Exploring Rich Features for Sentiment Analysis with Various Machine Learning Models

PRINCETON UNIVERSITY

Shuyang Li '16, Advisor: Xiaoyan Li

IEEE Undergraduate Research Conference

Introduction and Motivation

- **Sentiment Analysis:** finding sentiment polarity from text documents
- **Applications:** business intelligence, identifying trends in public sentiment

> [–] **swingdancetraining** 170 points 6 months ago
> Tailor Store's fabric quality and manufacturing is far better than Modern Tailor's, with a slight (if any) price increase. At least in my experience. Your mileage may vary.

> [–] **bigBastardDude** 385 points 6 months ago
> Thermapen instant read thermometer. If you like to cook, this is great. Zero second guessing if something is at the right temp.

Challenges

- **Sarcasm:** "Last year, the benevolent studio gods gave us *Digimon*, and this year, they bestow *Max Keeble's Big Move* on delighted moviegoers across the country"
- **Qualifying Statements:** "though good-looking, its lavish sets...can do little to compensate for the emotional wasteland"
- **Complicated Summarization:** "this is not a great motion picture but, considering how bad most January releases are…it's passable"

Workflow

1. Base Corpus

Original Corpus
Pang & Lee 2004

IMDB INTERNET MOVIE DATABASE

Subjective Corpus
- Subjective sentences only
- Contrasting opinions
- Summary sentences

Summary Corpus
- Summary sentences only
- 1 sentence per document

2. Feature Extraction and Selection

POS Tagging → Stemming **or** Lemmatization → **Unigram Extraction** (Negation); **Bigram Extraction** (Stopwords); **Semantic Features** (Polarity, Purity)

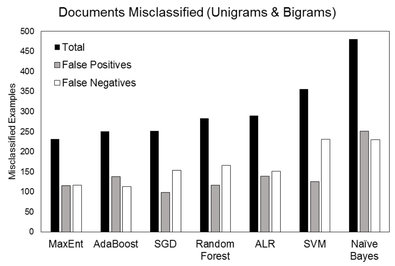
3. Classification

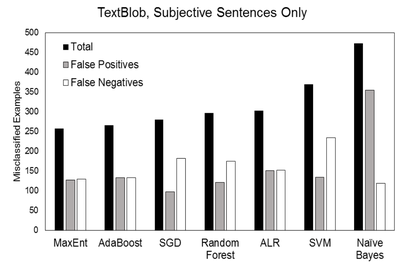
Singular Classifiers
- Naïve Bayes
- **Maximum Entropy**
- SVM (SVC)
- Stochastic Gradient Descent

Ensemble Classifiers
- **AdaBoost**
- Random Forest
- Additive Logistic Regression

Results

Analysis Method	MaxEnt	AdaBoost	SGDC
Full Corpus	0.8840	0.8745	0.8740
SAP	0.8545	0.8555	0.8470
SAF, f = 2	0.7720	0.6395	0.7710
AF	0.8310	0.8210	0.8295
POS SVM	0.8340	0.7035	0.8285
OpinionFinder	0.8290	0.8235	0.8155
TextBlob	**0.8710**	**0.8670**	**0.8600**
Manual Labeling	**0.8820**	**0.8705**	**0.8680**
Manual Summary	0.8200	0.7670	0.8145





Corpus	Additions	MaxEnt	AdaBoost	SGDC
Original	None	0.8840	**0.8745**	0.8740
Subj.		0.8820	0.8705	0.8680
Summary		0.8200	0.7670	0.8145
Original	Summary	0.8895	0.8690	**0.8920**
Subj.		0.8960	0.8725	0.8570
Original	Summary + Rich Features	0.8970	0.8640	0.8550
Subj.		**0.8980**	0.8515	0.8570

Experimental Design

- **4 singular classifiers** and **3 ensemble classifiers**
- **2 feature selection** methods: frequency cutoff and mutual information
- **4 part-of-speech-based rules** for subjectivity analysis
- **2 external subjectivity analysis utilities** (TextBlob, OpinionFinder)
- **Manual labeling** of subjective and summary sentences
- **Limited negation scope** for unigram generation
- Combinations of **7 rich (aggregate) feature classes**, in numerical and binarized forms

Feature Extraction Example

Original:
it is the perfect christmas film, flaws and all

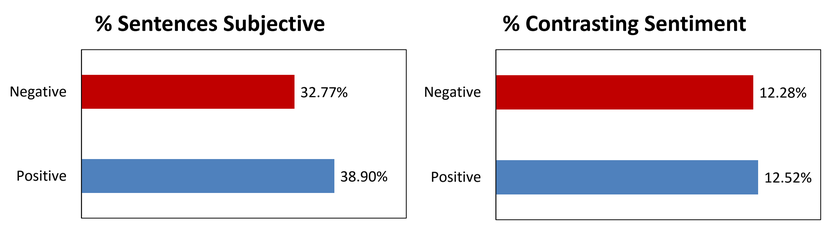
Part-of-speech tagging:
it is_V the perfect_A christmas_N film_N, flaws_N and all

Unigrams:
christmas (N)
flaw (N)
perfect (A)
film (N)

Bigrams:
the perfect
flaws and
it is
christmas film
and all
is the

Aggregate features:
Word polarity: 0.0323
Word subjectivity x polarity: 0.0266
Word purity: 0.7170

Manual Subjectivity Labeling



Objective (Removed): his love (kirstin scott thomas , mission impossible) was severely injured in a plane crash , and eventually died in a cave .

Subjective: lengthy and lousy are two words to describe the boring drama the english patient .

Contrasting: the only redeeming qualities about this film are the fine acting of fiennes and dafoe and the beautiful desert cinematography . *

Summary: other than these , the english patient is full of worthless scenes of boredom and wastes entirely too much film . ***

Conclusion and Discussion

- With simple unigrams AdaBoost is best; Maximum Entropy performs best on complex feature sets
- Accuracy can be improved with the usage of ideal subjectivity analysis (manual labeling)
- Performance is further boosted with addition of rich features (aggregate and summary)
- Potential to further improve accuracy with better **negation handling**, **topic analysis**, and **domain-specific lexicons**