Sentiment Analysis on Twitter Data using Support Vector Machine

Somiya Rani*1, Shobha Bhatt*2

*1 Assistant Professor, Department of Computer Science & Engineering, AIACTR, Delhi, India
*2 Student

Abstract—Social media is a popular platform for people to convey or share their messages, comments, reviews, blogs etc with other people. Twitter is a famous social networking platform that allows people to express their opinions in form of tweets. These tweets are short messages with a maximum length of 140 characters. This restriction on the length of tweets limits a user to be precise and accurate to express their sentiments. Sentiment analysis of tweets has become a most popular field of data analytics and mining. Sentiment analysis in simple terms is defined as a process of analyzing the data and classifying it into a category i.e, positive and negative. Sentiment analysis is mainly done using machine learning approach. In our work, we have used Support Vector Machine Classifier. The work implemented using python and its libraries like NLTK, Numpy, Scikit-learn, Scipy. We created our own datasets using Twitter API. Performance evaluation is done in terms of accuracy.

Keywords—Sentiment Analysis, Opinion Mining, Machine Learning, Support Vector Machine.

I. INTRODUCTION

Sentiment analysis is a process of analyzing the data like tweets, comments, reviews, blogs about a particular topic, product and utilizing for making specific kind of decision. Sentiment analysis is generally performed at three levels. The first is sentence-level sentiment analysis, in which a particular sentence is analyzed. Second is document-level sentiment analysis. At this level, instead of analyzing a particular sentence, we take the whole document for analysis. A document is nothing but a collection of sentences. Last is the aspect based sentiment analysis in which specific aspects of entities are considered in the sentence. Sentiment analysis can be done using different approaches among which the machine learning is the popular one. Machine learning techniques are categorized into supervised and unsupervised techniques. The supervised machine learning techniques utilize predefined dataset to give the end result. This is not the case for unsupervised machine learning techniques. The predefined dataset is absent in unsupervised techniques [1].

Sentiment Analysis can be categorized into various research fields such as subjectivity detection, sarcasm detection, fake review detection, text summarization etc [2]. Sentiment analysis can be useful in decision-making process such as whether the launch of a product is successful in the market or not, which version of the product is liked by customers. Sarcasm, fake reviews, fake star ratings are some common challenges for sentiment analysis and a lot of research has been done to overcome these challenges.

A number of techniques for supervised learning have been examined in past years but none of them has been proven to perform consistently well in all the domains. We surveyed many papers and most them proved that the Support Vector Machine(SVM) classifier can easily outperform the other machine learning techniques.

We have used SVM in this paper which has shown the promising results. In section 2, we have discussed the related work and how different researchers carried out their experiment. In section 3, we elaborated our proposed system. This section discusses the datasets we used, steps we chose for pre-processing and how the classifier will work. Next, we discuss the how the experiment is performed and the libraries we required for it. The statistics about the data is given. The result has been discussed in section 5. we compared accuracy achieved for two different datasets and gave a graphical representation of it.

II. LITERATURE REVIEW

Twitter is a popular social media platform that allows people to communicate using short tweets and texts. The Maximum length of a tweet is restricted to 140 characters. This restriction leads to the use of informal and less consistent language as the user have to convey his/her sentiment in only one or two sentences.

A lot of research has been done in this field in the past years. Various machine learning techniques have been employed and have shown promising results.

The authors in [3], proposed methodologies for sentiment analysis of opinions on the web forum. They used syntactic and stylistic feature sets for classification. In addition to these feature sets, they utilized genetic algorithm and information gain for feature selection. Their system design has two steps: the first was to select the feature sets and the second was to perform feature selection. Their system was able to perform analysis on opinions in English and Arabic language. They developed an Entropy Weighted Genetic Algorithm for a better selection of features to get improved and high accuracy. U.S. & Middle Eastern web forum postings were used as a dataset. 95.55% accuracy was achieved.

In [4], the authors used pre-labeled twitter dataset to train the SVM classifier. They used hashtags to retrieve latest tweets for their dataset. They took 3000 tweets for each class.
i.e., positive, negative and neutral. So the whole dataset they utilized for their work carried 9000 tweets. They used Twitter hashtags to determine the polarity of the tweet. To examine the accuracy of their proposed techniques, they used a random list of 100 tweets labeled manually. 80 out of 100 tweets were classified accurately giving an accuracy of 85%.

In [5], the authors presented an optimized SVM in addition to two other machine learning techniques namely Naive Bayes classifier (NB) and SVM to show comparative results of their study. They used three different datasets i.e., movie dataset, twitter dataset from twitter API and gold dataset from amazon.com to show the performance comparison of proposed technique on each dataset. The results showed that the optimized SVM performed better than other two classifiers. They achieved maximum accuracy of 72.74%, 74.73% and 76.92% for gold, movie and twitter dataset respectively with optimized SVM. However, NB classifier showed 69.10%, 74.55% and 76.67% for gold, movie and twitter dataset respectively and 72.74%, 74.73%, and 76.92% with SVM for gold, movie and twitter dataset respectively.

The authors in [6] analyzed product reviews from twitter. They used a prelabeled dataset of 18340 tweets and an unlabeled testing dataset of 1000 tweets. To find the accuracy of unlabeled test data, they used NB, Maximum Entropy (ME), SVM and WordNet in their Paper. Sentiment analysis utilizing WordNet gave the maximum accuracy of 89.9%. The NB classifier can perform better than ME classifier and SVM with unigram Model. The NB classifier showed 88.2% of accuracy while the ME classifier and SVM showed 83.8% and 85.5% of accuracy respectively.

III. METHODOLOGY

In figure 1, we presented a flow chart of our proposed system. The dataset, steps involved in preprocessing, feature model used and how the classifier will work have been discussed.

A. Dataset

The first phase of our system is to collect the tweets for training dataset. We have used two different polarity datasets having tweets belonging to one of the two classes (positive and negative). The tweets in the datasets have been extracted using Twitter API of Twitter Developer Platform.

The first twitter dataset (dataset1) have total 1100 tweets partitioned into training dataset1 with 1000 tweets and testing dataset1 with 500 tweets.

The second twitter dataset (dataset2) have total 15,662 tweets partitioned into training dataset2 with 10,662 tweets and testing dataset2 with 5000 tweets.

B. Preprocessing

The second phase of the proposed methodology is to preprocess the stored tweets. A tweet may contain vague information, which makes it difficult for the system to find the polarity. Therefore, we need to normalize the fetched data through pre-processing steps. To make it suitable for learning algorithms, the feature set size need to be decreased. Hence, we applied the following series of pre-processing steps.

- **Handle Removal:** Every user of Twitter has a unique username. This unique name in twitter is called a handle. Twitter handle is represented using the symbol ‘@’. Anything that is directed towards a user can be indicated by ‘@’ followed by the username. Twitter Handle does not add any information to the analysis process. So we have removed handle during preprocessing.

- **Hashtag Removal:** Hashtags are used to name subject or phrase that are currently on the most trending topics. A hashtag is represented by a symbol ‘#’. We removed this symbol while preprocessing and kept the name of the subject or the phrase as it is.

- **URL Removal:** Tweet feeds may contain URL. These URL does not make any contribution to the sentiment analysis. So we have also removed these URLs from the tweets while pre-processing step.

- **Punctuation Removal:** Punctuations that does not provide any kind of information about the sentiments of the text have to be removed. Here, we have not included the punctuations like a question mark, exclamation mark that are important from classification view.

- **Stop Word Removal:** Prepositions, pronouns, conjunctions have no specific meaning so we have removed them as well.

- **Tagging:** Suitable Part-of-Speech tag is marked to each word in the sentence. Some common POS categories are noun, pronoun, verb, adverb, adjective, conjunction, interjection, and prepositions.

- **Stemming:** Stemming is a process of converting a word into its base or root form. For example, words like saw, seen will be converted to see. There are
many algorithms available for stemming. We used Porter stemmer. It is used for term normalization.

- **Tokenization**: Tokenization is a process of dividing a sentence into chunks. After normalizing the tweet, it will be partitioned into chunks.

**C. Feature Extraction**

After pre-processing of raw data is done, the features from the tweets will be extracted. The aspect( adjective) will be extracted from the dataset in this phase. In our work, we have used unigram model. This unigram model will take out the adjective and separates it from the tweet. Then this feature will be utilized for determining the polarity of the tweet.

**D. Classification**

The next phase of the proposed methodology is classification phase. In this phase, the normalized data obtained from the preprocessing phase is given as an input to SVM classifier. The classifier is trained with the training dataset.

SVM is a supervised machine learning algorithm that tries to find an optimal hyperplane that best separates the two classes. Each feature is plotted as a data point in an n-dimensional space. The features that are closest to the street(margin around separating hyperplane) are called as support vector. The classifier will take the test dataset as an input and test it by comparing with the trained dataset utilizing the lexicon of positive and negative words. The output of the classifier is the polarity of tweets. We implemented this classifier in Python and its libraries.

**IV. EXPERIMENTATION**

We have used two labeled datasets(dataset1 and dataset2) in our work. Both datasets have been used separately with the same set of features to make a comparative analysis. For the purpose of classification, the classifier needs to be trained using training dataset. For this reason, the dataset is partitioned into training and testing dataset. Statistics about both datasets is given table 1. Further, we partitioned the training into two classes i.e., positive and negative. The training dataset1 of 1000 tweets partitioned into the positive and negative files, each carrying 500 tweets. While training dataset2 of 10,662 tweets partitioned in 5,331 tweets in both positive as well as the negative file. Statistics about training dataset1 and training dataset2 is given table 2 and table 3. Files in both the dataset are stored in CSV format. All these tweets are collected using twitter API. We then used consumer key, consumer secret, access token and access token secret to extracting the tweets for our datasets.

We built a lexicon of 6,614 words carrying both positive and negative words, as given in table 5. The SVM classifier utilize these lexicons while making the classification of tweets into positive and negative class.

<table>
<thead>
<tr>
<th>Twitter Datasets</th>
<th>Type</th>
<th>Total number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter Dataset1</td>
<td>Training Dataset1</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>Test Dataset1</td>
<td>100</td>
</tr>
<tr>
<td>Twitter Dataset2</td>
<td>Training Dataset2</td>
<td>10,662</td>
</tr>
<tr>
<td></td>
<td>Test Dataset2</td>
<td>5000</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th></th>
<th>Training Dataset 1</th>
<th>Number of tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive.csv</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>Negative.csv</td>
<td>500</td>
</tr>
</tbody>
</table>

**TABLE IV**

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Lexicon</th>
<th>Total number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive Words</td>
<td>2,753</td>
</tr>
<tr>
<td>2</td>
<td>Negative Words</td>
<td>3,865</td>
</tr>
<tr>
<td>3</td>
<td>Stop Words</td>
<td>338</td>
</tr>
</tbody>
</table>

We have performed this experiment using python libraries. The libraries used are NLTK, scipy, sci-kit-learn, pandas, and numpy. NLTK is a platform that allows us to work with human language data. Various text processing libraries supported by NLTK for tokenization, stemming, tagging etc. Scipy is scientific computing library for science, mathematics,
and engineering. Scikit-learn is python tool to perform classification, regression, clustering and various other data analytical operations. Numpy is python package for scientific computations.

V. RESULTS

To test the accuracy of our system, we used a random, manually labeled dataset of 100 tweets. These tweets are manually labeled into positive and negative class. These tweets in test dataset will be then given as an input to the trained classifier. 51 tweets out of 100 in test dataset 1 have been correctly classified for training dataset 1. 358 tweets out of 500 in test dataset 2 have been classified accurately by the classifier. So, the overall accuracy achieved for the first dataset is 50.85% whereas 71.60% is achieved for the second dataset. A graphical comparison of the accuracy for both datasets is given in figure 2.

![Accuracy comparison of the classifier](image)

**TABLE V**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy of SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter dataset 1 of 1000 tweets</td>
<td>50.85%</td>
</tr>
<tr>
<td>Twitter dataset 2 of 10,662 tweets</td>
<td>71.60%</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

In this paper, sentiment analysis on twitter data using Support Vector Machine has been done. The system is implemented using Python libraries. The proposed model, first collect the data from twitter which then go through the preprocessing and classification stage. We have used unigram+stop words feature set in this experiment. The experiment with dataset1 showed better results compared to dataset2. We found that size of a dataset greatly impacts on the accuracy of the classifier. While we created a simple model, a number of approaches can be developed to enhance the system accuracy. More features can be used to perform more accurate classification.

REFERENCES


