indices – Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.5	0.52	0.93	0.83	0.74 0.89	0.75	0.03	0.72	0.85	0.72
SGB	0.54	0.52	0.96	0.85	0.76 0.91	0.75	0.01	0.81	0.85	0.74
RF	0.40	0.40	0.94	0.81	0.72 0.88	0.75	0.09	0.71	0.82	0.81
kNN	0.58	0.56	0.96	0.86	0.77 0.92	0.75	0.005	0.82	0.86	0.76
SVM	0.55	0.60	0.92	0.84	0.75 0.90	0.75	0.02	0.71	0.87	0.76

Table Q.5: Classification results (CohMetrix-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
CNCAdd	100	CNCTempx	100	SYNSTRUTt	100	SMCAUSwn	100	SMCAUSwn	100
WRDADJ	92	DRPVAL	88	SMCAUSwn	642	SYNSTRUTt	90	WRDFRQa	94
PCCNCz	86	WRDADJ	73	CNCTempx	565	SYNSTRUTa	87	WRDAOAc	90
SMCAUSlsa	77	WRDAOAc	69	DRPVAL	552	PCCNCz	84	SYNSTRUTt	89
LSASS1d	75	SYNSTRUTt	63	DRGERUND	528	WRDIMGc	77	PCCNCz	84
WRDIMGc	75	CRFANP1	59	DRVP	406	WRDFRQa	74	DRPVAL	82
CRFANP1	71	SMCAUSwn	56	WRDFRQa	377	CNCTempx	72.2	SYNSTRUTa	80
WRDAOAc	70	SMCAUSlsa	51	WRDADJ	276	SMCAUSlsa	71	CNCTempx	79
DRVP	67	WRDVERB	48	CNCADC	195	DRVP	70	WRDIMGc	76
WRDMEAc	67	CNCAdd	45	LSASS1d	184	DRINF	70	WRDADJ	64
WRDVERB	66	SYNSTRUTa	43	SMCAUSlsa	173	DRPVAL	64	WRDMEAc	58

Table Q.6: CohMetrix Indices chosen by classifier as significant (CohMetrix-Boruta-PS).

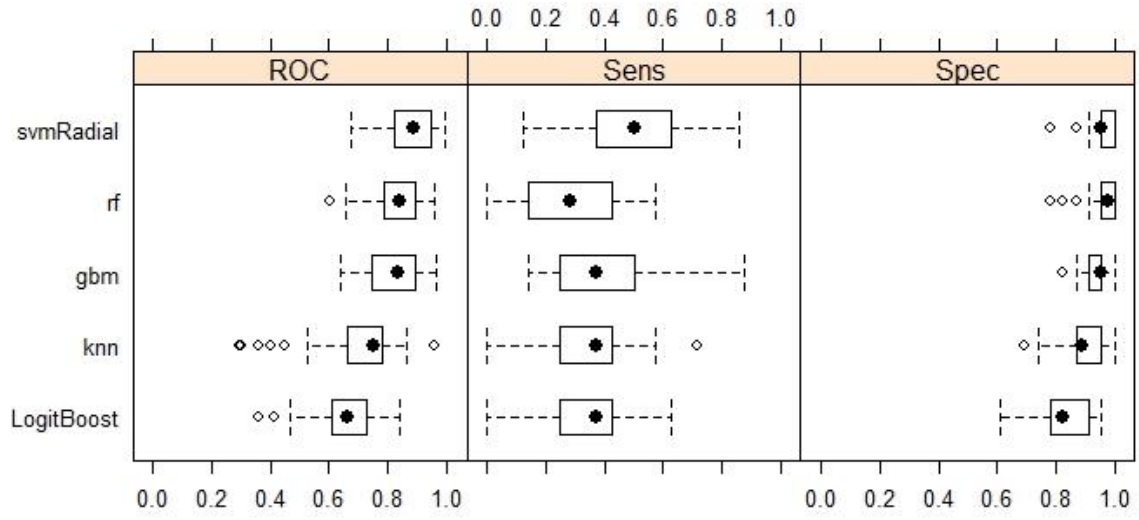


Figure Q.6: ROC, Sensitivity and Specificity for classifiers (CohMetrix-Boruta-PS).

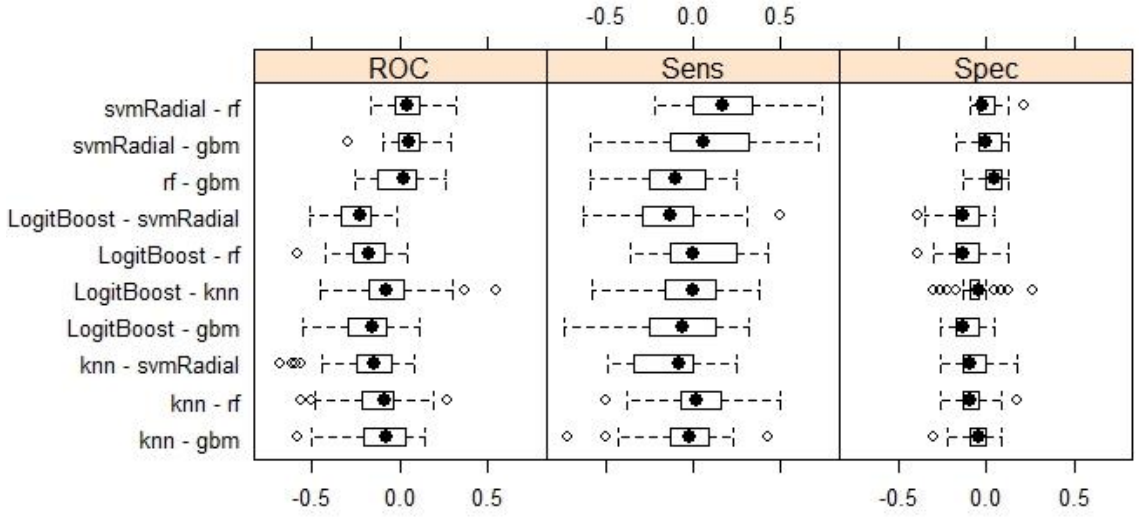


Figure Q.7: Performance compared between classifiers (CohMetrix-Boruta-PS).

Boruta feature selection on Coh-Metrix Indices – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.32	0.68	0.64	0.66	0.51 0.78	0.5	0.01	0.65	0.66	0.66
SGB	0.44	0.76	0.68	0.72	0.57 0.83	0.5	0.001	0.70	0.73	0.72
RF	0.68	0.84	0.84	0.84	0.70 0.92	0.5	5.818e-07	0.84	0.84	0.84
kNN	0.72	0.84	0.60	0.72	0.57 0.83	0.5	0.001	0.67	0.78	0.72
SVM	0.48	0.76	0.72	0.74	0.59 0.85	0.5	0.0004	0.73	0.75	0.74

Table Q.7: Classification results (CohMetrix-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
LSAPP1d	100	WRDADJ	100	LSAPP1d	100	DRVP	100	SYNSTRUTt	100
DRVP	52	LSAPP1d	72	WRDPOSin	48	SYNSTRUTt	903	LSAPP1d	89
WRDFRQa	47	DRVP	67	DRPP	43	LSAPP1d	818	DRVP	83
SYNSTRUTt	45	DRPP	55	DRVP	42	DRPP	521	DRPP	68
DRPP	33	WRDFRQa	48	WRDADJ	29	WRDPOSin	458	WRDPOSin	58
WRDADJ	30	SYNSTRUTt	45	WRDHYPnv	25	WRDADJ	258	WRDHYPnv	44
WRDPOSin	28	WRDPOSin	30	SYNSTRUTt	18	WRDHYPnv	76	WRDADJ	34
WRDHYPnv	0	WRDHYPnv	0	WRDFRQa	0	WRDFRQa	0	WRDFRQa	0
WRDMEAc	67	CNCAdd	45	LSASS1d	184	DRINF	70	WRDADJ	64
WRDVERB	66	SYNSTRUTa	43	SMCAUSlsa	173	DRPVAL	64	WRDMEAc	58

Table Q.8: CohMetrix Indices chosen by classifier as significant (CohMetrix-Boruta-MP).

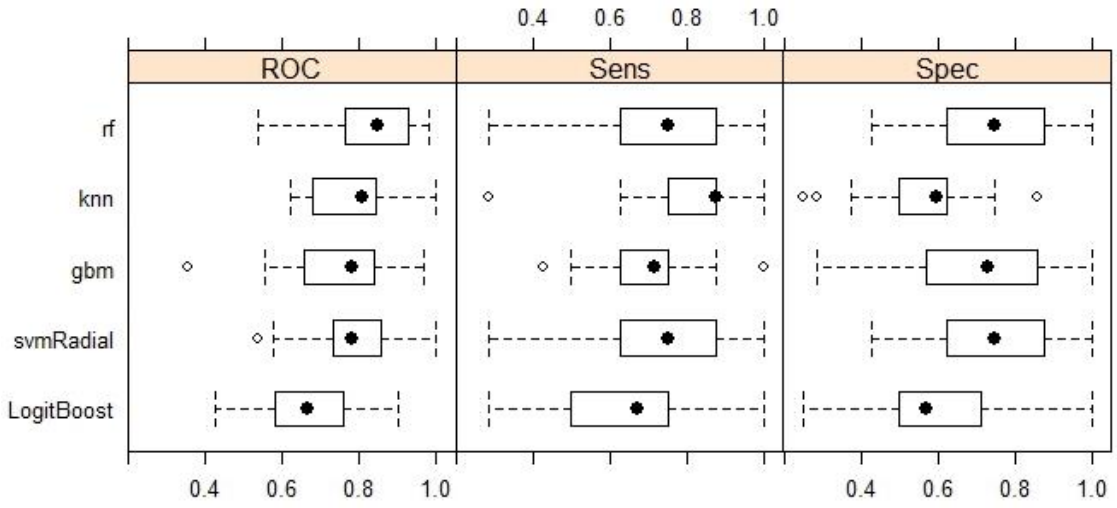


Figure Q.7: ROC, Sensitivity and Specificity for classifiers (CohMetrix-Boruta-MP).

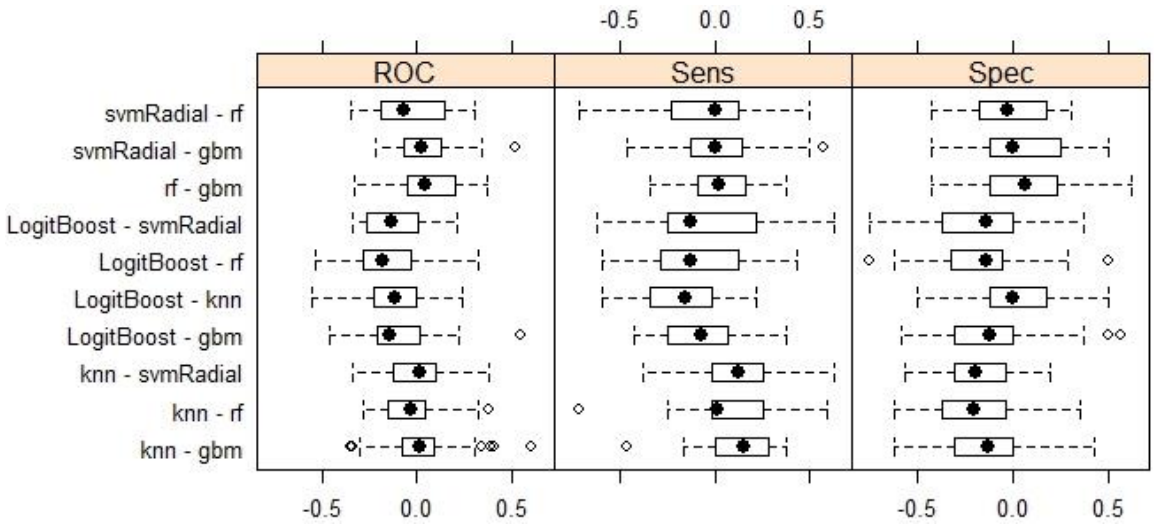


Figure Q.8: Performance compared between classifiers (CohMetrix-Boruta-MP).

Classification results for IG selected Coh-Metrix Indices

IG feature selection on Coh-Metrix Indices – Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.29	0.56	0.76	0.71	0.61 0.79	0.75	0.85	0.43	0.84	0.66
SGB	0.21	0.40	0.81	0.71	0.61 0.79	0.75	0.85	0.41	0.80	0.60
RF	0.41	0.36	0.97	0.82	0.73 0.89	0.75	0.06	0.81	0.82	0.66
kNN	0.42	0.40	0.96	0.82	0.73 0.89	0.75	0.06	0.76	0.82	0.68
SVM	0.23	0.28	0.92	0.76	0.66 0.84	0.75	0.46	0.53	0.79	0.60

Table Q.9: Classification results (CohMetrix-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
DESPL	100	WRDAOAc	100	CNCTempx	100	SYNSTRUTt	100	SYNSTRUTa	100
SYNSTRUTt	827	SYNSTRUTt	93	SMCAUSwn	63	LSAPP1d	78	SMCAUSwn	985
SMCAUSwn	785	DESPC	88	DESPC	63	DESPL	74	SYNSTRUTt	833
DESPC	617	CNCTempx	82	WRDAOAc	57.8	SMCAUSwn	66.8	DESPL	795
SYNSTRUTa	52	WRDHYPnv	74	LSAPP1d	44	DESPC	54	LSAPP1d	57
PCNARz	366	LSAPP1d	70	SYNSTRUTt	40.7	SYNSTRUTa	52	WRDAOAc	221
PCNARp	357	SMCAUSwn	69	DESPL	35	PCNARz	33	DESPC	195
CNCTempx	338	DESPL	68	PCNARz	34	PCNARp	33.4	WRDHYPnv	11
LSAPP1d	255	PCNARz	58	WRDHYPnv	33	CNCTempx	29	CNCTempx	81
WRDHYPnv	28	SYNSTRUTa	39	SYNSTRUTa	22	WRDAOAc	24	PCNARz	11
WRDAOAc	0	PCNARp	0	PCNARp	0	WRDHYPnv	0	PCNARp	0

Table Q.10: CohMetrix Indices chosen by classifier as significant (CohMetrix-IG-PS).

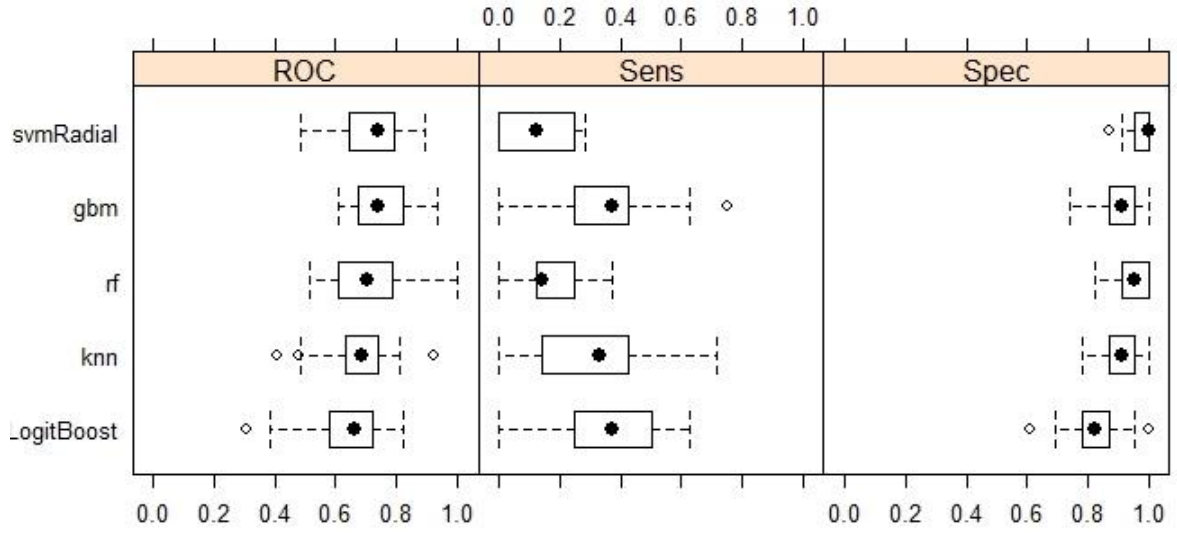


Figure Q.9: ROC, Sensitivity and Specificity for classifiers (CohMetrix-IG-PS).

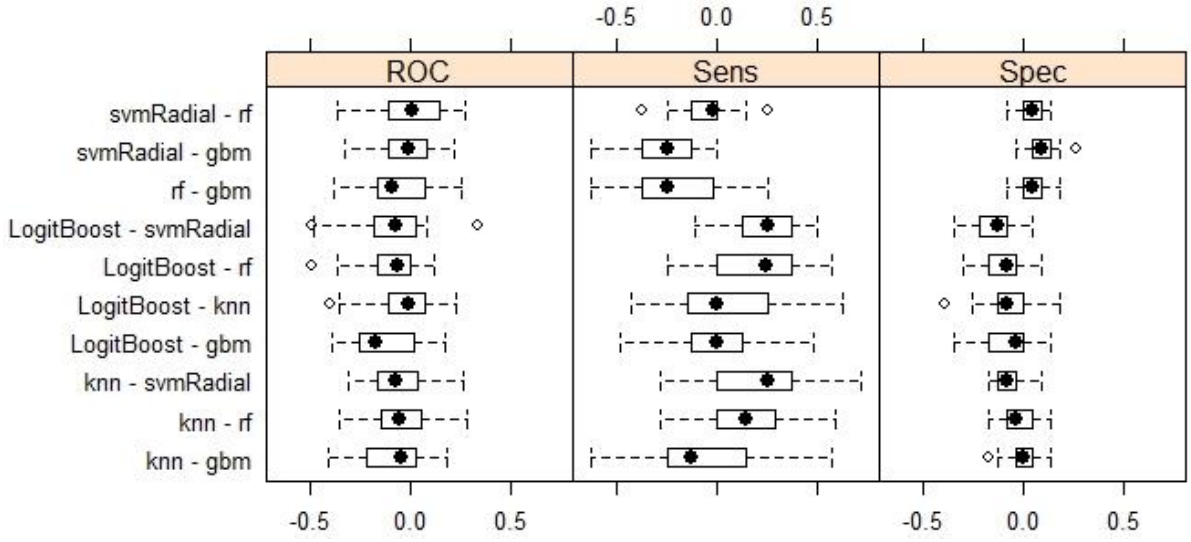


Figure Q.10: Performance compared between classifiers (CohMetrix-IG-PS).

IG feature selection on Coh-Metrix Indices – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.32	0.72	0.66	0.66	0.51 0.78	0.5	0.01	0.64	0.68	0.66
SGB	0.64	0.84	0.80	0.82	0.68 0.91	0.5	2.807e-06	0.80	0.83	0.82
RF	0.36	0.60	0.76	0.68	0.53 0.80	0.5	0.007	0.71	0.65	0.68
kNN	0.32	0.76	0.56	0.66	0.51 0.78	0.5	0.01	0.63	0.70	0.66
SVM	0.40	0.60	0.80	0.70	0.53 0.82	0.5	0.0033	0.75	0.66	0.70

Table Q.10: Classification results (CohMetrix-IG-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
LSAPP1d	100	WRDADJ	100	LSAPP1d	100	LSAPP1d	100	LSAPP1d	100
SYNSTRUTt	87	LSAPP1d	72	DRPP	87	SYNSTRUTt	851	SYNSTRUTt	96
WRDMEAc	83	DRVP	67	WRDFRQa	64	DRINF	721	DRINF	95
DRPP	81	DRPP	55	WRDMEAc	49	WRDMEAc	64	DRPP	81
DRINF	66	WRDFRQa	48	SYNSTRUTt	30	WRDFRQa	622	WRDMEAc	69
PCNARz	60	SYNSTRUTt	45	DRINF	27	DRPP	522	PCNARp	23
PCNARp	60	WRDPOSin	30	PCNARz	10	PCNARp	18	PCNARz	23
WRDFRQa	0	WRDHYPnv	0	PCNARp	0	PCNARz	0	WRDFRQa	0

Table Q.12: CohMetrix Indices chosen by classifier as significant (CohMetrix-IG-MP).

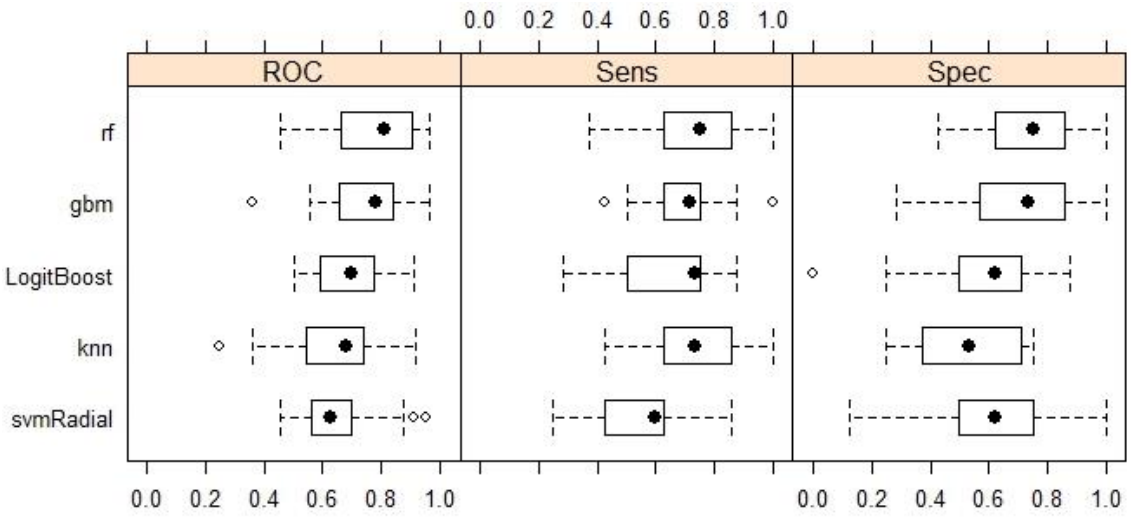


Figure N.11: ROC, Sensitivity and Specificity for classifiers (CohMetrix-IG-MP).

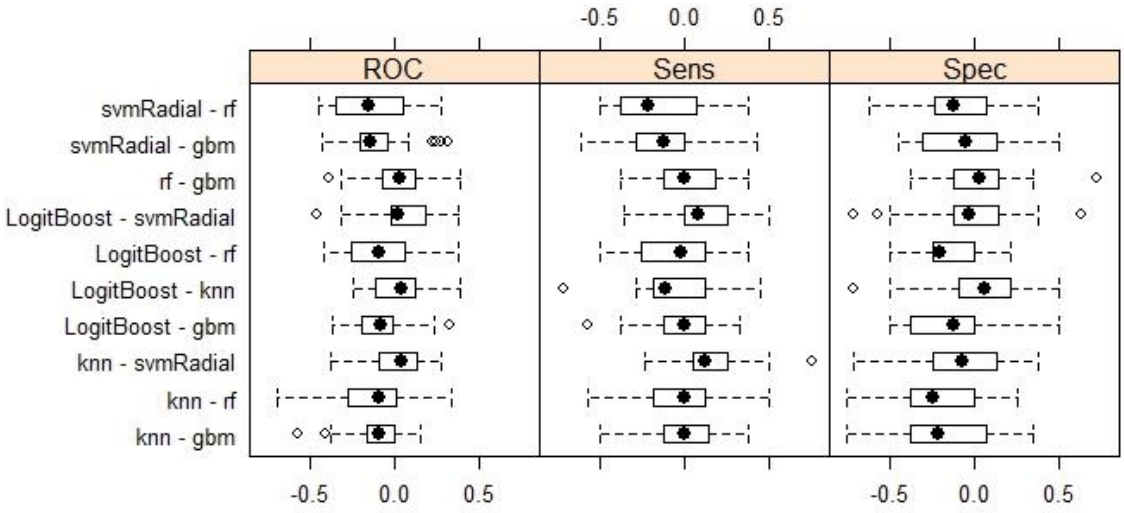


Figure Q.12: Performance compared between classifiers (CohMetrix-IG-MP).

APPENDIX R

Classification results for PCA selected LIWC variables

PCA feature selection on LIWC variables- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.23	0.43	0.80	0.71	0.62 0.78	0.75	0.87	0.41	0.81	0.61
SGB	0.52	0.52	0.94	0.84	0.75 0.90	0.75	0.02	0.76	0.85	0.73
RF	0.48	0.48	0.94	0.83	0.74 0.89	0.75	0.03	0.75	0.84	0.71
kNN	0.26	0.28	0.93	0.77	0.69 0.84	0.75	0.31	0.58	0.80	0.61
SVM	0.46	0.44	0.96	0.83	0.74 0.89	0.75	0.03	0.78	0.83	0.70

Table R.1: Classification results (LIWC-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
auxverb	100	compare	100	article	100	adj	100	Authentic	100
adj	88	WPS	94	interrog	66.4	Authentic	97	relativ	85
Authentic	81	adj	69.7	auxverb	65	Tone	88	auxverb	85
compare	80	adverb	66	Authentic	53	compare	87.3	negate	77
relativ	80	negate	63	relativ	40	relativ	85	verb	76
function.	79	cogproc	63	affect	37	auxverb	80	adj	73
verb	76	Tone	61	adj	32	article	79	cogproc	68
Tone	69	focuspresent	59	function.	31	negate	76	function.	66
percept	64	number	57	WPS	30	function.	71	percept	64
ipron	63	ppron	53	percept	29	ipron	70	focuspresent	64

Table R.2: LIWC variables chosen by classifier for (LIWC-PCA-PS).

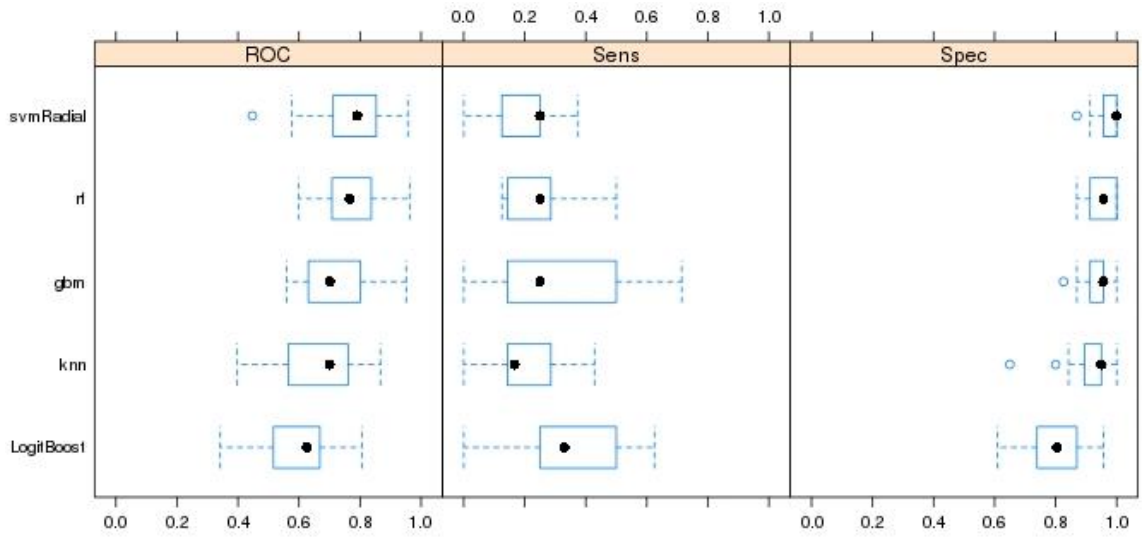


Figure R.1: ROC, Sensitivity and Specificity for classifiers (LIWC-PCA-PS).

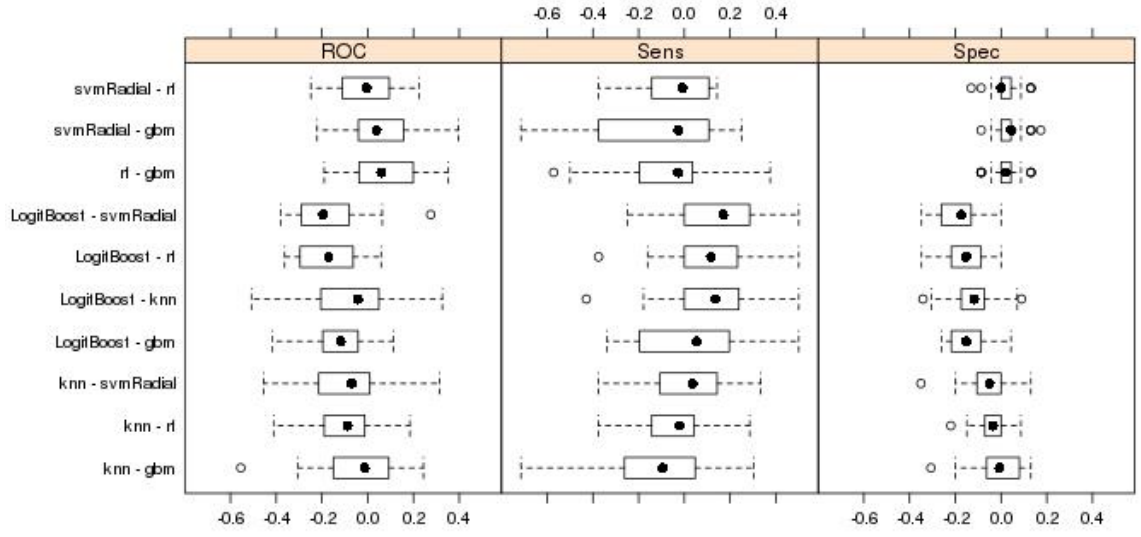


Figure R.2: Performance compared between classifiers (LIWC-PCA-PS).

PCA feature selection on LIWC variables- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.48	0.68	0.80	0.74	0.59, 0.85	0.5	0.0004	0.77	0.71	0.74
SGB	0.72	0.92	0.80	0.86	0.73 0.94	0.5	1.049e-07	0.82	0.90	0.86
RF	0.44	0.72	0.72	0.72	0.57 0.83	0.5	0.001	0.72	0.72	0.72
kNN	0.66	0.64	0.68	0.66	0.51 0.78	0.5	0.01	0.66	0.65	0.66
SVM	0.64	0.80	0.84	0.82	0.68 0.91	0.5	2.807e-06	0.83	0.80	0.82

Table R.3: Classification results (LIWC-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Authentic	100	ipron	100	Authentic	100	Authentic	100	Authentic	100
relativ	86	Tone	99	Tone	56	adj	92	auxverb	91
adj	84	Authentic	88	auxverb	46	Tone	88	relativ	89
compare	84	Analytic	83	adj	46	Tone.1	88	Tone.1	84
prep	81	function.	83	prep	43	prep	87	Tone	84
Tone	74	negate	81	Tone.1	43	ipron	85	prep	84
Tone.1	74	adj	79	compare	33	ipron.1	85	ipron.1	79
ipron.1	73	WPS	77	relativ	31	compare	85.1	ipron	79
ipron	73	focuspresent	77	percept	25	relativ	77	verb	76
focuspresent	65.9	compare	74	verb	24	auxverb	72	focuspresent	73
Analytic	63	interrog	68	focuspresent	24	verb	67	interrog	67
auxverb	60	number	62	WPS	23	Analytic	56	Analytic	63

Table R.4: LIWC variables chosen by classifier as significant (LIWC-PCA-MP)

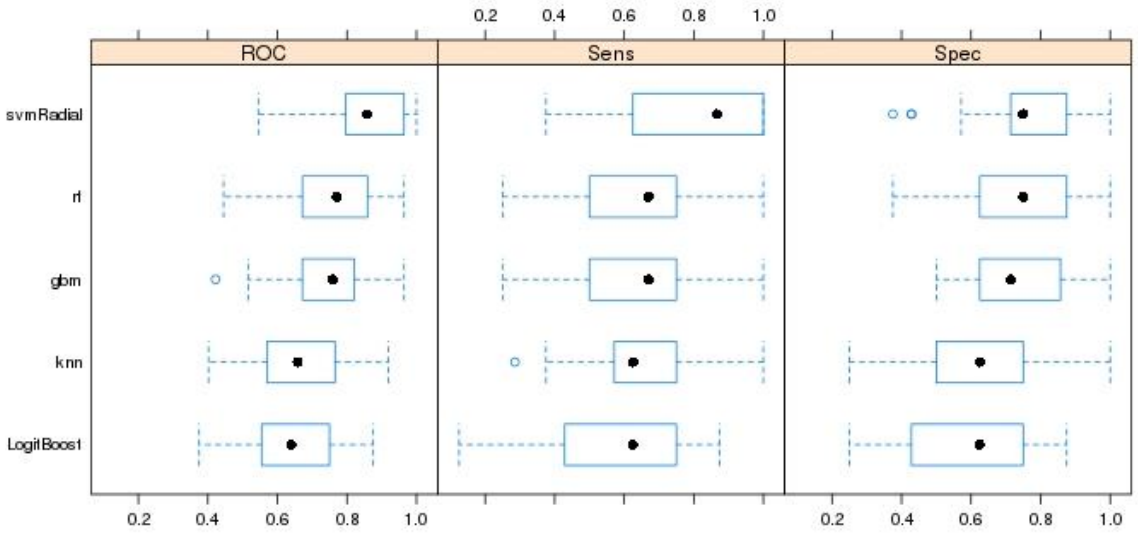


Figure R.3: ROC, Sensitivity and Specificity for classifiers (LIWC-PCA-MP).

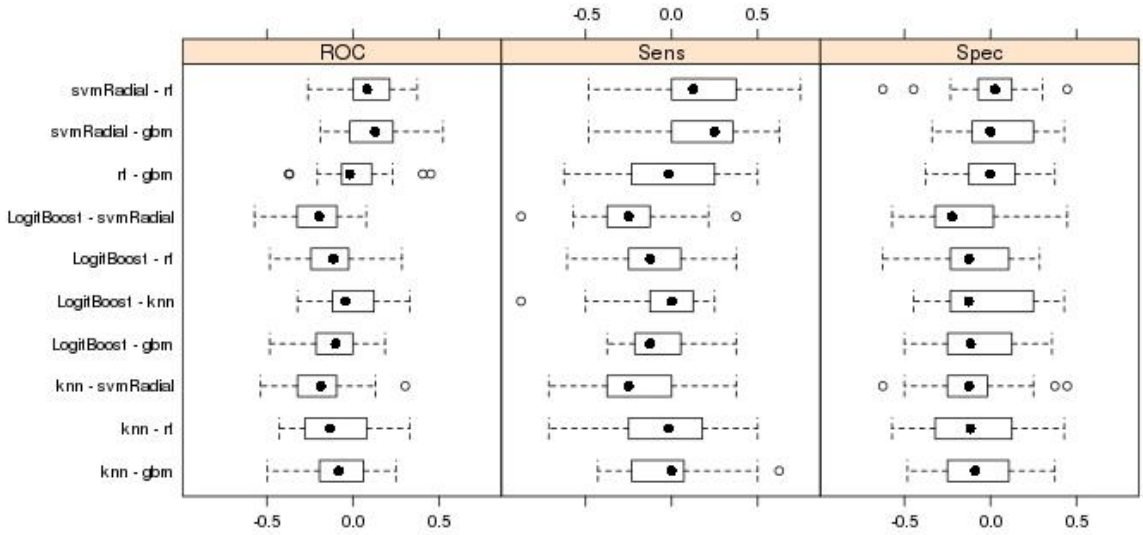


Figure R.4: Performance compared between classifiers (LIWC-PCA-MP).

Classification results for Boruta selected LIWC variables

Boruta feature selection on LIWC variables- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.42	0.81	0.71	0.63 0.79	0.75	0.85	0.42	0.81	0.62
SGB	0.43	0.52	0.89	0.80	0.71 0.87	0.75	0.14	0.61	0.85	0.70
RF	0.51	0.44	0.98	0.85	0.76 0.91	0.75	0.01	0.91	0.84	0.71
kNN	0.26	0.32	0.90	0.76	0.66 0.84	0.75	0.46	0.53	0.80	0.61
SVM	0.54	0.44	1	0.86	0.76 0.91	0.75	0.01	1	0.83	0.72

Table R.5: Classification results (LIWC-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Imagery	100	Avg.Sent. Length	100	Temporal.Imm..Ratio	100	Temporal.Imm .Ratio	100	Temporal.Imm..Ratio	100
Temporal.Imm. .Ratio	95	Temporal .Imm..Ratio	998	Imagery	784	Imagery	981	Pleasantness	977
Pleasantness	82	Imagery	769	Verb.Qty	769	Content.Word. Diversity	765	Imagery	938
Content.Word. Diversity	79	Avg.Word. Length	744	Modal.Verb .Ratio	754	Pleasantness	586	Content.Word .Diversity	72
Avg.Sent. Length	39	Content.Word.Diversity	678	Pleasantness	67	Modal.Verb. Ratio	543	Sensory.Ratio	543
Sentence.Qty	37	Modal.Verb .Ratio	664	Avg.Sent Length	649	Verb.Qty	524	Avg.Sent. Length	391
Pausality	33	Group.Ref	616	Avg.Word. Length	524	Sensory.Ratio	453	Verb.Qty	339
Word.Qty	33	Sensory.Ratio	589	Sensory. Ratio	48.6	Avg.Sent. Length	392	Modal.Verb. Ratio	329
Verb.Qty	33	Verb.Qty	56	Content.Word.Diversity	402	Sentence.Qty	37	Sentence.Qty	313
Modal.Verb. Ratio	31	Pleasantness	533	Group.Ref	257	Pausality	37	Avg.Word. Length	263
Sensory.Ratio	31	Pausality	475	Lexical.Diversity	229	Avg.Word.Length	253	Word.Qty	236

Table R.6: LIWC variables chosen by classifier as significant (LIWC-Boruta-PS).

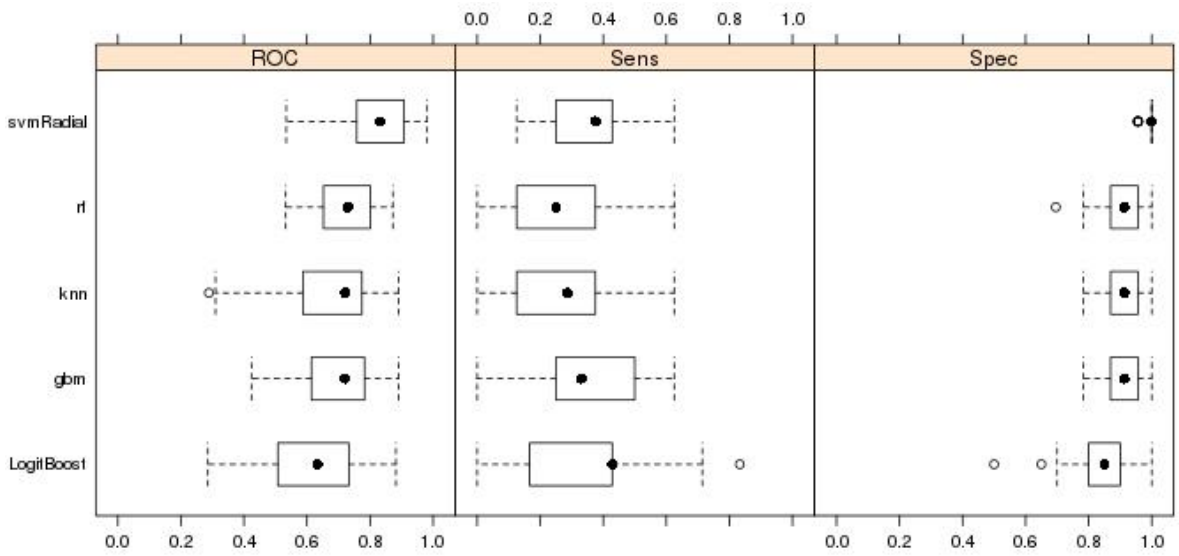


Figure R.6: ROC, Sensitivity and Specificity for classifiers (LIWC-Boruta-PS).

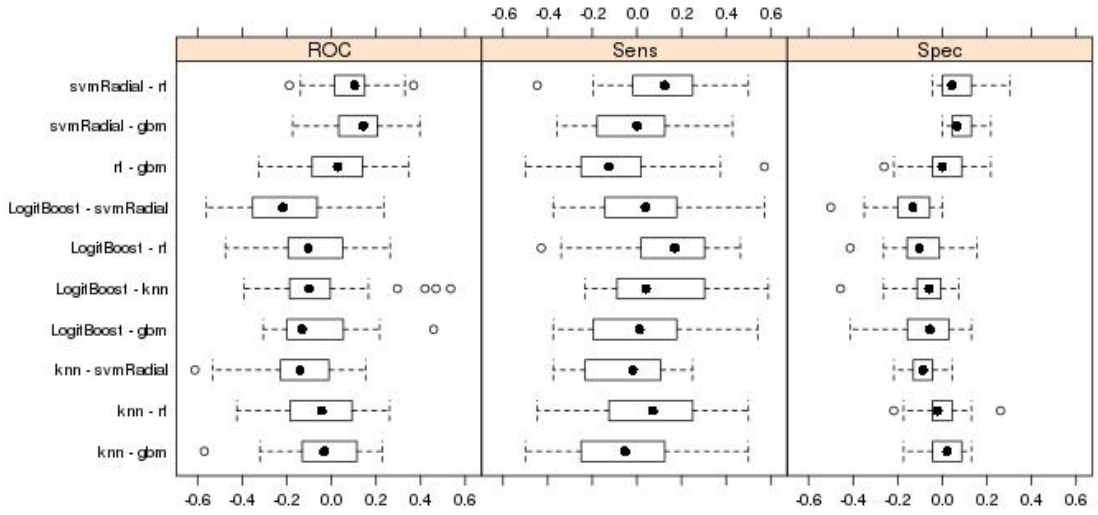


Figure R.7: Performance compared between classifiers (LIWC-Boruta-PS).

Boruta feature selection on LIWC variables- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.4	0.62	0.77	0.70	0.57 0.80	0.5	0.0005	0.73	0.67	0.70
SGB	0.6	0.76	0.84	0.80	0.66 0.89	0.5	1.193e-05	0.82	0.77	0.80
RF	0.4	0.60	0.80	0.70	0.55 0.82	0.5	0.0033	0.75	0.66	0.70
kNN	0.24	0.68	0.56	0.62	0.47 0.75	0.5	0.05	0.60	0.63	0.63
SVM	0.48	0.64	0.84	0.74	0.59 0.85	0.5	0.0004	0.80	0.70	0.74

Table R.7: Classification results (LIWC-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Authentic	100	Authentic	100	Tone	100	Authentic	100	ipron	100
ipron	70	Tone	98	Authentic	82	adj	92	Authentic	93
relativ	60	relativ	64	relativ	43	Tone	88	adj	83
prep	60	Analytic	63	ipron	22	prep	87	relativ	76
auxverb	44	ipron	58	auxverb	19	ipron	85	auxverb	76
Tone	41	auxverb	55	adj	15	relativ	77	Tone	36
adj	16	prep	53	interrog	15	auxverb	72	prep	35
interrog	0	adj	28	Analytic	10	Analytic	56	interrog	17
Analytic	0	interrog	0	prep	0	interrog	0	Analytic	0

Table R.8: LIWC variables chosen by classifier as significant (LIWC-Boruta-MP).

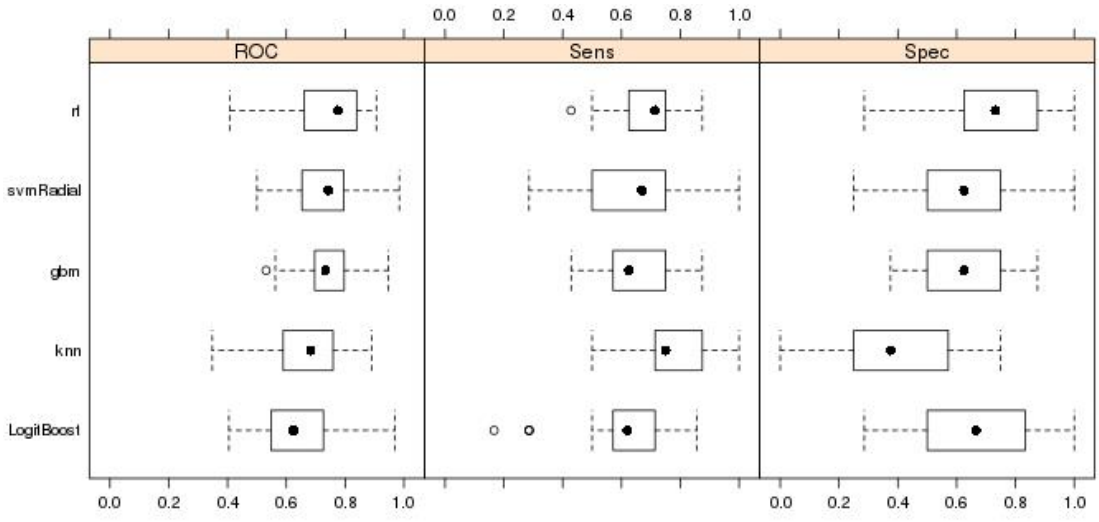


Figure R.7: ROC, Sensitivity and Specificity for classifiers (LIWC-Boruta-MP).

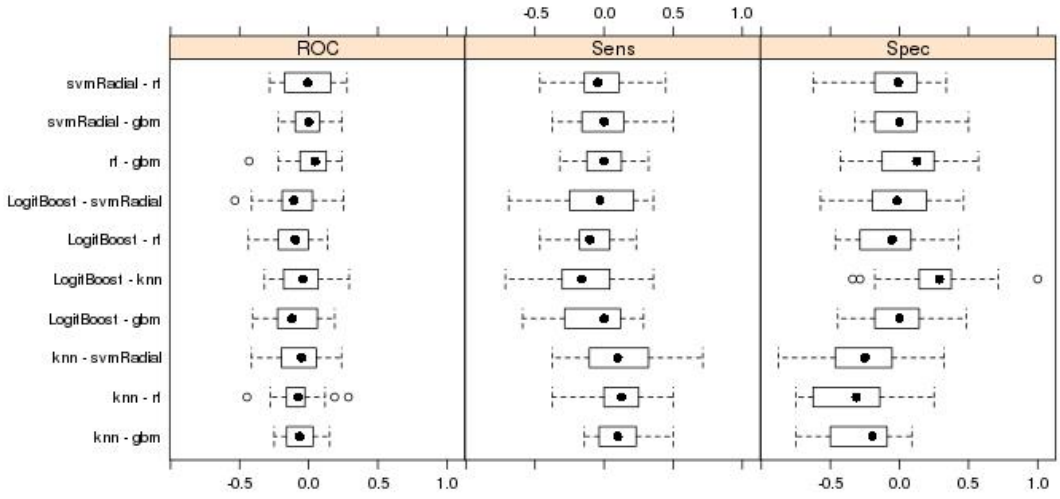


Figure R.8: Performance compared between classifiers (LIWC-Boruta-MP).

Peer Set: Classification results for IG selected LIWC variables

IG feature selection on LIWC variables- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.17	0.22	0.91	0.74	0.66 0.81	0.75	0.62	0.47	0.78	0.57
SGB	0.04	0.28	0.76	0.64	0.54 0.73	0.75	0.99	0.28	0.76	0.52
RF	0.28	0.32	0.92	0.77	0.67 0.84	0.75	0.37	0.57	0.80	0.62
kNN	- 0.017	0.04	0.94	0.72	0.62 0.80	0.75	0.79	0.20	0.75	0.49
SVM	- 0.019 4	0	0.98	0.74	0.64 0.82	0.75	0.64	0	0.75	0.49

Table R.9: Classification results (LIWC-IG-PS).

APPENDIX S

Classification results for PCA selected Linguistic Based Cues (LBCs)

PCA feature selection on LBCs- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.19	0.48	0.73	0.67	0.57 0.76	0.75	0.97	0.37	0.81	0.60
SGB	0.46	0.44	0.96	0.83	0.74 0.89	0.75	0.03	0.78	0.83	0.70
RF	0.36	0.36	0.94	0.80	0.71 0.87	0.75	0.14	0.69	0.81	0.65
kNN	0.28	0.40	0.86	0.75	0.65 0.83	0.75	0.55	0.50	0.81	0.63
SVM	0.42	0.40	0.96	0.82	0.73 0.89	0.75	0.06	0.76	0.82	0.68

Table S.1: Classification results (LBCs-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Imagery	100	Avg.Sent. Length	100	Temporal.Imm..Ratio	100	Temporal.Imm .Ratio	100	Temporal.Imm..Ratio	100
Temporal.Imm. .Ratio	95	Temporal .Imm..Ratio	998	Imagery	784	Imagery	981	Pleasantness	977
Pleasantness	82	Imagery	769	Verb.Qty	769	Content.Word. Diversity	765	Imagery	938
Content.Word. Diversity	79	Avg.Word. Length	744	Modal.Verb .Ratio	754	Pleasantness	586	Content.Word .Diversity	72
Avg.Sent. Length	39	Content.Wo rd.Diversity	678	Pleasantne ss	67	Modal.Verb. Ratio	543	Sensory.Ratio	543
Sentence.Qty	37	Modal.Verb .Ratio	664	Avg.Sent Length	649	Verb.Qty	524	Avg.Sent. Length	391
Pausality	33	Group.Ref	616	Avg.Word. Length	524	Sensory.Ratio	453	Verb.Qty	339
Word.Qty	33	Sensory.Ra tio	589	Sensory. Ratio	48.6	Avg.Sent. Length	392	Modal.Verb. Ratio	329
Verb.Qty	33	Verb.Qty	56	Content.Wo rd.Diversity	402	Sentence.Qty	37	Sentence.Qty	313
Modal.Verb. Ratio	31	Pleasantne ss	533	Group.Ref	257	Pausality	37	Avg.Word. Length	263
Sensory.Ratio	31	Pausality	475	Lexical.Div ersity	229	Avg.Word.Len gth	253	Word.Qty	236

Table S.2: LBCs chosen by classifier for (LBCs-PCA-PS).

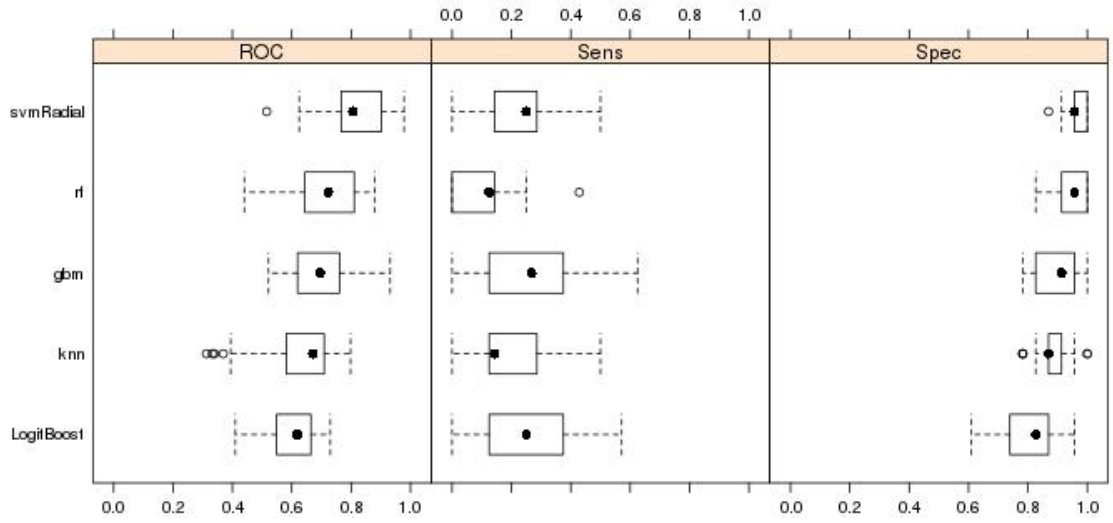


Figure S.1: ROC, Sensitivity and Specificity for classifiers (LBCs-PCA-PS).

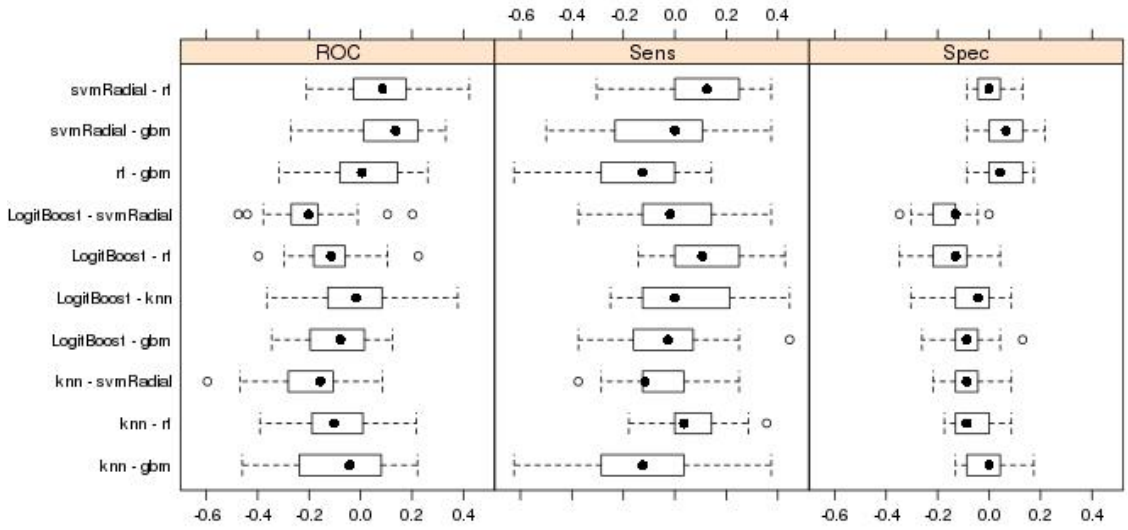


Figure S.2: Performance compared between classifiers (LBCs-PCA-PS).

PCA feature selection on LBCs- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.36	0.72	0.64	0.68	0.53 0.80	0.5	0.007	0.66	0.69	0.68
SGB	0.28	0.60	0.68	0.64	0.49 0.77	0.5	0.03	0.65	0.62	0.64
RF	0.40	0.68	0.72	0.70	0.55 0.82	0.5	0.0033	0.70	0.69	0.70
kNN	0.08	0.40	0.68	0.54	0.39 0.68	0.5	0.33	0.55	0.53	0.54
SVM	0.40	0.84	0.56	0.70	0.55 0.82	0.5	0.0033	0.65	0.77	0.70

Table S.3: Classification results (LBCs-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Lexical. Diversity	100	Modal.Verb. Ratio	100	Content.Word.Diversity	100	Content.Word.Diversity	100	Affect	100
Avg.Word. Length	95	Group.Ref	986	Group.Ref	836	Imagery	72	Content.Word.Diversity	84
Imagery	80	Imagery	666	Verb.Qty	581	Verb.Qty	66	Pleasantness	61
Content.Word. Diversity	74	Content.Word.Diversity	564	Avg.Sent. Length	58	Avg.Word.Length	43	Avg.Word. Length	59
Verb.Qty	73	Avg.Sent. Length	566	Affect	58	Group.Ref	42	Emotiveness	51
Avg.Sent. Length	59	Verb.Qty	554	Modal.Verb. Ratio	571	Modal.Verb. Ratio	42	Imagery	48
Emotiveness	55	Emotiveness	473	Imagery	511	Emotiveness	37	Verb.Qty	46
Group.Ref	52	Lexical. Diversity	399	Avg.Word. Length	393	Avg.Sent. Length	35	Pausality	44
Word.Qty	43	Affect	321	Emotiveness	38	Pausality	33	Sentence.Qty	42
Pausality	40	Pleasantness	293	Pleasantnes	318	Lexical.Diversity	32	Function.Word.Diversity	38
Pleasantness	39	Modifier.Qty	262	Lexical. Diversity	234	Sentence.Qty	30	Avg.Sent. Length	35
Function.Word .Diversity	39	Pausality	263	Pausality	112	Modifier.Qty	29	Lexical. Diversity	34
Modal.Verb. Ratio	38	Avg.Word.Length	236	Sentence. Qty	72	Pleasantness	5	Group.Ref	29

Table S.4: LBCs chosen by classifier as significant (LBCs-PCA-MP).

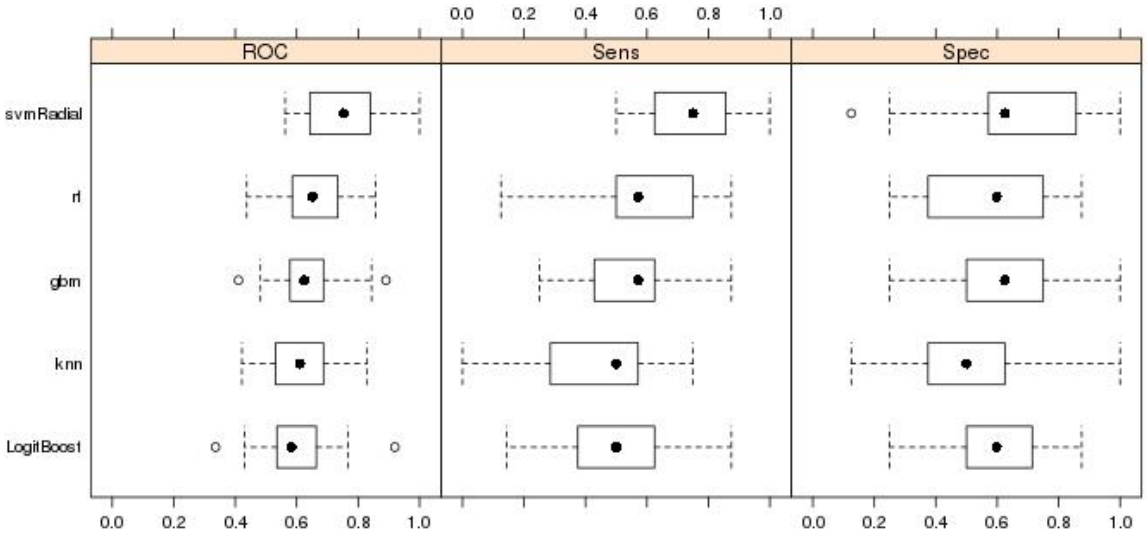


Figure S.3: ROC, Sensitivity and Specificity for classifiers (LBCs-PCA-MP).

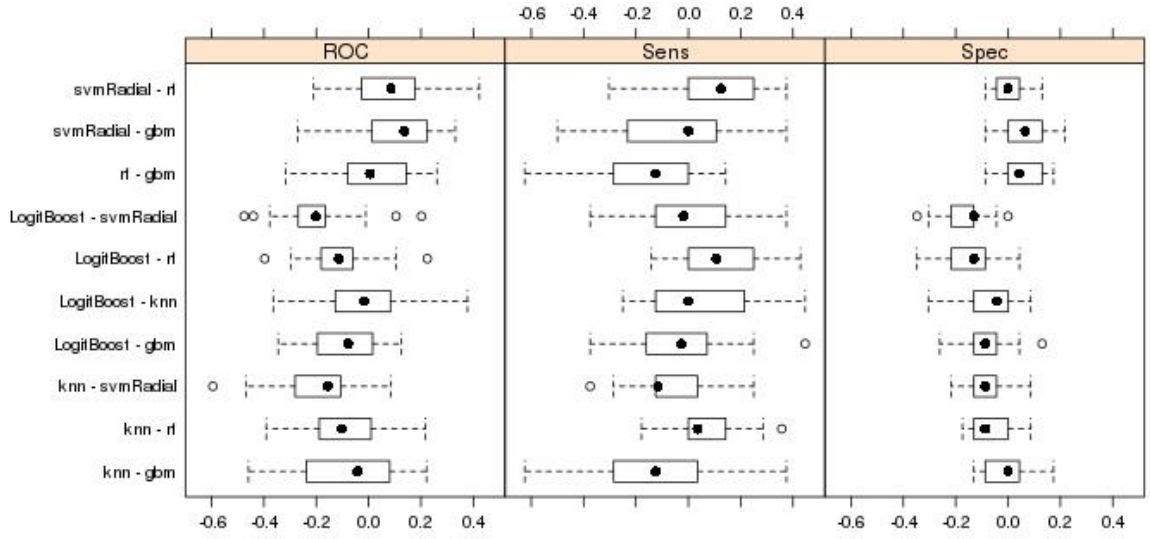


Figure S.4: Performance compared between classifiers (LBCs-PCA-MP).

Peer Set: Classification results for Boruta selected LBC

Boruta feature selection on LBCs- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.38	0.44	0.90	0.79	0.69 0.86	0.75	0.21	0.61	0.83	0.67
SGB	0.32	0.56	0.78	0.73	0.63 0.81	0.75	0.72	0.46	0.84	0.67
RF	0.45	0.43	0.95	0.82	0.74 0.88	0.75	0.033	0.76	0.83	0.69
kNN	0.18	0.15	0.98	0.77	0.67 0.86	0.75	0.35	0.75	0.77	0.56
SVM	0.13	0.16	0.94	0.75	0.65 0.83	0.75	0.55	0.50	0.77	0.55

Table S.5: Classification results (LBCs-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
passive.verb. ratio	100	Temporal. Imm..Ratio	100	Temporal. Imm..Ratio	100	Temporal. Imm..Ratio	100	Temporal. Imm..Ratio	100
Temporal .Imm..Ratio	77	passive.verb .ratio	71	passive.verb .ratio	85	passive.verb. ratio	69	passive.verb .ratio	94
Verb.Qty	50	Modal.Verb. Ratio	52	Modal.Verb. Ratio	75	Modal.Verb. Ratio	47	Modal.Verb .Ratio	60
Modal.Verb. Ratio	34	Verb.Qty	48	Verb.Qty	63	Sentence.Qty	39	Sentence.Qty	49
Pausality	15	Group.Ref	37	Group.Ref	42	Pausality	36	Group.Ref	49
Group.Ref	14	Modifier.Qty	9	Modifier.Qty	17	Word.Qty	32	Word.Qty	46
Modifier.Qty	13	Pausality	6	Pausality	12	Verb.Qty	31	Pausality	45
Word.Qty	9	Word.Qty	3	Word.Qty	2	Group.Ref	25	Function.Word.Diversity	43
Function.Word .Diversity	6	Function.Word.Diversity	1	Sentence. Qty	1	Modifier.Qty	21	Modifier.Qty	29
Sentence.Qty	0	Sentence. Qty	0	Function.Word.Diversity	0	Function.Word.Diversity	0	Verb.Qty	0

Table S.6: LBCs chosen by classifier as significant (LBCs-Boruta-PS).

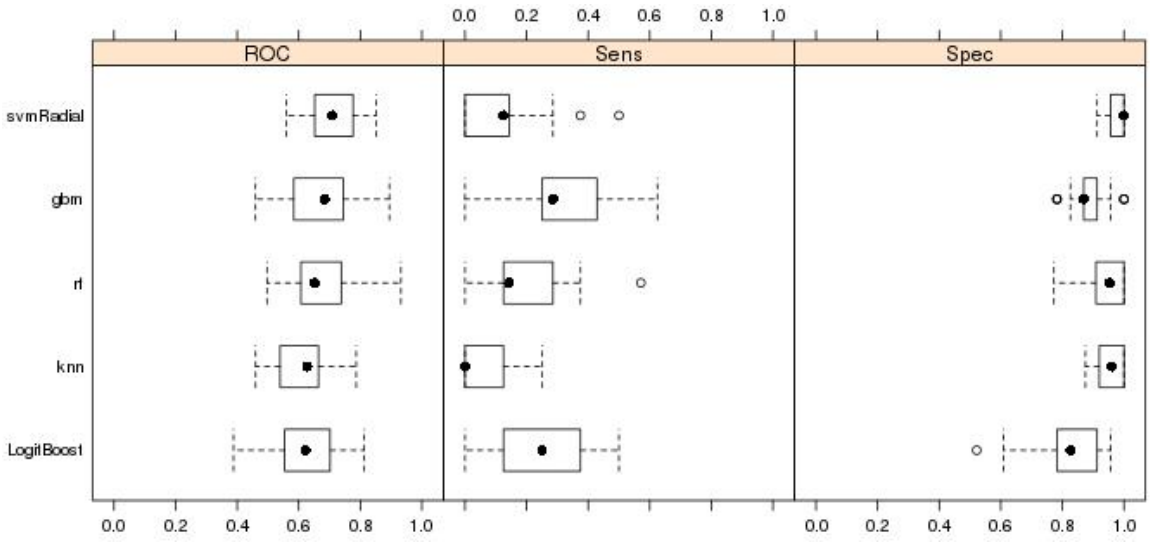


Table S.6: LBCs chosen by classifier as significant (LBCs-Boruta-PS).

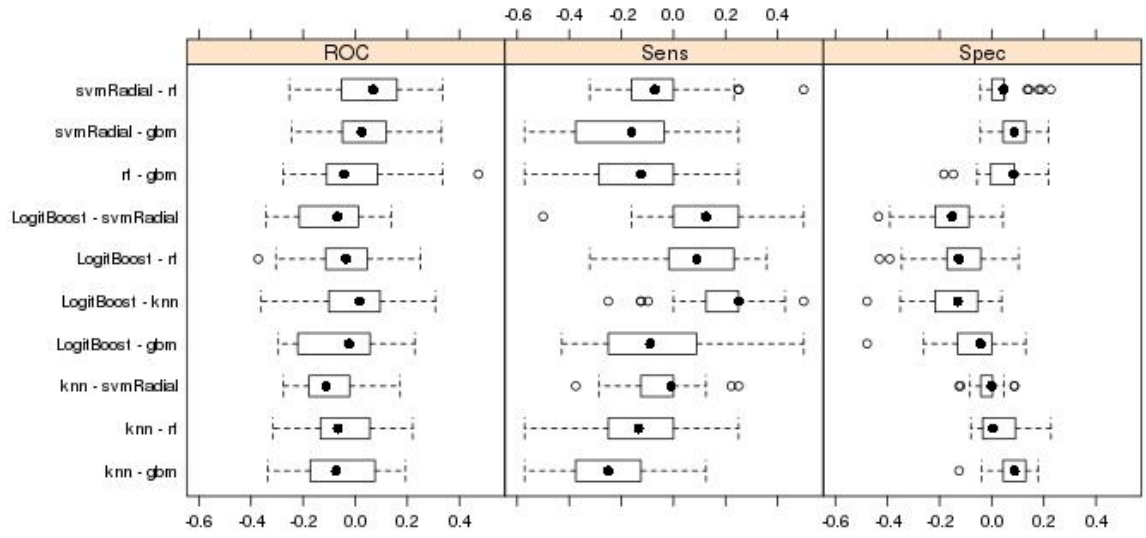


Figure S.8: Performance compared between classifiers (LBCs-Boruta-MP).

Boruta feature selection on LBCs- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.16	0.60	0.56	0.58	0.43 0.71	0.5	1	0.57	0.58	0.58
SGB	0.40	0.68	0.72	0.70	0.55 0.82	0.5	0.0033	0.70	0.69	0.70
RF	0.48	0.68	0.80	0.74	0.59 0.85	0.5	0.0004	0.77	0.71	0.74
kNN	0.24	0.64	0.60	0.62	0.47 0.75	0.5	0.05	0.61	0.62	0.62
SVM	0.28	0.68	0.60	0.64	0.49 0.77	0.5	0.03	0.62	0.65	0.64

Table S.9: Classification results (LBCs-IG-PS).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Content.Word.Diversity	100	Other.Ref	100	Content.Word.Diversity	100	Other.Ref	100	Content.Word.Diversity	100
Group.Ref	73	Content.Word.Diversity	60	Modal.Verb.Ratio	66	Content.Word.Diversity	92	Modal.Verb.Ratio	59
Other.Ref	13	Group.Ref	14	Other.Ref	30	Group.Ref	27	Group.Ref	45
Modal.Verb.Ratio	0	Modal.Verb.Ratio	0	Group.Ref	0	Modal.Verb.Ratio	0	Other.Ref	0

Table S.10: LBCs chosen by classifier as significant (LBCs-IG-PS).

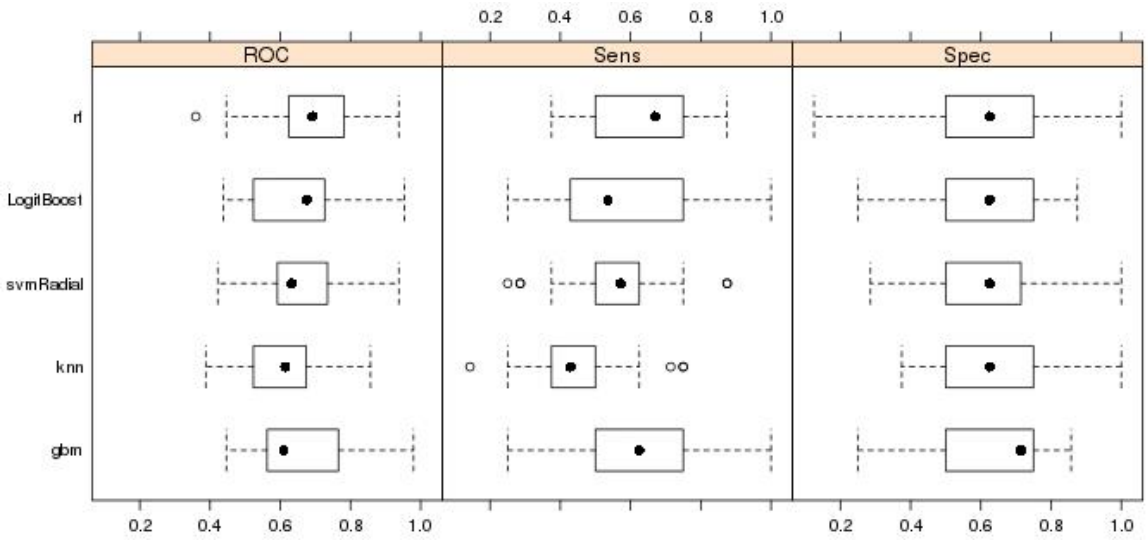


Figure S.9: ROC, Sensitivity and Specificity for classifiers (LBCs-IG-PS).

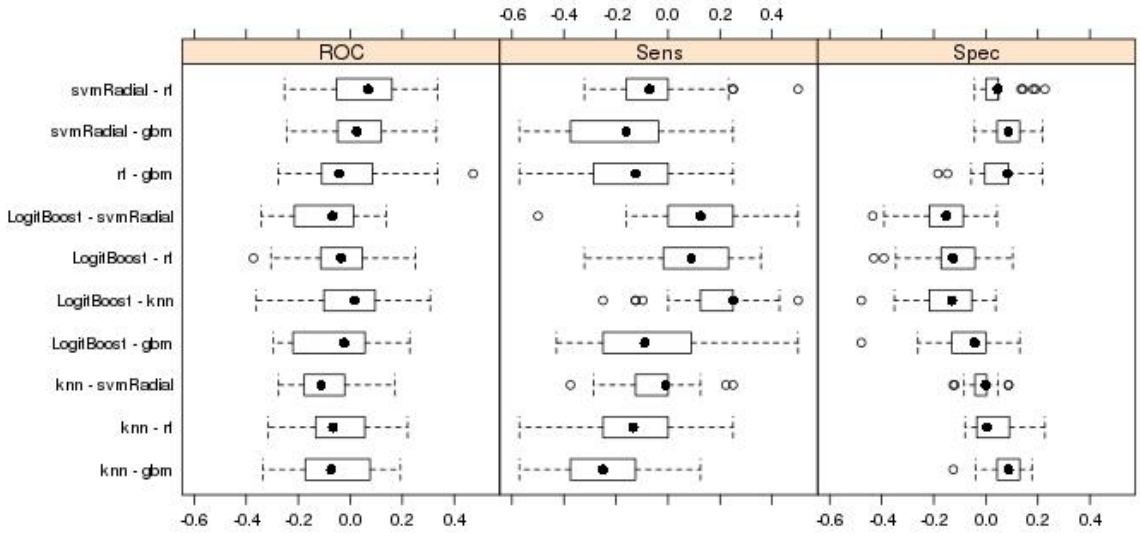


Figure S.10: Performance compared between classifiers (LBCs-IG-PS).

Peer Set: Classification results for IG selected LBCs

IG feature selection on LBCs- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.10	0.16	0.92	0.73	0.63 0.81	0.75	0.72	0.40	0.76	0.54
SGB	- 0.011	0.20	0.78	0.64	0.54 0.73	0.75	0.99	0.23	0.75	0.49
RF	0.10	0.24	0.85	0.70	0.60 0.78	0.75	0.89	0.35	0.77	0.54
kNN	0.05	0.12	0.92	0.72	0.62 0.80	0.75	0.79	0.33	0.76	0.52
SVM	0	0	1	0.75	0.65 0.86	0.75	0.55	0	0.75	0.50

Table S.9: Classification results (LBCs-IG-PS).

Only Temporal.Imm..Ratio for IG Peer Set

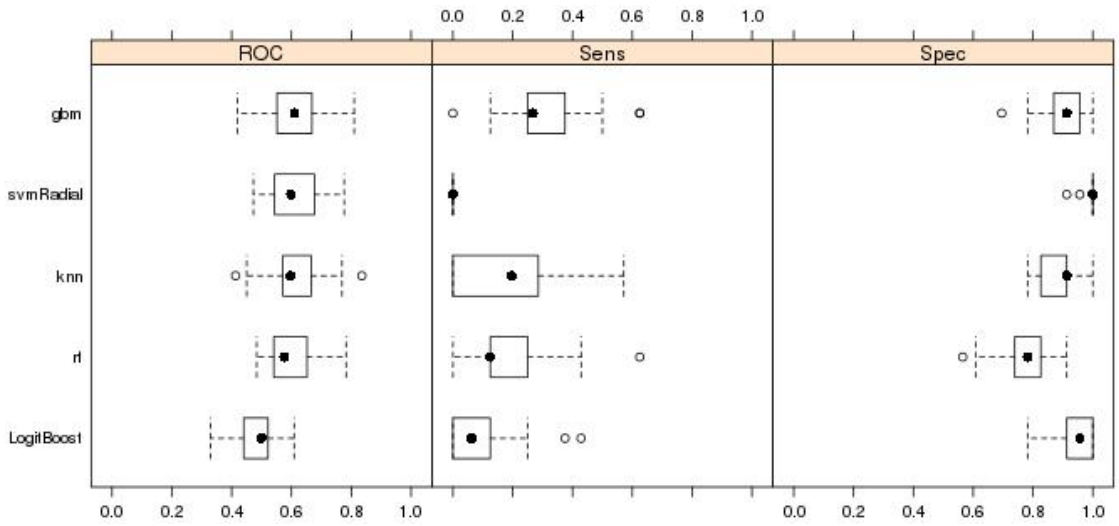


Figure S.9: ROC, Sensitivity and Specificity for classifiers (LBCs-IG-PS).

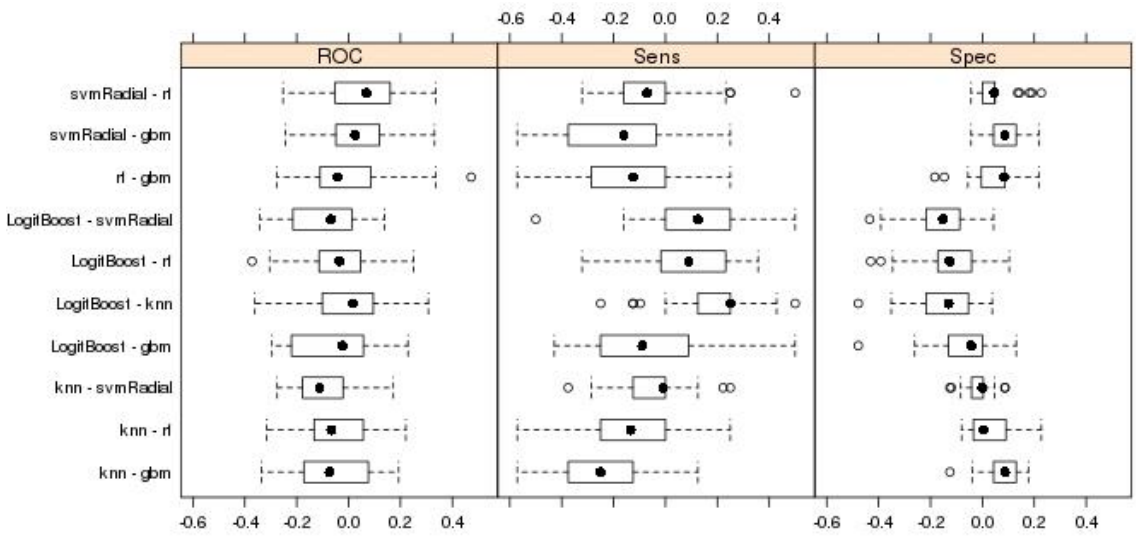


Figure S.10: Performance compared between classifiers (LBCs-IG-PS).

Matched Pair: Classification results for IG selected Topics

No features chosen as significant by IG

APPENDIX T

Peer Set: Classification results for PCA selected Topics

PCA feature selection on Topics- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.17	0.40	0.77	0.68	0.58 0.77	0.75	0.95	0.37	0.79	0.58
SGB	0.40	0.48	0.89	0.79	0.69 0.86	0.75	0.21	0.60	0.83	0.68
RF	0.28	0.28	0.94	0.78	0.68 0.85	0.75	0.28	0.63	0.80	0.61
kNN	0.35	0.45	0.88	0.77	0.67 0.86	0.75	0.35	0.56	0.83	0.66
SVM	0.40	0.40	0.94	0.81	0.72 0.88	0.75	0.09	0.71	0.82	0.67

Table T.1: Classification results (Topics-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Topic 24	100	Topic 24	100	Topic 24	100	Topic 24	100	Topic 24	100
Topic 3	81	Topic 3	69	Topic 3	89	Topic 3	783	Topic 5	72
Topic 5	52	Topic 0	65	Topic 0	65	Topic 5	637	Topic 3	72
Topic 4	48	Topic 22	35	Topic 22	46	Topic 4	572	Topic 4	68
Topic 23	47	Topic 9	29	Topic 4	37	Topic 23	483	Topic 14	49
Topic 0	32	Topic 5	28	Topic 5	32	Topic 0	248	Topic 15	46
Topic 1	23	Topic 17	23	Topic 9	29	Topic 1	237	Topic 1	40
Topic 10	22	Topic 23	20	Topic 7	29	Topic 22	222	Topic 7	37
Topic 14	17	Topic 15	16	Topic 10	19	Topic 14	188	Topic 10	36
Topic 7	17	Topic 4	15	Topic 15	8	Topic 7	139	Topic 23	31

Table T.2: Topics chosen by classifier for (Topics-PCA-PS).

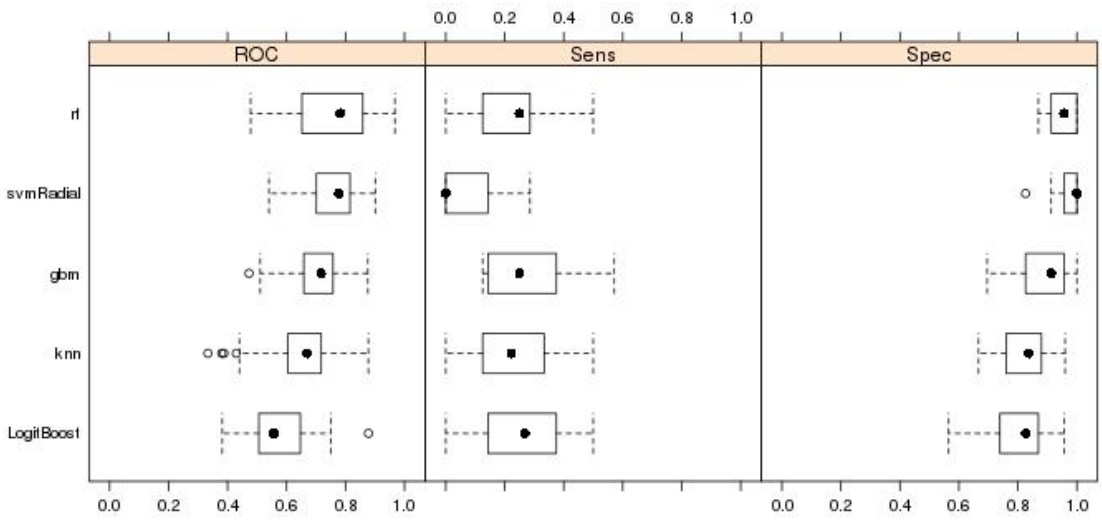


Figure T.1: ROC, Sensitivity and Specificity for classifiers (Topics-PCA-PS).

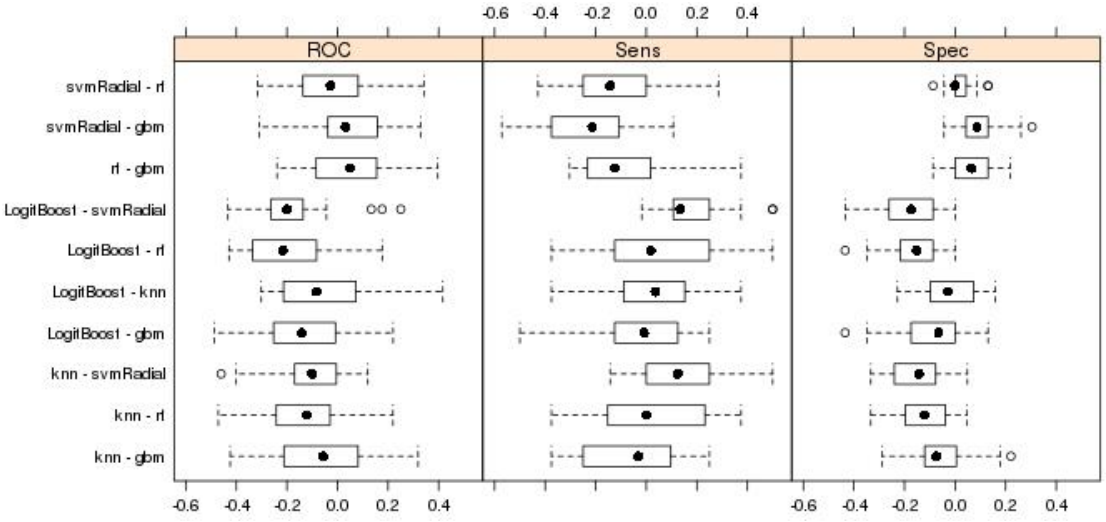


Figure T.2: Performance compared between classifiers (Topics-PCA-PS).

Classification results for PCA selected Topics

PCA feature selection on LIWC variables- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.32	0.64	0.68	0.66	0.51 0.78	0.5	0.01	0.66	0.65	0.66
SGB	0.32	0.68	0.64	0.66	0.51 0.78	0.5	0.01	0.65	0.66	0.66
RF	0.48	0.80	0.68	0.74	0.59 0.85	0.5	0.0004	0.71	0.77	0.74
kNN	0.35	0.75	0.60	0.67	0.50 0.81	0.5	0.01	0.65	0.70	0.67
SVM	0.4	0.68	0.72	0.7	0.55 0.82	0.5	0.0033	0.70	0.69	0.70

Table T.3: Classification results (Topics-PCA-MP).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Topic 24	100	Topic 24	100	Topic 3	100	Topic 24	100	Topic 24	100
Topic 3	697	Topic 23	43	Topic 24	99	Topic 3	92	Topic 3	81
Topic 0	65	Topic 3	28	Topic 5	34	Topic 5	52	Topic 23	49
Topic 22	352	Topic 4	21	Topic 4	30	Topic 4	49	Topic 4	43
Topic 9	297	Topic 0	17	Topic 0	28	Topic 0	42	Topic 5	42
Topic 5	288	Topic 17	16	Topic 15	27	Topic 23	37	Topic 0	31
Topic 17	238	Topic 15	16	Topic 9	25	Topic 9	27	Topic 11	19
Topic 23	203	Topic 9	14	Topic 12	17	Topic 14	24	Topic 7	19
Topic 15	164	Topic 10	11	Topic 10	16	Topic 17	20	Topic 9	14
Topic 4	157	Topic 7	11	Topic 17	11	Topic 11	18	Topic 10	11
Topic 11	111	Topic 5	8	Topic 7	10	Topic 10	18	Topic 14	10
Topic 1	119	Topic 6	6	Topic 23	7	Topic 6	17	Topic 1	10
Topic 7	107	Topic 14	4	Topic 6	7	Topic 7	6	Topic 17	6

Table T.4: Topics chosen by classifier as significant (Topics-PCA-MP).

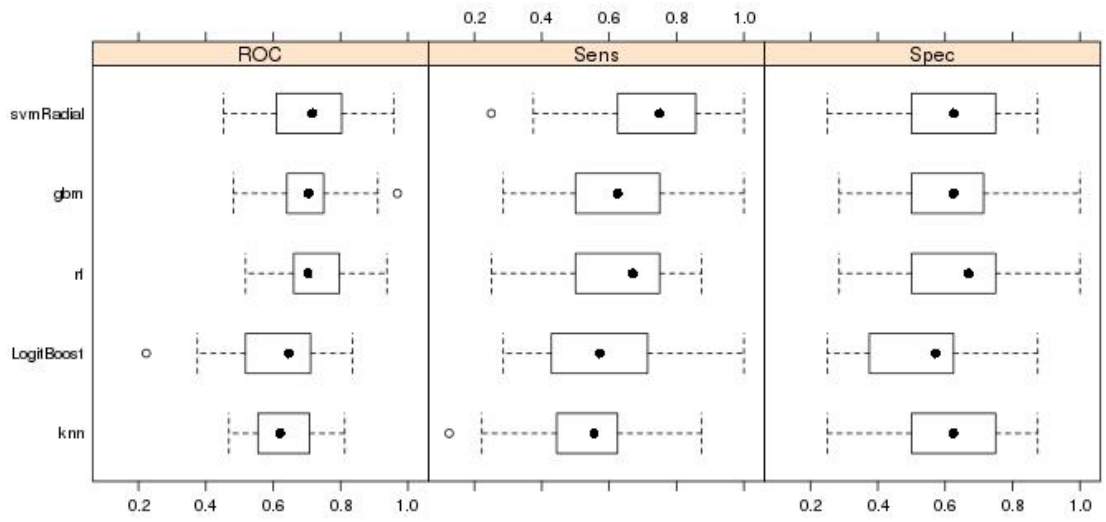


Figure T.3: ROC, Sensitivity and Specificity for classifiers (Topics-PCA-MP).

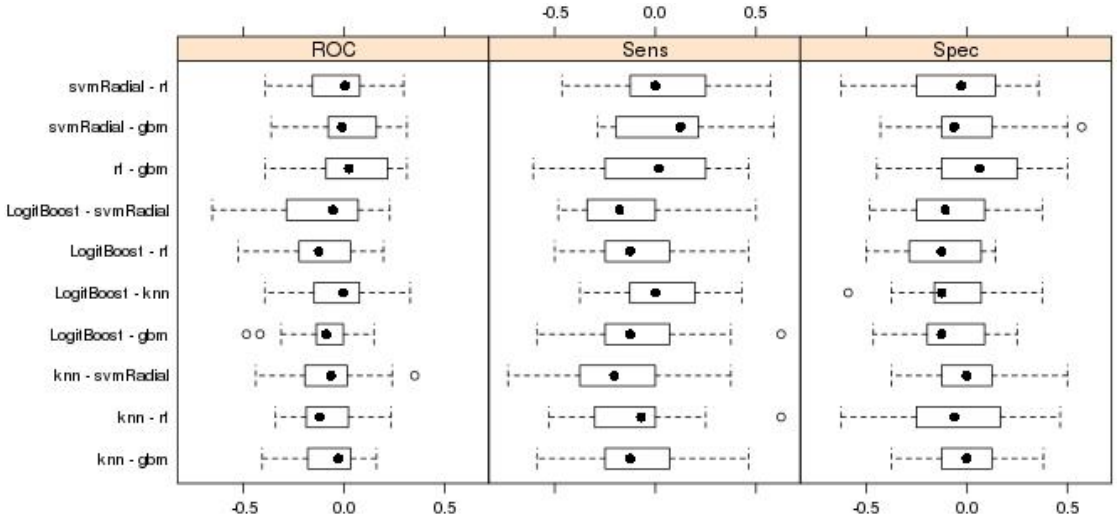


Figure T.4: Performance compared between classifiers (Topics-PCA-MP).

Classification results for Boruta selected Topics

Boruta feature selection on Topics- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.38	0.40	0.93	0.80	0.69 0.88	0.75	0.18	0.66	0.82	0.66
SGB	0.42	0.44	0.93	0.81	0.72 0.88	0.75	0.09	0.68	0.83	0.68
RF	0.27	0.25	0.96	0.79	0.68 0.87	0.75	0.26	0.71	0.79	0.60
kNN	0.24	0.30	0.91	0.76	0.67 0.83	0.75	0.46	0.52	0.79	0.60
SVM	0.25	0.23	0.96	0.78	0.70 0.85	0.75	0.23	0.70	0.79	0.60

Table T.5: Classification results (Topics-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Topic 24	100	Topic 24	100	Topic 24		Topic 24	100	Topic 24	100
Topic 3	77	Topic 18	33	Topic 3		Topic 3	71	Topic 3	75
Topic 5	72	Topic 22	23	Topic 12	100	Topic 18	51	Topic 18	63
Topic 18	72	Topic 5	11	Topic 5	79	Topic 5	37	Topic 5	59
Topic 4	62	Topic 3	10	Topic 22	51	Topic 4	23	Topic 4	51
Topic 12	4	Topic 4	10	Topic 4	44	Topic 12	2	Topic 22	17
Topic 22	0	Topic 12	0	Topic 18	44	Topic 22	0	Topic 12	0

Table T.6: Topics variables chosen by classifier as significant (Topics-Boruta-PS).

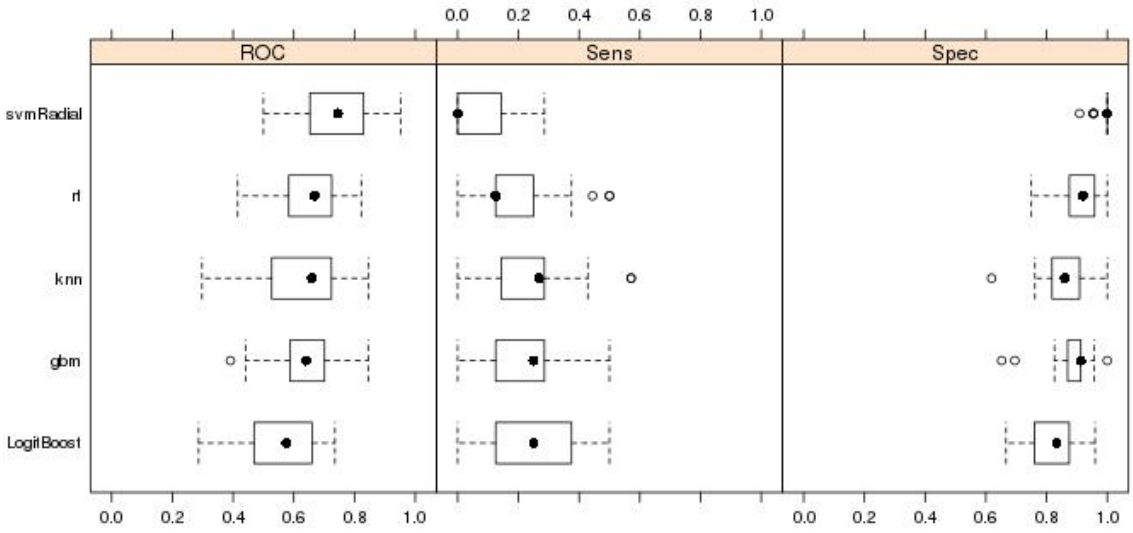


Figure T.6: ROC, Sensitivity and Specificity for classifiers (Topics-Boruta-PS).

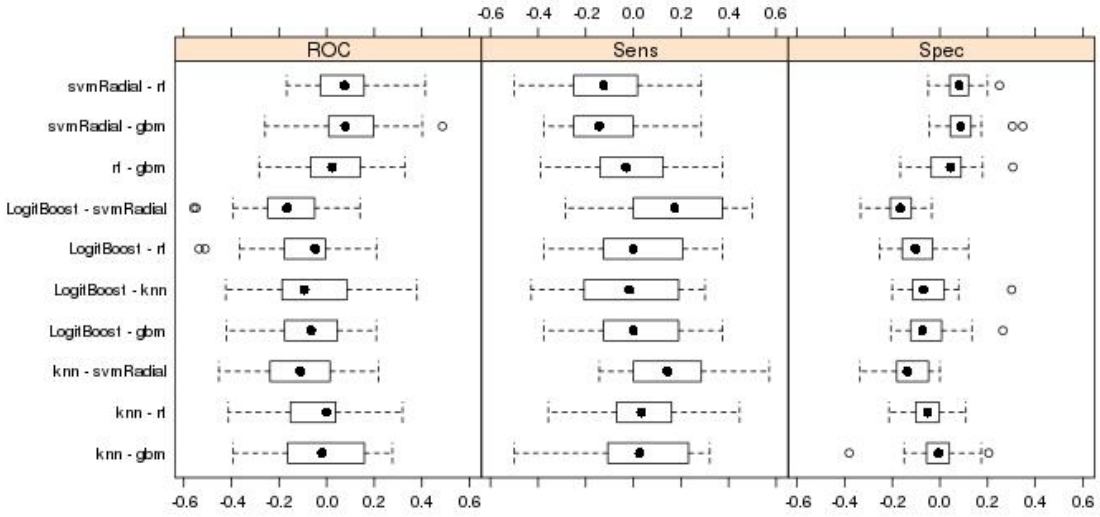


Figure T.7: Performance compared between classifiers (Topics-Boruta-PS).

Boruta feature selection on Topics– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.4	0.60	0.80	0.7	0.55 0.82	0.5	0.0033	0.75	0.66	0.70
SGB	0.48	0.80	0.68	0.74	0.59 0.85	0.5	0.0004	0.71	0.77	0.74
RF	0.4	0.75	0.65	0.70	0.53 0.83	0.5	0.008	0.68	0.72	0.70
kNN	0.36	0.68	0.68	0.68	0.53 0.80	0.5	0.007	0.68	0.68	0.68
SVM	0.24	0.60	0.64	0.62	0.47 0.75	0.5	0.05	0.62	0.61	0.62

Table T.7: Classification results (Topics-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Topic3	100	Topic24	100	Topic24	100	Topic24	100	Topic24	100
Topic24	88	Topic3	37	Topic3	58	Topic3	50	Topic3	56
Topic4	0	Topic4	34	Topic21	56	Topic21	31	Topic21	34
Topic21	0	Topic21	0	Topic4	0	Topic4	0	Topic4	0

Table T.8: Topics chosen by classifier as significant (Topics-Boruta-MP).

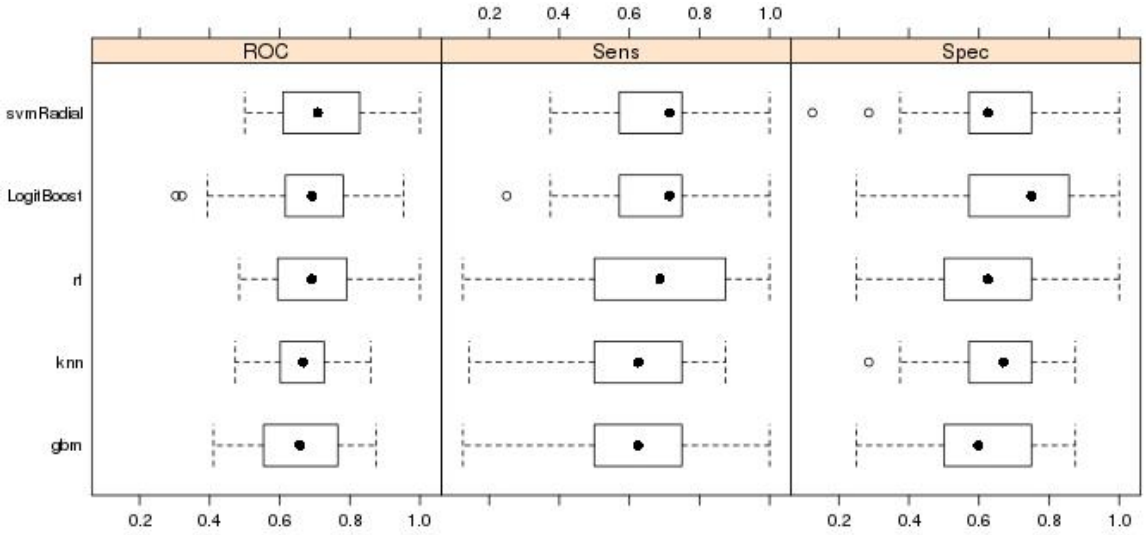


Figure T.7: ROC, Sensitivity and Specificity for classifiers (Topics-Boruta-MP).

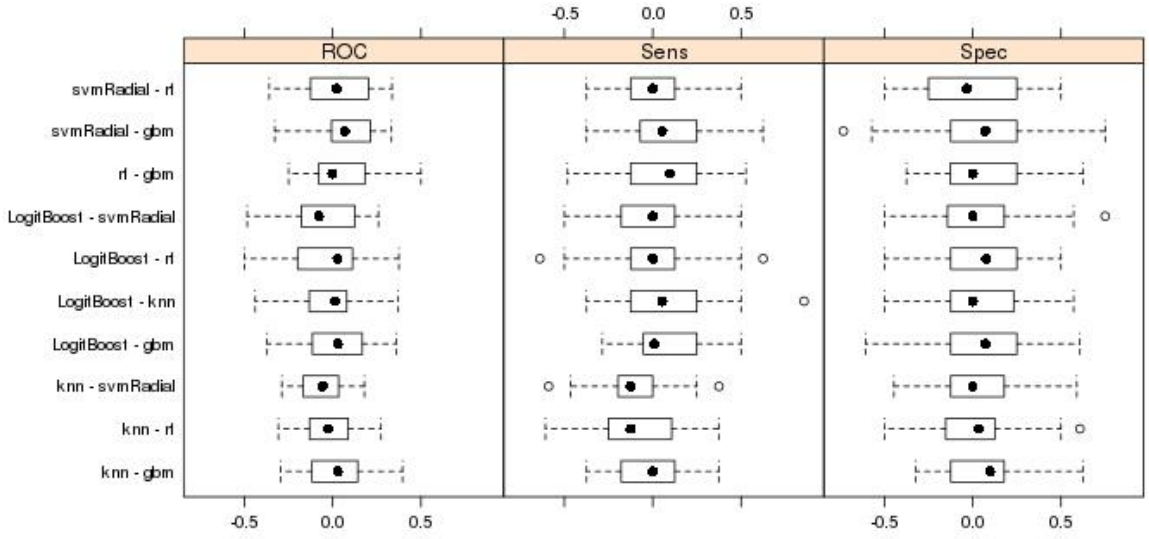


Figure T.8: Performance compared between classifiers (Topics-Boruta-MP).

Peer Set: Classification results for IG selected Topics

IG feature selection on Topics - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.15	0.12	0.98	0.77	0.69 0.83	0.75	0.32	0.71	0.77	0.55
SGB	0.12	0.16	0.93	0.74	0.64 0.82	0.75	0.64	0.44	0.77	0.54
RF	0.11	0.40	0.72	0.64	0.52 0.74	0.75	0.99	0.32	0.78	0.56
kNN	0.002 1	0.10	0.90	0.70	0.62 0.77	0.75	0.93	0.25	0.75	0.50
SVM	0.125 5	0.10	0.98	0.76	0.68 0.84	0.75	0.38	0.75	0.76	0.54

Table T.9: Classification results (Topics-IG-PS).

Only Topic 24 for IG Peer Set

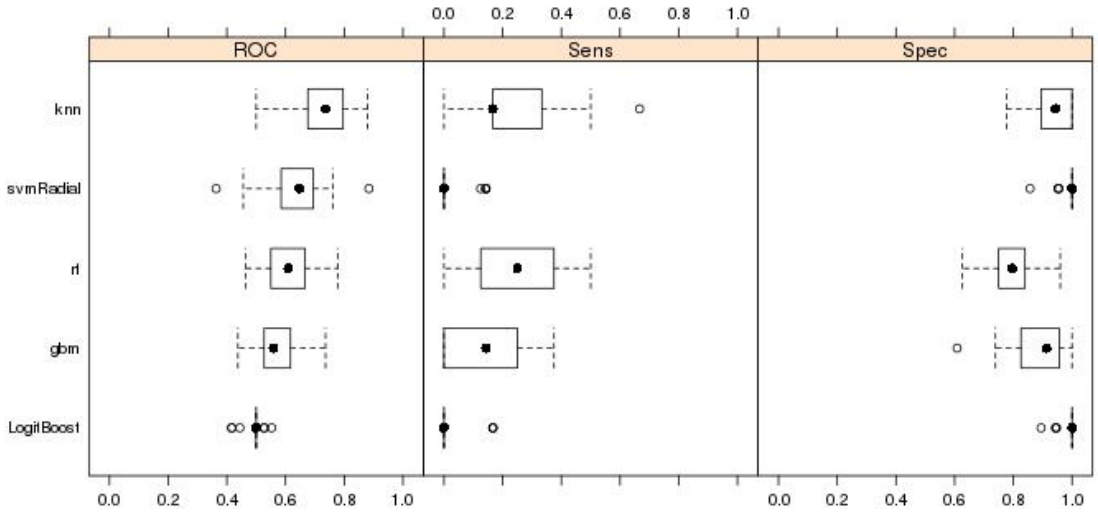


Figure T.9: ROC, Sensitivity and Specificity for classifiers (Topics-IG-PS).

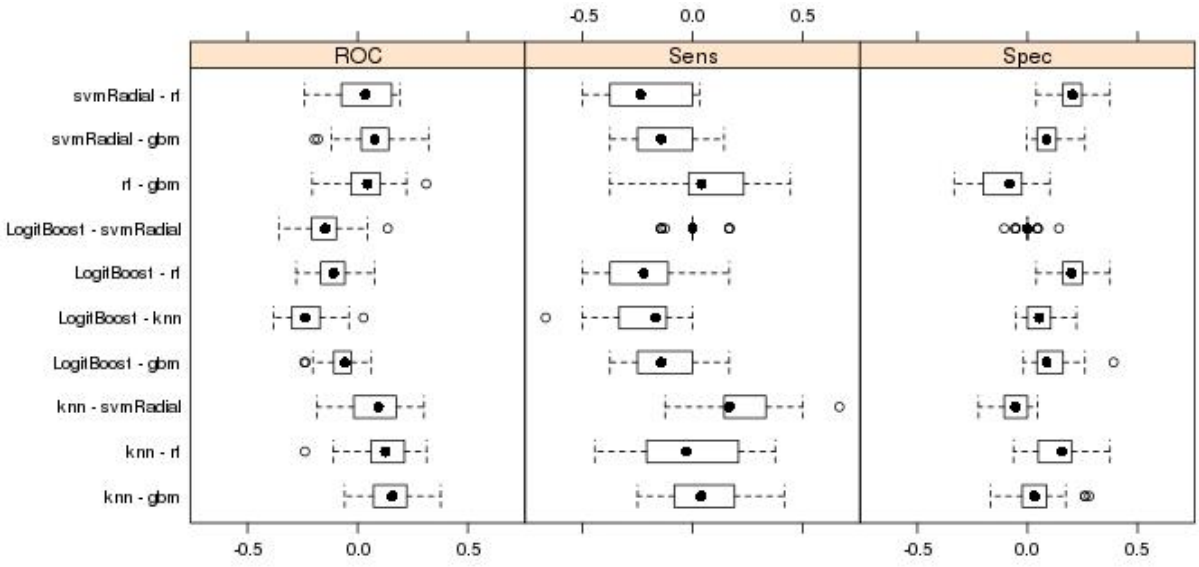


Figure T.10: Performance compared between classifiers (Topics-IG-PS).

IG feature selection on Topics– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.32	0.80	0.52	0.66	0.51 0.78	0.5	0.01	0.62	0.72	0.66
SGB	0.32	0.64	0.68	0.66	0.51 0.78	0.5	0.01	0.66	0.65	0.66
RF	0.25	0.50	0.75	0.62	0.45 0.77	0.5	0.07	0.66	0.60	0.62
kNN	0.2	0.57	0.62	0.60	0.48 0.70	0.5	0.04	0.60	0.59	0.60
SVM	0.3	0.56	0.73	0.65	0.61 0.76	0.5	0.01	0.68	0.62	0.35

Table T.10: Classification results (Topics-IG-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Topic 24	100	Topic.24	100	Topic.24	100	Topic.24	100	Topic.24	100
Topic.4	0	Topic.4	0	Topic.4	0	Topic.4	0	Topic.4	0

Table T.11: Topic chosen by classifier as significant (Topics-IG_MP).

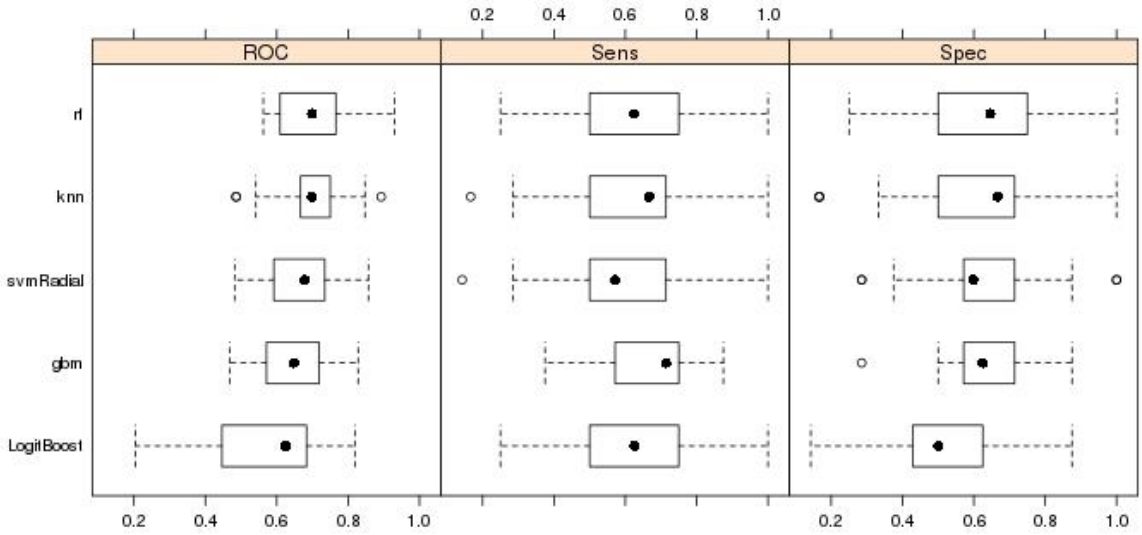


Figure T.11: ROC, Sensitivity and Specificity for classifiers (Topics-IG-MP).

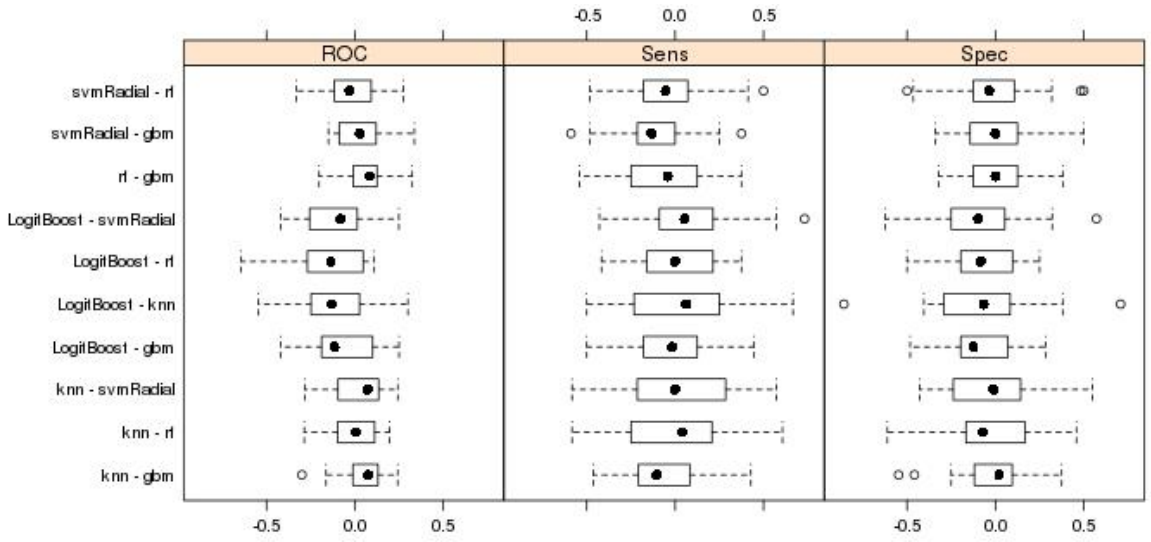


Figure T.12: Performance compared between classifiers (Topics-IG-MP).

APPENDIX U

Classification results for PCA selected concepts

PCA feature selection on Concepts- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.35	0.64	0.76	0.70	0.63 0.81	0.75	0.72	0.47	0.86	0.70
SGB	0.46	0.48	0.93	0.82	0.73 0.89	0.75	0.06	0.70	0.84	0.70
RF	0.22	0.16	1	0.79	0.69 0.86	0.75	0.21	1	0.78	0.58
kNN	0.43	0.46	0.92	0.80	0.72 0.87	0.75	0.08	0.66	0.84	0.69
SVM	0.42	0.48	0.90	0.80	0.71 0.87	0.75	0.14	0.63	0.84	0.69

Table U.1: Classification results (Concepts-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
ability.noun	100	events.noun	100	event.noun	100	assets.noun	100	event.noun	100
assets.noun	96	accounting. noun	90	assets.noun	90	based.verb	90	equity.noun	97
required.verb	91	ability.noun	85	provision. noun	90	provision. noun	86	ability.noun	92
event.noun	91	provision. noun	84	rate.noun	89. 1	required.verb	82	requires.verb	82
based.verb	90	payments. noun	80	ability.noun	83	accounting. noun	82	purchase. noun	78
accounting. noun	88	changes.noun	75	changes. noun	79	event.noun	81	required.verb	77
restrictions. noun	87	regulations. noun	71	including. verb	71	requiring.verb	80	include.verb	76
determined. verb	81	restrictions. noun	66	accounting. noun	70	ability.noun	80	requiring.verb	75
circumstance s.noun	77	rate.noun	65	events.noun	67	determined. verb	77	including.verb	75

Table U.2: Concepts chosen by classifier for (Concepts-PCA-PS).

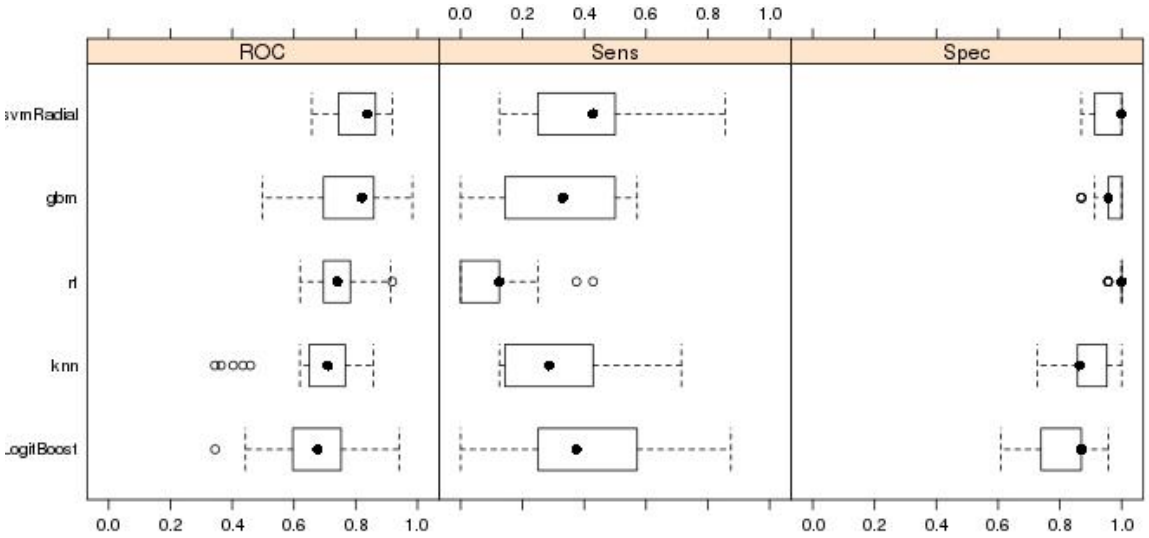


Figure U.1: ROC, Sensitivity and Specificity for classifiers (Concepts-PCA-PS).

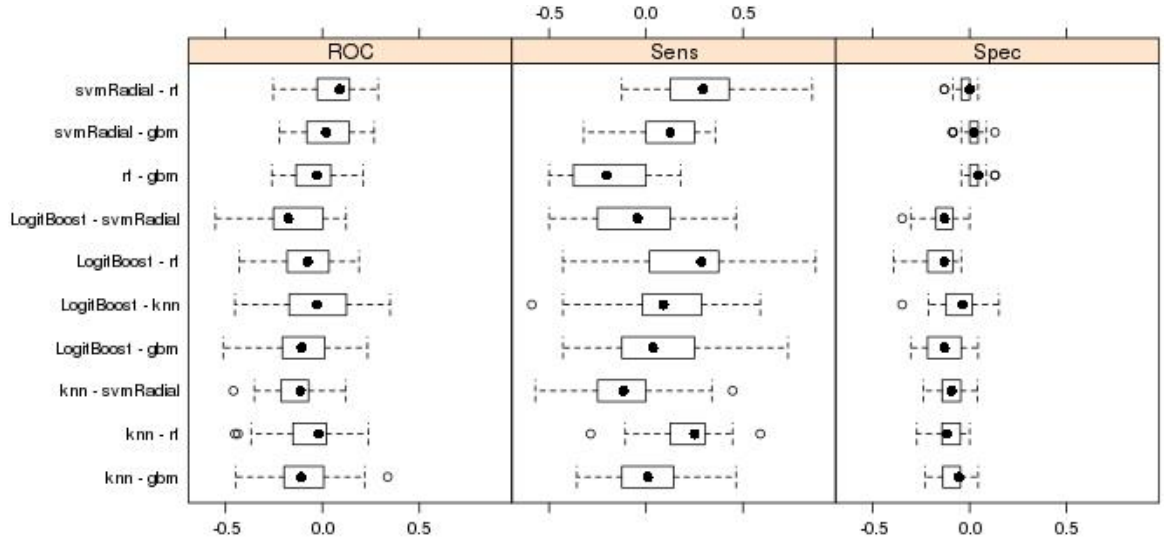


Figure U.2: Performance compared between classifiers (Concepts-PCA-PS).

PCA feature selection on Concepts- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.52	0.84	0.68	0.76	0.61 0.86	0.5	0.00015	0.72	0.80	0.76
SGB	0.52	0.92	0.60	0.76	0.61 0.86	0.5	0.00015	0.69	0.88	0.76
RF	0.48	0.72	0.76	0.74	0.59 0.85	0.5	0.0004	0.75	0.75	0.74
kNN	0.08	0.48	0.60	0.54	0.39 0.68	0.5	0.33	0.54	0.53	0.54
SVM	0.4	0.68	0.72	0.70	0.55 0.82	0.5	0.0033	0.70	0.69	0.70

Table U.3: Classification results (Concepts-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
agreements.noun	100	payments.noun	100	based.verb	100	ability.noun	100	agreements.noun	100
ability.noun	85	payment.noun	91	losses.noun	88	purchase.noun	99.4	purchase.noun	94
based.verb	81	compete.verb	84	required.verb	82	fee.noun	95	ability.noun	91
fee.noun	78	required.verb	77	liabilities.noun	73	agreements.noun	93	required.verb	84
purchase.noun	75	compensation.noun	75	ability.noun	71	required.verb	89	loss.noun	72
required.verb	74	Agreements.noun	71	loss.noun	67	based.verb	86	based.verb	70
restrictions.noun	68	action.noun	69	agreements.noun	60	payment.noun	81.7	fee.noun	69
loss.noun	66	ability.noun	52	purchase.noun	60.4	expense.noun	79	activities.noun	67.7
activities.noun	66	rates.noun	52	payment.noun	53	determined.verb	79	payment.noun	67
provide.verb	65	activities.noun	49	portion.noun	53	liability.noun	79	liability.noun	66.4

Table U.4: Concepts chosen by classifier as significant (Concepts-PCA-MP).

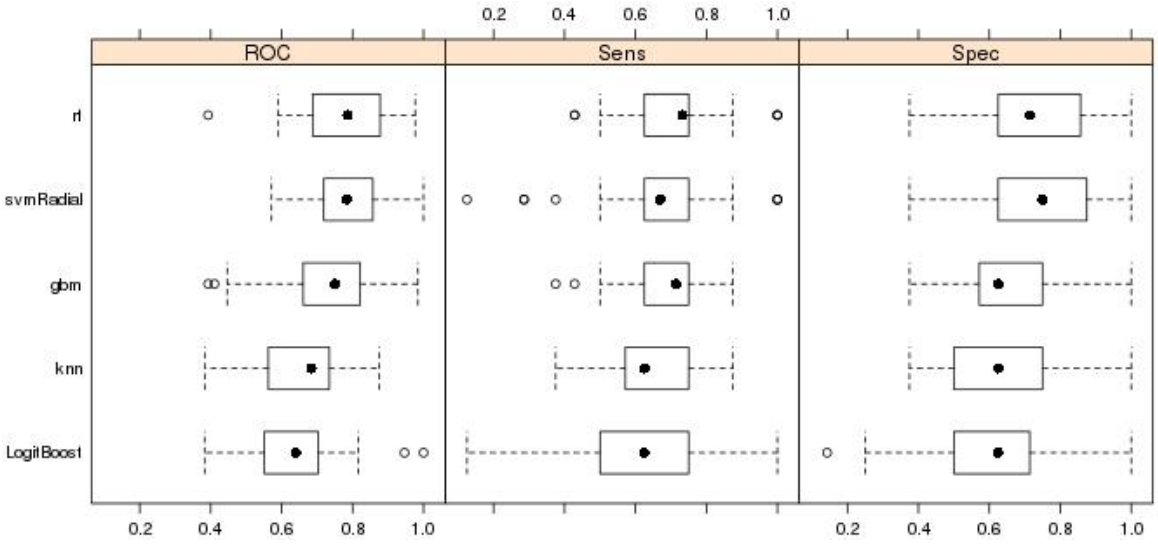


Figure U.3: ROC, Sensitivity and Specificity for classifiers (Concepts-PCA-MP).

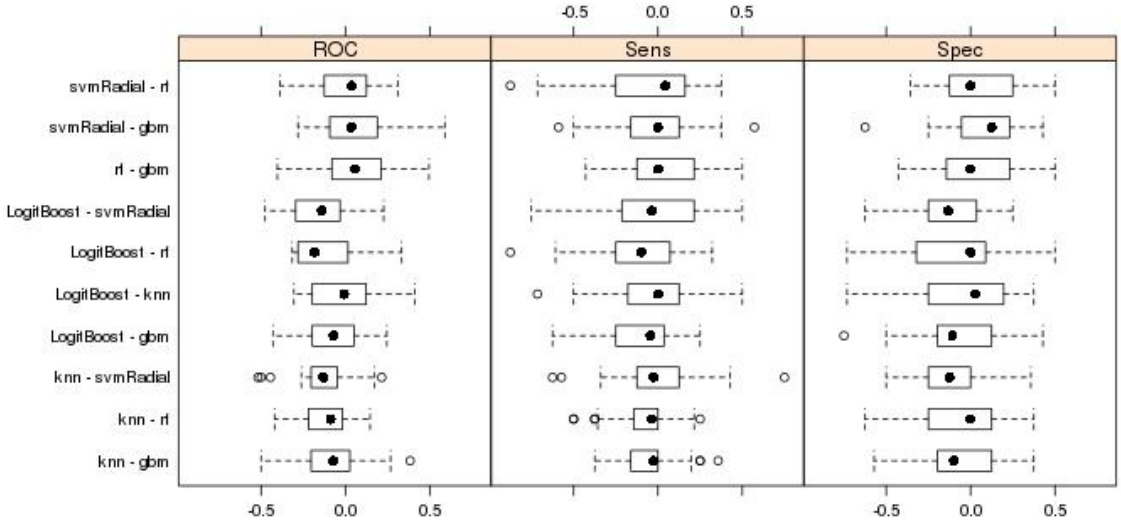


Figure U.4: Performance compared between classifiers (Concepts-PCA-MP).

Classification results for Boruta selected Concepts

Boruta feature selection on concepts- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.40	0.56	0.85	0.78	0.68 0.85	0.75	0.28	0.56	0.85	0.70
SGB	0.50	0.52	0.93	0.83	0.74	0.75	0.03	0.72	0.85	0.72
RF	0.57	0.52	0.97	0.86	0.77 0.92	0.75	0.005	0.86	0.86	0.74
kNN	0.008	0.05	1	0.76	0.68 0.83	0.75	0.391	1	0.76	0.52
SVM	0.59	0.60	0.94	0.85	0.78 0.91	0.75	0.002	0.78	0.87	0.77

Table U.5: Classification results (Concepts-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
acquisition.noun	100	accounts.noun	100	ended.verb	100	continued.verb	100	employee.noun	100
obtained.verb	96	expenses.noun	88	stockholder.noun	87.3	performed.verb	94	performed.verb	93
ended.verb	94.1	stockholder.noun	88	employee.noun	80	obtained.verb	91	ended.verb	89
acquired.verb	93	obtained.verb	72	arranged.verb	76	Acquisition noun	90	acquisition.noun	87
companies.noun	80	acquired.verb	70	accounts.noun	71.7	event.noun	88	obtained.verb	85
event.noun	76	results.noun	68	obtained.verb	68	ended.verb	83	relates.verb	85
purchase.noun	76	employee.noun	63	expenses.noun	64	put.verb	81	acquired.verb	83
performed.verb	73	ended.verb	59	losses.noun	64	acquired.verb	79	put.verb	80
improve.verb	71	acquisition.noun	59	acquisition.noun	63	entered.verb	78	event.noun	76
put.verb	70	event.noun	54	companies.noun	61	companies.noun	73	continued.verb	71

Table U.6: Concepts chosen by classifier as significant (Concepts-Boruta-PS).

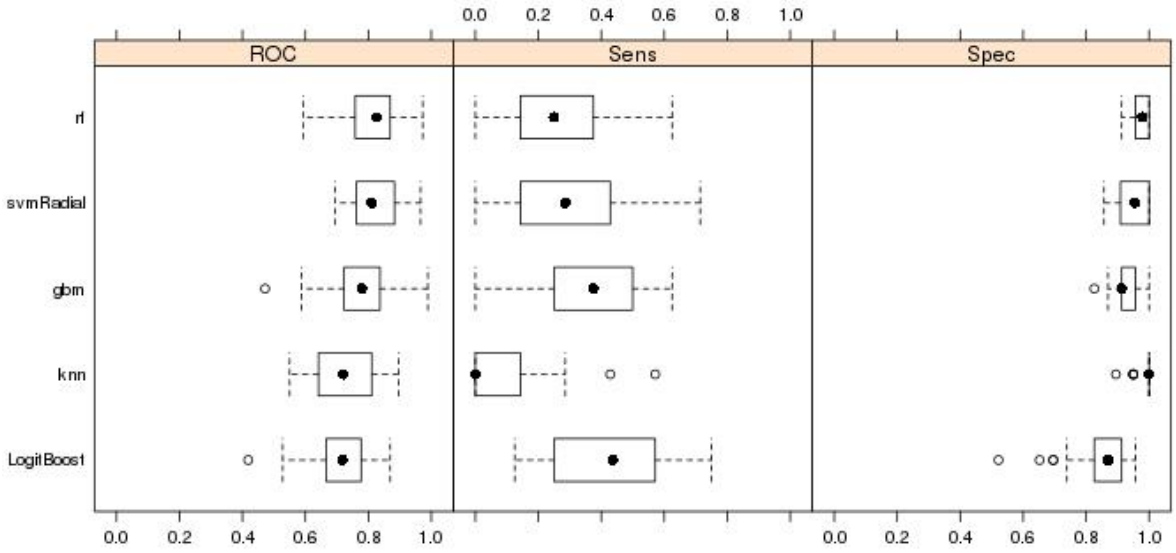


Figure U.6: ROC, Sensitivity and Specificity for classifiers (Concepts-Boruta-PS).

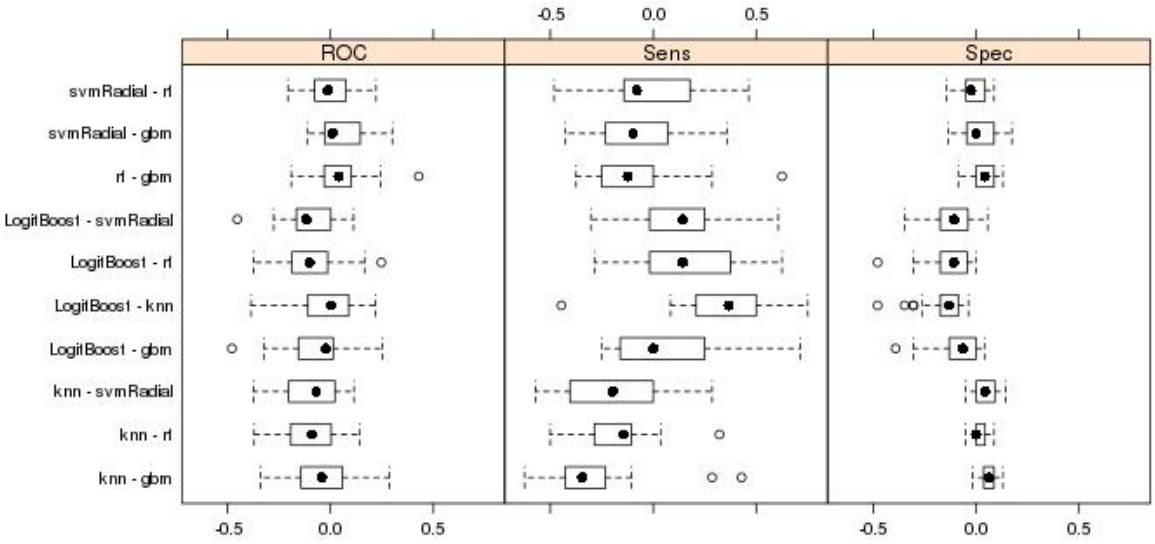


Figure U.7: Performance compared between classifiers (Concepts-Boruta-PS).

Boruta feature selection on Concepts – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.4	0.76	0.64	0.70	0.55 0.82	0.5	0.0033	0.67	0.72	0.70
SGB	0.48	0.80	0.68	0.74	0.59 0.85	0.5	0.0004	0.71	0.77	0.74
RF	0.4	0.72	0.68	0.70	0.55 0.82	0.5	0.0033	0.69	0.70	0.70
kNN	0.45	0.57	0.88	0.72	0.609 0.828	0.5	8.302e-05	0.83	0.67	0.72
SVM	0.46	0.70	0.76	0.73	0.60 0.83	0.5	0.000197	0.75	0.71	0.73

Table U.7: Classification results (Concepts-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
acquisition.noun	100	acquisition.noun	100	companies.noun	100	event.noun	100	acquisition.noun	100
companies.noun	99	employee.noun	96	expenses.noun	100	ended.verb	95	event.noun	98
event.noun	98	expenses.noun	94	ended.verb	95	employee.noun	80	companies.noun	95
ended.verb	89	event.noun	75	acquisition.noun	60	entered.verb	77	ended.verb	85
entered.verb	86	continued.verb	62	event.noun	55	companies.noun	72	entered.verb	84
acquired.verb	84	entered.verb	56	collect.verb	54	accounts.noun	70	accounts.noun	79
Accounts noun	80	consisting.verb	50	arranged.verb	54	credit.noun	69	acquired.verb	76
improve.verb	66	ended.verb	48	credit.noun	51	acquisition.noun	67	credit.noun	73
credit.noun	63	improve.verb	47	Accounts.noun	49	expenses.noun	62	expenses.noun	65
expenses.noun	63	accounts.noun	42	entered.verb	46	acquired.verb	60	improve.verb	60
continued.verb	59	acquired.verb	39	acquired.verb	43	continued.verb	57	continued.verb	44

Table U.8: Concepts chosen by classifier as significant (Concepts-Boruta-MP).

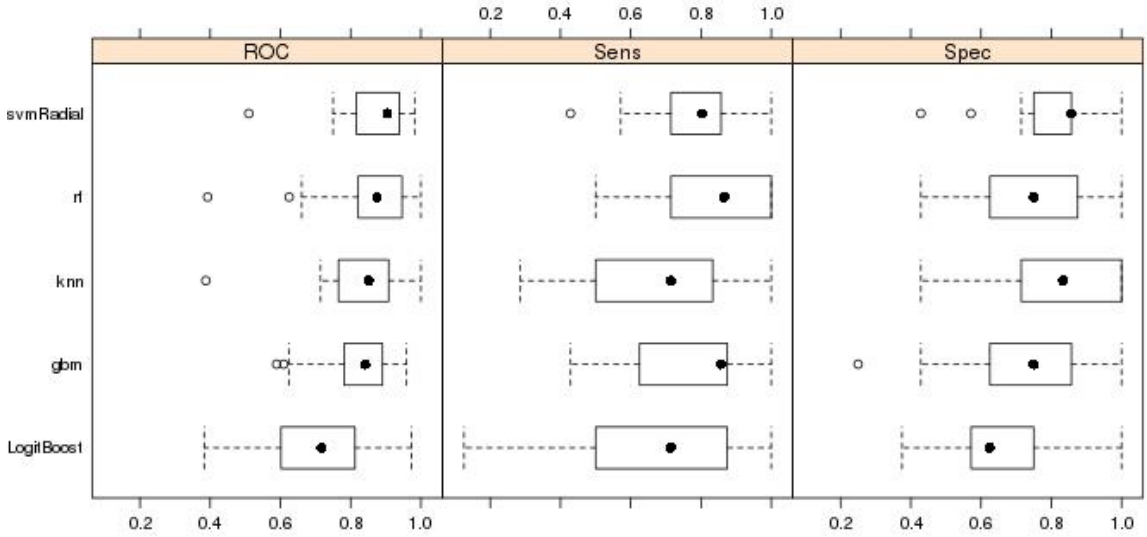


Figure U.7: ROC, Sensitivity and Specificity for classifiers (Concepts-Boruta-MP).

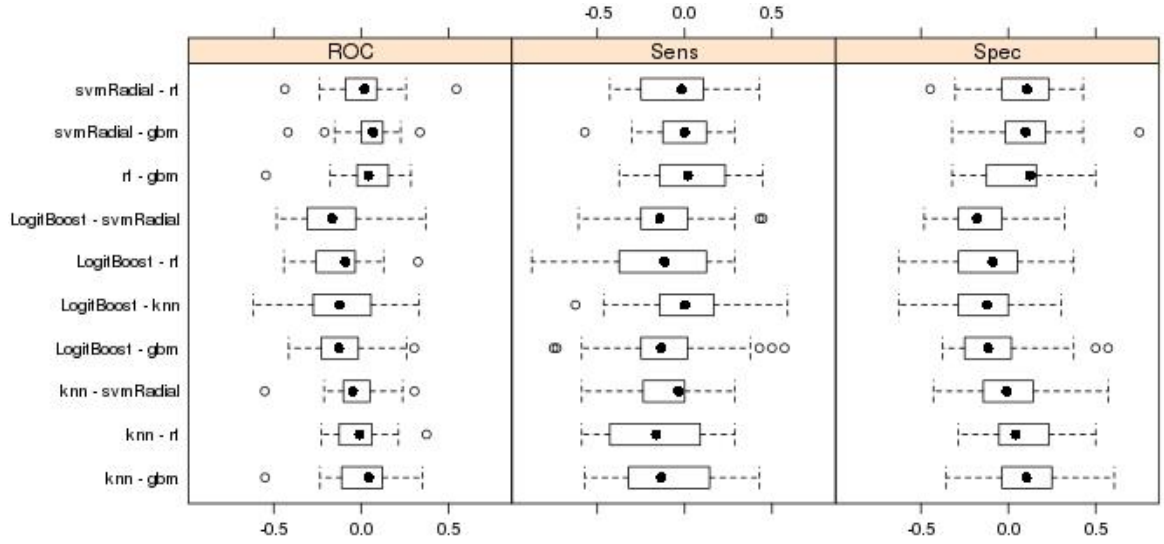


Figure U.8: Performance compared between classifiers (Concepts-Boruta-MP).

Classification results for IG selected concepts

IG feature selection on Concepts - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.49	0.56	0.90	0.82	0.73	0.75	0.06	0.66	0.86	0.73
SGB	0.55	0.60	0.92	0.84	0.75 0.90	0.75	0.02	0.71	0.87	0.76
RF	0.61	0.60	0.96	0.87	0.79 0.90	0.75	0.002	0.83	0.87	0.78
kNN	0.38	0.40	0.93	0.80	0.71 0.86	0.75	0.12	0.66	0.82	0.66
SVM	0.49	0.43	0.97	0.84	0.76 0.90	0.75	0.01	0.86	0.83	0.70

Table U.9: Classification results (Concepts-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
obtain.verb	100	acquisition.noun	100	acquisition.noun	100	obtain.verb	100	obtaining.verb	100
obtaining.verb	97	continued.verb	83	obtain.verb	63	obtaining.verb	98	obtain.verb	99
performed.verb	87	event.noun	79	mainly.Adv	57	acquisition.noun	91.2	obtained.verb	85
Acquisition.noun	81	performed.verb	76	event.noun	57	obtained.verb	87	acquisition.noun	81.5
obtained.verb	80	obtain.verb	69	obtaining.verb	54	event.noun	83	improving.verb	79
mainly.Adv	76	obtaining.verb	61	continued.verb	52	continued.verb	77	performed.verb	73
event.noun	72	mainly.Adv	59	performed.verb	50	improving.verb	75.6	offset.verb	68
failures.noun	70	offset.verb	48	offset.verb	50.1	mainly.Adv	72	event.noun	67.1
improving.verb	70	obtained.verb	44	stockholder.noun	49	offset.verb	71	mainly.Adv	65
continued.verb	69	stockholder.noun	41	obtained.verb	48	performed.verb	69	continued.verb	63

Table U.10: Concepts chosen by classifier as significant (Concepts-IG-PS).

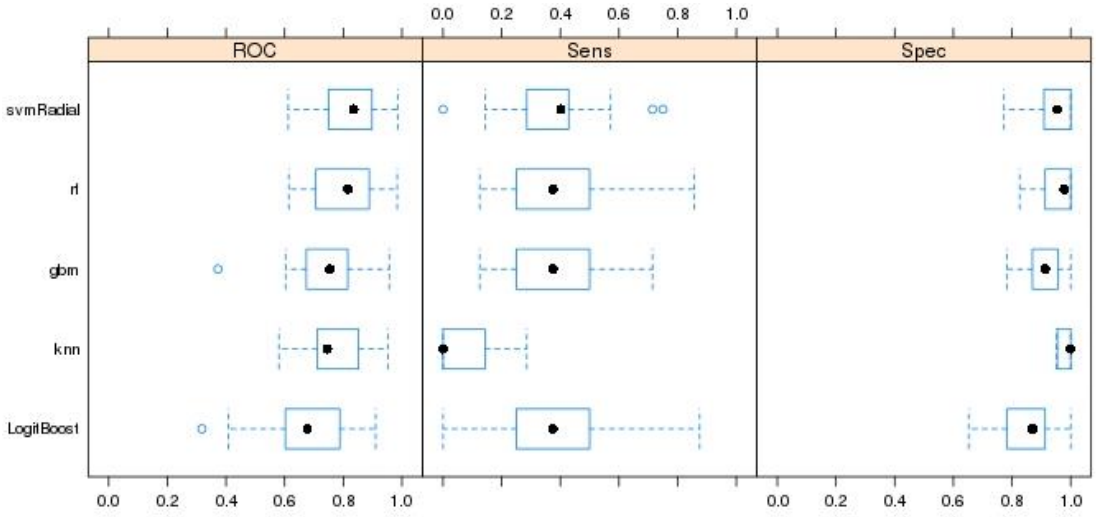


Figure U.9: ROC, Sensitivity and Specificity for classifiers (Concepts-IG-PS).

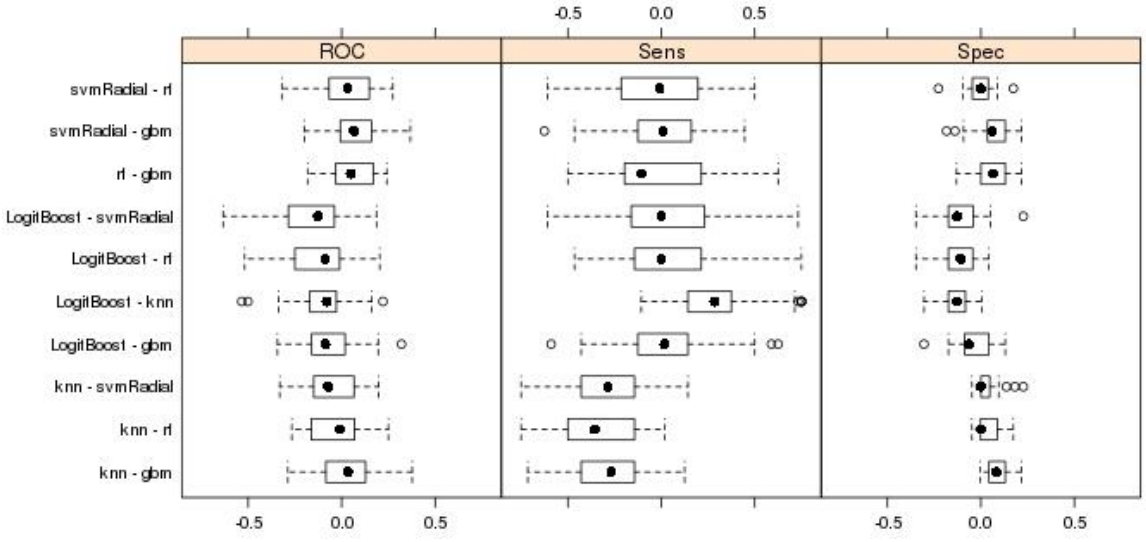


Figure U.10: Performance compared between classifiers (Concepts-IG-PS).

IG feature selection on Concepts – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.48	0.68	0.80	0.74	0.59 0.85	0.5	0.0004	0.77	0.71	0.74
SGB	0.56	0.80	0.76	0.78	0.64 0.88	0.5	4.511e-05	0.76	0.79	0.78
RF	0.44	0.68	0.76	0.72	0.57 0.83	0.5	0.001	0.73	0.70	0.72
kNN	0.3	0.66	0.63	0.65	0.51 0.76	0.5	0.01	0.64	0.65	0.65
SVM	0.36	0.70	0.66	0.68	0.55 0.79	0.5	0.003	0.67	0.68	0.68

Table U.10: Classification results (Concepts-IG-MP).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
obtaining. verb	100	acquisition. noun	100	acquisition. noun	100	acquisition. noun	100	acquisition. noun	100
obtained.verb	86	standards. noun	66	results.noun	87	obtaining.verb	91	obtaining.verb	83
acquisition. noun	84. 8	improve.verb	63	improve.verb	59	obtained.verb	82. 6	acquired.verb	78
acquired. verb	72	results.noun	62	acquired. verb	57	acquired.verb	66	obtained.verb	70
Improving .verb	64. 3	acquired.verb	56	obtaining. verb	57	division.noun	64	improving. verb	59
discounts. noun	58	obtaining.verb	468	obtained. verb	55	standards. noun	54	division.noun	54. 5
division.noun	51	division.noun	408	standards. noun	49	discounts. noun	54	improve.verb	50
standards. noun	51	discounts. noun	247	division. noun	44	Improving .verb	53	results.noun	43
precedent. noun	47	obtained.verb	231	discounts. noun	41	restoration. noun	50	discounts. noun	40
results.noun	47	improving. verb	23	improving. verb	34	improve.verb	43	precedent. noun	39

Table U.11: Concepts chosen by classifier as significant (Concepts-IG-MP).

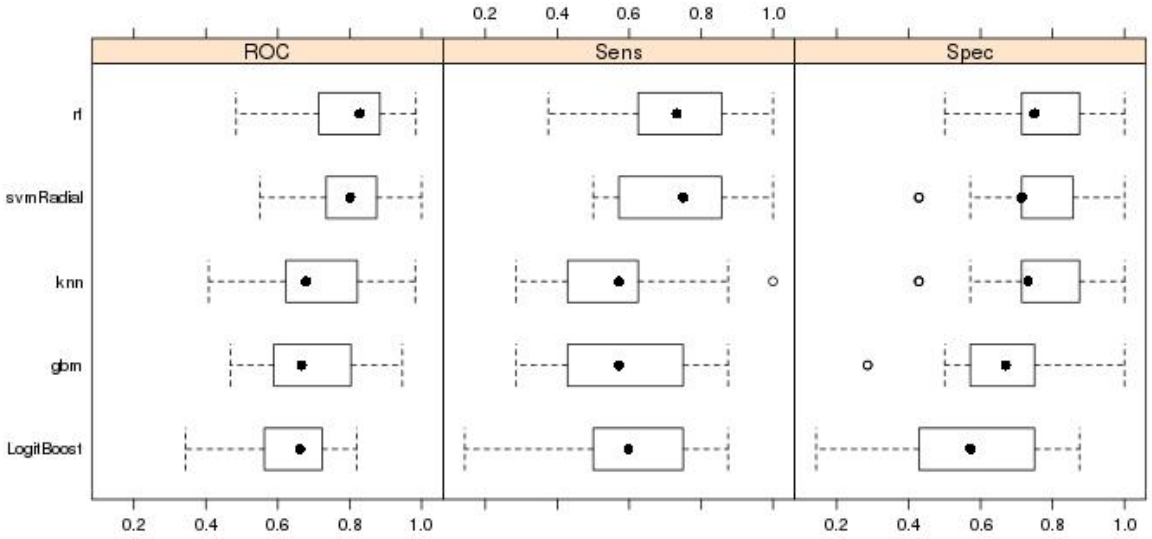


Figure U.11: ROC, Sensitivity and Specificity for classifiers (Concepts-IG-MP).

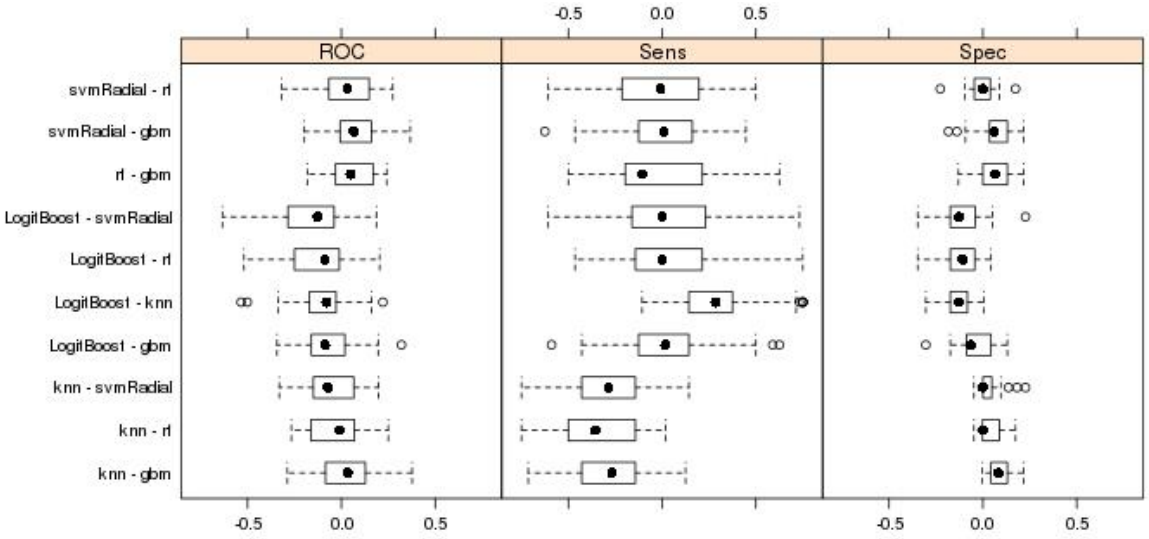


Figure U.12: Performance compared between classifiers (Concepts-IG-MP).

Classification results for LSA chosen concepts

LSA chosen concepts: Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.92	1	0.96	0.97	0.91 0.99	0.75	2.18e-09	0.89	1	0.98
SGB	0.92	0.96	0.97	0.97	0.91 0.99	0.75	2.18e-09	0.92	0.98	0.96
RF	0.94	0.92	1	0.98	0.93 0.99	0.75	1.94e-10	1	0.94	0.96
kNN	0.91	0.88	1	0.97	0.91 0.99	0.7	2.18e-09	1	0.96	0.94
SVM	0.89	0.92	0.97	0.96	0.90 0.98	0.75	1.822e-08	0.92	0.97	0.94

Table U.12: Classification results (LSAConcepts-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
result	100	result	100	result	100	result	100	result	100
operating	85	operating	10	operations	43	operations	85	operations	85
operations	84	may	9	operating	23	operating	76	operating	82
management	78	certain	4	managemen	18	management	74	management	79
amount	71	management	4	net	6	net	64	net	68
net	70	operations	4	amount	6	may	63	amount	67
may	68	costs	3	may	5	certain	62	may	65
certain	67	net	2	certain	4	amount	61	certain	63
cash	57	cash	2	costs	3	cash	59	cash	58
interest	55	required	2	cash	2	interest	52	interest	56
total	52	amount	1	total	2	costs	49	total	54
revenue	49	production	1	increased	2	increased	46	increased	53
costs	49	higher	1	revenue	1	required	44	revenue	45

Table U.13: LSA Concepts chosen by classifier as significant (LSAConcepts-PS).

LSA chosen concepts: Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.92	1	0.92	0.96	0.86 0.99	0.5	1.133e-12	0.92	1	0.96
SGB	0.96	0.96	1	0.98	0.89 0.99	0.5	4.53e-14	1	0.96	0.98
RF	1	1	1	1	0.92 1	0.5	8.88e-16	1	1	1
kNN	0.76	1	0.76	0.88	0.75 0.95	0.5	1.622e-08	0.80	1	0.88
SVM	0.72	0.84	0.88	0.86	0.73 0.94	0.5	1.04e-07	0.87	0.84	0.86

Table U.14: Classification results (LSAConcepts-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
value	100	value	100	value	100	value	100	value	100
result	88	result	2	result	41	costs	84	result	86
costs	85	costs	1	costs	33	result	80	costs	82
may	67	management	0.98	may	17	may	73	may	77
required	58	interest	0.55	required	12	certain	64	cash	63
cash	57	control	0.44	certain	52	amount	62	certain	61
amount	56	total	0.29	amount	57	interest	59	amount	61
total	55	amount	0.25	interest	34	required	57	required	58
certain	55	may	0.22	cash	30	cash	56	interest	57
revenue	54	operations	0.22	total	22	total	54	total	56
interest	53	adverse	0.21	operations	13	revenue	48	revenue	52
operations	41	required	0.21	management	10	securities	41	securities	51

Table U.15: LSA Concepts chosen by classifier as significant (LSAConcepts-MP).

APPENDIX V

Classification results for PCA selected Custom Dict.

PCA feature selection on Custom Dict.– Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.19	0.44	0.76	0.68	0.58 0.77	0.75	0.95	0.37	0.80	0.60
SGB	0.18	0.36	0.81	0.70	0.60 0.78	0.75	0.89	0.39	0.79	0.58
RF	0.32	0.32	0.94	0.79	0.69 0.86	0.75	0.21	0.66	0.80	0.63
kNN	0.10	0.16	0.92	0.73	0.63 0.71	0.75	0.72	0.40	0.76	0.54
SVM	0	0	1	0.75	0.65 0.83	0.75	0.55	0	0.75	0.50

Table V.1: Classification results (Custom Dict.-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Positivity Freq	100	positivity Freq	100	positivity Freq	100	Positivity Freq	100	Positivity Freq	100
Constraining	53	Constraining	48	Constraining	29	Forward Looking_Freq	51.44	Constraining	37
Uncert1_Freq	44	Litiguous	33	Uncert1 Freq	29	Constraining	51.07	Uncert1 Freq	31
Litiguous	38	Forward Looking_Freq	14	Modal.Weak	26	Litiguous	46.83	Modal.Weak	29
Modal.Weak	37	Negativity Freq	14	ForwardLooking_Freq	23	Modal.Weak	37.39	Forward Looking_Freq	28
Forward Looking_Freq	37	Uncert1_Freq	7	Litiguous	4	Uncert1_Freq	25.55	Litiguous	15
Negativity Freq	0	Modal.Weak	0	negativity Freq	0	Negativity Freq	0	Negativity Freq	0

Table V.2: Custom Dict. chosen by classifier for (Custom Dict.-PCA-PS).

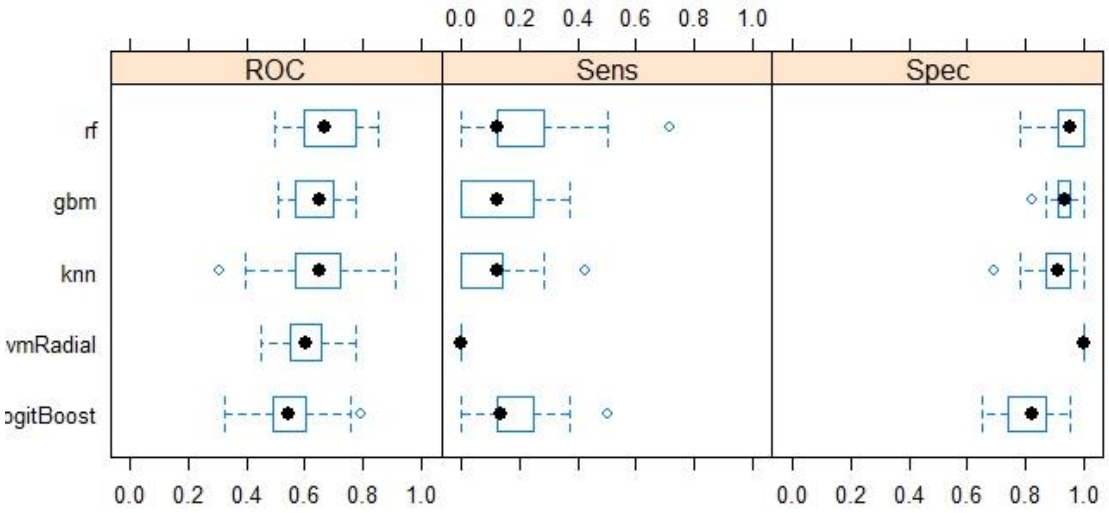


Figure V.1: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-PCA-PS).

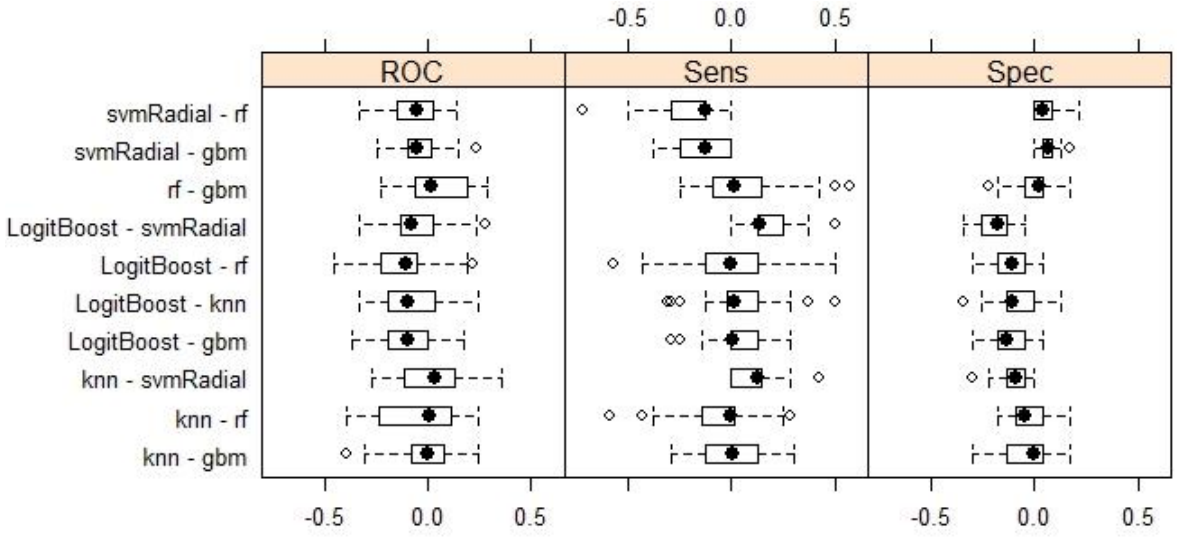


Figure V.2: Performance compared between classifiers (Custom Dict.-PCA-PS).

PCA feature selection on Custom Dict.– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.12	0.60	0.52	0.56	0.41 0.70	0.5	0.23	0.55	0.56	0.56
SGB	0.17	0.62	0.54	0.58	0.46 0.70	0.5	0.09	0.57	0.59	0.58
RF	0.32	0.76	0.56	0.66	0.51 0.78	0.5	0.01	0.63	0.70	0.66
kNN	0.14	0.45	0.68	0.57	0.44 0.68	0.5	0.14	0.59	0.55	0.57
SVM	0.23	0.70	0.53	0.61	0.48 0.73	0.5	0.04	0.60	0.53	0.61

Table V.3: Classification results (Custom Dict.-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Modal.Weak	100	Litiguous	100	Litiguous	100	Uncert1_Freq	100	positivity_Freq	100
Forward Looking_Freq	91	Forward Looking_Freq	69	Positivity Freq	83	Litiguous	84	Uncert1_Freq	95
Litiguous	73	Negativity Freq	62	Constraining	34	Modal.Weak	73	Litiguous	88
Uncert1_Freq	68	Positivity Freq	59	Modal.Weak	32	Negativity Freq	72	Constraining	87
positivity_Freq	64	Uncert1_Freq	37	Forward Looking_Freq	30	Positivity Freq	47	Negativity Freq	61
negativity req	29	Modal.Weak	22	Uncert1_Freq	10	Constraining	23	Modal.Weak	57
Constraining	0	Constraining	0	Negativity Freq	0	Forward Looking_Freq	0	Forward Looking_Freq	0

Table V.4: Custom Dict. chosen by classifier as significant (Custom Dict.-PCA-MP).

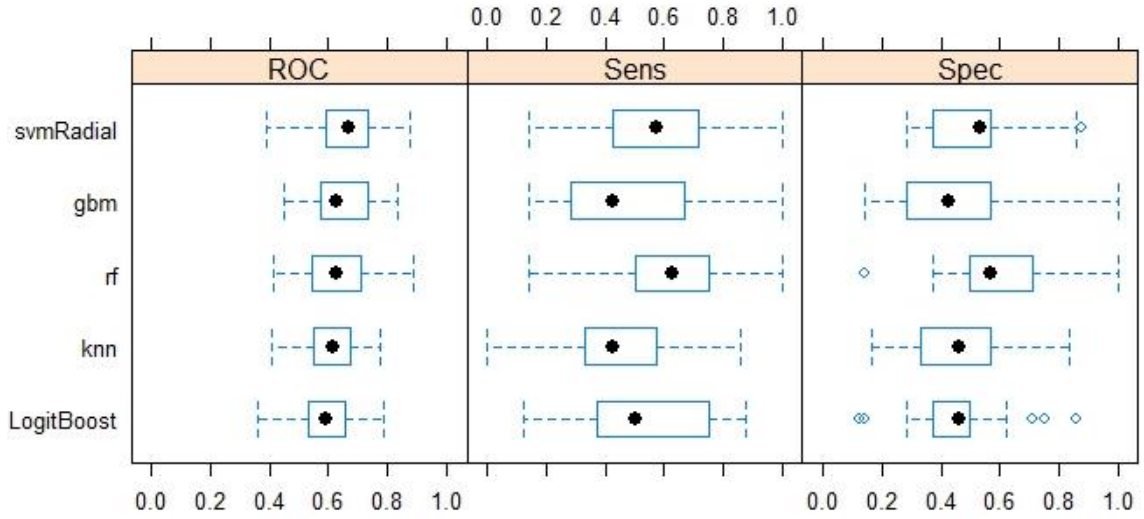


Figure V.3: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-PCA-MP).

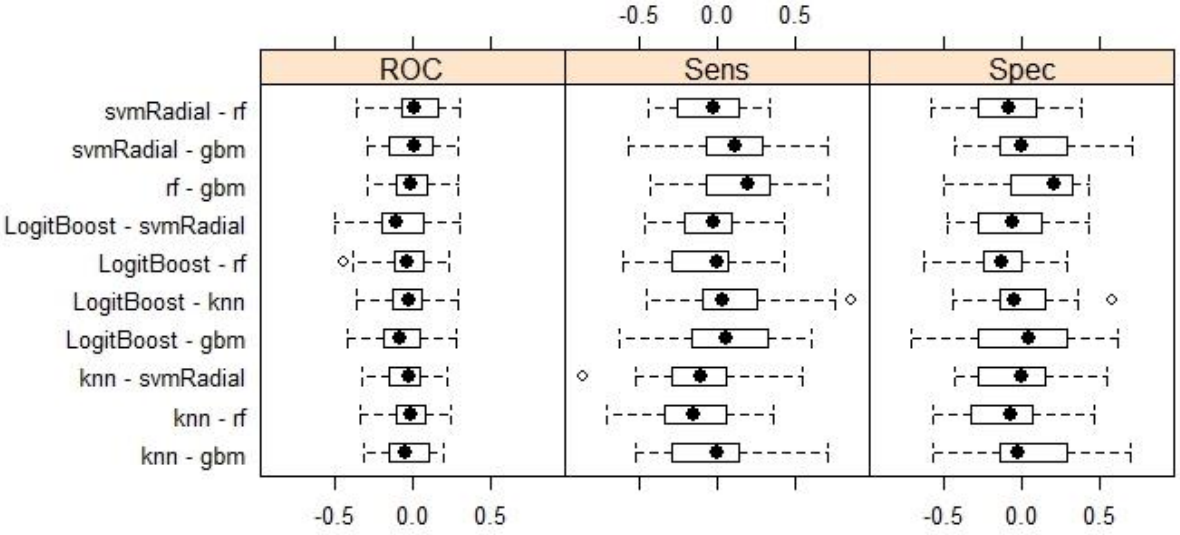


Figure V.4: Performance compared between classifiers (Custom Dict.-PCA-MP).

Classification results for Boruta selected Custom Dict.

Boruta feature selection on Custom Dict.– Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.25	0.40	0.84	0.7	0.63 0.81	0.75	0.72	0.45	0.81	0.62
SGB	0.36	0.36	0.84	0.72	0.62 0.80	0.75	0.79	0.42	0.80	0.60
RF	0.36	0.32	0.97	0.81	0.72 0.88	0.75	0.09	0.80	0.81	0.64
kNN	0.1	0.20	0.90	0.72	0.63 0.80	0.75	0.77	0.40	0.77	0.55
SVM	0.24	0.20	0.98	0.78	0.71 0.85	0.75	0.19	0.77	0.78	0.59

Table V.5: Classification results (Custom Dict.-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Positivity Freq	100	Positivity Freq	100	Positivity Freq	100	Positivity Freq	100	Positivity Freq	100
Forward Looking_Freq	70	Negativity Freq	47	Negativity Freq	72	Forward Looking_Freq	67	Constraining	46
Uncert1_Freq	59	Modal.Weak	42	Uncert1_Freq	57	Modal.Weak	47	Modal.Strong	36
Modal.Strong	54	Litiguous	36	Modal.Weak	51	Uncert1_Freq	43	Forward Looking_Freq	35
Constraining	44	Uncert1_Freq	19	Constraining	41	Constraining	42	Litiguous	32
Modal.Weak	41	Constraining	18	Forward Looking_Freq	41	Litiguous	34	Uncert1_Freq	31
Litiguous	39	Forward Looking_Freq	0	Risk	27	Modal.Strong	33	Modal.Weak	28
Risk	35			Litiguous	21	Negativity Freq	6	Risk	19
Negativity Freq	0			Modal.Strong	0	Risk	0	Negativity Freq	0

Table V.6: Custom Dict. chosen by classifier as significant (Custom Dict.-Boruta-PS).

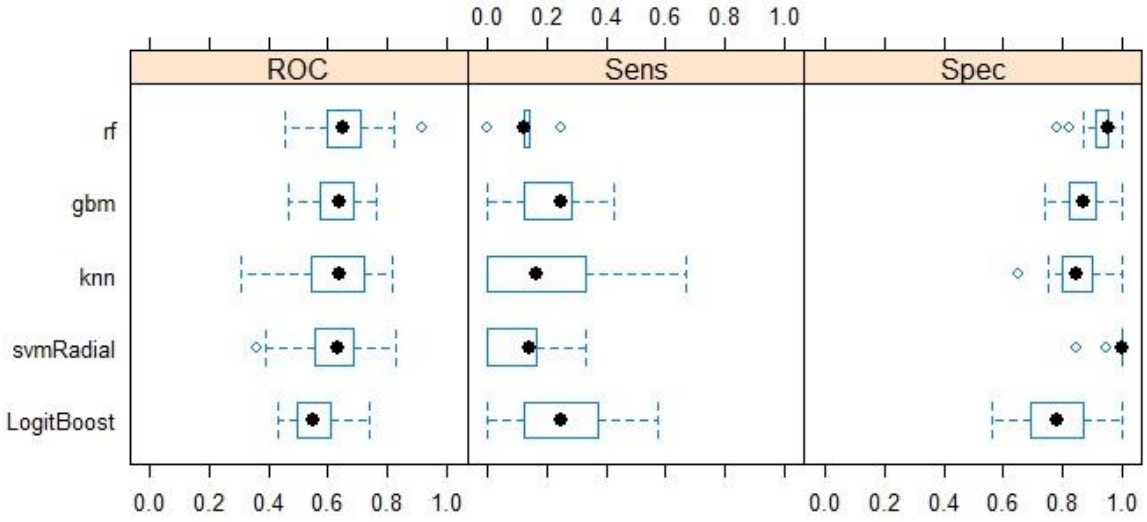


Figure V.5: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-Boruta-PS).

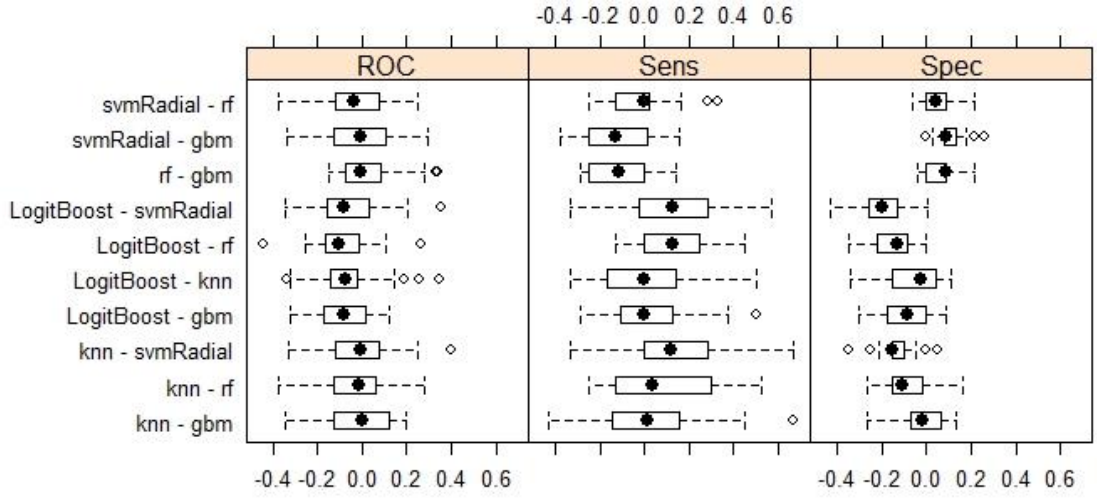


Figure V.6: Performance compared between classifiers (Custom Dict.-Boruta-PS).

Boruta feature selection on Custom Dict.– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.28	0.80	0.48	0.64	0.49 0.77	0.5	0.03	0.60	0.70	0.64
SGB	0.28	0.72	0.56	0.64	0.49 0.77	0.5	0.03	0.62	0.66	0.64
RF	0.24	0.64	0.60	0.62	0.47 0.75	0.5	0.05	0.61	0.62	0.62
kNN	0.7	0.86	0.53	0.7	0.56 0.81	0.5	0.001	0.65	0.80	0.70
SVM	0.6	0.84	0.76	0.80	0.66 0.89	0.5	1.193e-05	0.77	0.82	0.80

Table V.7: Classification results (Custom Dict.-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
Passive	100	Passive	100	Passive	100	Passive	100	Passive	100
Modal.Strong	88	Constraining	67	Risk	65	Risk	57	positivity_Freq	75
Risk	46	Risk	57	Constraining	59	Modal.Strong	52	Modal.Weak	67
Uncert1_Freq	41	negativity_Freq	45	positivity_Freq	57	positivity_Freq	49	Modal.Strong	53
Uncert2_Freq	41	Modal.Weak	41	Modal.Weak	39	Constraining	25.8	ForwardLooking_Freq	53
positivity_Freq	36	Modal.Strong	36.6	negativity_Freq	35	Uncert1_Freq	23	Uncert1_Freq	49
Forward Looking_Freq	30	Forward Looking_Freq	34	Forward Looking_Freq	34	Uncert2_Freq	23	Negativity Freq	0
Negativity Freq	18	Uncert2_Freq	21	Modal.Strong	31	Modal.Weak	20		

Table V.8: Custom Dict. chosen by classifier as significant (Custom Dict.-Boruta-MP).

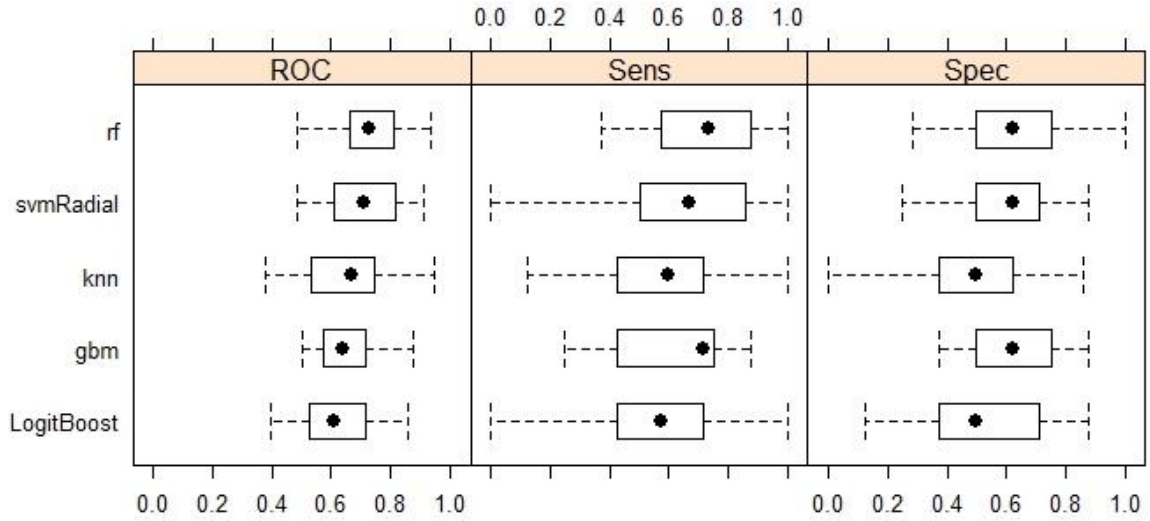


Figure V.7: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-Boruta-MP).

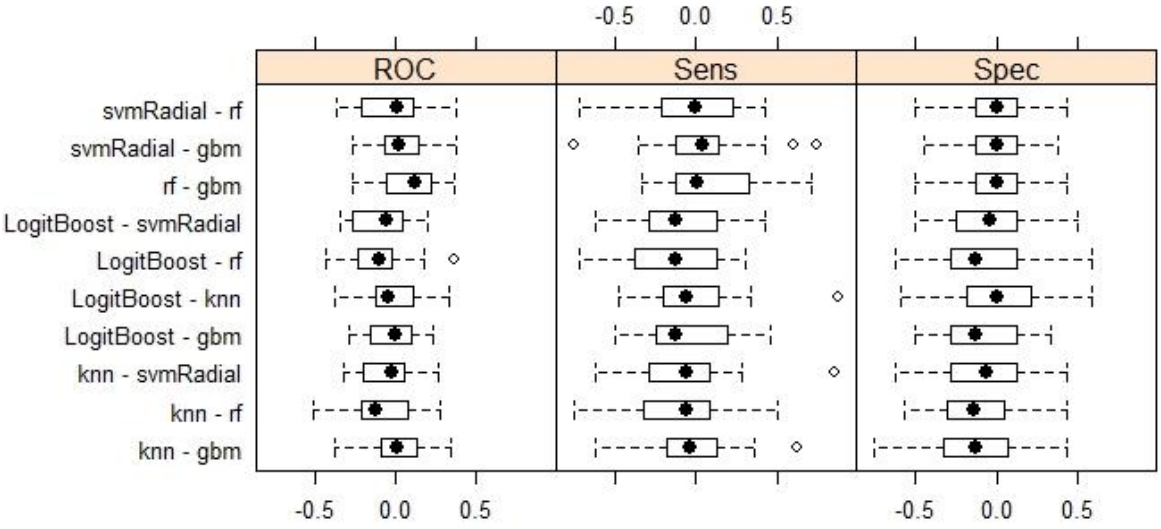


Figure V.8: Performance compared between classifiers (Custom Dict.-Boruta-MP).

Classification results for IG selected Custom Dict.

IG feature selection on Custom Dict.– Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.21	0.43	0.79	0.70	0.61 0.78	0.75	0.91	0.40	0.80	0.61
SGB	0.19	0.25	0.90	0.74	0.66 0.81	0.75	0.62	0.47	0.78	0.58
RF	0.22	0.28	0.90	0.75	0.65 0.83	0.75	0.55	0.50	0.79	0.59
kNN	0.11	0.11	0.97	0.76	0.68 0.82	0.75	0.46	0.57	0.77	0.54
SVM	0.09	0.10	0.96	0.75	0.67 0.81	0.75	0.54	0.50	0.76	0.53

Table V.9: Classification results (Custom Dict.-IG-PS).

Peer Set – Variable Importance									
Logistic Regression		***Stochastic Grad. Boosting***		***Random Forest***		***kNN***		***Support Vector Machine***	
Positivity Freq	100	Positivity Freq	100	Positivity Freq	100	Positivity Freq	100	Positivity Freq	100
Forward LookingFreq	43	Negativity Freq	57	Uncert1_Freq	48	Forward Looking_Freq	51	Forward LookingFreq	54
Uncert1_Freq	35	ForwardLooking_Freq	45.6	Negativity Freq	25	Uncert1_Freq	29	Uncert1_Freq	48
Negativity Freq	0	Uncert1_Freq		ForwardLooking_Freq	0	Negativity Freq	0	Negativity Freq	0

Table V.10: Custom Dict. chosen by classifier as significant (Custom Dict.-IG-PS).

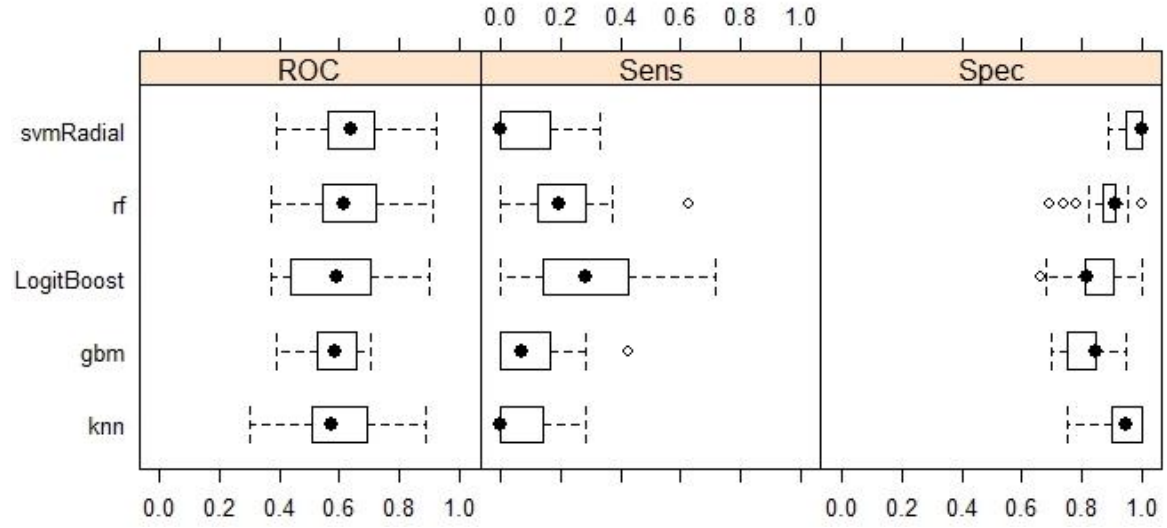


Figure V.9: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-IG-PS).

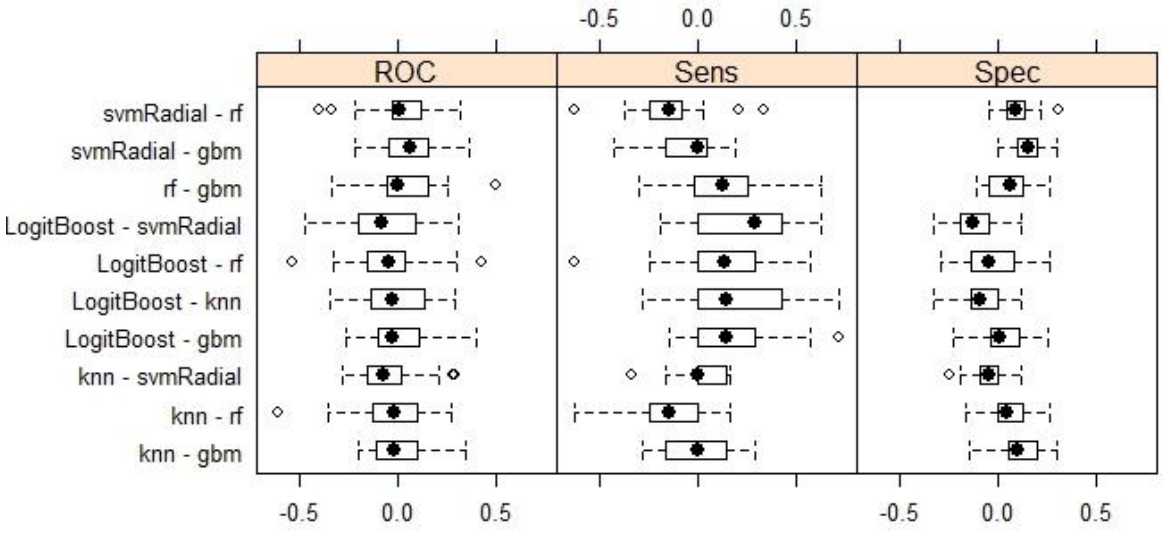


Figure V.10 : Performance compared between classifiers (Custom Dict.-IG-PS).

IG feature selection on Custom Dict.– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	0.24	0.56	0.68	0.62	0.47 0.75	0.5	0.05	0.63	0.60	0.62
SGB	0.24	0.52	0.72	0.62	0.47 0.75	0.5	0.05	0.65	0.60	0.62
RF	0.16	0.60	0.56	0.58	0.43 0.71	0.5	0.16	0.57	0.58	0.58
kNN	0.22	0.54	0.68	0.61	0.49 0.72	0.5	0.03	0.63	0.60	0.61
SVM	0.2	0.66	0.53	0.60	0.46 0.72	0.5	0.07	0.58	0.61	0.60

Table V.11: Classification results (Custom Dict.-IG-MP).

Peer Set – Variable Importance									
Logistic Regression		***Stochastic Grad. Boosting***		***Random Forest***		***kNN***		***Support Vector Machine***	
Forward Looking Freq	100	Negativity Freq	100	Uncert1 Freq	100	Negativity Freq	100	Positivity Freq	100
Uncert1 Freq	66	Forward Looking Freq	89	Negativity Freq	62	Uncert1 Freq	78	Forward Looking_ Freq	59
Negativity Freq	21	Positivity Freq	73	Positivity Freq	62	Forward Looking Freq	77	Uncert1 Freq	27
Positivity Freq	0	Uncert1 Freq	0	Forward Looking_Freq	0	Positivity Freq	0	Negativity Freq	0

Table V.12: Custom Dict. chosen by classifier as significant (Custom Dict.-IG_MP).

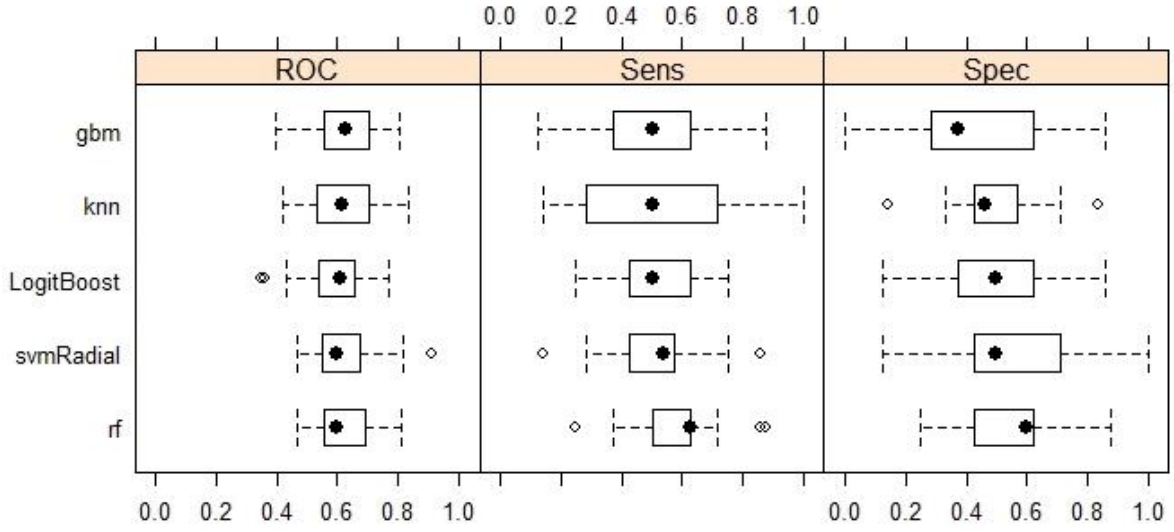


Figure V.11: ROC, Sensitivity and Specificity for classifiers (Custom Dict.-IG-MP).

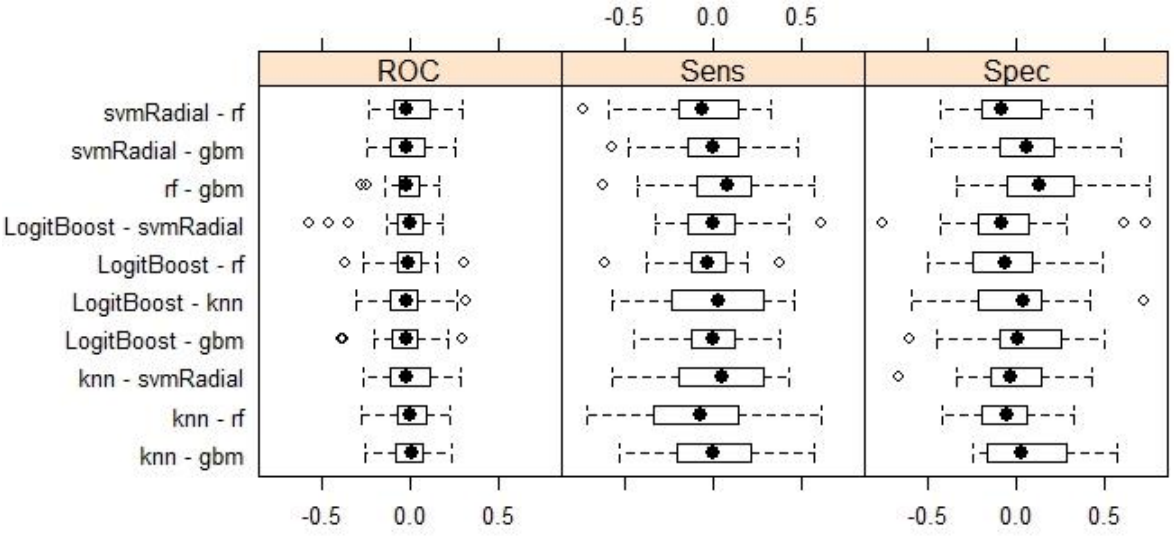


Figure V.12: Performance compared between classifiers (Custom Dict.-IG-MP).

APPENDIX W

Classification results for PCA selected Keywords

PCA feature selection on Keywords- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.961	1	3.355e-13	1	1	1
kNN	1	1	1	1	0.961	1	3.355e-13	1	1	1
SVM	1	1	1	1	0.961	1	< 2.2e-16	1	1	1

Table W.1: Classification results (Keywords-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
also	100	also	100	also	100	also	100	also	100
capital	99	capital	67	capital	78	capital	99	capital	98
result	91	result	0	result	55	result	90	result	91
management	76	offset	0	management	16	managemen	75	management	75
rate	74	system	0	stockholder	12	tax	71	net	71
net	73	fair	0	tax	8	rate	71	rate	71
tax	71	accounting	0	certain	7	net	70	tax	69.2
certain	68	results	0	net	6	certain	67	certain	67
accounting	66	earnings	0	accounting	5	results	66	accounting	64
results	64	management	0	rate	4	accounting	65	results	61
primarily	57	higher	0	results	2	unsecured	57	unsecured	55
failures	55	impact	0	losses	2	failures	54	failures	54
unsecured	55	compared	0	primarily	1	primarily	53	unfavorable	53

Table W.2: Keywords chosen by classifier for (Keywords-PCA-PS).

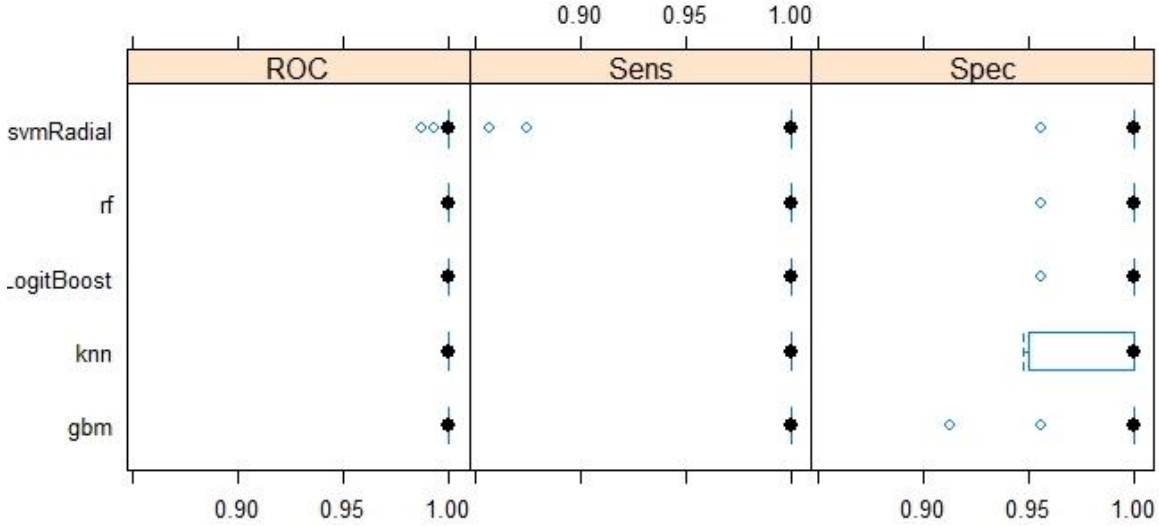


Figure W.1: ROC, Sensitivity and Specificity for classifiers (Keywords-PCA-PS).

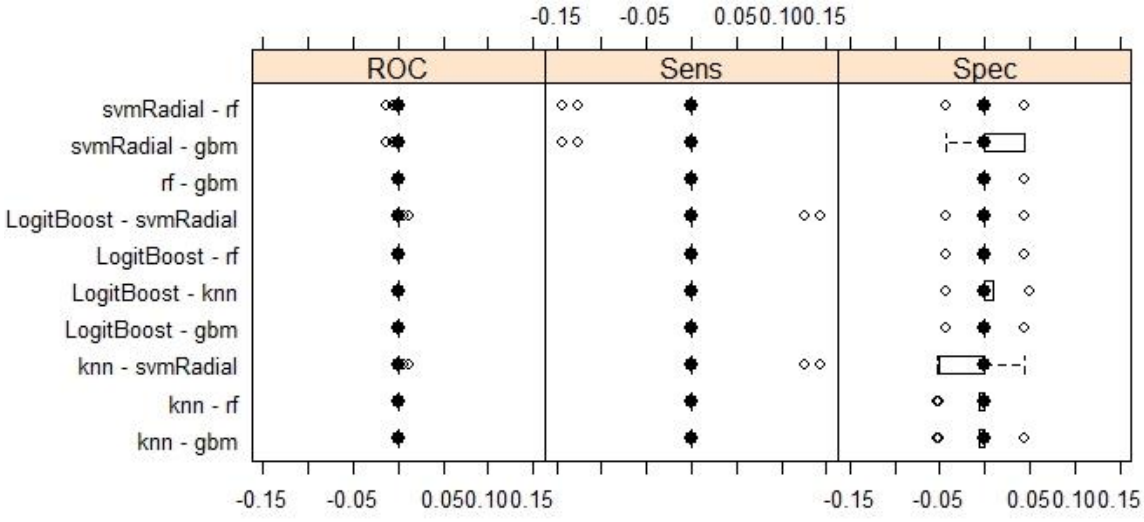


Figure W.2: Performance compared between classifiers (Keywords-PCA-PS).

PCA feature selection on Keywords- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
SGB	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
RF	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
kNN	0.97	1	0.97	0.98	0.92 0.99	0.5	<2e-16	0.97	1	0.98
SVM	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1

Table W.3: Classification results (Keywords-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
also	100	also	100	growth	100	also	100	growth	100
growth	100	growth	90	also	78	growth	98	also	99
result	89	results	2	result	35	result	86	result	86
results	84	general	2	results	22	results	83	rate	83
rate	83	result	1	rate	17	general	82	results	82
general	80.3	fair	0	accounting	9	rate	80	general	81
due	77	net	0	general	9	accounting	71.4	due	72.3
tax	73	date	0	required	8	date	70.4	tax	71
accounting	72	average	0	primarily	7	due	70	accounting	70.8
compared	72	accounting	0	certain	6	tax	69.5	compared	68
primarily	70	flows	0	flat	2	primarily	69	primarily	67
certain	65	believe	0	compared	1	compared	67	certain	62
related	64	impact	0	related	1	required	65.9	required	62

Table W.4: Keywords chosen by classifier as significant (Keywords-PCA-MP).

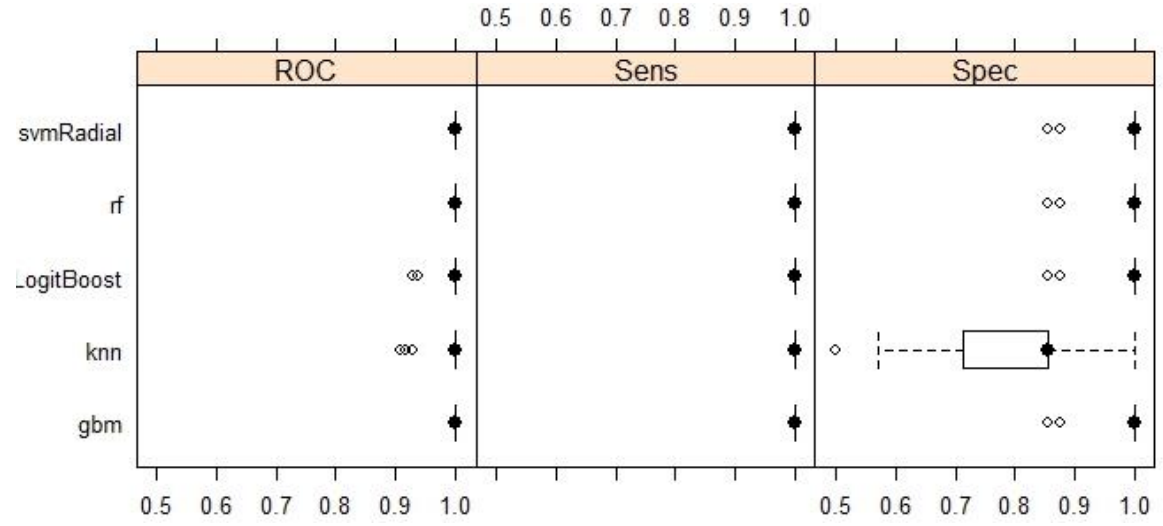


Figure W.3: ROC, Sensitivity and Specificity for classifiers (Keywords-PCA-MP).

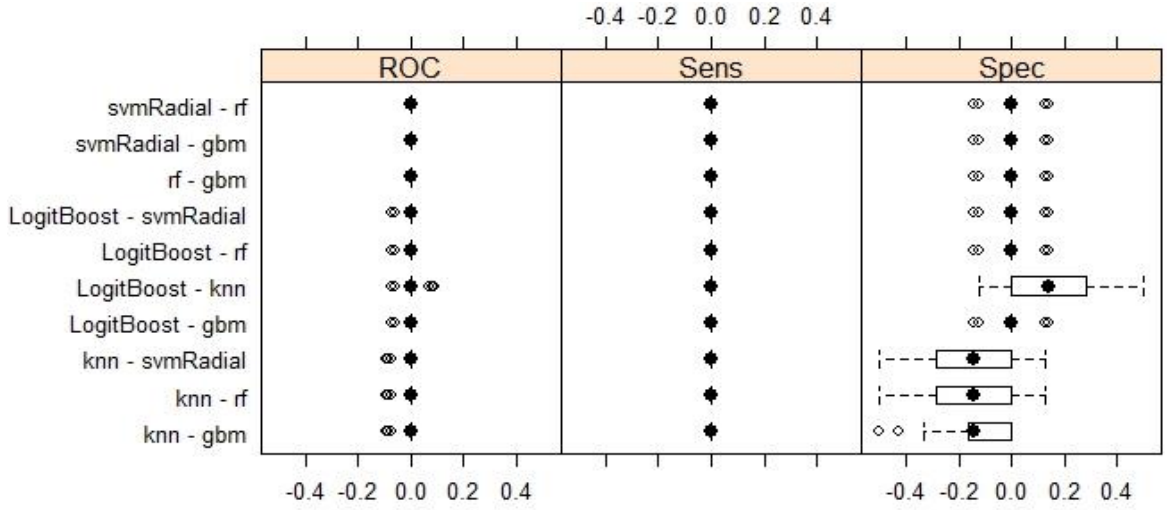


Figure W.4: Performance compared between classifiers (Keywords-PCA-MP).

Classification results for Boruta selected keywords

Boruta feature selection on Keywords.– Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SVM	1	1	1	1	0.961	0.75	3.355e-13	1	1	1

Table W.5: Classification results (Keywords-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
companies	100	capital	1.00E+02	capital	1.00E+02	capital	1.00E+02	capital	1.00E+02
growth	99	companies	4.88E+01	growth	4.25E+01	companies	9.94E+01	companies	1.00E+02
capital	99	growth	3.82E+01	companies	4.00E+01	growth	9.90E+01	growth	9.97E+01
operations	83.2	expected	1.15E+00	operations	6.14E+00	operations	8.30E+01	operations	8.41E+01
operating	81	Manage-ment	5.69E-01	operating	2.98E+00	operating	8.25E+01	operating	8.13E+01
management	73.8	results	4.12E-01	tax	6.48E-01	Management	7.89E+01	Management	7.59E+01
tax	73	operating	3.24E-01	expected	5.39E-01	tax	7.42E+01	expected	7.17E+01
certain	70	compared	2.17E-01	stockholder	4.93E-01	expected	6.99E+01	tax	7.01E+01

Table W.6: Keywords chosen by classifier as significant (Keywords-Boruta-PS).

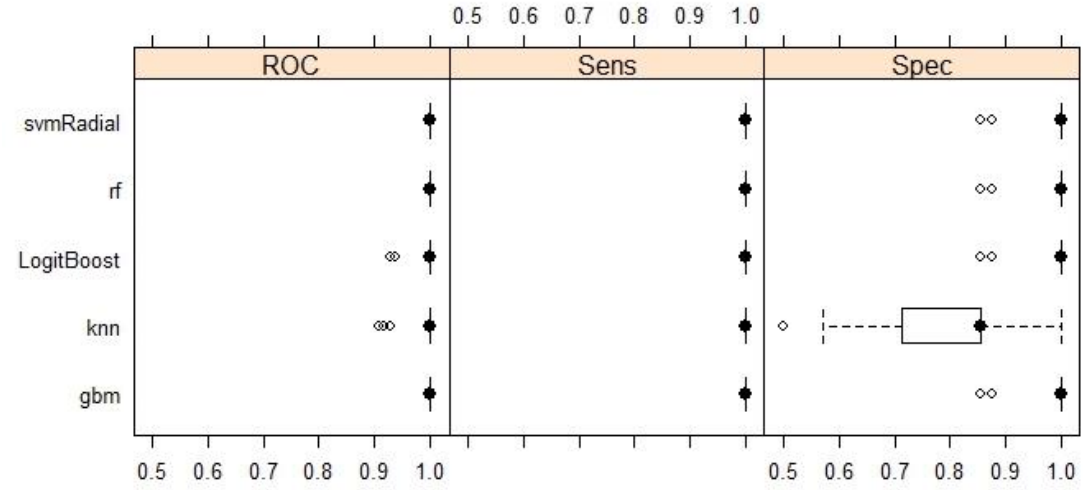


Figure W.5: ROC, Sensitivity and Specificity for classifiers (Keywords-Boruta-PS).

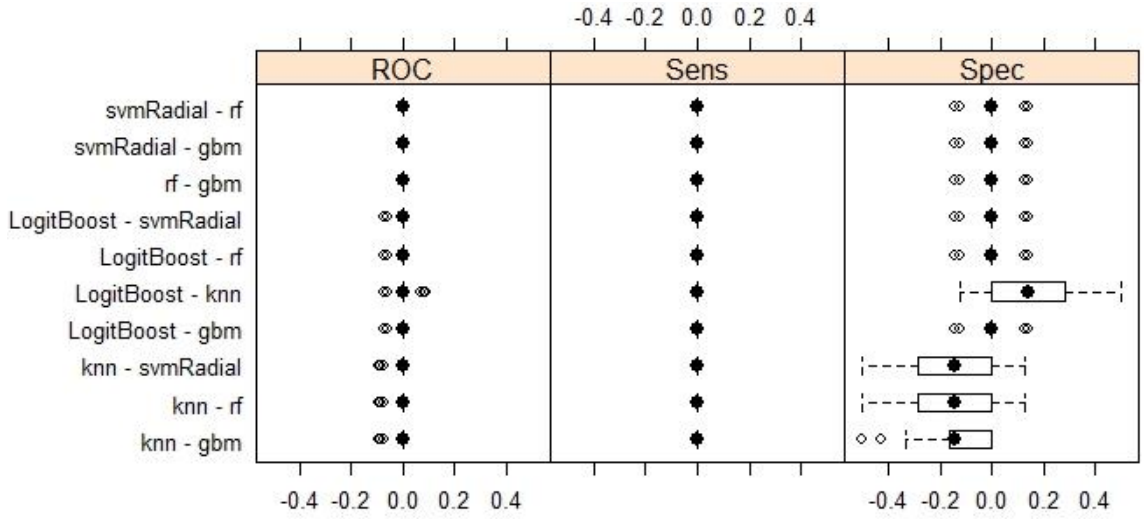


Figure W.6: Performance compared between classifiers (Keywords-Boruta-PS).

Boruta feature selection on Keywords.– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.91 1	0.5	8.882e-16	1	1	1
SGB	1	1	1	1	0.91 1	0.5	8.882e-16	1	1	1
RF	1	1	1	1	0.91 1	0.5	8.882e-16	1	1	1
kNN	1	1	1	1	0.91 1	0.5	8.882e-16	1	1	1
SVM	1	1	1	1	0.91 1	0.5	8.882e-16	1	1	1

Table W.7: Classification results (Keywords-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
companies	100	companies	1.00E +02	growth	100	growth	1.00E +02	companies	1.00E +02
growth	1.00E +02	growth	2.73E +01	companies	89	companies	9.91E +01	growth	1.00E +02
results	9.16E +01	results	1.83E +00	result	14	result	9.00E +01	result	8.65E +01
result	9.02E +01	accounting	1.52E +00	results	11	results	8.46E +01	results	8.26E +01
primarily	8.17E +01	acquisition	1.26E +00	primarily	3	accounting	7.85E +01	tax	7.51E +01
tax	8.00E +01	primarily	8.52E -01	accounting	2	primarily	7.76E +01	primarily	7.15E +01
accounting	7.74E +01	result	7.83E -01	certain	1	compared	7.75E +01	accounting	6.96E +01
certain	7.54E +01	unfavorable	3.58E -01	acquisition	0.82	tax	7.72E +01	compared	6.84E +01
compared	7.24E +01	demand	2.79E -01	compared	0.65	certain	6.92E +01	certain	6.69E +01
unsecured	6.67E +01	certain	2.60E -01	unfavorable	0.48	unsecured	6.36E +01	demand	6.07E +01

Table W.8: Keywords chosen by classifier as significant (Keywords-Boruta-MP).

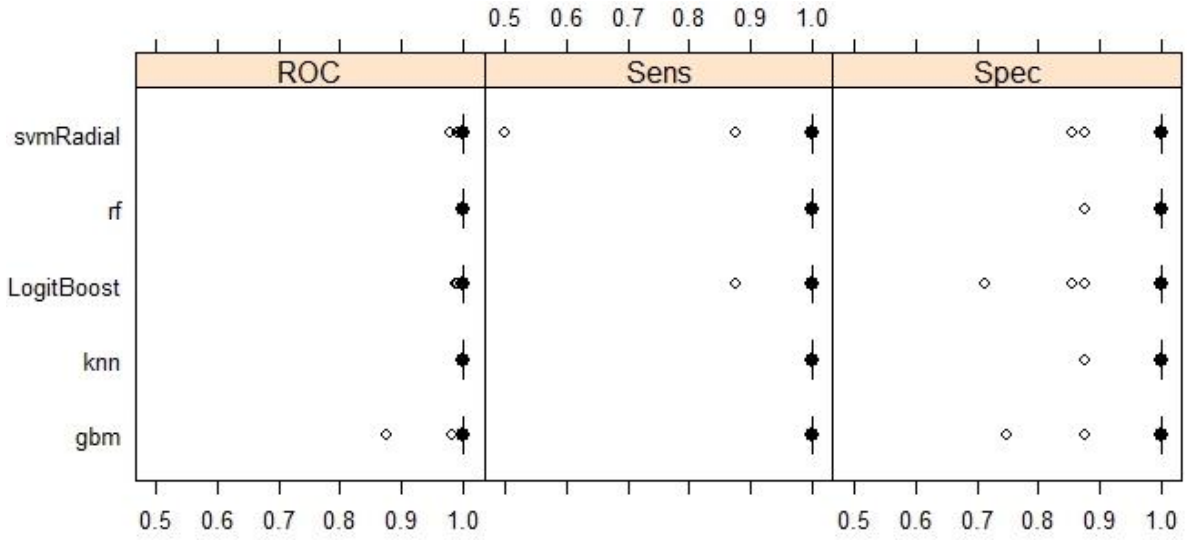


Figure W.7: ROC, Sensitivity and Specificity for classifiers (Keywords-Boruta-MP).

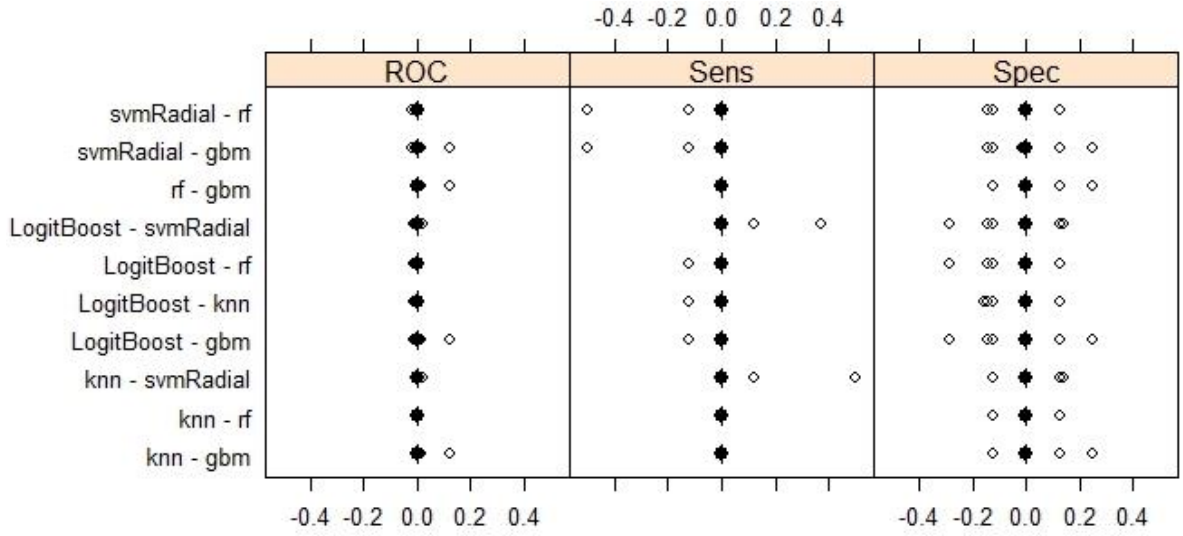


Figure W.8: Performance compared between classifiers (Keywords-Boruta-MP).

Peer Set: Classification results for IG selected Keywords

IG feature selection on Keywords.– Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
SVM	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1

Table W.9: Classification results (Keywords-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
also	100	also	1.00E +02	also	1.00E +02	also	100	also	100
capital	9.91E +01	companies	7.01E +01	capital	5.93E +01	growth	9.85E +01	capital	9.97E +01
companies	9.88E +01	capital	5.86E +01	companies	4.40E +01	capital	9.85E +01	companies	9.88E +01
growth	9.78E +01	growth	3.65E -01	growth	4.02E +01	companies	9.85E +01	growth	9.84E +01
result	9.02E +01	result	2.16E -01	result	1.71E +01	result	9.03E +01	result	9.42E +01
operating	8.22E +01	net	1.63E -01	operations	5.91E +00	operating	8.44E +01	operating	8.40E +01
operations	7.79E +01	certain	1.25E -01	operating	2.36E +00	operations	8.39E +01	operations	8.10E +01
Manage ment	7.72E +01	expected	8.54E -02	tax	1.67E +00	Manage ment	7.72E +01	Manage ment	7.76E +01
tax	7.29E +01	operations	5.56E -02	accounting	1.29E +00	tax	7.20E +01	tax	7.24E +01

Table W.10: Keywords chosen by classifier as significant (Keywords-IG-PS).

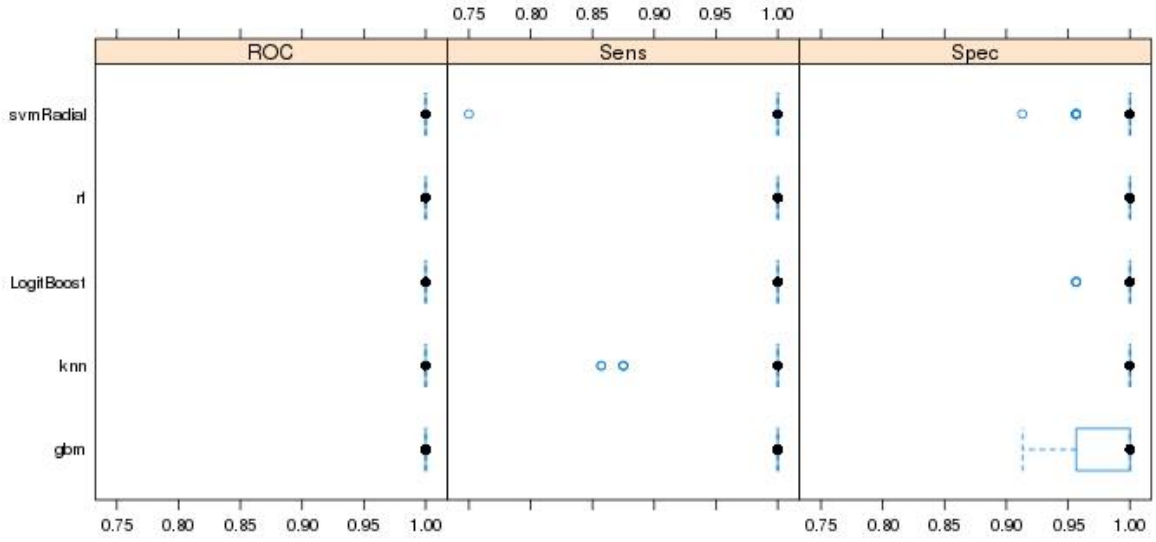


Figure W.9: ROC, Sensitivity and Specificity for classifiers (Keywords-IG-PS).

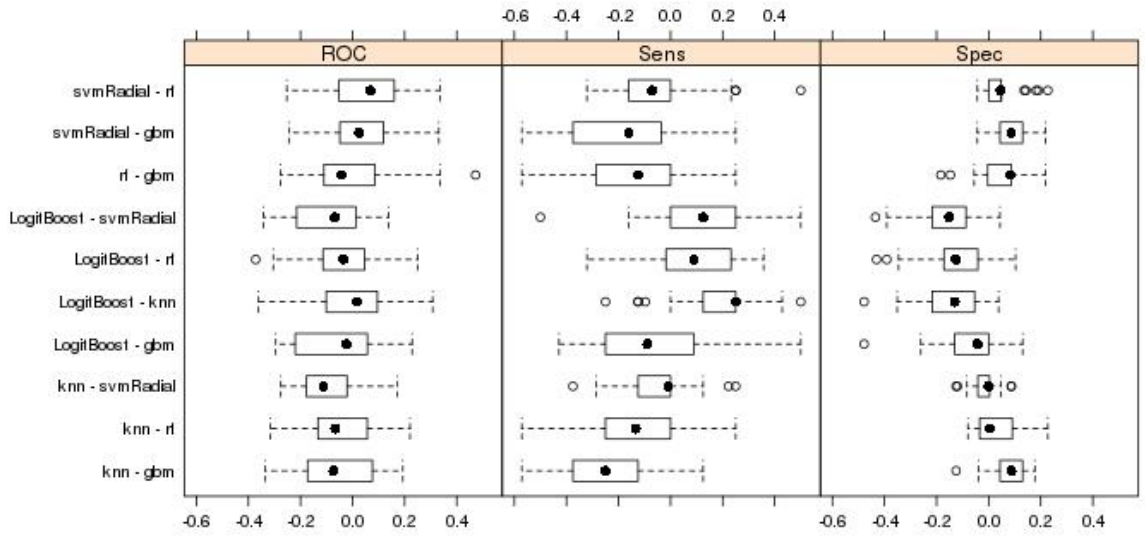


Figure W.10: Performance compared between classifiers (Keywords-IG-PS).

Boruta feature selection on Custom Dict.– Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.92 1	0.5	8.8822-16	1	1	1
SGB	1	1	1	1	0.92 1	0.5	8.8822-16	1	1	1
RF	1	1	1	1	0.92 1	0.5	8.8822-16	1	1	1
kNN	0.88	1	0.88	0.94	0.83 0.98	0.5	1.85e-11	0.89	1	0.94
SVM	1	1	1	1	0.92 1	0.5	8.8822-16	1	1	1

Table W.11: Classification results (Keywords-IG-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
also	100	also	1.00E+02	also	1.00E+02	also	100	also	100
capital	9.91E+01	companies	7.01E+01	capital	5.93E+01	growth	9.85E+01	capital	9.97E+01
companies	9.88E+01	capital	5.86E+01	companies	4.40E+01	capital	9.85E+01	companies	9.88E+01
growth	9.78E+01	growth	3.65E-01	growth	4.02E+01	companies	9.85E+01	growth	9.84E+01
result	9.02E+01	result	2.16E-01	result	1.71E+01	result	9.03E+01	result	9.42E+01
operating	8.22E+01	net	1.63E-01	operations	5.91E+00	operating	8.44E+01	operating	8.40E+01
operations	7.79E+01	certain	1.25E-01	operating	2.36E+00	operations	8.39E+01	operations	8.10E+01
Manage-ment	7.72E+01	expected	8.54E-02	tax	1.67E+00	Manage-ment	7.72E+01	Manage-ment	7.76E+01

Table W.12: Keywords chosen by classifier as significant (Keywords-IG_MP).

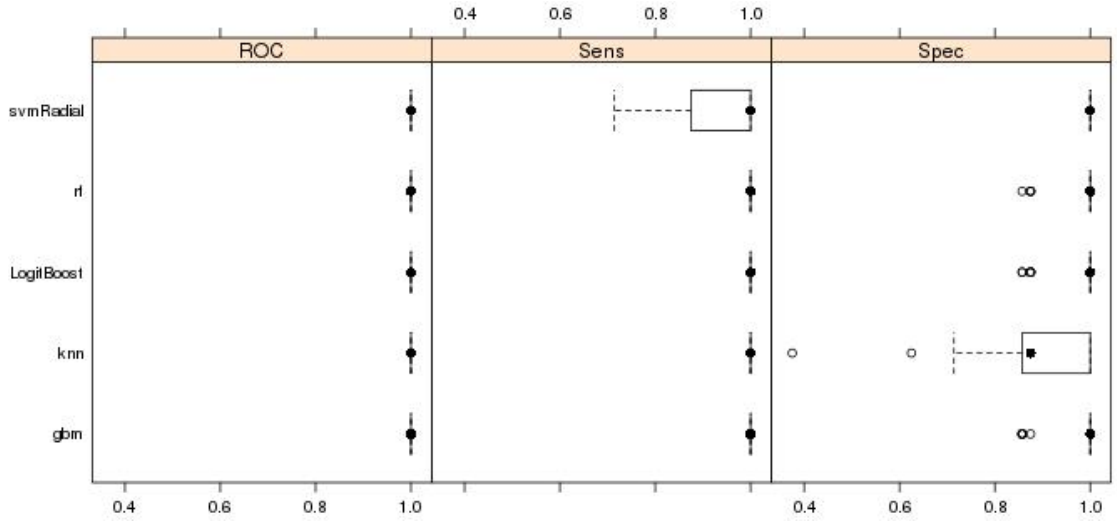


Figure W.11: ROC, Sensitivity and Specificity for classifiers (Keywords-IG-MP).

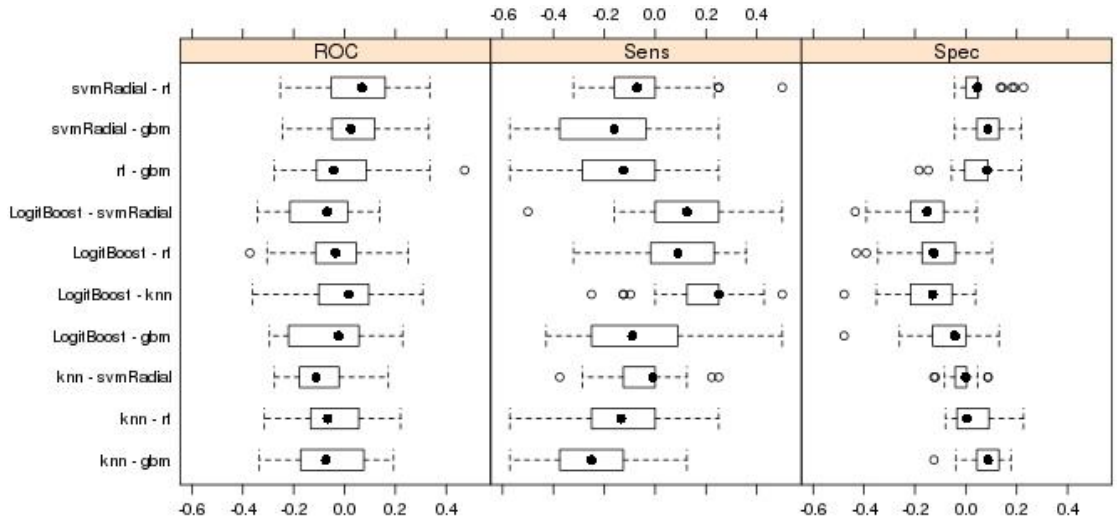


Figure W.12: Performance compared between classifiers (Keywords-IG-MP).

Classification results for PCA selected Keywords- Rutherford

PCA feature selection on Rutherford (Keywords)- Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SVM	1	1	1	1	0.961	0.75	3.355e-13	1	1	1

Table W.13: Classification results (Rutherford (Key.)-PCA-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
years	1.00E+02	significant	1.00E+02	financial	100	years	100	company	100
company	9.97E+01	years	8.67E+01	years	92	financial	100	years	100
increase	9.94E+01	financial	6.64E+01	company	79	significant	100	increase	99
significant	9.94E+01	company	5.13E+01	significant	63	increase	99	financial	99
financial	9.94E+01	capital	1.24E+00	capital	59	capital	99	capital	99
capital	9.87E+01	interest	1.79E-01	increase	55	company	99	significant	99
growth	9.81E+01	risk	5.26E-02	result	25	growth	98	growth	98
include	9.02E+01	loss	4.66E-02	growth	23	include	91	include	89
result	8.96E+01	rate	3.02E-02	include	8	result	90	result	88
operating	8.05E+01	trading	2.58E-02	operations	8	operations	80	operating	81
operations	7.82E+01	net	2.42E-02	turnover	3	management	79	operations	76

Table W.14: Rutherford (Key.) chosen by classifier for (Rutherford (Key.)-PCA-PS).

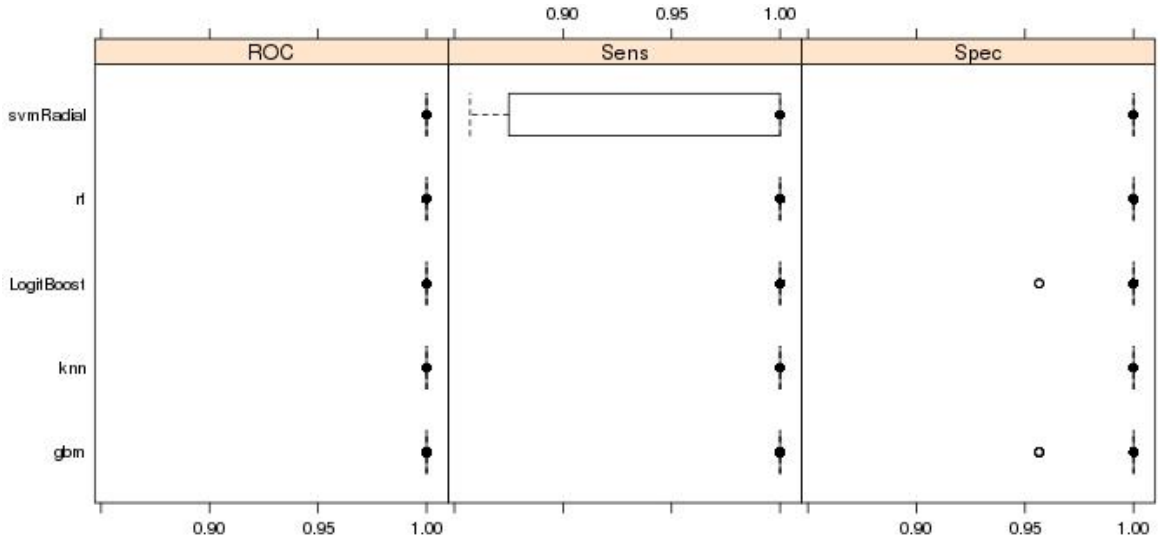


Figure W.13: ROC, Sensi. and Speci. for classifiers (Rutherford (Key.)-PCA-PS).

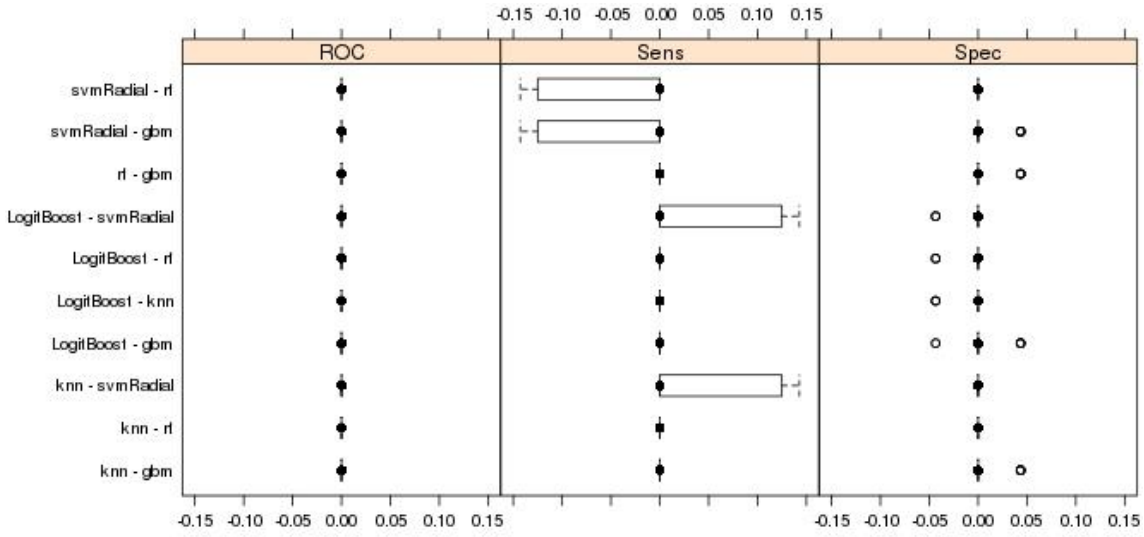


Figure W.14: Performance compared between classifiers (Rutherford (Key.)-PCA-PS).

PCA feature selection on Rutherford (Keywords)- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
SGB	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
RF	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
kNN	0.72	1	0.72	0.86	0.73 0.94	0.5	1.049e-07	0.78	1	0.86
SVM	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1

Table W.15: Classification results (Rutherford (Key.)-PCA-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
years	100	significant	100.00	years	100.00	years	100.00	years	100
company	100	company	76.04	company	98.93	company	99.06	company	99.06
increase	99.04	years	73.80	significant	64.35	significant	99.06	significant	99.06
significant	99.04	increase	46.44	growth	60.46	increase	98.11	growth	98.11
growth	99.04	sale	0.51	increase	48.82	growth	98.11	increase	98.11
include	87.5	exchange	0.41	include	27.71	result	90.03	include	93.19
result	85.41	tax	0.34	result	14.78	include	88.72	result	90.43
rate	83.53	risk	0.31	rate	10.06	rate	78.47	rate	80.12
new	77.29	result	0.23	loss	7.70	new	78.38	loss	75.32
make	72.34	loss	0.16	new	6.66	loss	73.54	new	73.38
loss	71.22	overall	0.14	make	5.13	total	70.49	tax	71.78
due	68.62	due	0.12	due	4.50	tax	69.65	total	71.05
number	68.34	item	0.11	cash	2.85	due	68.26	due	69.6

Table W.16: Rutherford (Key.) chosen by classifier as significant (Rutherford (Key.)-PCA-MP).

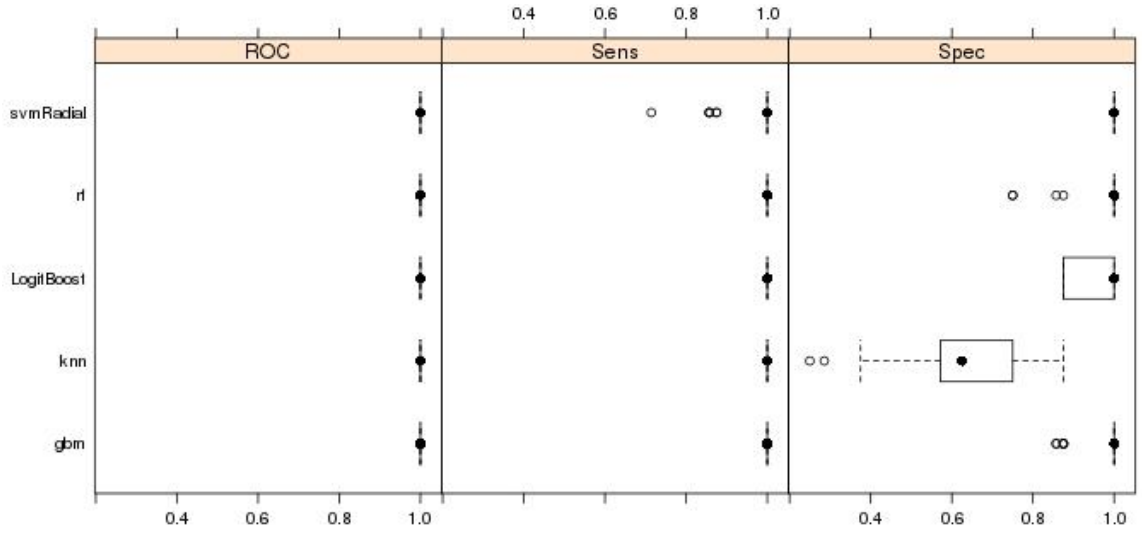


Figure W.15: ROC, Sensi. and Speci. for classifiers (Rutherford (Key.)-PCA-MP).

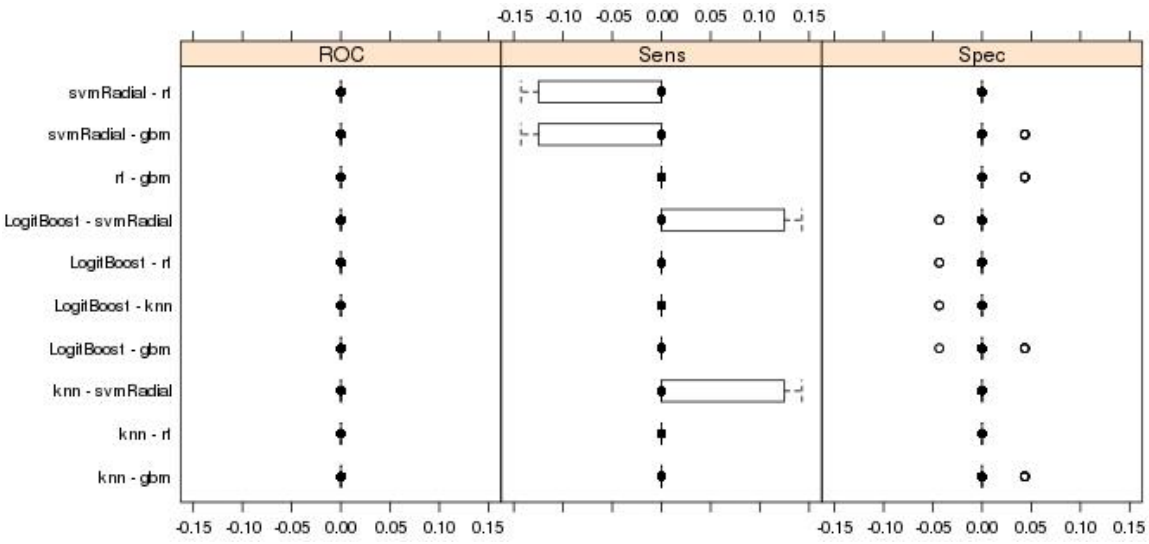


Figure W.16: Performance compared between classifiers (Rutherford (Key.)-PCA-MP).

Classification results for Boruta selected Keywords Rutherford

Boruta feature selection on Rutherford (Keywords)- Peer set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.961	0.75	3.355e-13	1	1	1
SVM	1	1	1	1	0.961	0.75	3.355e-13	1	1	1

Table W.17: Classification results (Rutherford (Key.)-Boruta-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
significant	100	significant	1.00E+02	years	1.00E+02	years	1.00E+02	years	1.00E+02
company	99.67	company	9.51E+01	company	9.44E+01	increase	9.97E+01	company	1.00E+02
financial	99.67	years	8.73E+01	significant	4.27E+01	significant	9.94E+01	financial	9.97E+01
increase	99.35	financial	8.08E+01	financial	3.92E+01	company	9.94E+01	increase	9.97E+01
capital	99.02	capital	7.90E-01	increase	2.88E+01	financial	9.91E+01	significant	9.91E+01
growth	97.72	loss	1.79E-01	capital	1.58E+01	growth	9.85E+01	capital	9.84E+01
include	90.85	operations	1.54E-01	growth	8.00E+00	capital	9.82E+01	growth	9.78E+01
result	90.19	rate	1.17E-01	result	3.33E+00	result	9.42E+01	include	8.98E+01
operating	78.78	revenue	9.11E-02	include	1.25E+00	include	9.27E+01	result	8.83E+01
operations	78.37	increase	7.41E-02	operating	5.80E-01	operating	8.49E+01	operations	7.80E+01
loss	72.55	net	2.66E-02	rate	3.19E-01	operations	8.35E+01	operating	7.79E+01
tax	72.37	interest	2.27E-02	operations	2.95E-01	Manage-ment	8.21E+01	Manage-ment	7.48E+01

Table W.18: Rutherford (Key.) chosen by classifier as significant (Rutherford (Key.)-Boruta-PS).

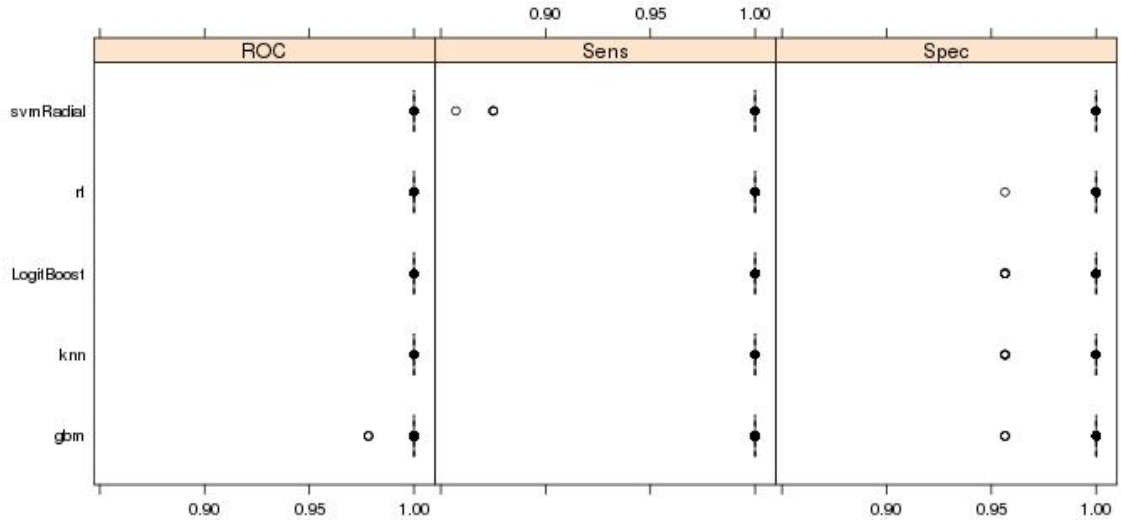


Figure W.17: ROC, Sensi. and Speci. for classifiers (Rutherford (Key.)-Boruta-PS).

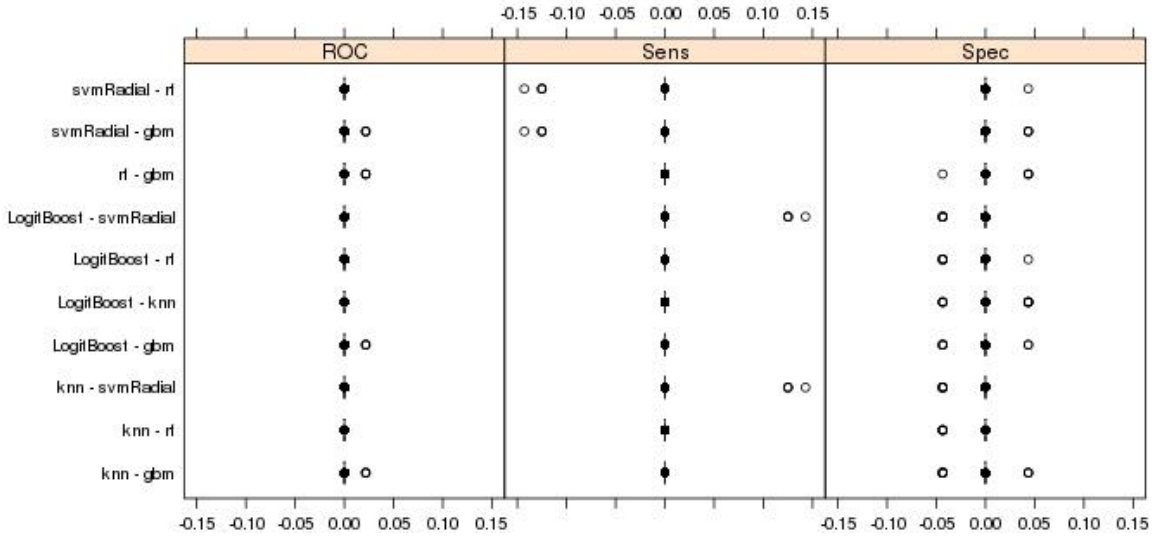


Figure W.18: Performance compared between classifiers (Rutherford (Key.)-Boruta-PS).

Boruta feature selection on Rutherford (Keywords)- Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
SGB	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
RF	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
kNN	0.96	1	0.96	0.98	0.89 0.99	0.5	4.53e-14	0.96	1	0.98
SVM	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1

Table W.19: Classification results (Rutherford (Key.)-Boruta-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
years	1.00E+02	significant	1.00E+02	years	1.00E+02	years	1.00E+02	significant	1.00E+02
company	9.90E+01	years	4.64E+01	company	6.92E+01	company	9.91E+01	years	1.00E+02
significant	9.90E+01	company	3.10E+01	significant	6.41E+01	significant	9.91E+01	company	9.90E+01
increase	9.81E+01	increase	1.69E+01	growth	5.90E+01	increase	9.81E+01	increase	9.81E+01
growth	9.81E+01	total	7.04E-01	increase	2.53E+01	growth	9.81E+01	growth	9.81E+01
include	8.56E+01	cash	1.86E-01	include	5.64E+00	include	8.79E+01	include	9.02E+01
result	8.54E+01	include	1.52E-01	result	3.73E+00	result	8.59E+01	rate	8.66E+01
rate	8.04E+01	tax	6.49E-02	loss	1.03E+00	rate	8.03E+01	result	8.53E+01
new	8.01E+01	make	5.34E-02	new	9.16E-01	new	7.80E+01	loss	7.61E+01
loss	7.57E+01	growth	4.67E-02	rate	8.89E-01	loss	7.14E+01	make	7.18E+01

Table W.20: Rutherford (Key.) chosen by classifier as significant (Rutherford (Key.)-Boruta-MP).

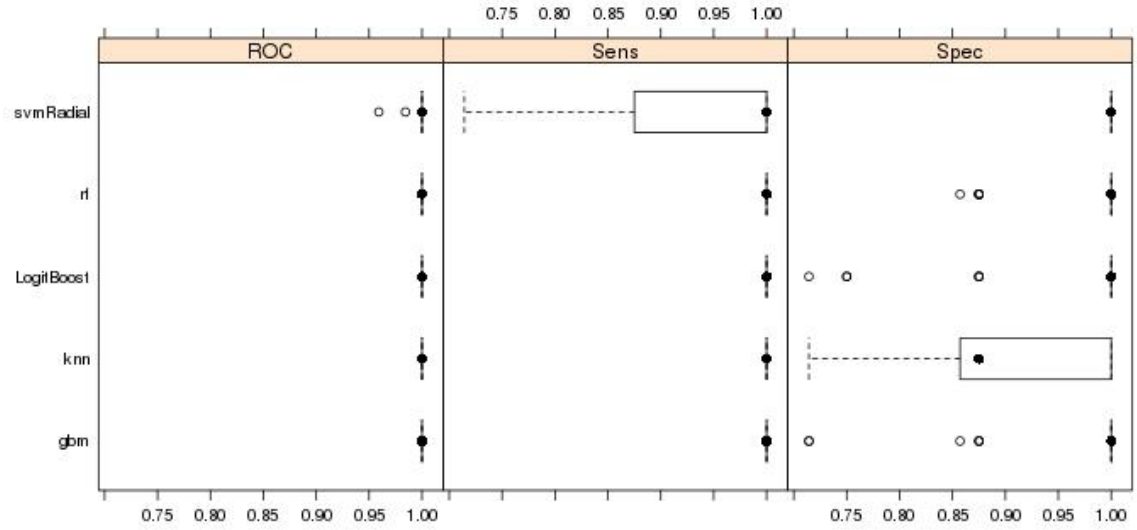


Figure W.19: ROC, Sensi. and Speci. for classifiers (Rutherford (Key.)-Boruta-MP).

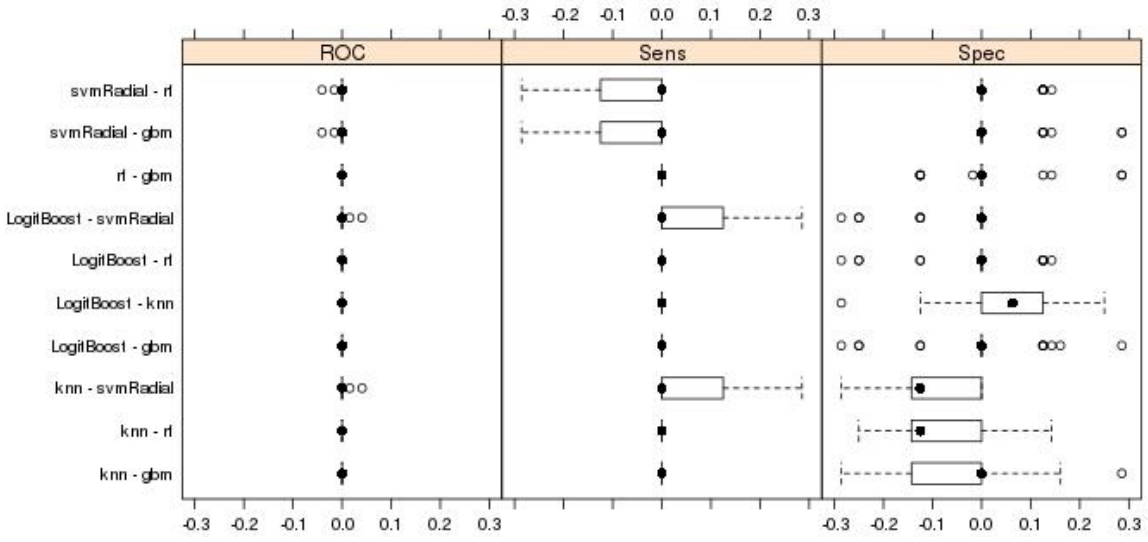


Figure W.20: Performance compared between classifiers (Rutherford (Key.)-Boruta-MP).

Classification results for IG selected Keywords Rutherford

IG feature selection on Keywords Rutherford - Peer Set										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
SGB	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
RF	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
kNN	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1
SVM	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1

Table W.21: Classification results (Rutherford (Key.)-IG-PS).

Peer Set – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
increase	1.00E+02	years	1.00E+02	company	1.00E+02	company	1.00E+02	company	1.00E+02
years	1.00E+02	company	7.86E+01	years	9.94E+01	years	1.00E+02	significant	1.00E+02
significant	1.00E+02	financial	6.45E+01	significant	5.89E+01	financial	9.97E+01	financial	9.94E+01
company	1.00E+02	significant	6.37E+01	increase	5.34E+01	significant	9.97E+01	growth	9.84E+01
financial	9.94E+01	capital	4.22E+01	financial	4.86E+01	increase	9.97E+01	capital	9.84E+01
capital	9.87E+01	increase	3.77E+01	capital	2.17E+01	capital	9.87E+01	result	9.14E+01
growth	9.81E+01	growth	1.28E+01	growth	1.62E+01	growth	9.78E+01	include	9.07E+01
result	9.24E+01	result	8.91E+00	include	4.64E+00	result	9.10E+01	operations	8.15E+01
include	9.02E+01	include	3.07E+00	result	3.86E+00	include	9.01E+01	operating	7.95E+01
operations	8.32E+01	operations	7.50E-01	operations	1.32E+00	operating	8.28E+01	management	7.25E+01

Table W.22: Rutherford (Key.) chosen by classifier as significant (Rutherford (Key.)-IG-PS).

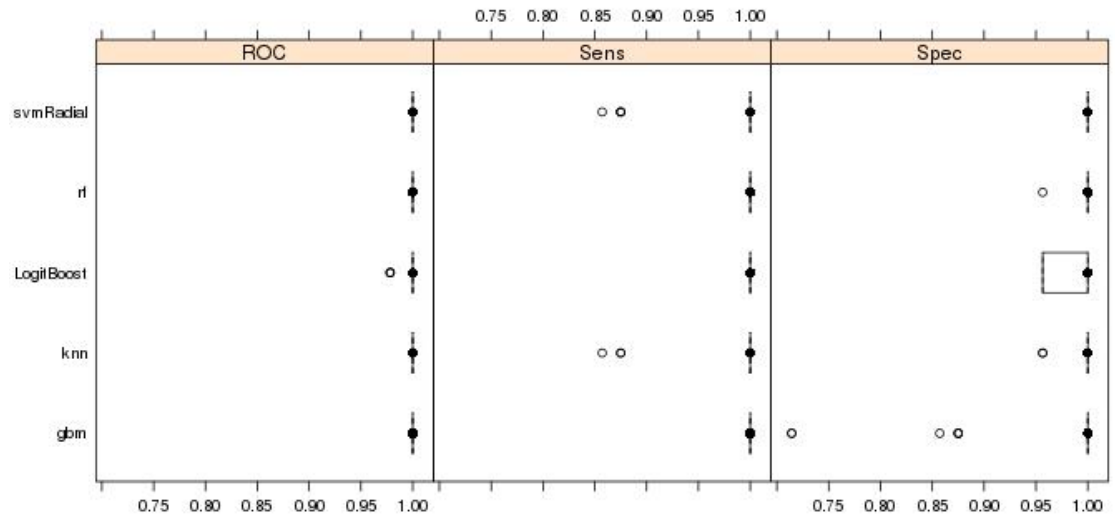


Figure W.21: ROC, Sensi. and Speci. for classifiers (Rutherford (Key.)-IG-PS).

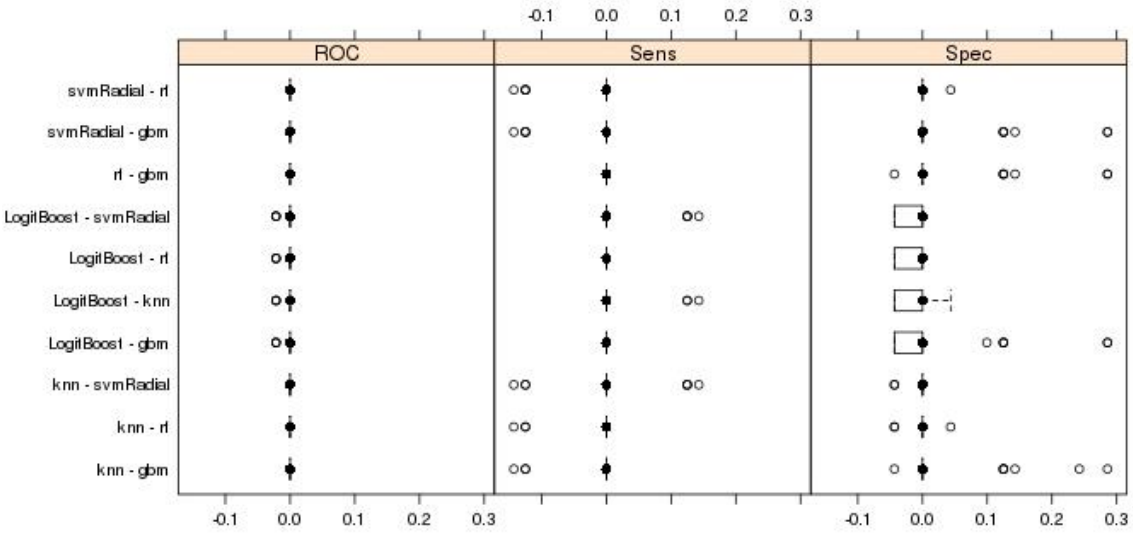


Figure W.22: Performance compared between classifiers (Rutherford (Key.)-IG-PS).

IG feature selection on Keywords Rutherford – Matched Pair										
Model	**Kappa**	**Sensitivity**	**Specificity**	**ACC**	**95% CI**	**NIR**	**P Value [ACC> NIR]**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
LR	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
SGB	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
RF	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1
kNN	0.92	1	0.92	0.96	0.86 0.99	0.5	1.133e-12	0.92	1	0.96
SVM	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1

Table W.23: Classification results (Rutherford (Key.)-IG-MP).

Matched Pair – Variable Importance									
LR		***SGB***		***RF***		***kNN***		***SVM***	
years	1.00E+02	years	1.00E+02	years	1.00E+02	years	1.00E+02	years	1.00E+02
significant	9.90E+01	significant	3.73E+01	significant	9.80E+01	company	1.00E+02	growth	9.91E+01
company	9.90E+01	company	3.45E+01	company	7.54E+01	significant	9.91E+01	company	9.81E+01
increase	9.81E+01	include	2.14E-01	increase	5.96E+01	growth	9.81E+01	significant	9.81E+01
growth	9.81E+01	number	1.84E-01	growth	5.12E+01	increase	9.81E+01	increase	9.81E+01
include	8.94E+01	new	1.41E-01	include	1.69E+01	result	8.67E+01	result	8.76E+01
result	8.44E+01	result	1.39E-01	rate	6.87E+00	include	8.61E+01	include	8.67E+01
rate	7.84E+01	risk	1.02E-01	result	5.76E+00	rate	8.29E+01	rate	8.14E+01
loss	7.64E+01	lower	8.75E-02	loss	5.14E+00	new	7.69E+01	new	7.83E+01
new	7.38E+01	interest	6.46E-02	new	2.17E+00	due	7.18E+01	loss	7.33E+01

Table W.24: Rutherford (Key.) chosen by classifier as significant (Rutherford (Key.)-IG_MP).

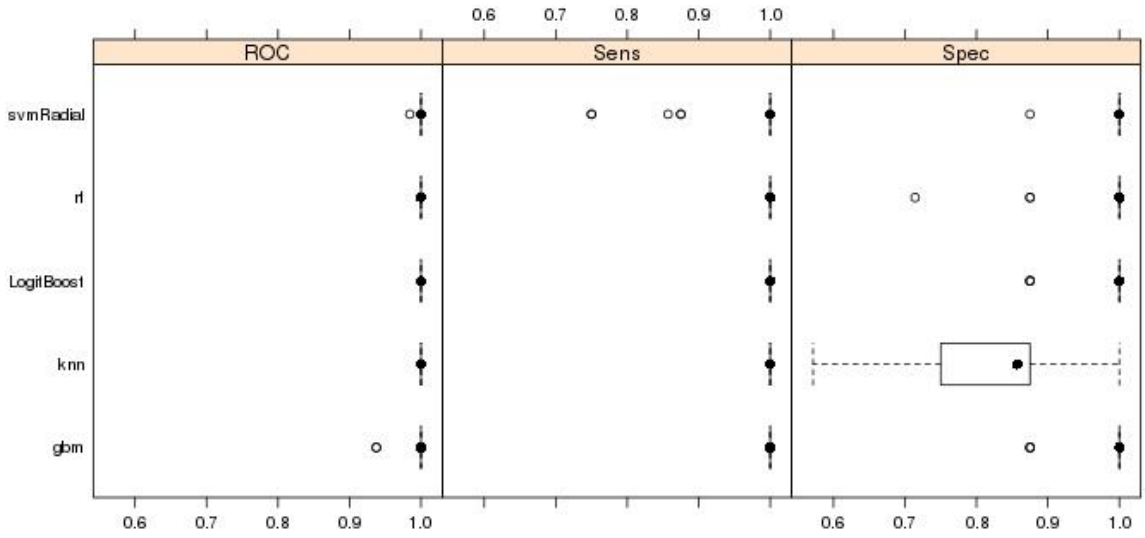


Figure W.23: ROC, Sensitivity and Specificity for classifiers (Rutherford (Key.)-IG-MP).

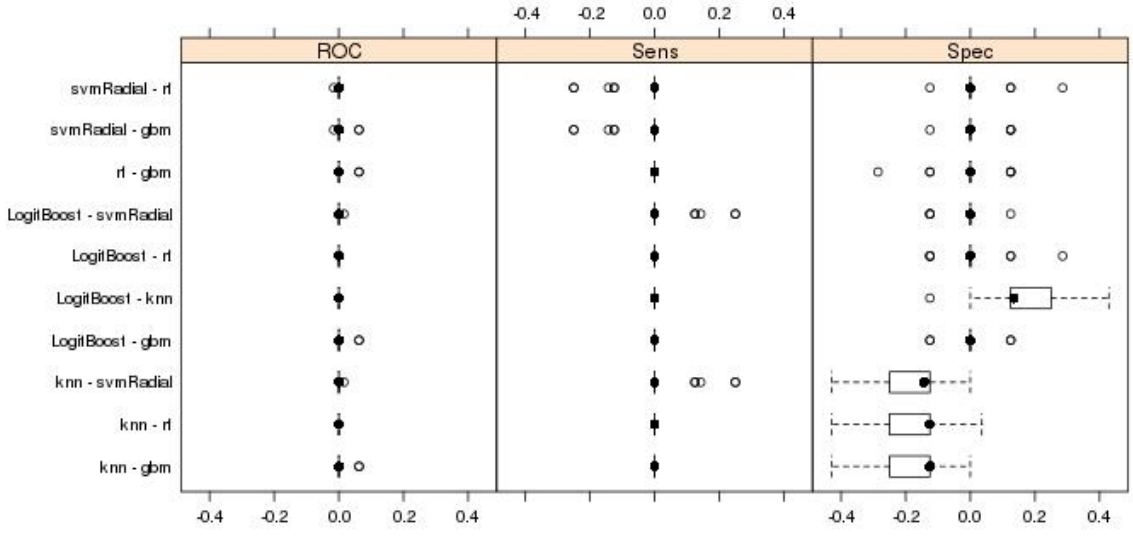


Figure W.24: Performance compared between classifiers (Rutherford (Key.)-IG-MP).

APPENDIX X

Table X.1: Best performing classifiers for the peer set data set up.

Table X.2: Best performing classifiers for the matched pair data set up.

Best Performing Classifiers on Document Representation Schemes - Peer Set

Doc Rep Scheme	Kappa	Sensi	Speci	ACC	95% CI	NIR	P Value [ACC> NIR]	Pos Pred Value	Neg Pred Value	Bal Acc	Feature Selection	Classifier
Unigrams	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1	PCA, Boruta	All
Bigrams	0.76	0.72	0.98	0.92	0.84, 0.96	0.75	1.23e-05	0.94	0.91	0.85	Boruta	SGB
Trigrams	0.64	0.60	0.97	0.88	0.80 0.93	0.75	0.001	0.88	0.88	0.78	Boruta	SGB
Coh-Metrix	0.55	0.60	0.92	0.84	0.75 0.90	0.75	0.02	0.71	0.87	0.76	Boruta	SVM
LIWC	0.54	0.44	1	0.86	0.76 0.91	0.75	0.01	1	0.84	0.72	Boruta	SVM
Custom Word List	0.36	0.32	0.97	0.81	0.72 0.88	0.75	0.09	0.80	0.81	0.64	Boruta	RF
LBCs	0.46	0.44	0.96	0.83	0.74 0.89	0.75	0.03	0.78	0.83	0.70	PCA	SGB
Topics	0.42	0.44	0.93	0.81	0.72 0.88	0.75	0.09	0.68	0.83	0.68	Boruta	SGB
Concepts	0.59	0.60	0.94	0.85	0.78 0.91	0.75	0.002	0.78	0.87	0.77	Boruta	SVM
LSA concepts	0.92	1	0.96	0.97	0.91 0.99	0.75	2.18e-09	0.89	1	0.98	No feature selection	LR
Keywords	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1	All	All
Keywords:-Rutherford	1	1	1	1	0.96 1	0.75	3.355e-13	1	1	1	All	All

Table X.1: Best performing classifiers for the peer set data set up.

Best Performing Classifiers on Document Representation Schemes – Matched Pair

Doc Rep Scheme	Kappa	Sensi	Speci	ACC	95% CI	NIR	P Value [ACC> NIR]	Pos Pred Value	Neg Pred Value	Bal Acc	Feature Selection	Classifier
Unigrams	1	1	1	1	0.96,1	0.75	3.35e-13	1	1	1	PCA, Boruta, IG	All
Bigrams	0.76	1	0.76	0.88	0.75 0.95	0.5	1.622e-08	0.80	1	0.88	IG	RF
Trigrams	0.68	0.84	0.84	0.84	0.70 0.92	0.5	5.818e-07	0.84	0.84	0.94	IG	RF
Coh-Metrix	0.68	0.84	0.84	0.84	0.70 0.92	0.5	5.818e-07	0.84	0.84	0.84	Boruta	RF
LIWC	0.72	0.92	0.80	0.86	0.73 0.94	0.5	1.049e-07	0.82	0.90	0.86	PCA	SGB
Custom Word List	0.6	0.84	0.76	0.80	0.66 0.89	0.5	1.193e-05	0.77	0.82	0.80	Boruta	SVM
LBCs	0.48	0.68	0.80	0.74	0.59 0.85	0.5	0.0004	0.77	0.71	0.74	Boruta	RF
Topics	0.48	0.80	0.68	0.74	0.59 0.85	0.5	0.0004	0.71	0.77	0.74	PCA/ Boruta	RF/SGB
Concepts	0.56	0.80	0.76	0.78	0.64 0.88	0.5	4.511e-05	0.76	0.79	0.78	IG	SGB
LSA concepts	0.96	0.96	1	0.98	0.89 0.99	0.5	4.53e-14	1	0.96	0.98	No feature selection	SGB
Keywords	1	1	1	1	0.92 1	0.5	8.8822-16	1	1	1	All	All (except kNN)
Keywords: Rutherford	1	1	1	1	0.92 1	0.5	8.882e-16	1	1	1	All	All (except kNN)

Table X.2: Best performing classifiers for the matched pair data set up.

APPENDIX Y

Figure Y.1: Code from the Caret package in R to setup up training and testing data for classifiers.

Figure Y.2: Model building parameters in Caret.

```
inTrainingG <- createDataPartition(data3$class, p = .75, list = FALSE)
training <- data3[ inTrainingG,]
testingG <- data3[-inTrainingG,]
ctrl <- trainControl (method = "repeatedcv", number = 10, repeats = 3,
classProbs = TRUE, summaryFunction = twoClassSummary)
svmGrid <- expand.grid(sigma=2, C=4)
```

Figure Y.1: Code from the Caret package in R to setup up training and testing data for classifiers.

Model Building parameters in Caret

createDataPartition: used to create a splits of a data set. For all the matrices under consideration 75% of the data will be used for model training and the remainder will be used for evaluating model performance.

trainControl: generates parameters that control how models are created, the main parameters being:-

method: The resampling method (only K-fold cross validation used)

number and repeats: number controls with the number of folds in K-fold cross-validation. repeats applied only to repeated K-fold cross-validation. For example if method = "repeatedcv", number = 10 and repeats = 3 then three separate 10-fold cross-validations are used as the resampling scheme.

summaryFunction: a function to compute alternate performance summaries.

selectionFunction: a function to choose the optimal tuning parameters.

classProbs: a logical value determining whether class probabilities should be computed for held-out samples during resample.

Train: This is the main function used to select model tuning parameters and estimate model performance using resampling. The train function has the following parameters:-

X: a matrix of predictors.

y: a numeric or factor vector of outcomes.

method: specifies the type of model to be used (logistic regression, svm, etc).

metric: a character string with values of "Accuracy", "Kappa", "RMSE" or "Rsquared".

“This value determines the objective function used to select the final model. For example,

Selecting "Kappa" makes the function select the tuning parameters with the largest value

of the mean Kappa statistic computed from the held-out samples” (Kuhn, 2015).

trControl *(see trainControl)*: takes a list of control parameters for the function. The type of resampling as well as the number of resampling iterations can be set using this list.

tuneLength: controls the size of the default grid of tuning parameters. For each model, train will select a grid of complexity parameters as candidate values.

preProcess: centering and scaling data

tuneGrid: can be used to define a specific grid of tuning parameters

predict: produces predicted values, obtained by evaluating the fitted model output from the train function onto the testing data set aside (25% of the data that was set aside with the *createDataPartition function).* The option type = "prob" is used to compute class probabilities from the model.

Figure Y.2: Model building parameters in Caret.

APPENDIX Z

Unigrams : PCA selected features	
data3Cluster$withinss	7357.975 8514.225
data3Cluster$betweenss	4477.8
table(data3Cluster$cluster)	1 2 87 321
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 22 65 64% 2 80 241

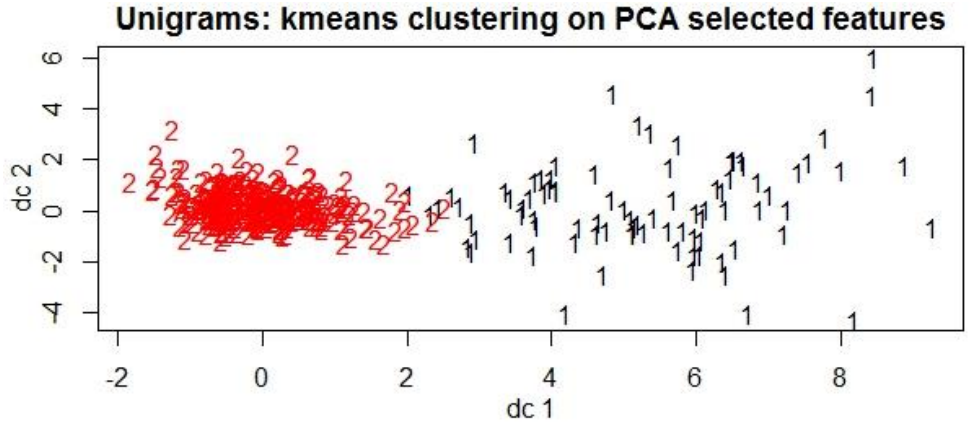


Figure Z.1: Clustering on unigrams based on PCA selected features.

Unigrams : Boruta selected features	
data3Cluster$withinss	1186.123 7252.985
data3Cluster$betweenss	3770.892
table(data3Cluster$cluster)	1 2 102 306
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 102 0 100% 2 0 306

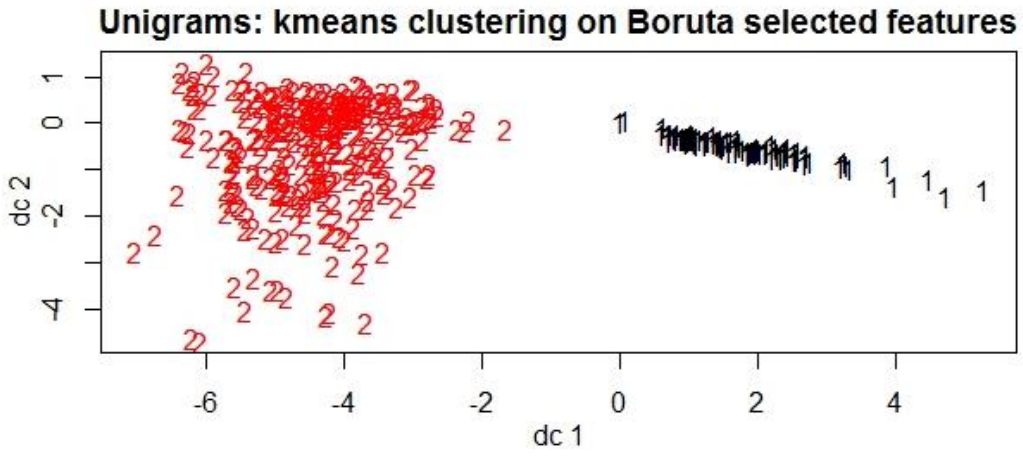


Figure Z.2: Clustering on unigrams based on Boruta selected features.

Unigrams : Information Gain selected features	
data3Cluster$withinss	449.2924 7553.7182
data3Cluster$betweenss	4206.989
table(data3Cluster$cluster)	1 2 102 306
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 102 0 100% 2 0 306

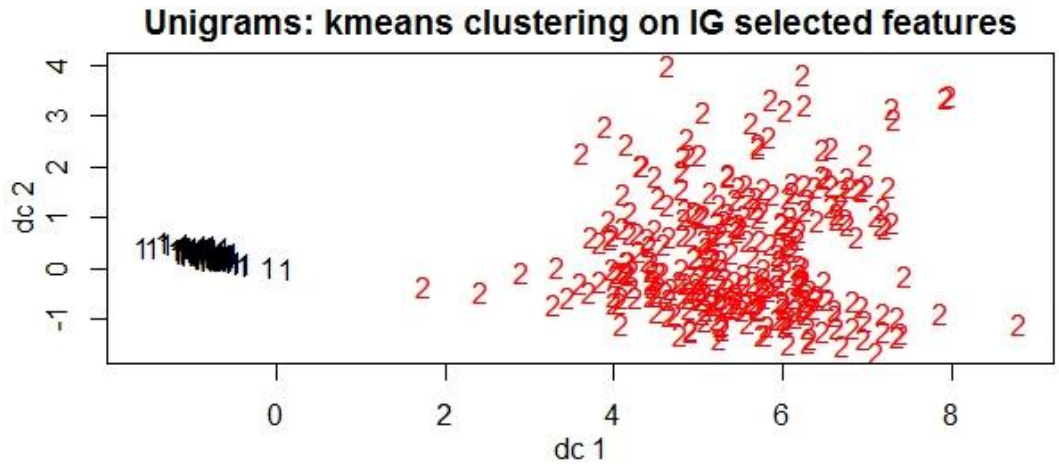


Figure Z.3: Clustering on unigrams based on IG selected features.

Bigrams : PCA selected features	
data3Cluster$withinss	4894.411 7557.029
data3Cluster$betweenss	7491.56
table(data3Cluster$cluster)	1 2 267 141
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 60 207 62% 2 42 99

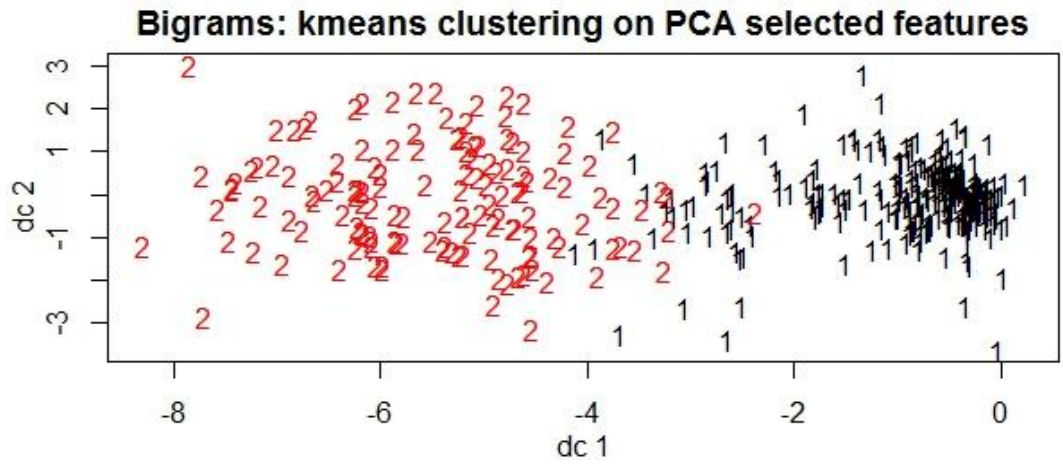


Figure Z.4: Clustering on bigrams based on PCA selected features.

Bigrams : Boruta selected features	
data3Cluster$withinss	4612.887 5656.538
data3Cluster$betweenss	719.5753
table(data3Cluster$cluster)	1 2 93 315
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 62 31 82% 2 40 275

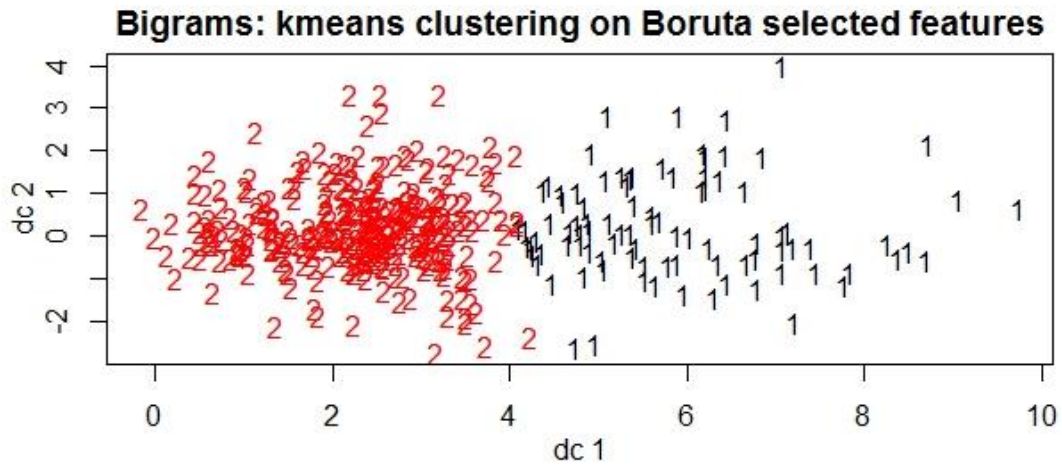
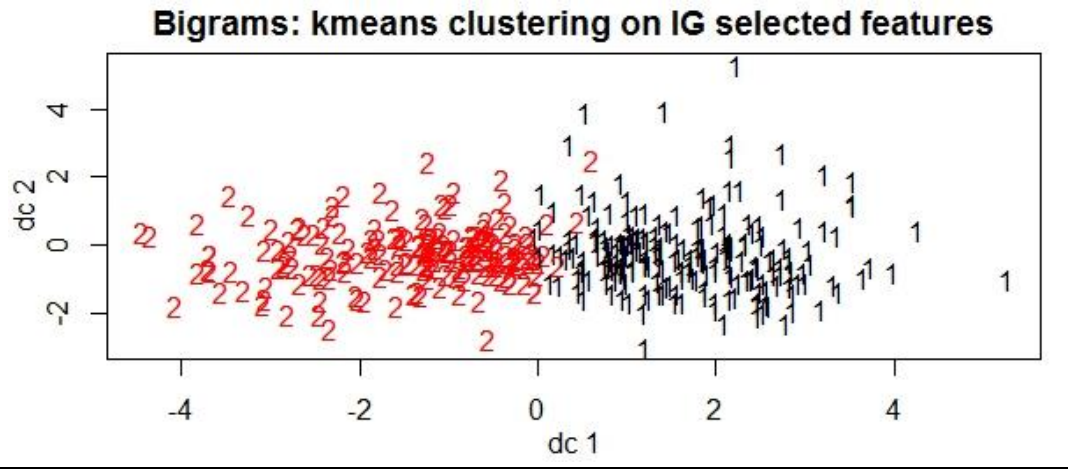


Figure Z.5: Clustering on bigrams based on Boruta selected features from 'f' and 'nf' reports.

Bigrams : IG selected features	
data3Cluster$withinss	5075.831 6392.460
data3Cluster$betweenss	741.7092
table(data3Cluster$cluster)	1 2 223 185
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 89 134 63% 2 13 172



.

Figure Z.6: Clustering on bigrams based on IG selected features.

Trigrams : PCA selected features	
data3Cluster$withinss	9960.655 3514.777
data3Cluster$betweenss	6874.568
table(data3Cluster$cluster)	1 2 139 269
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 42 97 62% 2 60 209

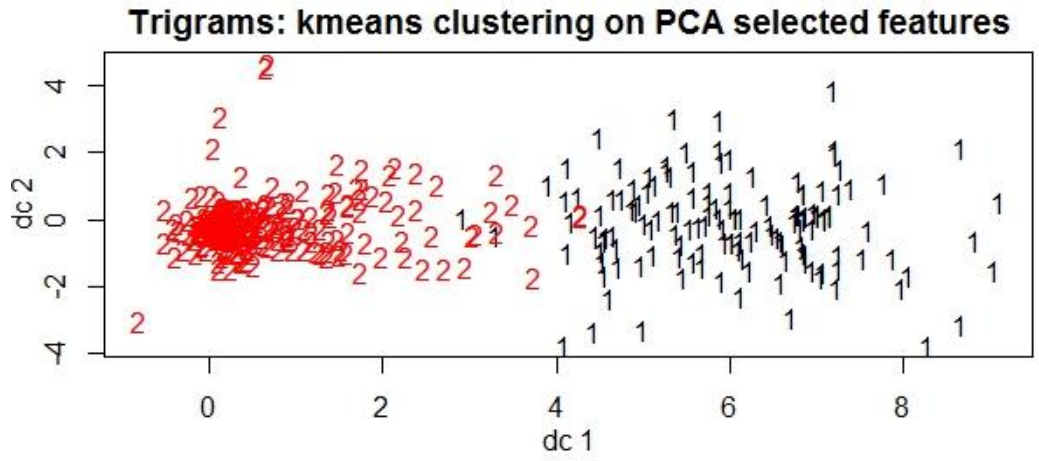


Figure Z.7: Clustering on trigrams based on PCA selected features.

Trigrams : Boruta selected features	
data3Cluster$withinss	6615.760 3147.411
data3Cluster$betweenss	818.828
table(data3Cluster$cluster)	1 2 311 97
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 58 253 72% 2 44 53

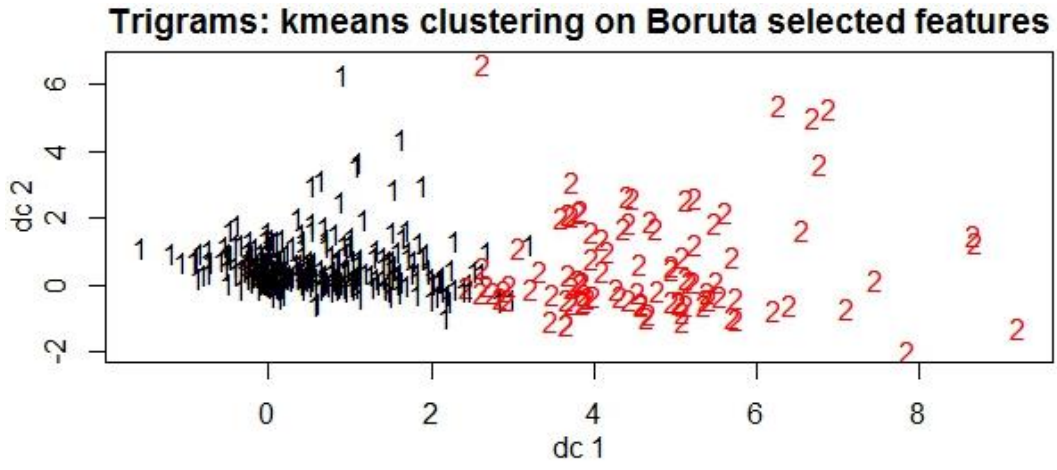


Figure Z.8: Clustering on trigrams based on Boruta selected features.

Trigrams : IG selected features	
data3Cluster$withinss	1934.480 4157.875
data3Cluster$betweenss	419.6444
table(data3Cluster$cluster)	1 2 81 327
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 8 73 59% 2 94 233

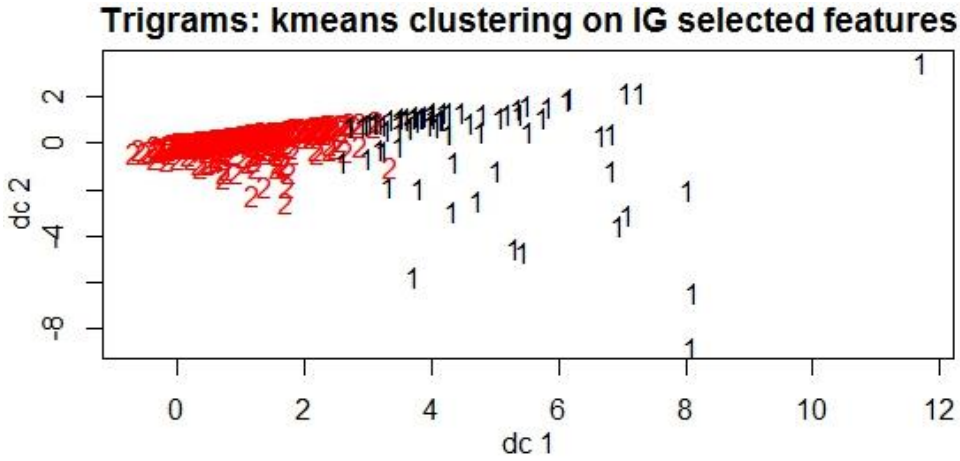


Figure Z.9: Clustering on trigrams based on IG selected features.

Coh-Metrix : PCA selected features	
data3Cluster$withinss	8437.976 4752.375
data3Cluster$betweenss	7159.65
table(data3Cluster$cluster)	317 91
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 83 234 62% 2 19 72

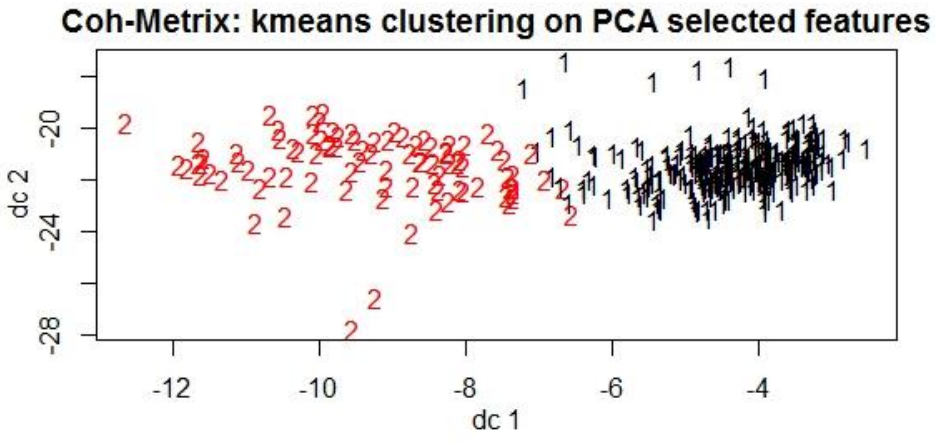


Figure Z.10: Clustering on Coh-Metrix Indices based on PCA selected features.

Coh-Metrix : Boruta selected features	
data3Cluster$withinss	8437.976 4752.375
data3Cluster$betweenss	7159.65
table(data3Cluster$cluster)	1 2 317 91
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 83 234 62% 2 19 72

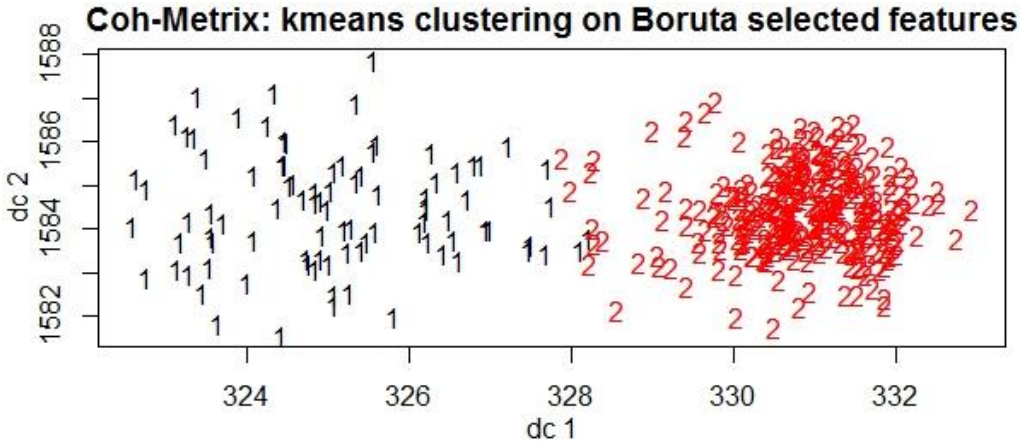


Figure Z.11: Clustering on Coh-Metrix Indices based on Boruta selected features.

Coh-Metrix : IG selected features	
data3Cluster$withinss	1531.954 2124.211
data3Cluster$betweenss	820.8344
table(data3Cluster$cluster)	1 2 189 219
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 57 132 56% 2 45 174

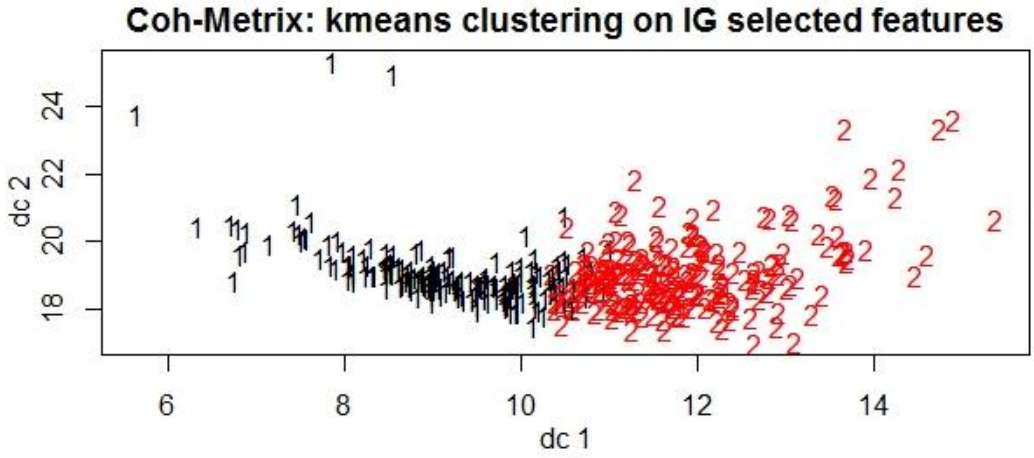


Figure Z.12: Clustering on Coh-Metrix Indices based on IG selected features.

LIWC : PCA selected features	
data3Cluster$withinss	7064.310 3003.757
data3Cluster$betweenss	2548.933
table(data3Cluster$cluster)	1 2 275 133
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 61 214 62% 2 41 92

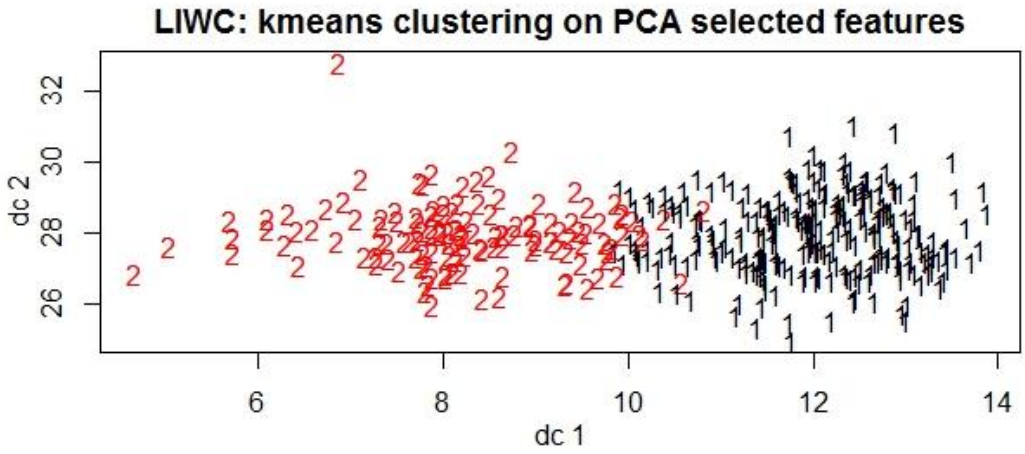


Figure Z.13: Clustering on LIWC variables based on PCA selected features.

LIWC : Boruta selected features	
data3Cluster$withinss	1613.441 2797.589
data3Cluster$betweenss	1693.971
table(data3Cluster$cluster)	1 2 263 145
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 54 209 62% 2 48 97

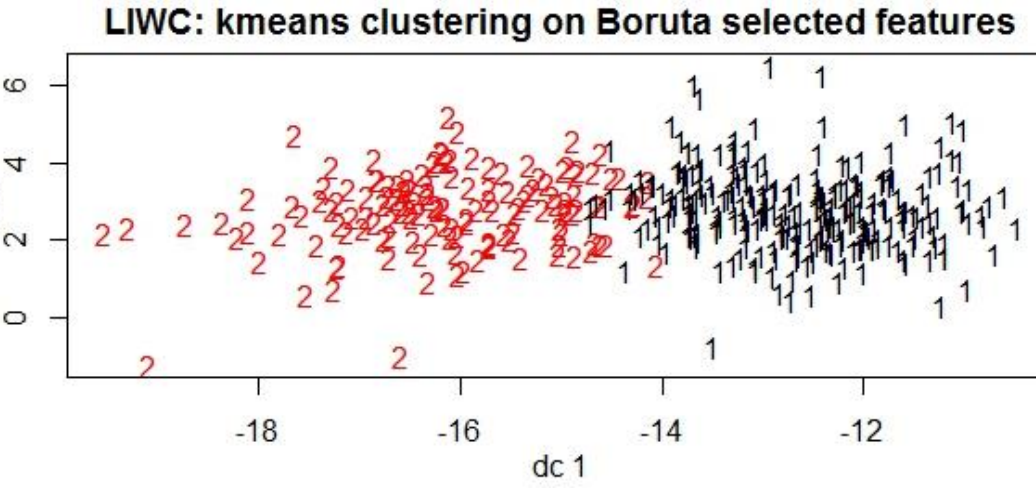


Figure Z.14: Clustering on LIWC variables based on Boruta selected features.

Custom Word List : All features	
data3Cluster$withinss	2891.8579 587.2146
data3Cluster$betweenss	590.9275
table(data3Cluster$cluster)	233 175
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 51 182 57% 2 51 124

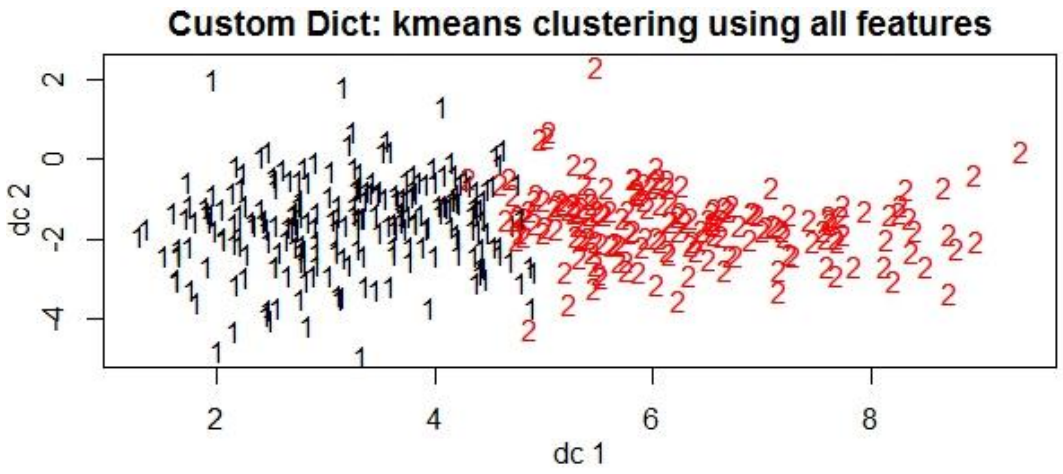


Figure Z.15: Clustering on Custom Dict. features from 'f' and 'nf' reports.

Figure Z.16: Clustering of concepts based on PCA selection.

Concept: Boruta selected features	
data3Cluster$withinss	5158.524 5429.679
data3Cluster$betweenss	1621.797
table(data3Cluster$cluster)	1 2 127 281
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 43 84 65% 2 59 222

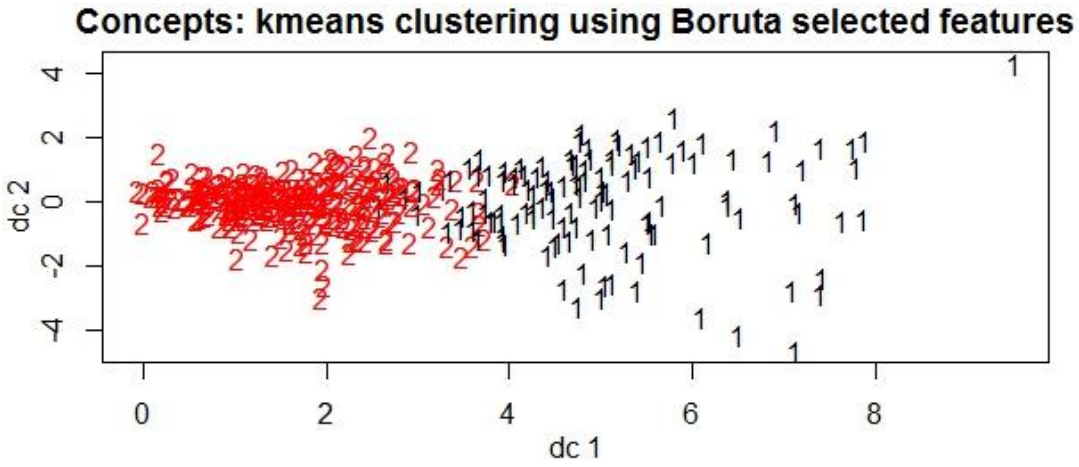
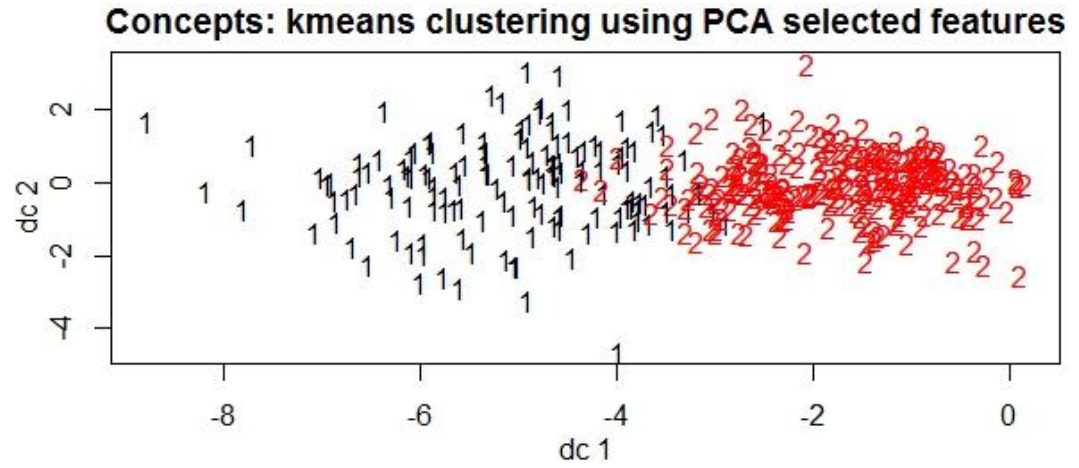


Figure Z.17: Clustering on concepts based on Boruta selected features.

Concept: PCA selected features	
data3Cluster$withinss	9077.678 5215.314
data3Cluster$betweenss	6057.009
table(data3Cluster$cluster)	1 2 145 263
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 40 105 59% 2 62 201



Concept: IG selected features	
data3Cluster$withinss	743.087 9917.04
data3Cluster$betweenss	1142.871
table(data3Cluster$cluster)	1 2 6 402
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 6 0 76% 2 96 306

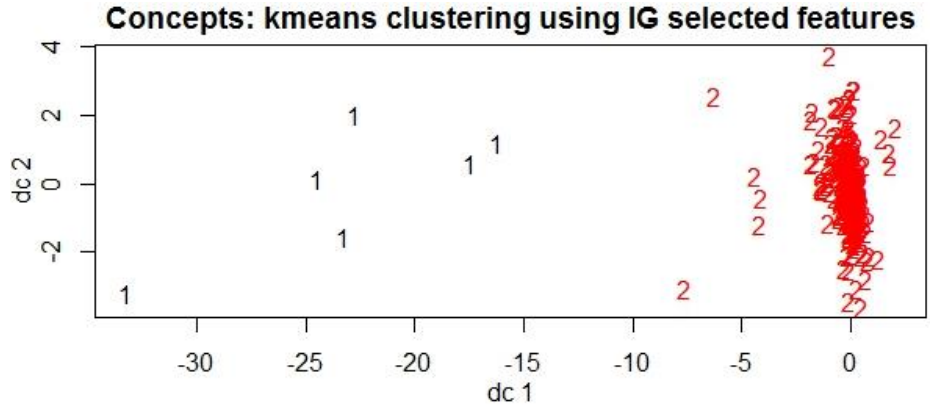


Figure Z.18: Clustering on concepts based on IG selected features.

Keywords: Boruta selected features	
data3Cluster$withinss	2607.481 6561.192
data3Cluster$betweenss	1413.327
table(data3Cluster$cluster)	1 2 101 307
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 1 1 0 99% 2 1 306

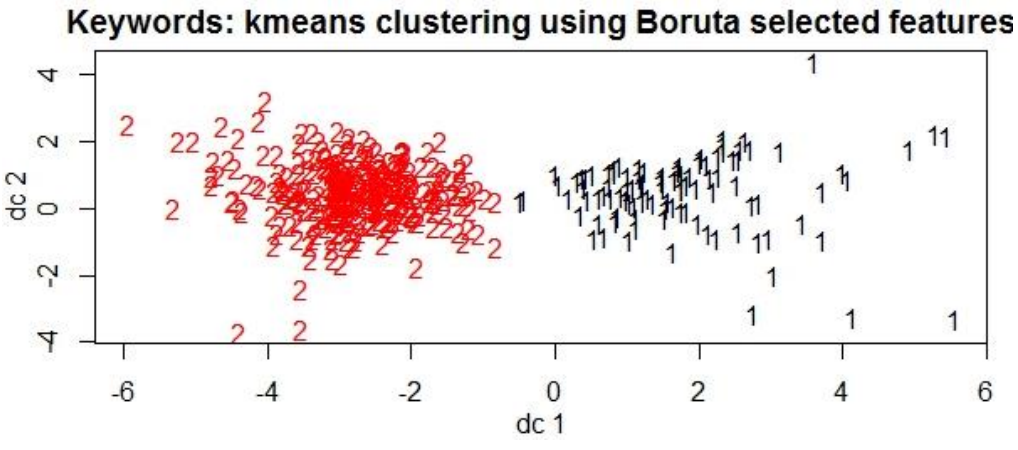


Figure Z.19: Clustering on keywords based on Boruta selected features.

Keywords: PCA selected features	
data3Cluster$withinss	7428.381 10083.747
data3Cluster$betweenss	2837.872
table(data3Cluster$cluster)	1 2 249 159
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 100 149 63% 2 2 157

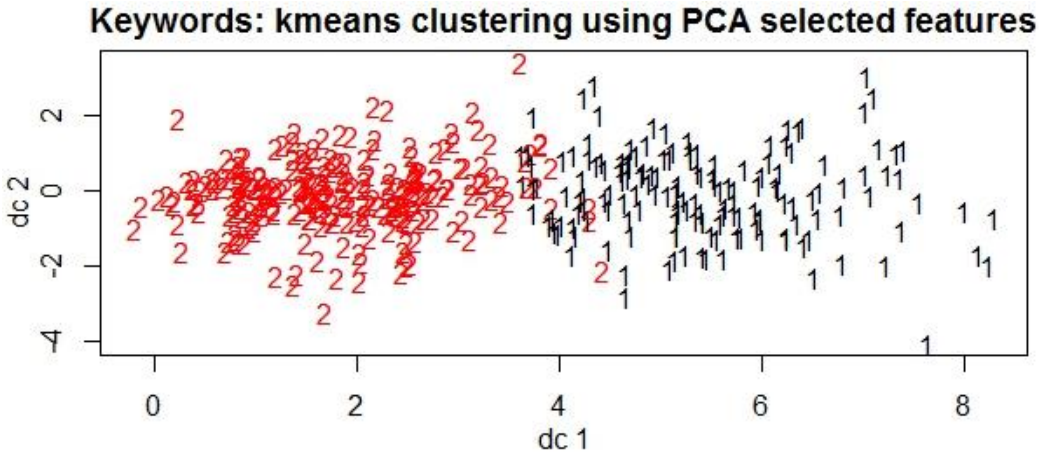


Figure Z.20: Clustering on keywords based on PCA selected features.

.

Keywords Rutherford: PCA selected features	
data3Cluster$withinss	13074.708 4666.405
data3Cluster$betweenss	2608.887
table(data3Cluster$cluster)	1 2 294 114
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 0 294 97% 2 102 12

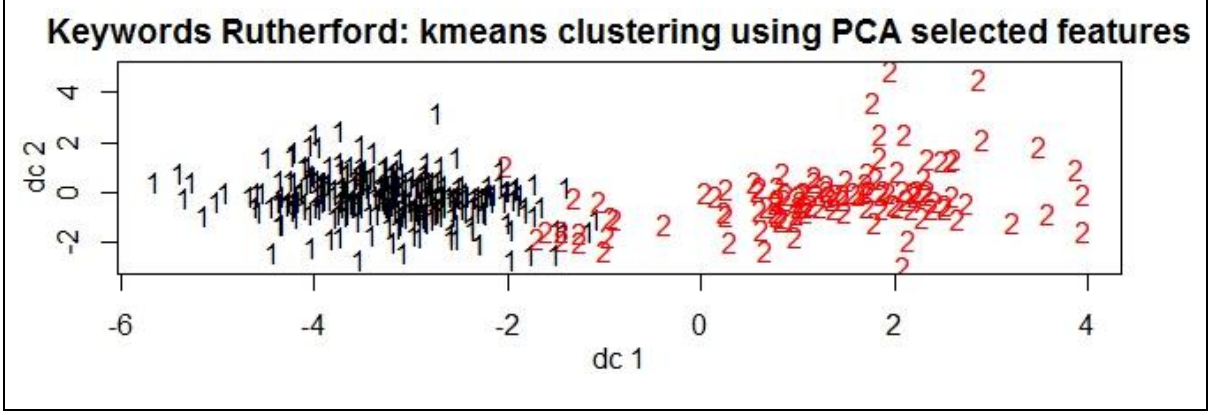


Figure Z.21: Clustering on keywords (Rutherford) based on PCA selected features.

Keywords Rutherford: IG selected features	
data3Cluster$withinss	2304.511 7388.122
data3Cluster$betweenss	2517.367
table(data3Cluster$cluster)	1 2 102 306
table(data3Cluster$cluster, data2$class)	f nf Correctly Clustered 1 102 0 100% 2 0 306

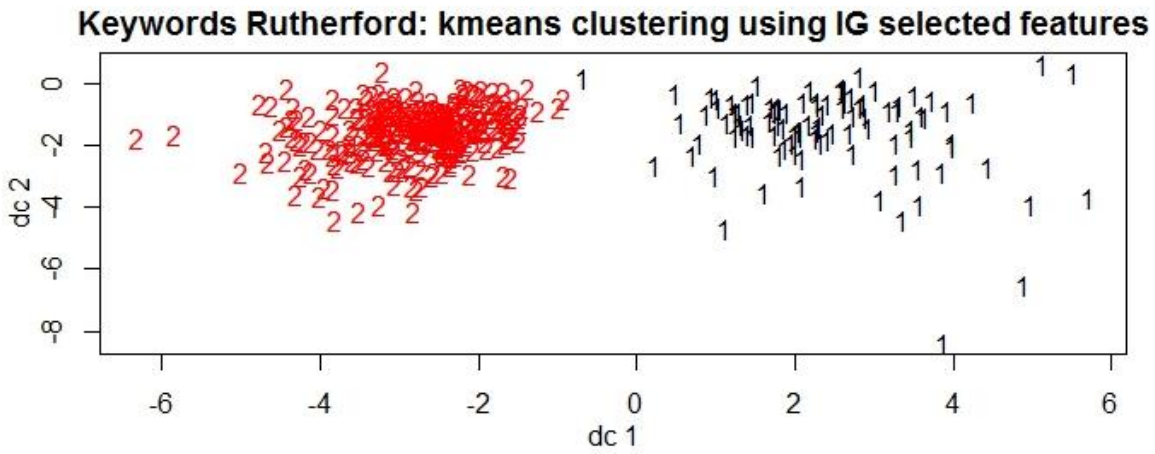


Figure Z.22: Clustering on keywords (Rutherford) based on IG selected features.