Exploration of Spam Reviews Detection

Yu Huang, Lingjie Zhang

Motivation

Consumers increasingly rate, review and research products and services online.

Motivation

http://myleott.com/prevalence_WWW2012_slides.pdf

Challenges

- **Recap:**
  - **Disruptive Opinion Spam**: uncontroversial
    - Content-based filtering
  - **Deceptive Opinion Spam**: hard to tell
    - Psycholinguistic failure
    - Unreliable human performance

Challenges

- **Limited qualified information**
  - User ID: grouped fake reviewers?
    - IP: no access
- **No labeled data**
  - Evaluation

Data Set → Classification → Human Evaluation

Overview

- Motivation
- Research Challenges
- Labeled data set
- Research questions and methodology
- Results and analysis
- Limitation

Labeled Data Set:

Deceptive Opinion Spam Corpus v1.4

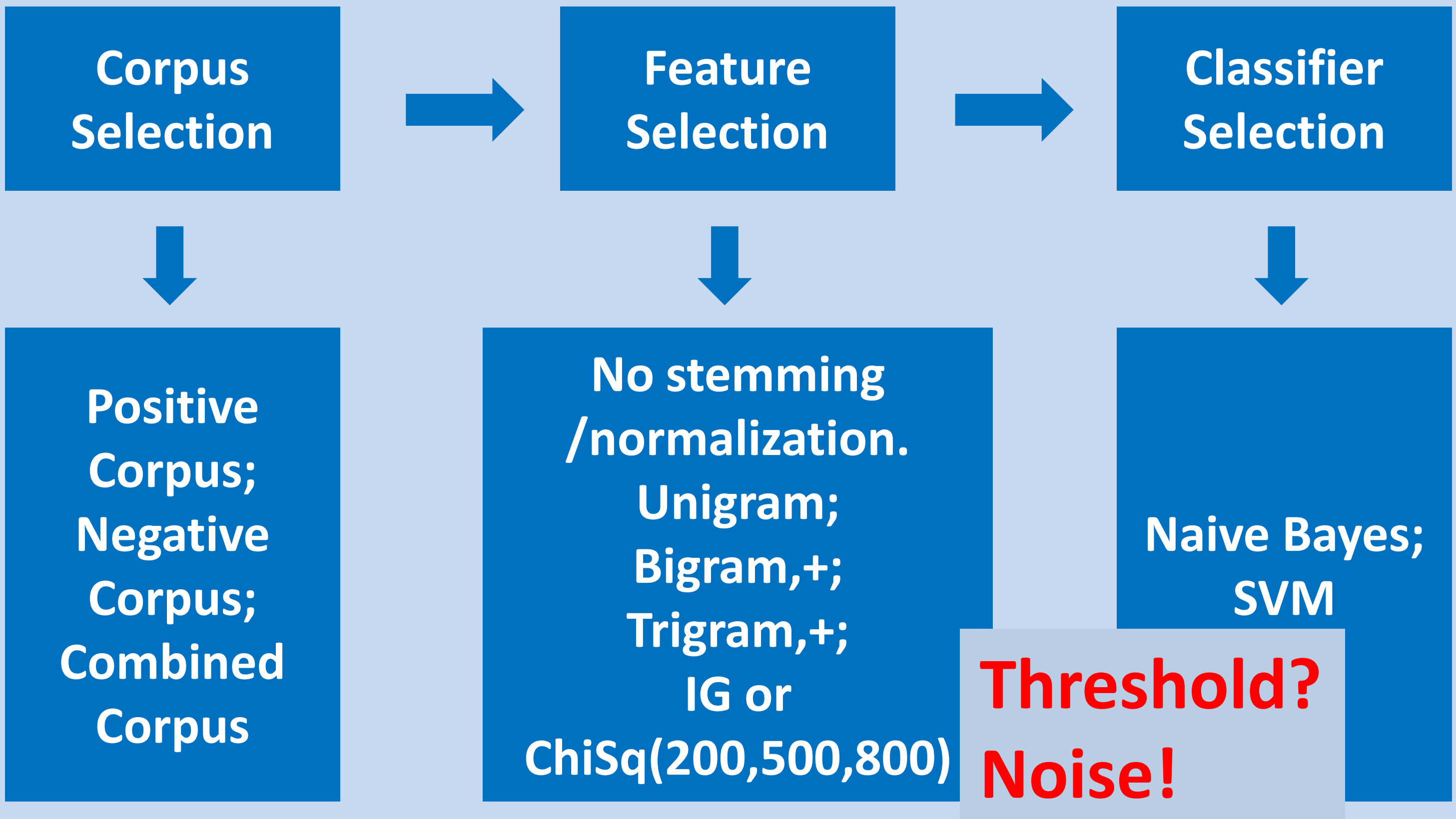
Previous work

	Positive	**Negative**
Truthful	Truthful: 400 TripAdvisor	Truthful: 400 Expedia, Hotels.com, Orbitz, Priceline
Deceptive	Deceptive: 400 Mechanical Turk	Deceptive: 400 Mechanical Turk

Research Questions

- Traditional preprocessing methods?
  - Stopwords removal
  - Normalization
  - Stemming
- Sentiment effects?
  - Positive corpus
  - Negative corpus
  - No previous sentiment information: complete corpus
- Feature selection and classifier model selection?

Methodology: experiment sets



Results Analysis

- Positive Corpus

Chi Square			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
NB	Unigram	200	0.861	0.823	0.841	0.83	0.868	0.848
		500	0.872	0.833	0.852	0.84	0.878	0.858
		800	0.859	**0.838**	0.848	**0.841**	0.863	0.852
	Bigram	200	0.856	0.82	0.838	0.827	0.863	0.845
	Bigram+	200	0.872	0.818	0.844	0.828	0.88	0.853
		500	**0.879**	0.833	**0.855**	0.841	0.885	**0.862**
		800	0.878	0.83	0.853	0.839	0.885	0.861
	Trigram	200	0.852	0.808	0.829	0.817	0.86	0.838
	Trigram+	200	0.762	0.81	0.785	0.797	0.748	0.772
		500	0.87	0.835	0.852	**0.841**	0.875	0.858
		800	0.876	0.83	0.852	0.838	**0.885**	0.86
SVM	Unigram	200	0.876	**0.845**	**0.86**	**0.85**	0.88	**0.865**
		500	0.855	0.84	0.847	0.843	0.858	0.85
		800	0.835	0.835	0.835	0.835	0.835	0.835
	Bigram	200	0.762	0.81	0.785	0.797	0.748	0.772
	Bigram+	200	0.856	0.82	0.838	0.827	0.863	0.845
		500	0.84	0.84	0.84	0.84	0.84	0.84
		800	**0.879**	0.833	0.855	0.841	**0.885**	0.862
	Trigram	200	0.846	0.823	0.834	0.827	0.85	0.838
	Trigram+	200	0.846	0.823	0.834	0.827	0.85	0.838
		500	0.841	0.83	0.835	0.832	0.843	0.837
		800	0.846	0.823	0.834	0.827	0.85	0.838

Information Gain			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
NB	Unigram	200	0.862	0.828	0.844	0.834	0.868	0.85
		500	0.869	0.833	0.851	0.839	0.875	0.857
		800	0.854	**0.835**	0.845	0.839	0.858	0.848
	Bigram	200	0.748	0.788	0.767	0.776	0.735	0.755
	Bigram+	200	0.869	0.813	0.84	0.824	0.878	0.85
		500	0.873	0.828	0.85	0.836	0.88	0.857
		800	**0.876**	0.833	**0.854**	**0.84**	**0.883**	**0.861**
	Trigram	200	0.609	0.785	0.686	0.697	0.495	0.579
	Trigram+	200	0.852	0.818	0.834	0.825	0.858	0.841
		500	0.871	0.83	0.85	0.838	0.878	0.857
		800	0.874	0.83	0.851	0.838	0.88	0.859
SVM	Unigram	200	0.882	0.838	**0.859**	**0.845**	**0.888**	**0.866**
		500	0.857	**0.84**	0.848	0.843	0.86	0.851
		800	0.838	0.825	0.831	0.828	0.84	0.834
	Bigram	200	0.769	0.79	0.779	0.784	0.763	0.773
	Bigram+	200	0.841	0.82	0.83	0.824	0.845	0.835
		500	0.837	0.823	0.83	0.826	0.84	0.833
		800	0.876	0.833	0.854	0.84	0.883	0.861
	Trigram	200	0.695	0.765	0.729	0.739	0.665	0.7
	Trigram+	200	0.852	0.818	0.834	0.825	0.858	0.841
		500	0.839	0.835	0.837	0.836	0.84	0.838
		800	0.845	0.818	0.831	0.823	0.85	0.836

Results Analysis

- Negative Corpus

Chi Square			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
NB	Unigram	200	0.819	0.803	0.811	0.806	0.823	**0.814**
NB	Unigram	500	0.813	0.805	0.809	0.807	0.815	0.811
NB	Unigram	800	0.807	0.783	0.794	0.789	0.813	0.8
NB	Bigram	200	0.714	0.755	0.734	0.74	0.698	0.718
NB	Bigram+	200	0.785	0.813	0.799	0.806	0.778	0.791
NB	Bigram+	500	**0.819**	0.803	**0.811**	0.806	**0.823**	**0.814**
NB	Bigram+	800	0.806	0.788	0.796	0.792	0.81	0.801
NB	Trigram	200	0.686	0.76	0.721	0.731	0.653	0.69
NB	Trigram+	200	0.776	**0.83**	0.802	**0.817**	0.76	0.788
NB	Trigram+	500	0.811	0.795	0.803	0.799	0.815	0.807
NB	Trigram+	800	0.808	0.788	0.797	0.793	0.813	0.802
SVM	Unigram	200	**0.833**	0.788	0.81	0.799	**0.843**	**0.82**
SVM	Unigram	500	0.83	0.795	**0.812**	0.803	0.838	**0.82**
SVM	Unigram	800	0.805	0.805	0.805	**0.805**	0.805	0.805
SVM	Bigram	200	0.723	0.743	0.732	0.735	0.715	0.725
SVM	Bigram+	200	0.808	0.798	0.803	0.8	0.81	0.805
SVM	Bigram+	500	0.798	0.78	0.789	0.785	0.803	0.794
SVM	Bigram+	800	0.798	0.763	0.78	0.773	0.808	0.79
SVM	Trigram	200	0.666	0.773	0.715	0.729	0.613	0.666
SVM	Trigram+	200	0.807	0.805	0.806	0.805	0.808	0.806
SVM	Trigram+	500	0.799	0.795	0.797	0.796	0.8	0.798
SVM	Trigram+	800	0.788	**0.808**	0.798	0.803	0.783	0.792

Information Gain			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
NB	Unigram	200	0.818	0.8	0.809	0.804	0.823	0.813
NB	Unigram	500	0.813	0.805	0.809	0.807	0.815	0.811
NB	Unigram	800	0.807	0.783	0.794	0.789	0.813	0.8
NB	Bigram	200	0.71	0.76	0.734	0.742	0.69	0.715
NB	Bigram+	200	0.786	0.815	0.8	0.808	0.778	0.792
NB	Bigram+	500	**0.819**	0.805	**0.812**	0.808	**0.823**	**0.815**
NB	Bigram+	800	0.806	0.788	0.796	0.792	0.81	0.801
NB	Trigram	200	0.686	0.76	0.721	0.731	0.653	0.69
NB	Trigram+	200	0.781	**0.83**	0.805	**0.819**	0.768	0.792
NB	Trigram+	500	0.814	0.81	0.812	0.811	0.815	0.813
NB	Trigram+	800	0.808	0.79	0.799	0.795	0.813	0.803
SVM	Unigram	200	0.831	0.8	**0.815**	**0.807**	0.838	**0.822**
SVM	Unigram	500	**0.832**	0.795	0.813	0.804	**0.84**	**0.822**
SVM	Unigram	800	0.805	**0.805**	0.805	0.805	0.805	0.805
SVM	Bigram	200	0.716	0.725	0.72	0.722	0.713	0.717
SVM	Bigram+	200	0.804	0.8	0.802	0.801	0.805	0.803
SVM	Bigram+	500	0.794	0.75	0.771	0.763	0.805	0.783
SVM	Bigram+	800	0.802	0.76	0.78	0.772	0.813	0.792
SVM	Trigram	200	0.661	0.765	0.709	0.721	0.608	0.659
SVM	Trigram+	200	0.81	0.798	0.804	0.8	0.813	0.806
SVM	Trigram+	500	0.79	0.773	0.781	0.778	0.795	0.786
SVM	Trigram+	800	0.787	0.785	0.786	0.786	0.788	0.787

Results Analysis

- Complete Corpus

Chi Square			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
		200	0.82	0.849	0.834	0.843	0.814	0.828
		500	0.824	0.844	0.834	0.84	0.82	0.83
	Unigram	800	**0.859**	0.838	0.848	0.841	**0.863**	**0.852**
	Bigram	200	0.71	0.758	0.733	0.74	0.69	0.714
		200	0.815	0.839	0.827	0.834	0.81	0.822
		500	0.817	0.846	0.831	0.84	0.81	0.825
	Bigram+	800	0.823	0.843	0.833	0.839	0.819	0.829
	Trigram	200	0.647	0.799	0.715	0.737	0.565	0.64
		200	0.845	**0.854**	**0.85**	**0.852**	0.844	0.848
		500	0.808	0.833	0.82	0.827	0.803	0.815
NB	Trigram+	800	0.818	0.834	0.826	0.831	0.815	0.823
		200	0.869	**0.871**	**0.87**	**0.871**	0.869	**0.87**
		500	0.873	0.854	0.863	0.857	0.876	0.867
	Unigram	800	**0.875**	0.848	0.861	0.852	**0.879**	0.865
	Bigram	200	0.758	0.803	0.78	0.79	0.744	0.766
		200	0.817	0.846	0.831	0.84	0.81	0.825
		500	0.868	0.863	0.865	0.863	0.869	0.866
	Bigram+	800	0.838	0.843	0.84	0.842	0.838	0.84
	Trigram	200	0.701	0.819	0.755	0.782	0.65	0.71
		200	0.855	0.865	0.86	0.863	0.854	0.859
		500	0.861	0.839	0.85	0.843	0.865	0.854
SVM	Trigram+	800	0.841	0.835	0.838	0.836	0.843	0.839

Information Gain			Truthful			Deceptive		
Apporach	Ngram	Feature-set Size	P	R	F	P	R	F
		200	0.821	0.839	0.83	0.835	0.818	0.826
		500	**0.824**	0.844	**0.834**	**0.84**	**0.82**	**0.83**
	Unigram	800	0.823	0.843	0.833	0.839	0.819	0.829
	Bigram	200	0.698	0.76	0.728	0.737	0.671	0.702
		200	0.81	0.846	0.828	0.839	0.801	0.82
		500	0.816	0.836	0.826	0.832	0.811	0.822
	Bigram+	800	0.821	0.844	0.832	0.839	0.816	0.828
	Trigram	200	0.642	0.804	0.714	0.738	0.553	0.632
		200	0.797	**0.851**	0.823	0.84	0.784	0.811
		500	0.804	0.835	0.819	0.828	0.796	0.812
NB	Trigram+	800	0.816	0.836	0.826	0.832	0.811	0.822
		200	0.867	0.866	0.867	**0.866**	0.868	0.867
		500	**0.878**	0.856	**0.867**	0.86	**0.881**	**0.87**
	Unigram	800	0.873	0.839	0.855	0.845	0.878	0.861
	Bigram	200	0.75	0.811	0.78	0.795	0.73	0.761
		200	0.859	0.861	0.86	0.861	0.859	0.86
		500	0.866	0.856	0.861	0.858	0.868	0.863
	Bigram+	800	0.854	0.841	0.848	0.844	0.856	0.85
	Trigram	200	0.69	0.823	0.751	0.781	0.631	0.698
		200	0.86	**0.866**	0.863	0.865	0.859	0.862
		500	0.859	0.848	0.853	0.85	0.861	0.855
SVM	Trigram+	800	0.85	0.829	0.839	0.833	0.854	0.843

Results Analysis

- IG vs. ChiSq
  - Almost the same
- SVM:
  - Unigram
- Naïve Bayes
  - Single polarity corpus: bigram+, larger feature set
  - Complete corpus: unigram
  - Lower requirement for complete corpus
- Best performance
  - Complete corpus
  - SVM
  - IG
  - 200
  - The simplest

Results Analysis

- Complete corpus
  - SVM, unigram, feature set size=200
  - Ranked by weights

Top 10		Bottom 10	
ChiSq	**IG**	**ChiSq**	**IG**
Upgraded	-	Luxury	luxury
River	Upgrade	Seemed	Vacation
-	&	Relax	Smelled
Attached	River	Vacation	Grand
Larger	Street	Smelled	Relax
&	Floor	Grand	Husband
Conference	Reviews	Hilton	Millennium
Reviews	Separate	Millennium	Once
Separate	Larger	Regency	Cleaned
floor	we	hotel	Hilton

Results Analysis

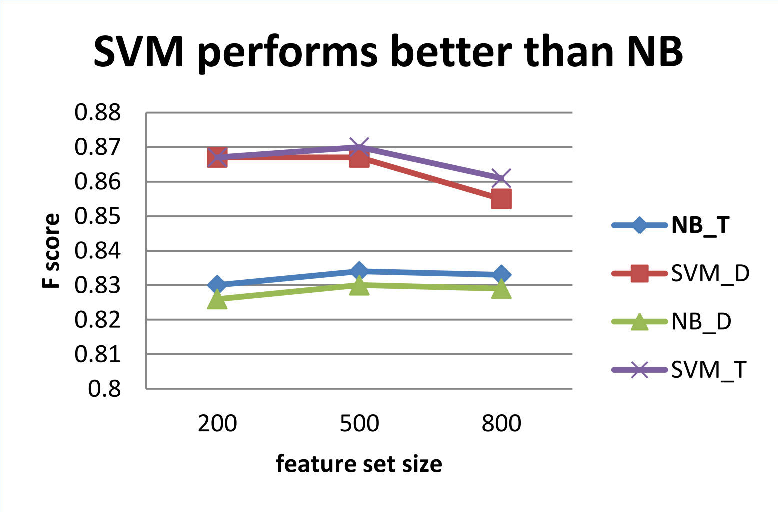
- Negative corpus
  - SVM, unigram, feature set size=200
  - Ranked by weights

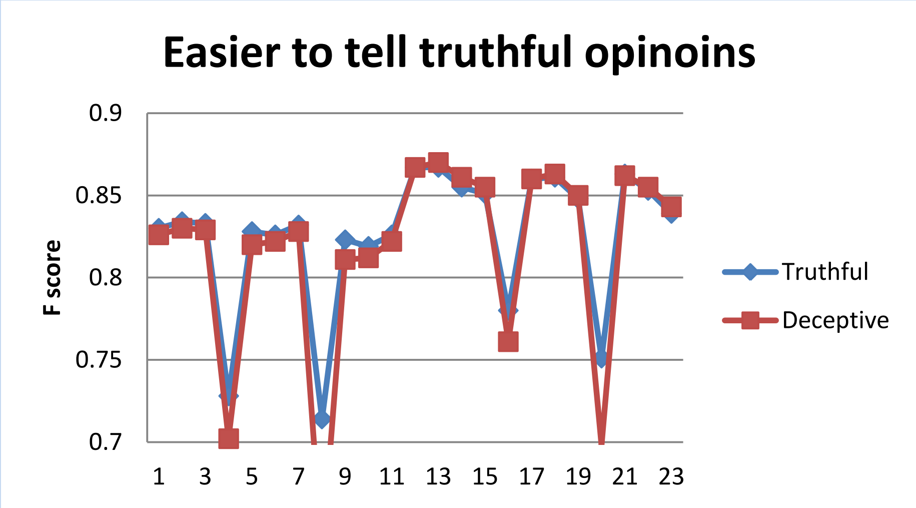
Spatial difficulty?!
Psychological distancing?!

Top 10		Bottom 10	
ChiSq	**IG**	**ChiSq**	**IG**
Star	Day	Luxury	Chicago
Covered	Star	Prices	Luxury
Conference	Conference	Seemed	Prices
Frequently	Frequently	Vacation	vacation
Prior	Covered	Decision	Seemed
Called	Prior	Website	Decision
Dated	-	Cleanliness	Cleanliness
Rate	Called	Hadn't	website
Trip	Dated	Sorely	settled
location	generally	Settled	Millennium

General Findings

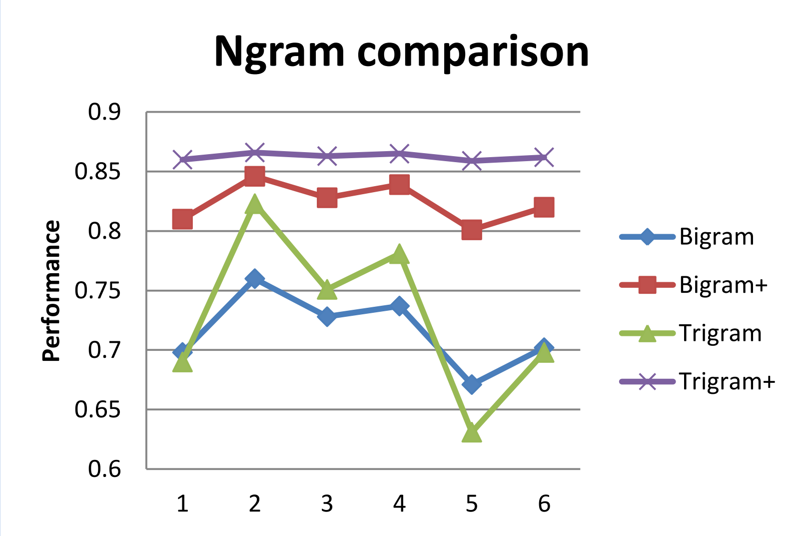
- SVM performs better than NB
- Truthful better than Deceptive

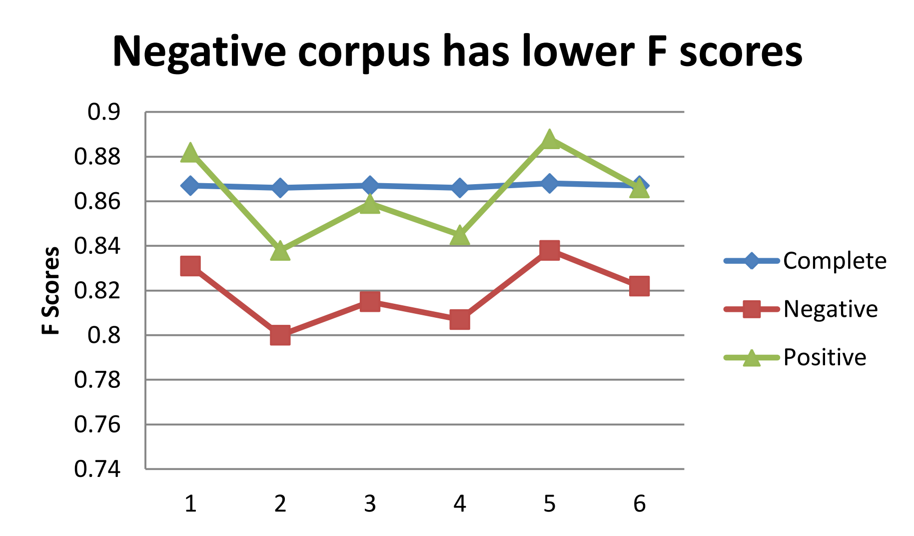




General Findings

- Ngram(bigram, trigram) only is a bad choice
- Negative corpus is more difficult to classify





Limitation

- Corpus
  - Size
  - Length: log normal?
- More effective feature sets
  - LIWC

Questions?

Thanks.

Yu Huang, Lingjie Zhang

backups

Classifier:Psycholinguistic deception detection

- Psycholinguistic deception detection
  - **L**inguistic **I**nquire and **W**ord **C**ount (Pennebaker et. al., 2007)
  - Counts instances of ~4500 keywords
  - Keywords are divided into 80 psycholinguistically meaningful dimensions across 4 broad groups
  - Create a feature for each of the 80 dimensions

Classifier: Psycholinguistic deception detection

- Psycholinguistic deception detection
  - Keywords are divided into 80 psycholinguistically meaningful dimensions across 4 broad groups
    - Linguistic processes
      - e.g., average number of words per sentence
    - Psychological processes
      - e.g., happy, feeling, eat, feeling
    - Personal concerns
      - e.g., job, cook, family
    - Spoken categories
      - e.g. umm, blah, yes

Motivation

- 87%
  - "Positive information I've read online has reinforced my decision to purchase a product or service recommended to me."
- 80%
  - "Negative information I've read online has made me change my mind about purchasing a product or service recommended to me."

http://www.coneinc.com/negativereviews-online-reverse-purchase-decisions