SAND2009-1347P

Latent Semantic Analysis and Classification Modeling in Applications for Social Movement Theory

Judy Spomer

March 6, 2009
Central Connecticut State University

Purpose of This Study

Develop statistical models to accurately classify text documents that are intended to influence the reader.

Outline

1. Overview of Social Movement Theory
    a) Framing Process
2. Global Warming Corpus
3. Text Preprocessing
    a) Term-Document Matrix
    b) Singular Value Decomposition and Latent Semantic Analysis
4. Exploratory Data Analysis
5. Preparation for Classification Modeling
6. Evaluation Metrics
7. Model 1: Framing vs. Non-Framing Classification
8. Model 2: Non-Framing vs. Diagnostic vs. Prognostic vs. Motivational Classification
9. Conclusions and Future Work

Overview of Social Movement Theory

Social Movement Theory (SMT)

An area of study in Social Science and Political Science that provides an analytical framework for understanding the factors involved in organized social action. [1,2]

Framing: The method by which an individual organizes and categorizes events, situations, and personal experiences. [3]

Framing Process: A key element of SMT, whereby communications are prepared with intent to influence perceptions and enlist help from others in order to address a social problem.

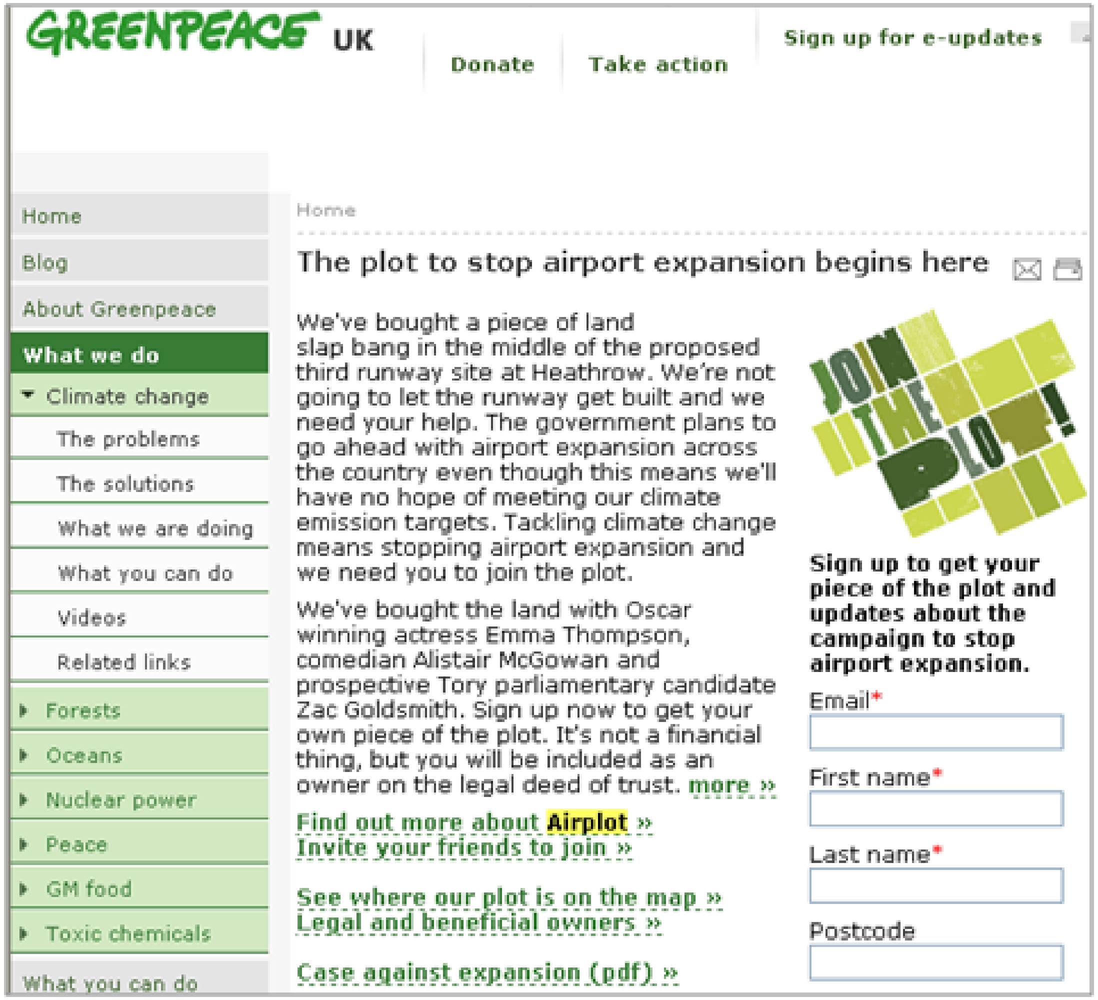
Frames that promote joining together with others to take action on a social issue are known as Collective Action Frames(CAF). The CAF process can be broken into three key tasks [4]:

1. ***Diagnostic:*** defines the problem, often places blame, and may describe how innocent victims are affected;

2. ***Prognostic:*** presents solutions or steps to resolve the issue; and

3. ***Motivational***: states an urgent need for action to address the problem, and invites others to join in ameliorative collective social actions.

Social Movement: Global Warming



From Greenpeace UK website, http://www.greenpeace.org.uk/climate/airplot, viewed February 2, 2009. Used with permission.

Example Diagnostic Text

Problem

No new coal – Stop Kingsnorth. In April 2008 the government will decide whether Kingsnorth in Kent will have the first new coal-fired power station in the UK for decades. Of all fuels, coal is the most polluting - even worse than burning oil or gas. Kingsnorth power station alone will release more CO2 each year than Ghana. It will not use carbon capture and storage technology, and so will contribute to climate change that is already hitting the world's poor first and hardest. For the UK to be encouraging the development of new coal-fired power stations, instead of promoting the switch to a low carbon future, is madness in an era of impending climate crisis. [6]

Example Diagnostic Text

Blame

No new coal – Stop Kingsnorth. In April 2008 the government will decide whether Kingsnorth in Kent will have the first new coal-fired power station in the UK for decades. Of all fuels, coal is the most polluting - even worse than burning oil or gas. Kingsnorth power station alone will release more CO2 each year than Ghana. It will not use carbon capture and storage technology, and so will contribute to climate change that is already hitting the world's poor first and hardest. For the UK to be encouraging the development of new coal-fired power stations, instead of promoting the switch to a low carbon future, is madness in an era of impending climate crisis. [6]

Example Diagnostic Text

Victims

No new coal – Stop Kingsnorth. In April 2008 the government will decide whether Kingsnorth in Kent will have the first new coal-fired power station in the UK for decades. Of all fuels, coal is the most polluting - even worse than burning oil or gas. Kingsnorth power station alone will release more CO2 each year than Ghana. It will not use carbon capture and storage technology, and so will contribute to climate change that is already hitting the world's poor first and hardest. For the UK to be encouraging the development of new coal-fired power stations, instead of promoting the switch to a low carbon future, is madness in an era of impending climate crisis. [6]

Example Prognostic Text

Solutions

Reduce emissions to avoid dangerous global warming: Scientists tell us that we must cut greenhouse gas emissions by at least 80% by 2050 to prevent global temperatures from rising more than 2º C over pre-industrial averages. Not only must global warming policy require such emissions reductions, but it must also ensure the U.S. adheres to this mandate by requiring periodic scientific review of progress toward sufficient emission reductions that will meet this goal. Legislation should direct EPA to adjust its regulatory process based on future scientific study and review of climate change to ensure that we meet measurable, intermittent emission reduction benchmarks between now and 2050 that will prevent a rise in global temperatures above dangerous levels. [7]

Example Motivational Text

Call to Action

Welcome to Climate Camp Australia. The camp for climate action will be five days of inspiring workshops & direct action aimed at shutting down the world's largest coal port in Newcastle, just north of Sydney. If you are concerned about climate change, and want real action instead of more hot air, then we encourage you to come, bring your friends and family and get involved. Whether you are old or young, a seasoned protestor or if you've never been to a protest in your life, if you share our passion for climate action, then climate camp is for you! We'd love for you to get involved and help make the camp as big, bold and effective as possible. Whatever your background, there is a role for you. Find out more about how you can get involved. [8]

Example Motivational Text

Invite Others

Welcome to Climate Camp Australia. The camp for climate action will be five days of inspiring workshops & direct action aimed at shutting down the world's largest coal port in Newcastle, just north of Sydney. If you are concerned about climate change, and want real action instead of more hot air, then we encourage you to come, bring your friends and family and get involved. Whether you are old or young, a seasoned protestor or if you've never been to a protest in your life, if you share our passion for climate action, then climate camp is for you! We'd love for you to get involved and help make the camp as big, bold and effective as possible. Whatever your background, there is a role for you. Find out more about how you can get involved. [8]

Global Warming Corpus

Global Warming Corpus:
6,531 Documents

Non-Framing: Abstracts from technical papers, conference presentations, and reviews.

Framing: Texts were gathered from web sites that support various social movements focused on the global warming issue.

Value	Proportion	%	Count
Framing		9.32	609
Non-Framing		90.68	5922

Value	Proportion	%	Count
Diagnostic		1.85	121
Prognostic		3.09	202
Motivational		4.38	286
Non-Framing		90.68	5922

Text Preprocessing

Text Preprocessing

1. ***Removal of Personal Identifying Information***
2. ***Document Classification***
   a) Non-Framing vs. Framing
   b) Non-Framing vs. Diagnostic vs. Prognostic vs. Motivational
3. ***Parsing the Text***
   a) Extract Terms and Noun Phrases
4. ***Part of Speech Tagging***
   a) Noun, Proper Noun, Verb, Adjective, Adverb, etc.
5. ***Stemming***
   a) Verbs and Nouns
6. ***Removal of Selected Terms***
   a) Non-Informative Parts of Speech: Conjunction, Preposition, Pronoun, Participle, etc.
   b) Stop Words: the, it, either, this, etc.

Term-Document Matrix

1. The sun rose and the sun set.
2. The moon rose.
3. The red rose rises from the rose bush.

	Document 1	Document 2	Document 3
and	1	0	0
bush	0	0	1
from	0	0	1
moon	0	1	0
red	0	0	1
rise (verb)	1	1	1
rose (adj)	0	0	1
rose (noun)	0	0	1
set	1	0	0
sun	2	0	0
the	2	1	2

Term Weighting [9]

$$\hat{a}_{ij} = \log_2(f_{ij} + 1)\left(1 + \sum_j \frac{(f_{ij}/g_i)\log_2(f_{ij}/g_i)}{\log_2(n)}\right)$$

where

f_{ij}	is the frequency of term *i* in document *j*
g_i	is the number of times that term *i* appears in the entire corpus
n	is the number of documents in the corpus

Weighted Term-Document Matrix

1. The sun rose and the sun set.
2. The moon rose.
3. The red rose rises from the rose bush.

	Document 1	Document 2	Document 3
and	0.4910	0.0000	0.0000
bush	0.0000	0.0000	0.4910
from	0.0000	0.0000	0.4910
moon	0.0000	0.4910	0.0000
red	0.0000	0.0000	0.4910
rise (verb)	0.3799	0.3799	0.3799
rose (adj)	0.0000	0.0000	0.4910
rose (noun)	0.0000	0.0000	0.4910
set	0.4910	0.0000	0.0000
sun	0.7782	0.0000	0.0000
the	0.6099	0.3848	0.6099

At this point:

- We have converted unstructured text into a structured format.
- We can represent each document as a vector of term weights.
- We can evaluate similarity between two documents by a method such as the cosine measure of distance between two vectors

But … there are problems:

- For the Global Warming corpus, the term-document matrix is high dimensional: ~23,000 terms by 6,531 documents.
- The term-document matrix is sparse.

Singular Value Decomposition (SVD) [11]

Given the M x N matrix, $\boldsymbol{T}$, of rank, r, there is a singular-value decomposition of $\boldsymbol{T}$ such that

$$\boldsymbol{T} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$$

Where

the eigenvalues $\lambda_1, \ldots, \lambda_r$ of $\boldsymbol{T}\boldsymbol{T}^{\mathsf{T}}$ are the same as the eigenvalues of $\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}$

For $1 \leq i \leq r$, let $\sigma_i = \sqrt{\lambda_i}$ with $\lambda_i \geq \lambda_{i+1}$. Then the M x N matrix $\boldsymbol{D}$ is composed by setting $\boldsymbol{D}_{ii} = \sigma_i$ for $1 \leq i \leq r$, and zero otherwise

Latent Semantic Analysis (LSA)

- **Still Have Problems:** Dimensionality & Synonymy.
- **The Solution is LSA [13]:** A method in text mining that applies a truncated SVD to the term-document matrix.
- **Truncated SVD:** The decomposition is reduced by eliminating k dimensions, beginning with the smallest values in $\boldsymbol{D}$. When the dimensionality is reduced in this manner, the reconstructed matrix, $\boldsymbol{UDV}^{\mathsf{T}}$, is the best rank-$k$ approximation of the original matrix.
- **Problems Solved!**

The V^T matrix: Documents by SVD Dimensions

Document	SVD_1	SVD_2	SVD_3	SVD_4	SVD_5	SVD_6	SVD_7	SVD_8	SVD_9	SVD_10
1	0.2946	-0.2187	0.0017	-0.0780	0.0734	-0.0680	0.1607	-0.0953	0.1435	-0.1967
2	0.3494	-0.3647	-0.0120	-0.0635	0.0585	-0.0540	0.0723	-0.0762	0.0332	-0.0325
3	0.4174	-0.0764	-0.0875	-0.0775	-0.3174	0.1287	0.1561	-0.1388	-0.1270	-0.0987
4	0.3831	-0.3359	-0.0739	-0.0255	0.0562	0.0586	-0.0429	-0.0484	0.1021	-0.1836
5	0.3305	-0.2595	-0.1324	-0.0585	-0.2358	-0.0076	0.1163	-0.2232	-0.1363	-0.0751
6	0.4936	-0.3039	-0.0029	0.1447	0.0061	0.0567	-0.0523	0.1218	0.0443	0.0768
7	0.2203	-0.2960	-0.0518	0.1304	0.0621	0.0429	-0.2700	0.1011	0.0485	-0.1004
8	0.3112	-0.3139	-0.0445	0.1329	0.0954	0.0346	-0.2820	0.1090	0.0620	-0.1081
9	0.5073	-0.3803	-0.0470	0.0461	-0.0868	-0.0205	-0.0545	0.0561	-0.1973	0.1343
10	0.3624	-0.3970	-0.0362	0.1053	0.0265	0.0046	-0.1409	0.1494	-0.1004	0.1582
11	0.2256	-0.2663	-0.0281	0.1466	0.0626	0.0188	-0.3207	0.1443	0.0319	-0.0468
12	0.3599	-0.1056	0.0519	-0.2765	-0.1094	-0.1145	0.1361	0.0959	-0.1363	0.0545
13	0.3470	-0.3089	0.0166	0.0650	0.0273	-0.0836	0.0016	0.1328	-0.0744	0.1511
14	0.4691	-0.1707	0.0826	-0.0965	0.0173	-0.1106	0.1916	0.0464	0.1060	-0.0683
15	0.3107	-0.4260	-0.0753	0.1351	0.0168	0.0276	-0.3293	0.1021	0.0247	-0.0241
16	0.3063	-0.1604	0.0186	-0.1269	-0.1505	-0.1151	0.0212	0.1700	-0.2796	0.2973
17	0.4499	-0.5062	-0.0362	-0.0751	-0.0733	-0.0842	0.1612	-0.1728	0.0144	-0.0660
18	0.2712	-0.3608	-0.0376	0.0237	-0.0361	-0.0634	-0.0010	-0.0545	-0.1639	0.1969
19	0.3698	-0.3011	-0.1071	0.0887	-0.0922	0.0907	-0.1064	-0.1614	0.0715	-0.1469
20	0.5529	-0.0254	-0.0320	0.0111	0.0402	0.2010	0.1600	-0.1268	0.1503	-0.0423

The *U* matrix: Terms by SVD Dimensions

Term	POS	SVD_1	SVD_2	SVD_3	SVD_4	SVD_5	SVD_6	SVD_7	SVD_8	SVD_9	SVD_10
arborist	Noun	0.0923	0.0638	0.0839	-0.1055	0.0720	-0.1932	-0.0968	-0.0971	0.0139	-0.0991
arc	Noun	0.1503	-0.0547	0.0636	-0.1504	-0.0109	-0.1390	0.0841	0.1707	-0.0003	0.0503
arc	Verb	0.1401	-0.1328	-0.0030	-0.0089	0.0174	-0.0797	0.0436	0.1711	-0.0398	0.1120
archaeological sites	NOUN_GROUP	0.0564	-0.0695	-0.0148	0.0222	0.0396	0.0107	-0.0842	0.0353	0.0301	-0.0302
archaic	Adj	0.1042	0.0513	0.0983	-0.0356	0.0705	-0.1415	0.0149	0.1001	0.1501	0.0157
archeological	Adj	0.1303	-0.1682	-0.0467	0.0419	0.0371	0.0217	-0.1427	0.0122	-0.0005	-0.0667
architect	Noun	0.1877	0.0550	0.0066	-0.0722	0.0096	-0.0256	0.0098	-0.0945	-0.0646	0.0103
architectural	Adj	0.1339	-0.1246	-0.0309	0.0146	0.0206	-0.0315	-0.0632	0.0169	-0.0181	-0.0280
architecture	Noun	0.1556	-0.1144	-0.0348	-0.0390	-0.0672	-0.0019	0.0315	-0.0177	-0.1171	0.0556
architecture	Prop	0.2016	0.0109	-0.0073	-0.1338	-0.2749	-0.0415	0.0188	-0.1290	-0.1837	0.0148
archive	Noun	0.2608	-0.2528	-0.0335	-0.0449	-0.1186	-0.0450	0.0793	0.0658	-0.2441	0.2191
archive	Verb	0.1441	-0.1862	-0.0415	0.0303	-0.0305	0.0033	-0.0370	0.0683	-0.2254	0.3137
arctic	Adj	0.1716	-0.1858	-0.0279	0.0807	0.0179	0.0412	-0.1052	0.0152	0.0223	-0.0308
arctic	Prop	0.2883	-0.2293	-0.0008	0.1852	0.0952	0.0498	-0.1580	0.0841	0.0663	0.1313
arctic biota	NOUN_GROUP	0.1207	-0.1754	-0.0566	0.0107	-0.0442	0.0352	-0.0403	-0.1069	-0.0052	-0.1431
arctic ocean	Prop	0.1428	-0.2386	-0.0411	0.0875	0.0529	0.0026	-0.1357	0.0569	-0.0011	0.0718
area	Noun	0.6187	-0.3814	-0.0792	0.0703	0.0756	0.0272	0.0172	-0.1589	0.1001	-0.1119
area index	NOUN_GROUP	0.1740	-0.2515	-0.0268	-0.0345	0.0692	-0.0239	0.0812	-0.0917	0.1150	-0.1642
areal	Adj	0.1774	-0.2783	-0.0279	0.1286	0.0849	0.0226	-0.1926	0.0606	-0.0391	0.1142
areal extent	NOUN_GROUP	0.0865	-0.1453	-0.0184	0.0752	0.0481	0.0213	-0.1384	0.0186	-0.0255	0.0279

Exploratory Data Analysis

Exploratory Data Analysis

The documents in the entire corpus were clustered using SVD dimension values as inputs. Clusters were profiled and named.

Proportion of Framing and Non-Framing Documents in each cluster

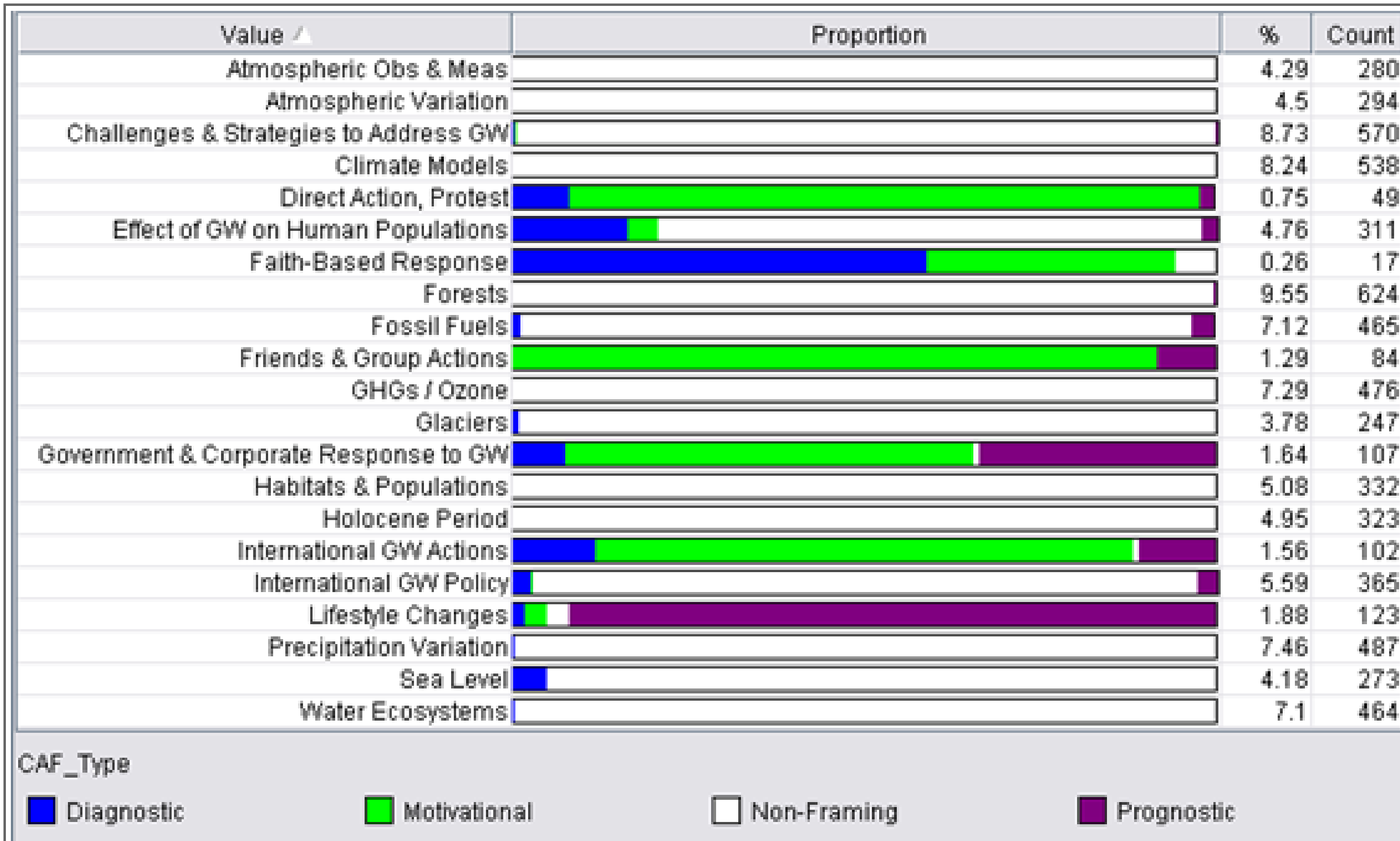
Value	Proportion	%	Count
Atmospheric Obs & Meas		4.29	280
Atmospheric Variation		4.5	294
Challenges & Strategies to Address GW		8.73	570
Climate Models		8.24	538
Direct Action, Protest		0.75	49
Effect of GW on Human Populations		4.76	311
Faith-Based Response		0.26	17
Forests		9.55	624
Fossil Fuels		7.12	465
Friends & Group Actions		1.29	84
GHGs / Ozone		7.29	476
Glaciers		3.78	247
Government & Corporate Response to GW		1.64	107
Habitats & Populations		5.08	332
Holocene Period		4.95	323
International GW Actions		1.56	102
International GW Policy		5.59	365
Lifestyle Changes		1.88	123
Precipitation Variation		7.46	487
Sea Level		4.18	273
Water Ecosystems		7.1	464

Framing_flag

Framing   Non-Framing

Exploratory Data Analysis

Proportion of Non-Framing, Diagnostic, Prognostic, and Motivational Documents in each cluster



Preparation for Classification Modeling

Training and Test Data Sets

- The corpus of documents was randomly split into a training data set of 4,358 documents and a test data set of 2,173 documents.
- Random selection was within document class in order to maintain class proportions for both data sets.

Training Data Set

Value	Proportion	%	Count
Diagnostic		2.04	89
Prognostic		3.03	132
Motivational		4.02	175
Non-Framing		90.91	3962

Test Data Set

Value	Proportion	%	Count
Diagnostic		1.47	32
Prognostic		3.22	70
Motivational		5.11	111
Non-Framing		90.2	1960

Scoring the Test Data Set

- Training Data Set was parsed and SVD performed without influence of Test Data Set.

- In order to validate the models, the Test Data Set was subsequently parsed and "folded into" the LSA space to obtain SVD values [12]. Each test document vector, $\boldsymbol{t}$, is mapped into the k-dimensional LSA space by:

$$t_k = D_k^{-1} U_k^T t$$

Defining Dummy Variables

Bivariate Analysis of SVD_23 for Non-Framing (NF) vs. Framing (F) Classification

NF	F	% of NF	% of F	5% Interval	Ratio Neg.	Ratio Neutral	Ratio Pos.	Dummy Variable Range
41	57	2.64%	14.39%	LOW -< -.143			5.45	1
74	22	4.77%	5.55%	-.143 -< -.105			1.16	
77	20	4.96%	5.05%	-.105 -< -.084		1.02		N
88	11	5.67%	2.77%	-.084 -< -.067	-2.04			2
86	12	5.54%	3.03%	-.067 -< -.055	-1.83			
83	16	5.35%	4.04%	-.055 -< -.046	-1.32			
83	11	5.35%	2.77%	-.046 -< -.036	-1.93			
83	14	5.35%	3.53%	-.036 -< -.028	-1.51			
83	15	5.35%	3.78%	-.028 -<-.017	-1.41			
79	18	5.09%	4.54%	-.017 -< -.007	-1.12			
83	15	5.35%	3.78%	-.007 -< .002	-1.41			
85	13	5.48%	3.28%	.002 -< .012	-1.67			
86	10	5.54%	2.52%	.012 -< .021	-2.20			
83	15	5.35%	3.78%	.021 -< .030	-1.41			
85	14	5.48%	3.53%	.030 -< .042	-1.55			
82	13	5.28%	3.28%	.042 -< .055	-1.61			
77	21	4.96%	5.30%	.055 -< .070		1.07		N
76	21	4.90%	5.30%	.070 -< .089		1.08		
72	26	4.64%	6.56%	.089 -< .113			1.41	3
45	52	2.90%	13.13%	.113 - HIGH			4.53	4

Note. There are 1,551 non-framing documents and 396 framing documents.

Evaluation Metrics

Evaluation Metrics for Dichotomous Model

Four measures: precision, recall, F_1 measure, and accuracy are often used to evaluate models that deal with text with a dichotomous target variable. [12]

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F_1 = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Number\ of\ Documents}$$

Evaluation Metrics for Polychotomous Model

- Precision, recall, F_1 measure, and accuracy are calculated for each class.

  Example: Motivational vs. Non-Motivational.

- Overall precision, recall, F_1 measure, and accuracy are calculated by macro-averaging. [20]

  Example: Overall precision = (Non-Framing precision + Diagnostic precision + Prognostic precision + Motivational precision) divided by 4.

Model 1: Framing vs. Non-Framing Classification

Purpose and Methods

Classify documents into one of two classes:

1. Framing
2. Non-Framing

Modeling Algorithms:

1. Classification and Regression Tree (CART)
2. Logistic Regression
3. Neural Network
4. Combination of Models

Model 1: CART

CART Model 1 Dummy Variables Confusion Matrix

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	196	17	213
Non-Framing	30	1,930	1,960
Total	226	1,947	2,173

CART Dummy Variables Evaluation Metrics

Precision	0.8673
Recall	0.9202
F_1 Measure	0.8929
Accuracy	0.9784

CART Model 1 SVD Variables Confusion Matrix

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	208	5	213
Non-Framing	107	1,853	1,960
Total	315	1,858	2,173

CART SVD Variables Evaluation Metrics

Precision	0.6603
Recall	0.9765
F_1 Measure	0.7879
Accuracy	0.9485

Model 1: Logistic Regression

Logistic Regression Model 1 Confusion Matrix

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	203	10	213
Non-Framing	44	1,916	1,960
Total	247	1,926	2,173

Logistic Regression Evaluation Metrics

Precision	0.8219
Recall	0.9531
F_1 Measure	0.8826
Accuracy	0.9751

Model 1: Neural Network

Neural Network Model 1 Confusion Matrix

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	206	7	213
Non-Framing	3	1,957	1,960
Total	209	1,964	2,173

Neural Network Evaluation Metrics

Precision	0.9856
Recall	0.9671
F_1 Measure	0.9763
Accuracy	0.9954

Model 1: Voting Models

Confusion Matrix Voting Model 1a
1 or More Models = "Framing"

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	209	4	213
Non-Framing	52	1,908	1,960
Total	261	1,912	2,173

Confusion Matrix Voting Model 1b
2 or More Models = "Framing"

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	203	10	213
Non-Framing	23	1,937	1,960
Total	226	1,947	2,173

Confusion Matrix Voting Model 1c
All 3 Models = "Framing"

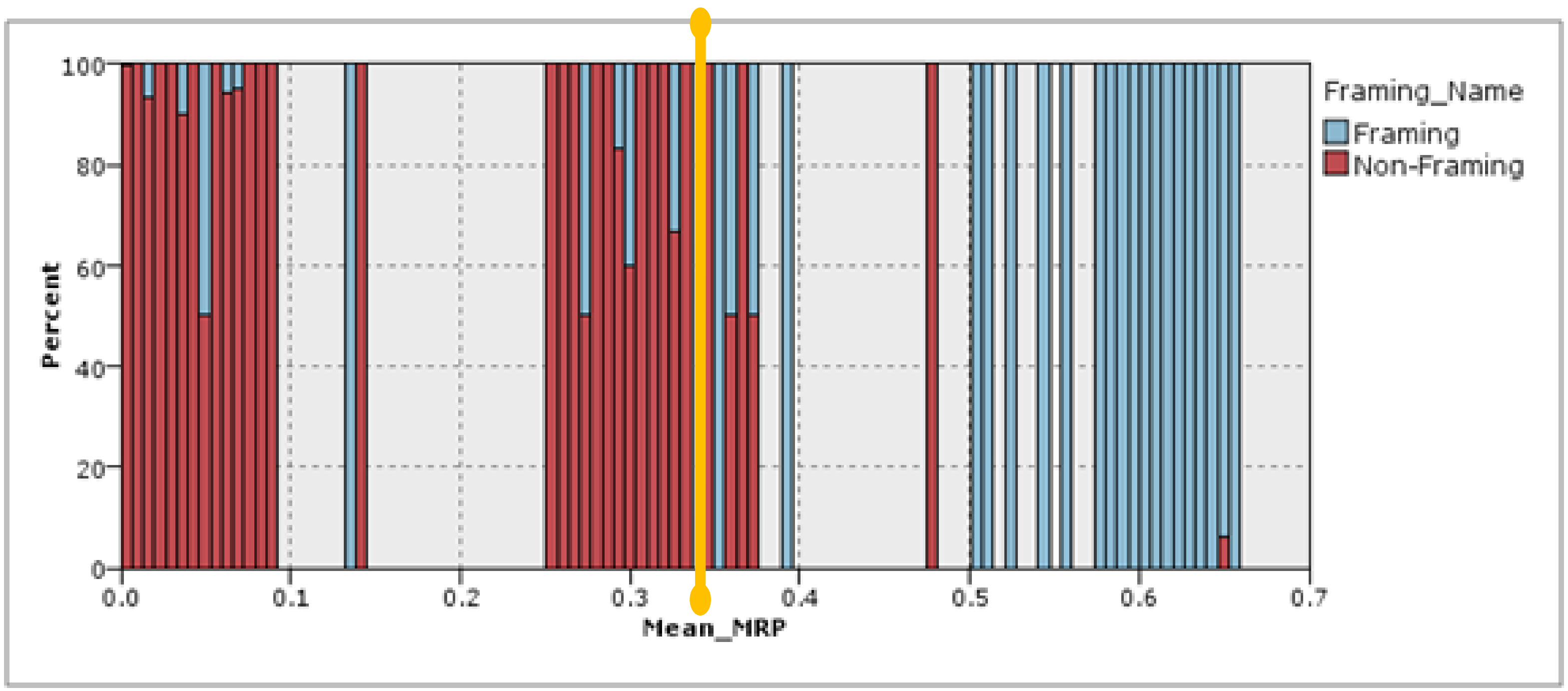
True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	193	20	213
Non-Framing	2	1,958	1,960
Total	195	1,978	2,173

Model 1: Voting Models

Evaluation Metrics	Voting 1a	Voting 1b	Voting 1c
Precision	0.8008	0.8982	0.9897
Recall	0.9812	0.9531	0.9061
F_1 Measure	0.8819	0.9248	0.9461
Accuracy	0.9742	0.9848	0.9899

Model 1: Mean Model Response Probability (MMRP) Model

$$MMRP = \frac{\sum(CART\ MRP, LogReg\ MRP, NNMRP)}{3} \qquad [14]$$



Model 1: Mean Model Response Probability (MMRP) Model

MMRP Model 1 Confusion Matrix

True Classification	Model Classification: Framing	Model Classification: Non-Framing	Total
Framing	198	15	213
Non-Framing	7	1,953	1,960
Total	205	1,968	2,173

MMRP Model 1
Evaluation Metrics

Precision	0.9659
Recall	0.9296
F_1 Measure	0.9474
Accuracy	0.9899

Model 1 Selection

Model 1 Candidates by Decreasing Accuracy

Model	Precision	Recall	F1 Measure	Accuracy
Neural Network	0.9856	0.9671	0.9763	0.9954
Mean MRP	0.9659	0.9296	0.9474	0.9899
Voting 1c	0.9897	0.9061	0.9461	0.9899
Voting 1b	0.8982	0.9531	0.9248	0.9848
CART (Dummy Variables)	0.8673	0.9202	0.8929	0.9784
Logistic Regression	0.8219	0.9531	0.8826	0.9751
Voting 1a	0.8008	0.9812	0.8819	0.9742

Model 2:
Non-Framing vs. Diagnostic vs. Prognostic vs. Motivational Classification

Purpose and Methods

Classify documents into one of four classes:

1. Non-Framing
2. Diagnostic
3. Prognostic
4. Motivational

Modeling Algorithms:

1. Classification and Regression Tree (CART) with Neural Network Model 1
2. Logistic Regression with Neural Network Model 1
3. Neural Network
4. Combination of Models

Model 2: CART

STEP 1:
A CART model was trained to classify just framing documents by framing task using dummy variables.

STEP 2:
The CART model was combined with Neural Network Model 1

CART Model 2 Confusion Matrix

True Classification	Model Classification: Non-Framing	Diagnostic	Prognostic	Motivational	Total
Non-Framing	1,935	10	13	2	1,960
Diagnostic	2	19	2	9	32
Prognostic	3	6	48	13	70
Motivational	0	7	5	99	111
Total	1,940	42	68	123	2,173

CART Model 2 Evaluation Metrics

Evaluation Metric	Non-Framing	Diagnostic	Prognostic	Motivational	Macro-Average
Precision	0.9974	0.4524	0.7059	0.8049	0.7401
Recall	0.9872	0.5938	0.6857	0.8919	0.7897
F1 Measure	0.9923	0.5135	0.6957	0.8462	0.7619
Accuracy	0.9862	0.9834	0.9807	0.9834	0.9834

Model 2: Logistic Regression

STEP 1:
A logistic regression model was trained to classify just framing documents by framing task using dummy variables.

STEP 2:
The logistic regression model was combined with Neural Network Model 1

Logistic Regression Model 2 Confusion Matrix

True Classification	Model Classification: Non-Framing	Diagnostic	Prognostic	Motivational	Total
Non-Framing	1,940	6	11	3	1,960
Diagnostic	2	18	1	11	32
Prognostic	4	3	50	13	70
Motivational	0	3	2	106	111
Total	1,946	30	64	133	2,173

Logistic Regression Model 2 Evaluation Metrics

Evaluation Metric	Non-Framing	Diagnostic	Prognostic	Motivational	Macro-Average
Precision	0.9969	0.6000	0.7813	0.7970	0.7938
Recall	0.9898	0.5625	0.7143	0.9550	0.8054
F1 Measure	0.9933	0.5806	0.7463	0.8689	0.7973
Accuracy	0.9880	0.9880	0.9844	0.9853	0.9864

Model 2: Neural Network

Neural Network Model 2 Confusion Matrix

True Classification	Model Classification: Non-Framing	Diagnostic	Prognostic	Motivational	Total
Non-Framing	1,954	2	3	1	1,960
Diagnostic	4	20	2	6	32
Prognostic	5	4	45	16	70
Motivational	0	2	2	107	111
Total	1,963	28	52	130	2,173

Neural Network Model 2 Evaluation Metrics

Evaluation Metric	Non-Framing	Diagnostic	Prognostic	Motivational	Macro-Average
Precision	0.9954	0.7143	0.8654	0.8231	0.8495
Recall	0.9969	0.6250	0.6429	0.9640	0.8072
F1 Measure	0.9962	0.6667	0.7377	0.8880	0.8221
Accuracy	0.9931	0.9908	0.9853	0.9876	0.9892

Model 2: Voting Model

$$Vote_c = \begin{cases} 1 \;\; if\; (1 * CVote_c + 2 * LVote_c + 3 * NVote_c)/2 \geq 2 \\ 0 \;\; otherwise \end{cases}$$

where

c	is the class (non-framing, diagnostic, prognostic, motivational)
$Vote_c$	is the vote tally for class c
$CVote_c$	is 1 if CART Model 2 classified the observation as c, is 0 otherwise
$LVote_c$	is 1 if Logistic Regression Model 2 classified the observation as c, is 0 otherwise
$NVote_c$	is 1 if Neural Network Model 2 classified the observation as c, is 0 otherwise

Model 2: Voting Model

Voting Model 2 Confusion Matrix

True Classification	Model Classification: Non-Framing	Diagnostic	Prognostic	Motivational	Total
Non-Framing	1,948	5	5	2	1,960
Diagnostic	2	22	2	6	32
Prognostic	4	4	47	15	70
Motivational	0	1	2	108	111
Total	1,954	32	56	131	2,173

Voting Model 2 Evaluation Metrics

Evaluation Metric	Non-Framing	Diagnostic	Prognostic	Motivational	Macro-Average
Precision	0.9969	0.6875	0.8393	0.8244	0.8370
Recall	0.9939	0.6875	0.6714	0.9730	0.8314
F1 Measure	0.9954	0.6875	0.7460	0.8926	0.8304
Accuracy	0.9917	0.9908	0.9853	0.9880	0.9890

Model 2 Selection

Model 2 F_1 Measure and Accuracy

	Document Class	CART 2b	Logistic Regression	Neural Network	Combination
F_1 Measure	Non-Framing	0.9923	0.9933	0.9962	0.9954
	Diagnostic	0.5135	0.5625	0.6667	0.6875
	Prognostic	0.6957	0.7463	0.7377	0.7460
	Motivational	0.8462	0.8689	0.8880	0.8926
Macro-Averaged F_1 Measure		0.7619	0.7973	0.8221	0.8304
Accuracy	Non-Framing	0.9862	0.9880	0.9931	0.9917
	Diagnostic	0.9834	0.9880	0.9908	0.9908
	Prognostic	0.9807	0.9844	0.9853	0.9853
	Motivational	0.9834	0.9853	0.9876	0.9880
Macro-Averaged Accuracy		0.9834	0.9864	0.9892	0.9890

Conclusions and Future Work

Conclusions

1. The accuracy of the methods was excellent.
   a) Dichotomous Models: ranged from 97.5% to 99.5%
   b) Polychotomous models: ranged from 98.3% to 98.9%
2. Latent Semantic Analysis techniques were shown to be effective in providing robust predictor variables for the classification models.
3. Leveraging Social Movement Theory was essential.

Future Work

1. Use this approach to identify "tone" (reasonable or rhetorical) in framing documents, thus singling out those that are more apt to be successful in recruiting [15]

2. Extend capability to identify framing in multiple languages by employing 3-way tensor decomposition

3. Use cross-validation methods to estimate prediction error.

QUESTIONS

References

1. Della Porta, D., & Diani, M. (1999). *Social movements: An introduction.* Oxford: Blackwell Publishers.
2. McAdam, D., McCarthy, J., & Zald, M. (1988). Social movements. In N. Smelser (Ed.), Handbook of sociology. Thousand Oaks, CA: Sage Publications.
3. Goffman, E. (1974). Frame analysis: An essay on the organization of experience. New York, NY: Harper & Row.
4. Snow, D., & Benford, R. (1988). Ideology, frame resonance and participant mobilization. International Social Movement Research , 1, 197-219.
5. Shao, G. (1994). Potential impacts of climate change on a mixed broadleaved-Korean pine forest stand: A gap model approach. *International-Geosphere-Biosphere-Program Workshop on the Application of Forest-Stand-Models-to-Global-Change-Issues.* Apeldoorn Netherlands: Kluwer Academic Publ.
6. World Development Movement. (2008). *No new coal - stop Kingsnorth*. Retrieved May 19, 2008, from World Development Movement Campaigns: http://www.wdm.org.uk/campaigns/climate/action/kingsnorth.htm
7. Sierra Club. (2008). Global warming policy solutions. Retrieved May 14, 2008, from http://www.sierraclub.org/energy/energypolicy/
8. Climate Camp. (2008). *Camp for climate action Australia.* Retrieved May 19, 2008, from http://www.climatecamp.org.au/
9. SAS Institute, Inc. (2003). Weighting methods. Text Miner Node . Cary, NC.

References

11. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes , 25, 259-284.
12. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York, NY: Cambridge University Press.
13. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science , 41 (6), 391-407.
14. Larose, D. (2006). Data mining methods and models. Hoboken, NJ: John Wiley & Sons.
15. Benjamin, D. (2007, December). Finding a reasonable tone. Retrieved January 5, 2009, from FrameWorks Institute: http://www.frameworksinstitute.org/framebytes.html

BACKUP SLIDES

Example Non-Framing Text

A gap-typed forest dynamic model KOPIDE was used to assess the dynamic responses of a mixed broadleaved-Korean pine forest stand to climate change in northeastern China. The GFDL climate change scenario was applied to derive the changes in environmental variables, such as 10 degrees C based DEGD and PET/P, which were used to implement the model. The simulation result suggests that the climate change would cause important changes in stand structure. Korean pine, the dominant species in the area under current climate conditions, would disappear under the GFDL equilibrium scenario. Oak and elm would become the dominant species replacing Korean pine, ash and basswood. Such a potential change in forest structure would require different strategies for forest management in northeastern China. [5]

Example Diagnostic Text

No new coal – Stop Kingsnorth. In April 2008 the government will decide whether Kingsnorth in Kent will have the first new coal-fired power station in the UK for decades. Of all fuels, coal is the most polluting - even worse than burning oil or gas. Kingsnorth power station alone will release more CO2 each year than Ghana. It will not use carbon capture and storage technology, and so will contribute to climate change that is already hitting the world's poor first and hardest. For the UK to be encouraging the development of new coal-fired power stations, instead of promoting the switch to a low carbon future, is madness in an era of impending climate crisis. [6]

Example Prognostic Text

Reduce emissions to avoid dangerous global warming: Scientists tell us that we must cut greenhouse gas emissions by at least 80% by 2050 to prevent global temperatures from rising more than 2º C over pre-industrial averages. Not only must global warming policy require such emissions reductions, but it must also ensure the U.S. adheres to this mandate by requiring periodic scientific review of progress toward sufficient emission reductions that will meet this goal. Legislation should direct EPA to adjust its regulatory process based on future scientific study and review of climate change to ensure that we meet measurable, intermittent emission reduction benchmarks between now and 2050 that will prevent a rise in global temperatures above dangerous levels. [7]

Example Motivational Text

Welcome to Climate Camp Australia. The camp for climate action will be five days of inspiring workshops & direct action aimed at shutting down the world's largest coal port in Newcastle, just north of Sydney. If you are concerned about climate change, and want real action instead of more hot air, then we encourage you to come, bring your friends and family and get involved. Whether you are old or young, a seasoned protestor or if you've never been to a protest in your life, if you share our passion for climate action, then climate camp is for you! We'd love for you to get involved and help make the camp as big, bold and effective as possible. Whatever your background, there is a role for you. Find out more about how you can get involved. [8]

SVD Example [12]

1. We ate.
2. He ate.

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$U = \begin{pmatrix} -0.816 & 0.000 \\ -0.408 & -0.707 \\ -0.408 & 0.707 \end{pmatrix} \quad D = \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 1.000 \end{pmatrix} \quad V^T = \begin{pmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{pmatrix}$$

- The documents are now represented as vectors (dimensions) of values which are not sparse – $\boldsymbol{V}^{\mathrm{T}}$
- The terms are likewise represented as vectors (dimensions) of values - $\boldsymbol{U}$

Model 1: Logistic Regression

NON_FRAMING[a]	B	Std. Error	Wald	df	Sig.	Exp(B)	95.0% Confidence Interval for Exp(B)	
							Lower Bound	Upper Bound
Intercept	-7.798	1.726	20.422	1	0.000			
SVD1_01=0	1.625	0.368	19.520	1	0.000	5.080	2.470	10.448
SVD2_01=0	1.009	0.464	4.736	1	0.030	2.743	1.105	6.804
SVD2_02=0	8.038	0.566	201.561	1	0.000	3095.514	1020.534	9389.402
SVD3_01=0	-1.836	0.410	20.047	1	0.000	0.159	0.071	0.356
SVD5_05=0	1.829	0.551	11.035	1	0.001	6.230	2.117	18.334
SVD5_06=0	2.500	0.575	18.943	1	0.000	12.188	3.953	37.580
SVD6_02=0	-1.224	0.361	11.499	1	0.001	0.294	0.145	0.596
SVD6_03=0	1.390	0.605	5.268	1	0.022	4.013	1.225	13.148
SVD6_05=0	1.700	0.652	6.808	1	0.009	5.475	1.527	19.636
SVD8_04=0	-1.766	0.381	21.453	1	0.000	0.171	0.081	0.361
SVD9_01=0	-1.396	0.535	6.803	1	0.009	0.248	0.087	0.707
SVD11_01=0	1.402	0.417	11.279	1	0.001	4.063	1.793	9.208
SVD12_03=0	-1.952	0.410	22.608	1	0.000	0.142	0.064	0.318
SVD22_01=0	1.341	0.364	13.591	1	0.000	3.823	1.874	7.799

[a]The reference category is 0.

Model 2: Logistic Regression

Motivational Class (Diagnostic is Reference Class)

	CAF_Name[a]	B	Std. Error	Wald	df	Sig.	Exp(B)	95.0% Confidence Interval for Exp(B) Lower Bound	Upper Bound
MOTIVATIONAL	Intercept	0.572	1.385	0.171	1	0.680			
	DPM_SVD2_02=0	-2.309	0.445	26.925	1	0.000	0.099	0.042	0.238
	DPM_SVD3_01=0	-1.530	0.565	7.325	1	0.007	0.217	0.071	0.656
	DPM_SVD5_01=0	-1.315	0.478	7.566	1	0.006	0.269	0.105	0.685
	DPM_SVD5_03=0	1.841	0.701	6.890	1	0.009	6.303	1.594	24.921
	DPM_SVD6_01=0	0.681	0.760	0.803	1	0.370	1.976	0.445	8.769
	DPM_SVD6_03=0	0.682	0.524	1.692	1	0.193	1.977	0.708	5.522
	DPM_SVD8_01=0	-1.819	0.509	12.758	1	0.000	0.162	0.060	0.440
	DPM_SVD8_03=0	1.596	0.621	6.613	1	0.010	4.932	1.462	16.641
	DPM_SVD9_02=0	1.087	0.469	5.377	1	0.020	2.966	1.183	7.434
	DPM_SVD10_01=0	1.506	0.439	11.764	1	0.001	4.511	1.907	10.669
	DPM_SVD11_02=0	-1.510	0.473	10.191	1	0.001	0.221	0.087	0.558
	DPM_SVD27_01=0	-1.248	0.502	6.189	1	0.013	0.287	0.107	0.767

Model 2: Logistic Regression

Prognostic Class (Diagnostic is Reference Class)

	CAF_Name[a]	B	Std. Error	Wald	df	Sig.	Exp(B)	95.0% Confidence Interval for Exp(B) Lower Bound	Upper Bound
PROGNOSTIC	Intercept	0.544	1.329	0.167	1	0.682			
	DPM_SVD2_02=0	-0.407	0.466	0.761	1	0.383	0.666	0.267	1.661
	DPM_SVD3_01=0	-1.580	0.662	5.694	1	0.017	0.206	0.056	0.754
	DPM_SVD5_01=0	-1.596	0.523	9.320	1	0.002	0.203	0.073	0.565
	DPM_SVD5_03=0	0.689	0.624	1.217	1	0.270	1.991	0.586	6.767
	DPM_SVD6_01=0	-3.088	0.622	24.617	1	0.000	0.046	0.013	0.154
	DPM_SVD6_03=0	1.722	0.564	9.335	1	0.002	5.598	1.854	16.900
	DPM_SVD8_01=0	-0.440	0.541	0.663	1	0.416	0.644	0.223	1.859
	DPM_SVD8_03=0	1.090	0.616	3.127	1	0.077	2.974	0.889	9.956
	DPM_SVD9_02=0	1.711	0.534	10.258	1	0.001	5.534	1.942	15.767
	DPM_SVD10_01=0	0.760	0.460	2.739	1	0.098	2.139	0.869	5.265
	DPM_SVD11_02=0	-0.372	0.482	0.597	1	0.440	0.689	0.268	1.773
	DPM_SVD27_01=0	0.113	0.555	0.041	1	0.839	1.119	0.377	3.325