Sentiment Analysis of Tweets and Movie Reviews

Noelle Harp, Debzani Deb
Department of Computer Science
Winston-Salem State University, Winston-Salem, NC

Purpose

Sentiment analysis (or, opinion mining), an important tool of E-commerce and business intelligence, emphasizes on classification of user-generated web comments (e.g., product reviews, forum discussions, and blogs) into positive and negative categories. The purpose of this study is to perform sentiment analysis via machine learning algorithms on two different data sets such as movie reviews and tweets and compare classification accuracy while varying different parameters.

Input Datasets

The twitter dataset (TD) consists of 200 tweets extracted from Edinburgh Twitter corpus [1], equally divided into 100 positive tweets and 100 negative tweets.

The movie review dataset (MRD) contains 2000 user-created movie reviews and is known as *Sentiment Polarity Dataset version 2.0* [2]. The reviews are equally partitioned into 1000 positive and 1000 negative reviews.

Acknowledgement

NSF RIA Award #1600864

WSSU RIP Award

Feedback @ http://compsci.wssu.edu/iBigCloud

Methodology

- WEKA [3], an open source data mining tool, was utilized for preprocessing, classification and sentiment analysis.
- The input data sets were imported in WEKA explorer and converted to WEKA supported ARFF format.
- **Preprocessing**: *StringToWordVector* filter from the package *weka.filters.unsupervised.attribute* was used for word parsing and tokenization. Stop-words (Frequently used words) such as conjunctions, prepositions, base verbs, etc. were removed.
- **Attribute Selection**: *AttributeSelection* filter from the *weka.filters.supervised.attribute* package is used to extract a set of best correlated words/attributes.
- **Classification**: Naïve Bayes (NB) [4] and K-nearest neighbors (KNN) [4] classifiers (with 1, 3, 5, and 10 neighbors) from WEKA were utilized to classify both data sets.
- Experiments were performed while vocabulary being All words (after preprocessing) vs. selected words (after attribute selection). Variations were made to training dataset by cross validating, making 66% or 80% of data available for training etc.
- Results are recorded as % of correctly classified tweets/reviews for unlabeled data.

References

1. Saša Petrović , Miles Osborne , Victor Lavrenko, The Edinburgh Twitter corpus, Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, p.25-26, June 06-06, 2010, Los Angeles, California.
2. Movie Review Data, https://www.cs.cornell.edu/people/pabo/movie-review-data/
3. Mark Hall , Eibe Frank , Geoffrey Holmes , Bernhard Pfahringer , Peter Reutemann , Ian H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explorations Newsletter, v.11 n.1, June 2009.
4. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

Results & Discussions

Twitter Data Set

	All Words	1086 Words		
	Cross Validation	Cross Validation	66% used for training	80% used for training
NB	53.50%	53.50%	50%	55%
KNN:1	51%	51%	51.50%	55%
KNN:3	48%	48%	55.90%	52.50%
KNN:5	49%	49%	51.50%	47.50%
KNN:10	45.50%	45.50%	50%	47.50%

Movie Review Data Set

	All Words	52 Words		
	Cross Validation	Cross Validation	66% used for training	80% used for training
NB	80.8%	79.50%	80.80%	**81.5%**
KNN:1	55.30%	68.90%	68.20%	68.25%
KNN:3	57.85%	70.80%	69.90%	69%
KNN:5	59.15%	70.55%	69.40%	69.50%
KNN:10	61.30%	71.15%	69.70%	68.75%

- Once preprocessed, the vocabulary for TD is 1351 words and for MRD is 1166 words.
- Our experiments on twitter dataset showed that there is no significant variation in classification accuracy while varying ML algorithms and other options. However, same set of experiments on the movie review data set revealed notable variations in classification accuracy.
- The best accuracy was achieved when filtering the datasets with stop words, using 80% as training data, and using the Naïve Bayes to process the data.
- In all cases, Naïve Bayes outperformed KNN algorithm.
- Focusing on important words generated better accuracy than considering all words in the web content.