UTILIZING TWITTER TO PERFORM AUTONOMOUS SENTIMENT ANALYSIS

Akanksha Srivastava[1], Mr. Sambhav Agarwal[2]

[1]M.Tech, Computer Science and Engineering, SR Institute of Management & Technology, Lucknow, India
[2]Associate Professor, Computer Science, and Engineering, SR Institute of Management & Technology, Lucknow

---***---

Abstract - *Applications in many domains make Sentiment Analysis an exciting area for study. The use of online polls and surveys to get feedback from the public regarding goods, current events, and societal or political issues are on the rise. The public and the stakeholders benefit from hearing the thoughts and feelings of the general public when important choices must be made. Opinion mining is the practice of gleaning insights from online sources including web search engines, blogs, micro-blogs, Twitter, and social networks to produce meaningful conclusions. Twitter's user base provides a wealth of material from which to get insight into the public's perspective. The massive volume of tweets as the unstructured text makes it challenging to physically delineate the information. Consequently, extracting and condensing the tweets from corpora calls for expert computational methodologies, which in turn necessitates familiarity with terms that convey emotion. Sentiment analysis from the unstructured text may be accomplished using a wide variety of computer methodologies, models, and algorithms. The vast majority are based on machine learning methods, namely the Bag-of-Words (BoW) representation. In this research, we used a lexicon-based strategy to automatically identify sentiment for tweets gathered from the Twitter public domain. To further investigate the efficacy of alternative feature combinations, we have used three distinct machine learning algorithms for the task of tweet sentiment identification: Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). Our results suggest that both NB with Laplace smoothing and SVM are successful in categorizing the tweets. The feature used for NB is unigram and Part-of-Speech (POS), while unigram is utilized for SVM.*

Key Words: Bag-of-Words, Lexicon, Machine Learning Algorithms, Laplace Smoothing, Part-of-Speech.

1. INTRODUCTION

It has been found via two separate polls of over 2000 American adults that 81% of Internet users (or 60% of Americans) have done product research online at least once and that 20% of Internet users (15% of Americans) prefer it on a certain day. We may claim that people's consumption of goods and services is not the only factor for their online information-seeking and opinion-sharing activities. The need for access to current political information is another critical factor to consider. At the moment, individuals may utilize email for political campaigns by sharing information and discussing candidates and issues online. The user trusts internet advice and suggestions since they deal mostly with an opinion. Despite the generally pleasant experiences of American Internet users during online product research, Horrigan [1] found that 58% of users reported experiencing missing, difficult-to-discover, confused, or overwhelming online information. Therefore, there is a significant need for improved information-access technologies to aid shoppers and researchers. Web 2.0 sites like blogs, message boards, and other kinds of social media have made it easier than ever for customers to voice their thoughts and views on the brands they use. In recent years, businesses have begun to acknowledge the power that user reviews have on shaping the perceptions of others and the standing of certain brands. Companies are beginning to watch social media to react to customer feedback and adjust their marketing, brand positioning, product development, and other strategies appropriately.

1.1. Opinion Mining and Sentiment Analysis

Extracting views from text is called "Opinion Mining" (OM). Viewpoint mining (OM) is a new field at the intersection of information retrieval, text mining, and computational linguistics that seeks to detect the opinion represented in natural language texts, as described by Pang et al. [3]. Opinion mining is a subfield of KDD that employs Natural Language Processing (NLP) and statistical machine learning methods to identify and distinguish between opinionated and factual content. Tasks in opinion mining include locating opinions, labeling them as favorable, negative, or neutral, determining where those opinions originated, and summarising them. To automatically extract a summary of an entity's opinion from a large body of the unstructured text is the primary goal of the Opinion Mining assignment.

Opinion Mining and Sentiment Analysis (SA) are two names for the same thing: the study of how people feel about something. An individual's thoughts, feelings, and impressions about a matter, as expressed in the form of an opinion, are deeply personal and confidential. Individuals, groups, and societies may benefit greatly from the advice and counsel of others throughout the decision-making process, as concluded by the work of Liu et al. [2]. To act swiftly and wisely, humans demand information that is both precise and brief. While making a choice, people often seek advice from friends, family, and experts for whom they have developed an opinion or point of view based on their own experiences, observations, conceptions, and beliefs (which may or may not be good or negative).

2. SENTIMENT TARGET IDENTIFICATION

Identifying sentiment (opinion) targets is a crucial part of SA work. The aim here might be anything from the subject of the statement to the object of that statement. Everyone involved in making and selling a product has to do a thorough evaluation of it in light of public and buyer feedback. Automatically identifying and extracting aspects mentioned in reviews is a key step in conducting a review comparison. Opinion mining and summarization, thus, rely heavily on product feature mining [10]. Sentiment analysis is a difficult field of study. This is because a system has to be able to discern evaluative expressions and some qualities that are not overtly present and need to be identified from the term semantic to correctly identify opinion targets in a phrase or document. Previous studies on the topic of sentiment target identification have shown that several Natural Language Processing (NLP) methods, including processing, Part-of-Speech tagging, noise reduction, feature selection, and classification, are all necessary stages in the extraction process.

3. METHODOLOGY

Research data collecting is more complex than it may seem since it requires drawing important and relevant inferences. Test data, subjective training data, and objective (neutral) training data are the three types of data that have been gathered. The Twitter API will be covered beforehand.

3.1. Twitter API

Developers may access Tweets, DMs, media, and other Twitter data using the Twitter API, which provides a collection of programming interfaces. Through the API, programmers may create products that communicate with the Twitter service and carry out actions like publishing Tweets, getting user information, and viewing trending topics, among other things. Different endpoints, authentication mechanisms, and use constraints apply to the API's several flavors, which include REST (Representational State Transfer), streaming, and advertising. A Twitter developer account and API keys (also known as access tokens) are prerequisites for interacting with the API.

3.2. Twython

Twython is a Python library for accessing the Twitter API. It provides a simple and convenient way for Python developers to interact with the Twitter platform and perform tasks such as posting Tweets, retrieving user information, and accessing timelines. Twython abstracts many of the complexities of the Twitter API and provides a simple, Pythonic interface for accessing the API's resources. To use Twython, you will need to obtain API keys or access tokens from a Twitter developer account, and then use these credentials to initialize a Twython client object, which you can use to make API requests. The library supports both REST and Streaming APIs and includes functionality for OAuth 1.0a and OAuth 2.0 authentication.

3.3. Data Preprocessing in Twitter

Data preprocessing in Twitter involves cleaning and transforming Twitter data into a format that is suitable for further analysis or modeling. This may include tasks such as:

1. Data Collection: Collect raw data from the Twitter API, such as tweets, user profiles, and trends.
2. Data Cleaning: Removing irrelevant information, correcting errors, handling missing values, and removing duplicates from the collected data.
3. Text Processing: Processing textual data from tweets, such as removing stop words, stemming, and converting text to lowercase.
4. Sentiment Analysis: Classifying tweets into positive, negative, or neutral sentiment categories.
5. Data Transformation: Converting the data into a format that is suitable for analysis, such as converting textual data into numerical representations.
6. Data Reduction: Reducing the dimensionality of the data, such as aggregating data by user or period.

These steps ensure that the data is in a clean, consistent, and usable format, and help improve the accuracy and reliability of any subsequent analysis or modeling.

3.4. Lexicon-Based Approach

The lexicon-based approach is a method used in sentiment analysis and opinion mining to classify the sentiment of a piece of text, such as a tweet, into positive, negative, or neutral categories. The approach involves using a predefined lexicon, or a list of words, that are associated with specific sentiments.

In a lexicon-based approach, the sentiment of a piece of text is determined by counting the number of words in the text that match words in the lexicon and then aggregating the sentiment scores associated with these words. The resulting sentiment score is then used to classify the text as positive, negative, or neutral.

There are many different lexicons available for use in sentiment analysis, each with its strengths and weaknesses. Some popular lexicons include SentiWordNet, the Harvard IV dictionary, and the AFINN lexicon.

The lexicon-based approach is simple to implement and has been widely used in sentiment analysis. However, it has some limitations, such as being limited to the words in the lexicon and not taking into account the context in which words are used. To overcome these limitations, other approaches such as machine learning and deep learning models have been developed.

3.5. SentiWordNet

SentiWordNet is a lexicon for sentiment analysis and opinion mining. It is a manually constructed, multi-word expression resource for the English language that provides sentiment scores for words and phrases.

SentiWordNet assigns sentiment scores to words based on three dimensions: positivity, negativity, and objectivity. Each word in the lexicon is associated with three sentiment scores, representing its positivity, negativity, and objectivity. The scores are based on the collective sentiment of words that are semantically similar to the word being scored.

SentiWordNet can be used as a resource in sentiment analysis and opinion mining to classify the sentiment of a piece of text into positive, negative, or neutral categories. To do this, the sentiment scores of the words in the text are aggregated to determine the overall sentiment of the text.

SentiWordNet has been widely used in sentiment analysis and has been shown to perform well in comparison to other lexicons and machine learning models. It is a valuable resource for researchers and practitioners in the field of sentiment analysis.

4. RESULTS AND ANALYSIS

4.1. Naive Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes' Theorem. It is a popular algorithm in the field of machine learning and is widely used for tasks such as text classification, sentiment analysis, and spam filtering.

The basic idea behind Naive Bayes is to use Bayes' Theorem to calculate the probability of a class (e.g., positive, negative, or neutral sentiment) given a set of features (e.g., words in a text). The algorithm assumes that the features are conditionally independent, meaning that the presence of one feature does not affect the presence of another feature. This is the "naive" part of the algorithm, hence its name.

There are several variants of the Naive Bayes algorithm, including the Multinomial Naive Bayes, Bernoulli Naive Bayes, and Gaussian Naive Bayes. Each variant is suited for different types of data and different classification tasks.

Naive Bayes is a fast and effective algorithm for text classification and sentiment analysis. It is simple to implement and requires little data preparation. However, its performance can be limited by the "naive" assumption of independence between features, which is not always accurate in practice. Despite this, Naive Bayes remains a popular and widely used algorithm in the field of text classification and sentiment analysis.

4.2. For Twitter Dataset

We investigate a wide range of characteristics that have a significant impact on sentiment analysis. We have made use of N-gram features such as unigrams (n = 1) and bigrams (n = 2), which are used often in a variety of text classifications including sentiment analysis. In the course of our research, we played around with boolean features using both unigrams and bigrams. Each n-gram feature has a boolean value that is connected with it. This value is set to true if and only if the corresponding n-gram appears in the tweet [12]. The many characteristics that we have employed are outlined in Table 1, along with the accuracy results obtained from each particular classifier. A comparison of this dataset with the one that Pang Lee et al. utilized for their research on movie reviews has been carried out here. According to what was found in Table 1, the classification accuracies that resulted from using unigrams as features gave better results in the case of tweets than movie reviews when we used the NB classifier with Laplace smoothing; however, when we used the MaxEnt classifier, the accuracy result of movie reviews was more than the tweets.

Table 1: Accuracy of tweets using different features

	Features	# of Features	Frequency or Presence	Naive Bayes		Maximum Entropy		Support Vector Machine	
				reviews	tweets	reviews	tweets	reviews	tweets
(1)	unigram	5989	presence	81.0 %	81.5%	80.4%	78.36%	82.9%	82.5%
(2)	bigram	19148	presence	77.3%	78.60%	77.4%	78.0%	77.1%	77.8%
(3)	unigram + bigram	25,748	presence	80.6%	80.92%	80.8%	79.78%	82.7%	81.6%
(4)	Unigram+POS	19061	presence	81.5 %	82.0%	81.2%	80.3%	81.9%	81.99%
(5)	Adjectives	1197	presence	77.0%	69.48%	77.7%	76.4%	75.1%	76.4%

Table 2: F1 score of MNB classifier

Class label	Precision(%)	Recall(%)	F1 score(%)
Positive	65.25%	20.51%	31.21%
Negative	77.41%	16.05%	27.20%
Neutral	80.48%	61.82%	69.93%

We investigate a wide range of characteristics that have a significant impact on sentiment analysis. We have made use of N-gram features such as unigrams (n = 1) and bigrams (n = 2), which are used often in a variety of text classifications including sentiment analysis. In the course of our research, we played around with boolean features using both unigrams and bigrams. Each n-gram feature has a boolean value that is connected with it. This value is set to true if and only if the corresponding n-gram appears in the tweet [12]. The many characteristics that we have employed are outlined in Table 1, along with the accuracy results obtained from each particular classifier. A comparison of this dataset with the one that Pang Lee et al. utilized for their research on movie reviews has been carried out here. According to what was found in Table 1, the classification accuracies that resulted from using unigrams as features gave better results in the case of tweets than movie reviews when we used the NB classifier with Laplace smoothing; however, when we used the MaxEnt classifier, the accuracy result of movie reviews was more than the tweets.

The effectiveness of POS features has been validated using sentiment analysis. As a general rule, adjectives are regarded as useful components for sentiment analysis since they serve as reliable indicators of a subject's feelings. Taking into account solely adjectives provides results that are comparable to those produced by employing unigrams and bigrams, as can be seen in Line (5) of the table displaying the results of our experiment. Line (4) of the table displaying the results demonstrates that when unigrams and POS are used as a feature, all three classifiers generate superior results. The first line of the table displaying the results demonstrates that using SVM with unigram as a feature yields the best result out of all the characteristics that were taken into consideration. The comprehensive findings of the MNB classifier may be seen in Table 2, which displays the F1 score. The Receiver Operating Characteristic (ROC) curve of the MNB classifier is shown in Figure 1. This curve is for tweets that have been manually annotated.

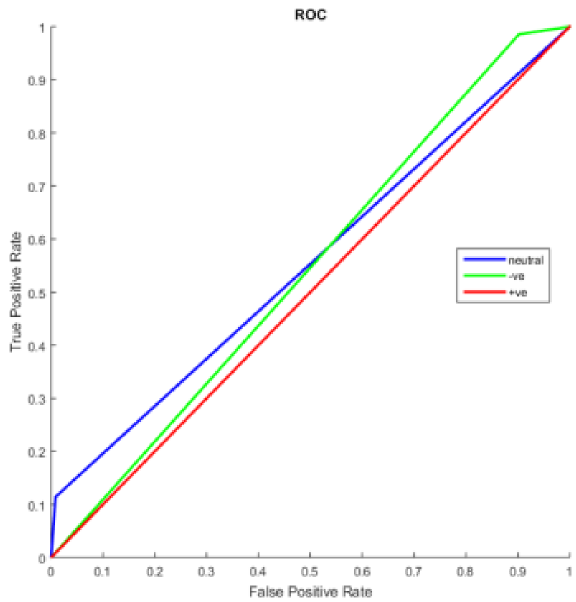
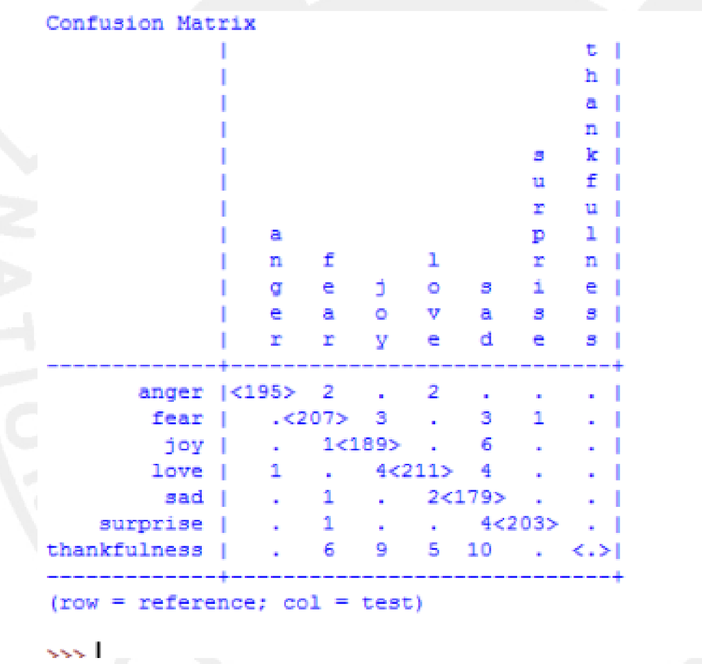


Figure-1: ROC curve of MNB classifier for tweets

4.3. Emotion Dataset

Hashtags are often used as a means for people to communicate their thoughts and feelings. Therefore, a satisfactory amount of feelings and sentiments may be gleaned from these hashtagged phrases. These hashtags have been included in our machine-learning algorithm to provide it with more data. Figure 2 depicts a snapshot of the confusion matrix for our emotion dataset's unigram features. Additionally, the F1 score of each class for the unigram feature is shown in this figure. Figure 3 shows the ROC curve that was generated by our classifier.



```
Confusion Matrix
             |                               t |
             |                               h |
             |                               a |
             |                               n |
             |                           s   k |
             |                           u   f |
             |                           r   u |
             |   a                       p   l |
             |   n   f       l           r   n |
             |   g   e   j   o   s       i   e |
             |   e   a   o   v   a       s   s |
             |   r   r   y   e   d       e   s |
-------------+---------------------------------+
       anger |<195>  2   .   2   .       .   . |
        fear |   .<207>  3   .   3       1   . |
         joy |   .   1<189>  .   6       .   . |
        love |   1   .   4<211>  4       .   . |
         sad |   .   1   .   2<179>      .   . |
    surprise |   .   1   .   .   4<203>      . |
thankfulness |   .   6   9   5  10       .  <.>|
-------------+---------------------------------+
(row = reference; col = test)

>>> |
```

Figure 2: Snapshot of emotion dataset

Table 3: Accuracy of emotion dataset using different features

Features	# of features	MNB classifier
Unigram	4635	95.0%
Bigram	17628	71.23%
Unigram+Bigram	35356	95.3%
POS	12443	92.9%
Adjective	1503	84.5%

Table 4: F1 score of MNB classifier for unigram feature

Class label	Precision(%)	Recall(%)	F1 score(%)
anger	99.48%	97.98%	98.72%
fear	94.95%	96.72%	95.82%
joy	92.19%	96.42%	94.25%
love	95.90%	95.90%	95.9%
sad	86.89%	98.35%	92.26%
surprise	99.5%	97.59%	98.53%

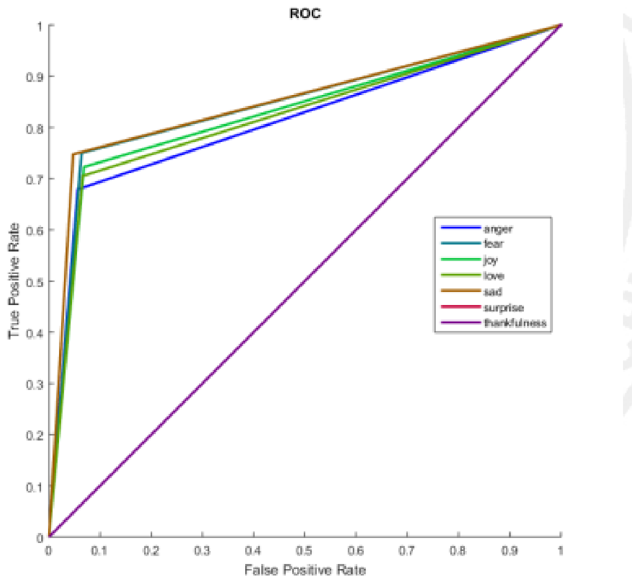


Figure 3: ROC curve of MNB classifier for an emotion data set

When compared to the data set that is generated by manually annotating tweets, we observed that constructing a dataset by automatically collecting tweets via the use of hashtags demonstrates a clear advantage. This was one of the findings of our experiment. This is because authors are accurate about their feelings, but the conventional method of annotating material requires annotators to infer the writers' feelings from the text, which is not possible to do accurately.

5. CONCLUSION

As part of our study, we looked at the difficulties of Sentiment Analysis and the many approaches used in this area. Identification of sentiment in social media data is notoriously challenging due to the data's richness and subtlety. To determine which characteristics are most useful for Sentiment Analysis, we experimented using tweets collected from the public domain. We have used Machine Learning and lexicon-based algorithms for SA. The goal of our project was to make the most efficient use of the SentiWordNet vocabulary to develop a Twitter Sentiment Analysis platform. Using the SentiWordNet lexicon, we obtained an accuracy of 75.20 percent for our dataset, although we observed that this number varied significantly from one area to the next. Because the current lexicon has a huge number of terms with their emotion score, it is lacking specific words that are common in a certain domain, it is preferable to construct a lexicon from the test corpus and use it for classification. Our model, which uses the Google search engine to determine a term's score utilizing pointwise mutual information, outperforms the SentiWordNet lexicon on our dataset and can deal with one of the difficulties of Sentiment Analysis—the unexpected shift from positive to negative sentiments.

REFERENCES

[1] C. Alm, D. Roth, and R. Sproat, "Emotions from the text: machine learning for text-based emotion prediction," in Proceedings of HLT and EMNLP. ACL, 2005, pp. 579–586.

[2] S. Aman and S. Szpakowicz, "Using Roget's thesaurus for fine-grained emotion recognition," in Proceedings of IJCNLP, 2008, pp. 296–302.

[3] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 27–29.

[4] M. D. Choudhury, S. Counts, and M. Gamon, "Not all moods are created equal! exploring human emotional states in social media," in Proceedings of ICWSM, 2012.

[5] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," The Journal of Machine Learning Research, vol. 9, pp. 1871–1874, 2008.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for Twitter: annotation, features, and experiments," in Proceedings of HLT: short papers, ser. HLT '11. Stroudsburg, PA, USA: ACL, 2011, pp. 42–47.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.

[8] G. Mishne, "Experiments with mood classification in blog posts," in Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access.

[9] S. Mohammad, "#emotional tweets," in Proceedings of the Sixth International Workshop on Semantic Evaluation. ACL, 7-8 June 2012, pp. 246–255.

[10] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: A novel rule-based approach to affect sensing from text," Natural Language Engineering, vol. 17, no. 1, pp. 95–135, 2011.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of EMNLP. ACL, 2002, pp. 79–86.

[12] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach." Journal of personality and social psychology, vol. 52, no. 6, pp. 1061–1086, 1987.

[13] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in Proceedings of the 2008 ACM

symposium on Applied computing. ACM, 2008, pp. 1556–1560.

[14] C. Strapparava and A. Valitutti, “Wordnet-affect: an affective extension of wordnet,” in Proceedings of LREC, vol. 4. Citeseer, 2004, pp. 1083– 1086.

[15] C. Strapparava and R. Mihalcea, “Semeval-2007 task 14: affective text,” in Proceedings of the 4th International Workshop on Semantic Evaluations, ser. SemEval ’07, 2007, pp. 70–74.

[16] R. Tokuhisa, K. Inui, and Y. Matsumoto, “Emotion classification using massive examples extracted from the web,” in Proceedings of COLING. ACL, 2008, pp. 881–888.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of HLT and EMNLP. ACL, 2005, pp. 347–354.

[18] I. Witten, E. Frank, and M. Hall, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2011.

[19] C. Yang, K. Lin, and H. Chen, “Emotion classification using web blog corpora,” in IEEE/WIC/ACM International Conference on Web Intelligence. IEEE, 2007, pp. 275–278.