International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 10 | Oct -2019 ISSN: 2590-1892

# Sentimental Analysis using Slang Words

**Ajay Kumar Phogat**

*Maharaja Surajmal Institute, New Delhi, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** Sentiment analysis is one of the best ways for a company or an organization to know the sentiment of a user, using a particular product or talking about a particular topic. This type of analysis gathers data before applying various pre-processing techniques and classifying what percentage of users are talking positively, negatively or neutrally about the product. In this paper we gather raw data by streaming it from the twitter servers where we search for a particular keyword used by the user. Now, we remove the unnecessary information which doesn't provide any idea about the sentiment of the user. We also train some of the data set to the system so that it can classify tweets as positive, negative or neutral and later provide it with test data to analyze the sentiment of the twitter user on some topic. We emphasize on gathering the twitter data. We use twitter as a source of the required data as it has become the most favorite platform for people to express their views. The main objective of this paper is to implement Opinion Mining using enhanced preprocessing steps which is the proposed solution. Preprocessing is vital as it involve hash tag priority and multilingual data. With the limited set of 140 characters, users on twitter use all kinds of emoticons, slangs to express them better in the limited set of available characters. This is the major problem of opinion mining but the results are more accurate as the data is more authentic. There are three classification techniques used for solving this purpose i.e. Naïve Bayes classification, SVM and Maximum Entropy. In Naive Bayes, models that assign class label to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Whereas Support Vector Machine (SVM) is a machine learning tool that is based on the idea of large margin data classification and Maximum Entropy is rooted in information theory, the meme seeks to extract as much information from a measurement as is justified by the data's signal-to-noise ratio.

***Key Words*:** Sentimental Analysis, Machine Learning, twitter, SVM, Naive Bayes, Maximum Entropy.

## 1.INTRODUCTION (

Today's generation is prone to social media. Social media is used to share information online. As it is becoming popular, everyone's opinion is digitized. [1]There has been a huge growth in the use of social media over recent years due to increase in broadband access, high street availability of powerful computers, and new website technologies that make content sharing easier. The most popular social media platforms are used by many millions of visitors. The number of persons worldwide who use internet has increase of roughly 40% since 1995 and reached a count of 3.2 billion.[3] With the advent of 21st century and huge boom in Internet, social media platforms like Blogs, Facebook, Twitter are major hub for users to express their opinions online. Since people post their comments and opinions on various topics, these sites have now become a rich source of information to analyze user behavior, product feedback, user intentions, lead generations. Plenty of time is spent by the companies to understand the problems and suggestions given by their customer. Over 32% of buyers rate an item on the web, over 33% composes surveys and about 88% trust online audits. In this way, audits assume a basic job in influencing the offers of a ware or a help. Each survey posted on the web comprises of the client's conclusions (positive or negative) and inclinations. Hence sentimental analysis plays a vital role in business since it is used to analyze the sentiments of the users. Comments or opinions are in the form of text which can be a big document like review of any product, or it can be small status messages. Techniques like Machine learning, artificial intelligence are now being used to analyze the opinions of the masses. These opinions can be used by companies for market research for need of a particular product or for checking public opinions on a newly launched product. Hence, Opinion Mining comes into the picture. It helps marketing strategist analyze the people's mindset about a product or an idea. It needs a large amount of data for analysis. Sentiments or opinions about an issue, person or event can be only as accurate as the number of persons participating. Hence, Data gathering and its classification are two very crucial tasks in opinion mining.

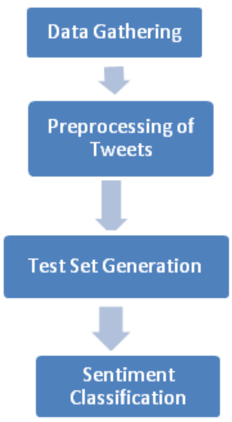


Fig 1: Block Diagram of Sentiment classification

### 1.1 Opinion Mining

Opinion mining is a process of looking on customer behavior towards a particular product, for example, constructing a system to analysis the comments and conversation around it. Opinion mining is very useful in product lunching and advertising and also determines which type of a product is popular and identifies which product like or dislike particular product features.

### 1.2 Social Media Platforms

Social media is now very popular in young generation people of our country. Social media platform includes facebook, twitter, blogs etc. Content sharing is now done through social media platforms. We can post text, images and videos through social media.

There are several thousands of social media platforms accessible, every with completely different functions, designs and user demographics. Before choosing that platform to use, accept what your audience is, and that platform would be most applicable. You will do that by consulting existing statistics of demographic use of social media and conjointly simply brooding about whether or not an audience can very have interaction with a media kind.

### 1.3 Classification Techniques

#### 1.3.1Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class label to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.[6] A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

This classifier is based on the Naive Bayes algorithm. Naive Bayes classifiers are parameterized by two probability distributions:

P(label): It gives the probability that an input will receive a label, with no prior information about the input's features.

P(feature_name=feature_value|label) gives the probability that a feature (feature_name) will receive a given value (feature_val) for a label (label).

If the classifier encounters an input with an unprecedented, then it will ignore the feature.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. "In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests".

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

#### 1.3.2 Probabilistic model

Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector

$$\mathbf{x} = (x_1, \dots, x_n)$$

representing some n features (dependent variables), it assigns to this instance probabilities $p(C_k \mid x_1, \dots, x_n)$ for each of k possible outcomes or classes.

The main problem with the above formulation is that if the number of features n is huge or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}.$$

In Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on $C$ and the values of the features $F_i$are given, so that the denominator is effectively constant. The numerator is equals to the joint probability model $p(C_k, x_1, \dots, x_n)$ which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
p(C_k, x_1, \dots, x_n) &= p(C_k)\ p(x_1, \dots, x_n \mid C_k) \\
&= p(C_k)\ p(x_1 \mid C_k)\ p(x_2, \dots, x_n \mid C_k, x_1) \\
&= p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k, x_1)\ p(x_3, \dots, x_n \mid C_k, x_1, x_2) \\
&= p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k, x_1)\ \dots p(x_n \mid C_k, x_1, x_2, x_3, \dots, x_{n-1})
\end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature $F_i$ is conditionally independent of every other feature $F_j$ for $j \neq i$, given the category $C$. This means that

$p(x_i \mid C_k, x_j) = p(x_i \mid C_k)$,

$p(x_i \mid C_k, x_j, x_k) = p(x_i \mid C_k)$,

$p(x_i \mid C_k, x_j, x_k, x_l) = p(x_i \mid C_k)$,

and so on, for $i \neq j, k, l$. Thus, the joint model can be expressed as

$$\begin{aligned} p(C_k|x_1,\ldots,x_n) &\propto p(C_k,x_1,\ldots,x_n) \\ &\propto p(C_k)\ p(x_1|C_k)\ p(x_2|C_k)\ p(x_3|C_k)\ \cdots \\ &\propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k)\,. \end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable $C$ is:

$$p(C_k|x_1,\ldots,x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

where the evidence $Z = p(\mathbf{x})$ is a scaling factor dependent only on $x_1,\ldots,x_n$, that is, a constant if the values of the feature variables are known.

**Constructing a classifier from the probability model**

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k).$$

**1.3.3 Support Vector Machine(SVM)**

SVM is a machine learning tool that is based on the idea of large margin data classification. The tool has strong theoretical foundation and the classification algorithms based on it give good generalization performance. Standard implementations, though provide good classification accuracy, are slow and do not scale well. Hence they cannot be applied to large-scale data mining applications. They typically need large number of support vectors. Hence the training as well as the classification times is high.

Support vector machines (SVMs) are shown to be extremely effective once used for ancient categorization of text. They're notable to be large-margin, rather than probabilistic classifiers, in distinction to Naive bayes and MaxEntropy. Within the two-category case, the fundamental idea behind the training procedure is to search out a hyper plane, described by vector ~w, that not only separates the document vectors in one category from those within the different, except for that the separation, or margin, is as massive as possible.

**1.3.4 Maximum Entropy**

Rooted in information theory, the mem seeks to extract as much information from a measurement as is justified by the data's signal-to-noise ratio. The max entropy classifier framework considers all of the probability distributions that are empirically consistent with the training data; and chooses the distribution with the highest entropy. It is parameterized by a set of weights used to combine the joint-features that are generated from a feature set by an "encoding". This encoding maps each (feature set, label) pair to a vector.

Entropy maximization with no testable information respects the universal "constraint" that the sum of the probabilities is one. Under this constraint, the maximum entropy discrete probability distribution is the uniform distribution,

$$p_i = \frac{1}{n}\ \text{for all}\ i\in\{1,\ldots,n\}.$$

**1.3 Pre-processing**

The raw data is filter and transform in readable form from which meaningful result can be extracted.

There are various feature extraction methods available for collecting relevant features from text. We chose to carry out the feature extraction process differently in two phases to extract relevant features.

The feature extraction process is conducted varying from course to fine level. In the first phase, twitter specific features are extracted. First of all, all the tweets are translated to a single language. English is used as the target language here. Then, emoticons, slang language, misspellings, email id, URLs etc are forced to preprocessing before feature extraction.

In the second phase, the stop words which do not contribute towards sentiment of the tweets are removed and the tweets are converted into feature vectors. This feature vector is then used in classification algorithm.

For preprocessing, following functions are implemented:

- Converting text to lower case
- Removing username and email
- Removing unnecessary URLs
- Removing symbols
- Removing white spaces
- Replace #word with word

Replace words with multiple occurrences with single word.

**1.4 LIERATURE REVIEW**

Sentiment analysis is an expanding area of Natural Language Processing with research work from document level classification to learning the polarity of words and phrases. Due to the character limitations on tweets, categorizing the sentiment of Twitter messages is close to sentence-level sentiment analysis; however, the informal and slang language used in tweets, as well as the very nature of the microblogging domain makes Twitter sentiment analysis a very difficult task. It's a much asked question how effectively the features and techniques used on well-formatted data will reflect in the microblogging area. In recent years, there have been various number of papers exploring Twitter sentiment and buzz (Jansen et al. 2009; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010; Davidov, Tsur, and Rappoport 2010). Many scientists have begun to research the use of part-of-speech characteristics but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has not been much investigation into the usefulness of current sentiment techniques developed on non-microblogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data (Pak and Paroubek 2010; Bifet and Frank 2010) exploit existing Twitter sentiment sites for collecting training data.. Researchers have also begun to explore various ways of automatically collecting training data. A large portion of the research in the sentiment classification using symbolic techniques makes use of available lexical resources. A lexical database WordNet [5] is also widely used and was also used by Kamps et al. to determine the sentiment content of a word along n-dimensions. They developed a distance metric on WordNet and calculated the semantic orientation of adjectives. WordNet database is made up of words which are synonyms of each other. Turney used bag-of-words approach for sentiment analysis. He used an unsupervised approach to learn the emotional orientation of words/phrases: positive vs. negative. Relationships between the individual words are not taken into consideration but the document is represented as a mere bunch of words. He found the polarity of a review based on the average semantic orientation of tuples, sentences or parts of sentences composed of different parts of speech, extracted from the review. Baroni et al designed a system using word spacing model formalism that overcomes the tediousness in task of lexical substitution. A word's overall distribution and local context is represented by this model. Coarse grained and fine grained approaches are also used to identify sentiments. In coarse grained approach, they performed binary classification of emotions and in fine grained approach emotions were classified into different levels. Knowledge based approach is found to be difficult due to the requirement of a large lexical database. Since social networking sites generate huge amount of data each moment, sentiment analysis has become tedious and error prone. Balahur et al produced EmotiNet, a representation of textual data that incorporates the structure and the semantics related to a specific domain. EmotiNet made use of the concept of Finite State Automata to find the emotional reactions triggered due to actions.

**B. Machine Learning Techniques**

Machine Learning techniques use a training set and a test set for classification. Training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. Few features that can be used for sentiment classification are Term Presence, Frequency, negation, n-grams and Part-of-Speech. These features can be used to find out the semantic orientation of words, phrases, sentences and that of documents. Semantic orientation is the polarity which may be either positive or negative.

## 2 COMPARATIVE ANALYSIS

After applying both the techniques (Naive Bayes and Maximum Entropy), we found that more the number of dependent features. This is unexpected as the main assumption of Naive Bayes is that the features are not dependent. Zhen Niu et al introduced a novel method in which efficient approaches are used for weight calculation, feature selection and classification. The new model is based on Bayesian algorithm. Here eights of the classifiers are regulated by using of representative and unique feature. 'Representative feature' is the information that represents a class and 'Unique feature' is the information that helps in distinguishing classes. Using those weights, they computed the probability of each classification and thus improved the Bayesian algorithm. Barbosa et al designed a 2-step automatic sentiment analysis method for classifying tweets. A noisy training set was used to reduce the labeling effort in classifiers. Firstly, tweets were categorized into subjective and objective comments. Following that, subjective tweets are classified as positive and negative tweets. Various machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews. Celikyilmaz et al developed a pronunciation dependent word clustering method for normalizing noisy tweets. In pronunciation based word clustering, words having similar pronunciation are clustered and assigned same tokens. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a decreased error rate. Pak et al created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a chance of error since emotions of tweets in training set are labeled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons. Xia et al used an ensemble framework for sentiment classification.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 10 | Oct -2019 ISSN: 2590-1892

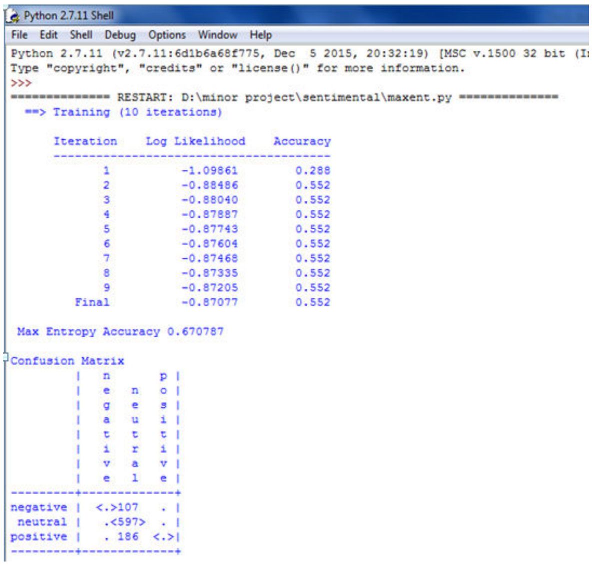


Fig 2: Output of Maximum Entropy

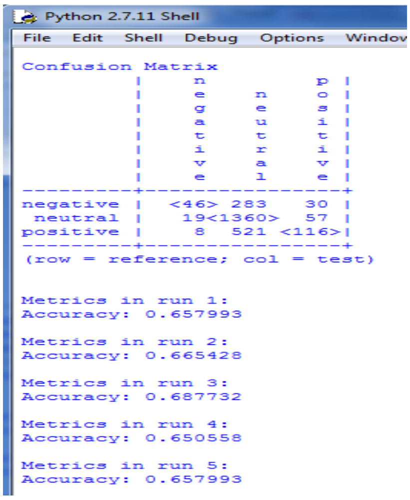


Fig 3: Output of Naive Bayes

## 2.1 RESULTS ANALYSIS

After applying both the techniques (Naive Bayes and Maximum Entropy), we found that more the number of iterations more the accuracy. At the same number of iteration count the percentage of accuracy is more observed Maximum Entropy technique. By increasing the number of iterations to train the data set. The accuracy of maximum entropy algorithm can be increased substantially although the time of computation is also increased

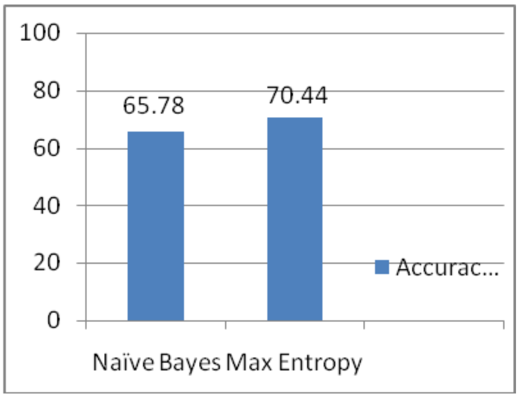


Fig 4. Comparative analysis of techniques

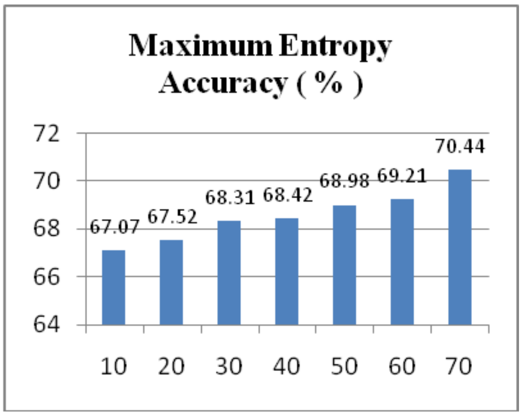


Fig 5. Different values obtained at different iterations

## 3. CONCLUSIONS

Many different Symbolic and Machine Learning techniques are used to identify sentiments from text. Machine Learning techniques being simpler and more efficient are applied for sentiment analysis in various domains. Opinion mining in Twitter is more accurate and helpful. But there are some issues while dealing with tweets such as text having multiple keywords, slangs and other misleading text data. To deal with these issues, performing feature extraction creates an efficient feature vector. For feature extraction, preprocessing is vital as quality of preprocessing affects the feature vector and the results. Here, preprocessing is done in two steps.

In the first step, symbols, URLs, slangs etc. are removed. In the next step, feature vector is created after removing all stop words. Classification accuracy of the feature vector is tested using different classifiers like Naïve Bayes, SVM and Maximum Entropy. All these classifiers have almost similar accuracy for the new feature vector. The quality of result is better as the preprocessed data is more

## REFERENCES

[1] Bongsug (Kevin) Chae; "Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter

and Twitter data for supply chain practice and research"; International Journal of Production Economics (2015).

[2] Ching-Chin Chern, Anthony J. T. Lee and Chih-Ping Wei; "Introduction to the special issue on Data analytics for marketing intelligence''; Springer-Verlag Berlin Heidelberg, 2014.

[3] Clyde Holsapple, Shih-Hui Hsiao and Ram Pakath; "Business Social Media Analytics: Definition,Benefits, and Challenges"; Twentieth Americas Conference on Information Systems, Savannah, 2014.

[4] Pravesh Kumar Singh and Mohd Shahid Husain; "Methodological Study Of Opinion Mining and sentiment analysis techniques"; International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014.

[5] Khan Farhan Hassan, Qamar Usman and Javed Younus; "SentiView:A visual sentiment analysis framework", International Conference on Information Society (i-Society), 10-12 Nov. 2014, London, UK.

[6] Fan Sun, Belatreche A, Coleman S, McGinnity, T.M. et al; "Preprocessing online financial text for sentiment classification: A natural language processing approach"; IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) at London, 2014

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up?: sentiment classification using machine learning techniques" In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10

[8] Vishwanathan, S.V.M.; Murty, M.N., "SSVM: a simple SVM algorithm," Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on , vol.3.

[9] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2) (2008) 1–135.

[10] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12) (2010) 2544–2558.

[11] Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999)

[12] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phraselevel sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics (2005)

[13] Hansen, L.K., Arvidsson, A., Nielsen, F.˚A., Colleoni, E., Etter, M.: Good friends, bad news — affect and virality in Twitter. Accepted for The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011) (2011)

[14] Akkaya, C., Conrad, A., Wiebe, J., Mihalcea, R.: Amazon Mechanical Turk for subjectivity word sense disambiguation. In: Proceedings of the NAACL HLT 2010 Workshop on Creating, Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics (2010) 195–203.

[15] Baudhuin, E.S.: Obscene language and evaluative response: an empirical study. Psychological Reports 32 (1973).

[16] Sapolsky, B.S., Shafer, D.M., Kaye, B.K.: Rating offensive words in three television program contexts. BEA 2008, Research Division (2008).

[17] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly, Sebastopol, California (June 2009).

[18] Biever, C.: Twitter mood maps reveal emotional states of America. The New Scientist 207(2771) (July 2010) 14.

[19] M. Berland and E.Charniak. Finding parts in very large corpora. In Proc. of the 37th ACL Conf., pages 57–64, 1999.

[20] S. Das and M. Chen. Yahoo! for anazon: Extracting market sentiment from stock message boards. In Proc. of the 8th APFA, 2001.

[21] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classificationof productreviews. InProc.ofthe12th Int.WWWConf., 2003.

[22] T.E.Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 1993.

[23] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL Conf., pages 174–181, 1997.

[24] M.Hearst. Direction-based textinterpretation asan information access refinement. Text-Based Intelligent Systems, 1992.

[25] B. Katz. From sentence processing to information access on the world wide web. In Proc. of AAAI Spring Symp. on NLP, 1997.

[26] H.Li and K. Yamanishi. Mining from open answers in questionnaire data. InProc. of the 7th ACMSIGKDD Conf. 2001.

[27] C. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[28] M.P.Marcus, B.Santorini and M.A.Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. Computational Linguistics, 19, 1993.

[29] G. A. Miller. Nouns in WordNet : A lexical inheritance system. Int. J. of Lexicography, 2(4):245–264, 1990. Also available from ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps.

[30] S. Morinaga, K. Yamanishi, K. Teteishi, and T. Fukushima. Mining product reputations on the web. In Proc. of the 8th ACM SIGKDD Conf., 2002.