International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-169
Volume: 6 Issue: 4 120 - 125

Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm

Vikas Malik
Department of Computer Science
Manav Institute of Technology & Management
Hisar, India
e-mail: vikasprof.om@gmail.com

Amit Kumar
Department of Computer Science
Manav Institute of Technology & Management
Hisar, India
e-mail: amitkrnayak10@gmail.com

***Abstract*—** Sentiment analysis the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Now a days the growth of social websites, blogging services and electronic media con-tributes huge amount of user give messages such as customer reviews, comments and opinions. Sentiment Analysis is an important term referred to collect information in a source by using NLP, computational linguistics and text analysis and to make decision by subjective information extracting and analyzing opinion, identifying positive and negative reviews measuring how positively and negatively an entity (public ,organization, product) is involved. Sentiment analysis is the area of study to analyze people's reviews, emotion, attitudes and emotion from written languages. We concentrate on field of different opinion classification techniques, performed on any data set. Now a days most popular approaches are Bag of words and feature extraction used by researchers to deal with sentiment analysis i.e used by politician, news groups, manufactures organization, movies, products etc.

Keywords- *Twitter, Sentiment Analysis (SA), Machine Learning, Naïve Bayes, Natural Language Processing (NLP).*

*****

I. INTRODUCTION

"Sentiment Analysis," otherwise known as "Opinion Mining," involves inference of sentiment (i.e., opinion) from text. For instance, movie reviews on Rotten Tomatoes are often positive or negative, product reviews on Amazon. Similarly do opinions underlie many tweets on Twitter.

Some words tend to have positive connotations (e.g., "love"), while some words tend to have negative connotations (e.g., "hate"). And so, if someone were to tweet "I love you", you might infer positive sentiment. And if someone were to tweet "I hate you", you might infer negative sentiment. Of course, individual words alone aren't always reliable, as "I do not love you" probably isn't a positive sentiment, but lets not worry about those cases. Some words, meanwhile, have neither positive nor negative connotations (e.g., "the").

A few years back, Dr. Minqing Hu and Prof. Bing Liu of the University of Illinois at Chicago kindly put together lists of 2006 positive words and 4783 negative words.

Opinion mining is not in itself a new research theme. Automated methods for content analysis have been increasingly used, and have increased at least 6 folds from 1980 to 2002 (Neuendorf, K. A. 2002. The Content Analysis Guidebook. Sage). The research theme is based in long established computer science disciplines, such as Natural Language Processing, Text Mining, Machine Learning and Artificial Intelligence, Automated Content Analysis, and Voting Advise Applications. However, according to Pang and Lee (2008), since 2001 we see a growing awareness of the problems and opportunities, and subsequently there have been literally hundreds of papers published on the subject.

II. SENTIMENT ANALYSIS

Opinion mining involves analysing opinions, sentiments or mentality of the writer from the written text. Opinion mining uses the concepts of NLP, data mining and machine learning to perform this task. This section involves analysing requirement for opinion mining. In the next segments, we concentrate on sentiment mining assignments and present a review.

A. Why Sentiment Analysis

Online opinions have indirect influence on the business of several e-commerce sites. Those sites market their products and the web users go through the reviews of the product before buying that product. Many organizations utilize opinion mining systems to track customer reviews of products sold online.

Opinion mining is an incredible way of maintaining focus on several business trends related to deals administration, status management and also advertising. Pattern prediction is also done using the opinion of the customers.

III. LITERATURE REVIEW

A. Introduction

Sentiment analysis is not a new task, it has been studied since 90s. However, in 2000s SA attracted the interest of scientists due to its significance in different scientific areas also SA had a many unstudied research questions. Moreover, the wide availability of opinionated data pushed research in this area on a new stage.

In other words, sentiment analyses deals with processing of opinionated text in order to extract and categorize opinions from certain document. The polarity of sentiment usually expressed in terms of positive or negative opinion (binary classification) [3],[4]. However, it can be multiclass classification [5],[6],[7], hence sentiment may have a neutral label or even broadened variation of labels like very positive, positive, neutral, negative, very negative, also labels can be associated with emotions like sad, anger, fearful, happy, etc.

Sentiment analysis is a developing area that arouses the interest of humans and especially organizations because SA can be used for decision making process. Individuals are no longer limited to ask opinions from friends about particular product or service they can freely find such information on the Internet. Furthermore, organizations may save time and money by avoiding of conducting surveys instead they can concentrate on processing opinions that can be obtained from the Web freely. Nevertheless, it is important to notice that sources that contain opinionated data are noisy sometimes, so it is important to extract the essential meaning from that information to use it further. SA uses different techniques and approaches for handling this challenging task.

.

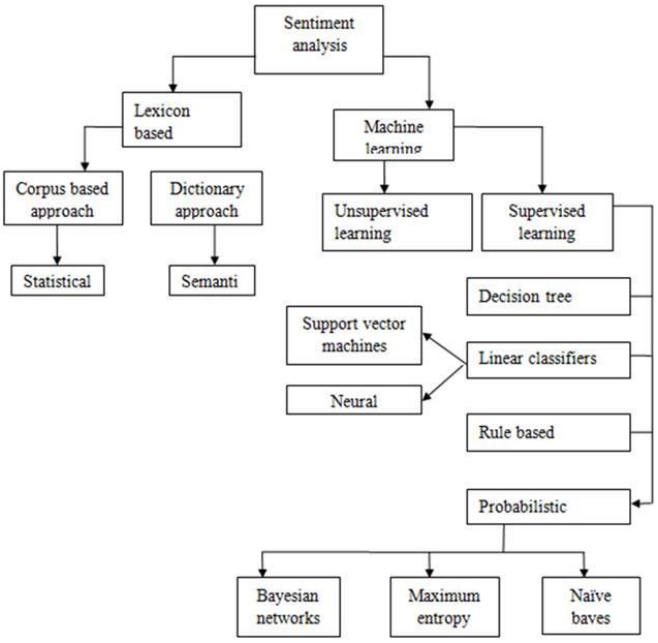

Fig:-1 Techniques for Sentiment Analysis

B. Lexicon Based Approach

The first technique that can be used for SA is the lexicon based method [2]. It uses a lexicon that consists of terms with respective sentiment scores to each term. The term can be associated with a single word, phrase or idiom [10]. The sentiment is defined based on the presence or absence of terms in the lexicon. The lexicon-based approach includes corpus-based approach and dictionary-based approach [9] that are discussed further.

Dictionary Based Approach

The main idea behind the dictionary-based approach is to use lexical databases with opinion words to extract sentiment from the document. Based on [1][11] a set of seed sentiment words (e.g. good, bad) with their polarities is collected by hand. At the beginning, this initial set does not have to be large, 30 opinion words is enough [12]. Next step is to use the polar words to enrich a set by looking up for respective synonyms and antonyms in a lexical database. Examples of such databases are WordNet [13], HowNet [14], SentiWordNet SenticNet, MPQA [15] etc. The look-up procedure is iterative. At each iteration the algorithm takes updated set of words (expanded set) and does search again until there will be no new words to include. In the end, a set of sentiment words can be reviewed with a purpose of deleting errors.

Hu and Liu [12] have focused their research on the classification of customer reviews, namely they extracted product features that contain sentiments, then classified sentences based on that features and as a result, the summary of the product reviews was composed. For example, if a review was about a camera, authors retrieved such features as picture quality and size of the camera, and using these features, the classification was made on positive and negative camera reviews. In order to assign a positive or negative tag for a sentence, first, researchers retrieved polar words from each review. In this case, adjectives were used. The prediction was based on the polarity of an adjective which had the same polarity as its synonyms and opposite to polarity of its antonyms. Polar words were utilized for searching their synonyms and antonyms with known orientation in WorldNet. Therefore, the orientation of polar words that appear in the review was identified. The method that was described in [12] showed good results, average accuracy constituted 84%. Hence, current method can be effective for prediction of adjective semantic orientations and sentence polarity.

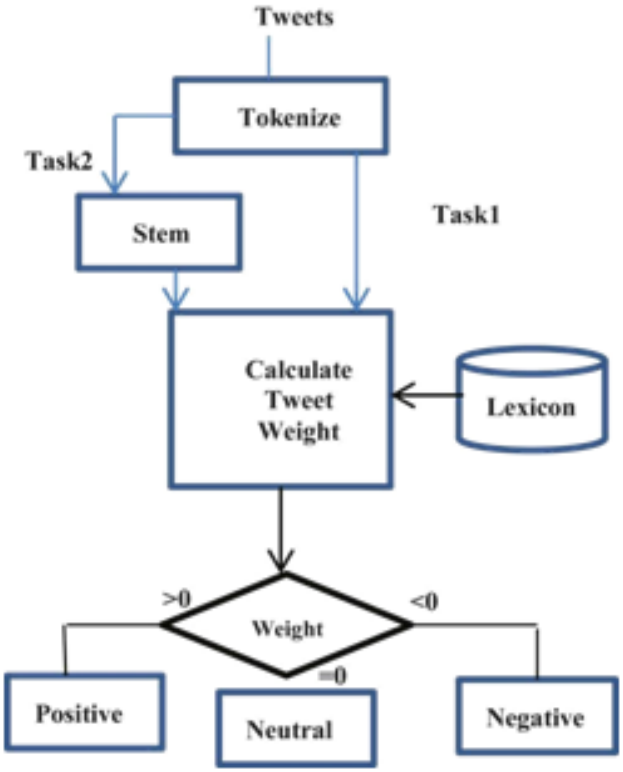

Fig:-2 Processing of Lexicon Based Approach

Kim and Hovy [16] investigated the sentiment of the text and its holder regarding a given topic. Authors of the research paper [16] have applied several classifiers. The first classifier was applied to each word in the sentence to get its polarity. The second classifier defined the polarity of the entire sentence expressed by opinion holder. In addition, the authors introduced the use of small initial list of seed words in a similar way as in [12] (adjective and verbs). This latter was extended by looking up for corresponding synonyms and antonyms in WorldNet. Authors mentioned that some synonyms/antonyms had neutral or even opposite orientation that makes them inappropriate to use. Moreover, the researchers emphasized the necessity of defining the strength of positiveness and negativeness of the words that would allow to eliminate ambiguous words. Kim and Hovy identified the four different regions in the sentence that are close to opinion holder and can contain sentiment. For determining the sentence orientation, authors developed three models. First model was based on the assumption that negatives cancel one another out [16]. Second and third models were the harmonic and geometric mean of the sentiment strengths in the particular region respectively. After con-ducting experiments it was concluded that the best results retrieved by using first model and region that starts from opinion holder to the end of the sentence.

Generally, the main drawback of dictionary-based approach is the inability to detect sentiment words with domain and context specific polarity orientations [2].

Corpus Based Approach

In [1] Bing Liu indicates that corpus-based approach can be applied in two cases. First case is an identification of opinion words and their polarities in the domain corpus using a given set of opinion words. The second case is for building a new lexicon within the particular domain from another lexicon using a domain corpus. The findings suggest that even if opinion words are domain-dependent it can happen that the same word will have opposite orientation depending on context.

The research conducted by Hazivassiloglou and McKeown [18] is prominent in the literature about corpus-based technique. Authors proposed a method that extracts semantic orientation of conjoined adjectives from the corpus. The technique is based on the usage of textual corpora and seed opinion words (adjectives). Special linguistic rules are applied to the corpora in order to discover opinion words with corresponding polarities. Authors assume that adjectives have the same polarity if they are joined by the conjunction and. However, the conjunction “but” is used for linking adjectives with opposite polarities. Additionally such conjunctions as or, either-or, neither-nor are used. Sometimes these rules are not applicable. Therefore, authors also predict the polarities of the conjoined adjectives to check whether the polarities are same or not, for this purpose log-linear regression model is used. After prediction stage, the graph is obtained that provides links between adjectives. Then clustering is carried out on the graph to divide adjectives into positive and negative subsets. To conclude, Hazivassiloglou and McKeown were able to achieve 90% precision.

As mentioned above, the same sentiment word can have different semantic orientation depending on the context. Ding et al. [19] proposed a method for finding the orientation of sentiment conveyed by reviewers. Authors have emphasized that some adjectives (mostly quantifiers, like long, short, etc.) are context-dependent and can change their polarities. Researchers consider sentiment words with their aspects in the sentence in order to identify the polarity of the product feature. Ding et al. use words, phrases, and idioms as an opinion lexicon. List of adjectives and adverbs is taken from [12] and extended by authors to include verbs and nouns. Moreover, they annotated around 1000 idioms that contain clearly expressed sentiment. After the lexicon is ready, they define the polarity score for each feature in the review sentence. To get the score for the whole sentence they sum up all the scores using proposed score function that gives better results than simple summation used in [12]. Additionally, authors applied several linguistic rules for handling negations and sentences that contain the conjunction “but”. Furthermore, the paper introduces a holistic approach to solve the problem of the

identifying polarity of context dependent sentiment words. For this purpose, three consistency techniques about connectivity are suggested [19]: intra-sentence conjunction technique, pseudo intra-sentence conjunction technique, and inter-sentence conjunction technique. To sum up, the authors report that the proposed approach is effective and gives better results than previously proposed methods.

The corpus-based method alone is less effective that dictionary-based method due to a limitation of words that are in the corpus. However, usage of this approach can help to construct domain and context specific lexicon. Overall, the performance of lexicon-based methods in terms of time complexity and accuracy heavily depend on the number of words in the dictionary, namely, performance decreases significantly with the exponential growth of the dictionary size[20].

C. Machine Learning Approach

The second technique that can be used for Sentiment analysis is machine learning that includes unsupervised and supervised machine learning methods that are explained below.

Unsupervised Machine Learning Methods

Unsupervised learning [9] approach uses unlabeled datasets in order to discover the structure and find the similar patterns from the input data. Unsupervised method is usually used when a collection of reliable annotated dataset is difficult, but collecting of unlabeled data is easier. It does not cause any difficulties when new domain-dependent data have to be retrieved.

Supervised Machine Learning Methods

Supervised machine learning methods assume the presence of labeled training data that are used for the learning process. The latter esti-mates the output from the input dataset, we re-fer to the case when the classifier defines the label the object belongs to. As training data set, labeled documents have to be used. Usu-ally, bag-of-words model [24] is employed to represent a document as a feature vector

$$d = (w_1, w_2 \ldots\ldots, w_{i\ldots\ldots} w_N) \quad (1)$$

, where N is set of all the unique terms in the training dataset and w_i is weight of the i_{th} term. To convert training dataset to a feature vector, vocabulary with N unique words has to be created from the training data. Further, any of feature models can be used for constructing a feature vector itself.

After the dataset is represented as a vector, it can be used by the classifier for learning and estimating labels. Different kind of methods can be used for training the classifier. Lets discuss some of them.

IV. SENTIMENT ANALYSIS USING NAÏVE BAYES CLASSIFIER

In this section we introduce the Naive Bayes Classifier, that makes a simplifying (naive) assumption about how the features interact and analyses performance of Naive Bayes algorithm on real world dataset.

A. Data Set

This data originally came from Crowdflower's Data for Everyone library.

(http://www.crowdflower.com/data-for-everyone).

As the original source says,

We looked through tens of thousands of tweets about the early August GOP debate in Ohio and asked contributors to do both sentiment analysis and data categorization. Contributors were asked if the tweet was relevant, which candidate was mentioned, what subject was mentioned, and then what the sentiment was for a given tweet. We've removed the non-relevant messages from the uploaded dataset.

I decided to only do sentiment analysis on this dataset, therefore I dropped the unnecessary columns, keeping only sentiment and text.

B. Processing Data Set

A tweet contain s a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features . The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points

- Remove all URLs (e.g. www.xyz .com), hash tags (e.g. topic), targets (@u ser-name) .
- Correct the spellings; sequence of repeated characters is to be handled .
- Replace all the emoticons with their sentiment .
- Remove all punctuations ,symbols, numbers .
- Expand Acronyms(Use a acronym dictionary).
- Remove Non - English Tweets.

Stop Word: Stop Words are words which do not contain important significance to be used in Search Queries. Usually these words are filtered out from search queries because they return vast amount of unnecessary information. (the, for, this etc.)

For using our machine learning algorithms we first need to

convert words to vectors. After that we can run our algorithms on that feature vector. We then feed this vectorised data to our algorithm.

C. Algorithm

The most common and simple method that is used for text classification is Naive Bayes. The model is based on Bayes theorem with the assumption that features are independent. Naive Bayes classifier defines the probability of the document belonging to a particular class. The advantages of the Bayes classifier are: simplicity of the implementation, learning process is quite fast, it also gives quite good results [4],[20], [21], [22]. However, naive assumption may cause a problem because in the real world features are dependent.

According to the idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint. The probability of the document belonging to a particular class [23] estimated as:

$$P(C|d,\lambda)=\frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_{cj}\exp[\sum_i \lambda_i f_i(c,d)]} \qquad (2)$$

Where, c is the class, d is the document to be classified, f is a weight of the i_{th} classification indicator f_i . Maximum Entropy classifier does not assume independence of features. Thus such classifier theoretically may outperform Naive Bayes. However, Maximum Entropy algorithm is more difficult to implement and learning process is slower.

D. Result

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

Accuracy = (TP+TN)/(TP+TN +FP+FN)

Precision = TP /(TP + FP)

Recall = TP /(TP + FN)

F1 = (2*Precision*Recall)/(Precision + Recall)

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances.

TABLE I. ACTUAL POSTIVE AND ACTUAL NEGATIVE

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

After analysing result of our machine learning model on testing set, we obtained following result:

TABLE II. CALCULATED TWEETS

True Positive (TP)	91
False Negative (FN)	139
False Positive (FP)	59
True Negative (TN)	790

Now we will see Accuracy, Precision, Recall and F-Score of our result:

TABLE III. FINAL RESLUT

Accuracy	81.64
Precision	60.6
Recall	39.56
F-Score	49.25

V. CONCLUSION

In this project I was curious how well NLTK and the Naïve Bayes Machine Learning algorithm performs for Sentiment Analysis. In my experience, it works rather well for negative comments. The problems arise when the tweets are ironic, sarcastic has reference or own difficult context.

Consider the following tweet: ”Muhaha, how sad that the Liberals couldn’t destroy Trump. Marching forward.” As you may already thought, the words sad and destroy highly influences the evaluation, although this tweet should be positive when observing its meaning and context.

To improve the evaluation accuracy, we need something to take the context and references into consideration. I will try to build an LSTM network, and benchmark its results compared to this NLTK machine learning implementation.

VI. REFRENCES

[1] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 1-167.
[2] Medhat, W., Hassan, A., Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.
[3] 3.Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).
[4] Btow, F., Schultze, F., Strauch, L. Semantic Search: Sentiment Analysis with Machine Learning Algorithms on German News.
[5] Pak, A., Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREc (Vol. 10, No. 2010).

[6] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In Advances in ICT for emerging regions (ICTer), 2012 International Conference on (pp. 182-188). IEEE.

[7] Hallsmar, F., Palm, J. (2016). Multi-class sentiment classification on twitter using an emoji training heuristic.

[8] Turney, P. D. (2002, July). Thumbs up or thumbs down?:semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

[9] Salas-Zrate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodrguez- Garca, M. ., Valencia-Garca, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computational and mathematical methods in medicine, 2017.

[10] Chiavetta, F., Bosco, G. L., Pilato, G. (2016). A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language.

[11] Hailong, Z., Wenyan, G., Bo, J. (2014, September). Machine learning and lexicon based methods for sentiment classification:A survey. In Web Information System and Application Conference (WISA), 2014 11th (pp. 262-265). IEEE.

[12] Hu, M., Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[13] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[14] Dong, Z., Dong, Q., Hao, C. (2010, August). Hownet and its computation of meaning. In Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (pp. 53-56). Association for Computational Linguistics.

[15] Musto, C., Semeraro, G., Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog posts. Information Filtering and Retrieval, 59.

[16] Kim, S. M., Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics

[17] Park, S., Kim, Y. (2016, June). Building thesaurus lexicon using dictionary-based approach for sentiment classification. In Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on (pp. 39-44). IEEE.

[18] 18.Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

[19] 19. Ding, X., Liu, B., Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240). ACM.

[20] Thakkar, H., Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. arXiv preprint arXiv:1512.01043.

[21] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

[22] Rothfels, J., Tibshirani, J. (2010). Unsupervised sentiment classification of En-glish movie reviews using automatic selection of positive and negative sentiment items. CS224N-Final Project.

[23] Tang, B., Kay, S., He, H. (2016). To-ward optimal feature selection in naive Bayes for text categorization. IEEE Transactions on Knowledge and Data Engineering, 28(9), 2508-2521.