Sentiment Analysis on Twitter Data using Support Vector Machines

Gaurav Vinod Bhambhani
Computer Science and Engineering Department
Institute of Technology, Nirma University
Ahmedabad, India
18BCE072@nirmauni.ac.in

Manish Pradeepkumar Gupta
Computer Science and Engineering Department
Institute of Technology, Nirma University
Ahmedabad, India
18BCE075@nirmauni.ac.in

Ishan Vijay Tewari
Computer Science and Engineering Department
Institute of Technology, Nirma University
Ahmedabad, India
18BCE080@nirmauni.ac.in

Abstract— **In this era of increasing dependency on digital media, a high amount of User data is available for researchers to mine and analyse. Due to this, sentiment analysis has been popular for a long time. A lot of research work is being done to analyse data and draw out systematic and valuable inferences from it.**

Performing sentiment analysis on textual data is a challenging task even though much research has been done. In this paper, we perform sentiment analysis on data obtained from a social media platform called Twitter in the form of text.

To perform this task, we will use the Supervised Machine Learning Algorithm called Support Vector Machine due to its popularity in performing classification and regression analysis using data analysis and pattern recognition.

Keywords— Sentiment Analysis, Support Vector Machine, Classification, Naïve Bayes, Maximum Entropy

I. Introduction

Analysis of sentiments of data means to identify whether the given data is positive, neutral or negative.

Sentiment Analysis is widely used in various industries in fields such as market research, understanding and enhancing user experience, monitoring social media platforms, etcetera.

Due to the variations in how sentiments are expressed in different individuals, it gets a little strenuous and tedious to analyse the data and find out the sentiment behind it.

With the advent of technology, people have moved towards social media to express their feelings, emotions and even to give their day to day life updates through various platforms such as Twitter, Facebook, Instagram and many more in the form of tweets, videos and images, stories, etcetera. Due to this, a large amount of practices are being carried out to mine the data from these platforms to get a complete analysis of the users sentiments.

Not only Machine Learning researchers, but also psychologists have started taking these things into account. For example, traditional psychology research is based on questionnaires and academic interviews, but many psychologists are turning to web media to try to analyse the data from the users view point.

Although sentiment analysis has been there for a long time and is widely investigated from different perspectives, it is still considered a challenge.

In this paper, we have tried to counter this challenge by using Support Vector Machines.

The rest of the paper is organized as follows,In Section 2, we discuss various other approaches to perform Sentiment analysis using various other Machine Learning algorithms. In Section 3 we talk about the proposed approach to perform sentiment analysis using Support Vector Machine, In Section 4 we talk about the results and experiments carried out. Finally Section 5 is the conclusion wherein we conclude the work done and discuss the future aspects of this domain.

II. Related Works

Sentiment analysis is fundamentally used to express the emotion of an individual. Applying it on textual data is one of the hot topics today. Current state-of-the-art methods usually focus on classifying them into two classes - positive, negative. This section discusses the various approaches adopted to detect the sentiment using the textual data of user reviews.

Many researches are working on the automated techniques of extraction and analysis of this textual data. In [1], the authors proposed a way to extract the pre-labelled data from twitter which is then used to train a SVM classifier. Hashtags were also incorporated to judge the polarity of the tweet. Upon testing, the accuracy obtained was 85%. The authors of [2] utilized J48 and MLP for classification using five different datasets. They used TP rate, FP rate, Precision, Recall, F-measure and ROC area. They observed that MLP performed best on all the datasets and their analysis showed that the Neutral Network also has the better learning capability. In [3] the authors introduced a novel approach to classify the sentiment of the tweets as either positive or negative. In their work distant supervision was used to discuss and compare the results of various Machine learning models for twitter sentiment analysis.

As per the experiments conducted, Machine Learning algorithms such as Naive Bayes, Maximum Entropy and SVM performed good with an accuracy of about 80% when trained with emotional tweets. In the paper [4], authors presented an application of Arabic sentiment analysis on twitter data by analyzing 1000 tweets using Naive Bayes and SVM. In [5], authors have proposed an efficient feature vector technique by dividing the feature extraction process in two steps. First the features are extracted and added to the feature vector, after that those features are removed from the tweets and the process is carried out again just like in case of text. The extracted features in the second steps are then added again to the feature vector. The performance of algorithms like Naive Bayes, SVM and Maximum Entropy are similar.

The various methods discussed above demonstrate the various approaches used for extracting the sentiment from textual data. Among all the Machine Learning models, SVM, Naive Bayes and Maximum Entropy gave better results and thus we have chosen SVM and applied to the textual data to perform sentiment analysis. More details are to follow in the next section.

III. Methodology

The workflow starts by importing the dataset, which is obtained from sentiment140 dataset[6] which utilizes the Twitter API for generation of the dataset. Then we have the data preprocessing module where we refine the data obtained from the dataset by removing unwanted data fields and performing processes such as tokenization, vectorization, normalization, lemmatization, and dividing the comments into bigrams and trigrams. After this the text classifier is trained on the refined text data, in the training phase.Class prediction is made on the test dataset to identify the sentiment of the comment, in the testing phase.

A. Dataset Used

In this paper, we have used the sentiment140 dataset [6]. It contains 1,600,000 tweets from which according to our systems capabilities, we have extracted 16000 random entries using uniform random sampling, so that we get the negative and positive samples in almost equal proportion, which consists of 8045 negative entries and 7955 positive entries. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiments .

The dataset consists of 6 fields from which we have only used 2 fields, most relevant to our task, which were the ‘Target’ field which consists of the polarity of the tweet (0 = negative, 4 = positive) and the ‘Tweet’ field which contains the text of the tweet.

B. Data Preprocessing

1. Tokenization : It is where we have converted our text data into tokens or small chunks of words before performing vectorization since it becomes easier to remove unnecessary tokens.

2. Normalization : It is where we remove unnecessary data or noise from our text data and clean our data as much as possible and make it consistent.

3. Vectorization : It is where we convert our data into vectors. We give our data in the form of text, a numerical representation. On performing vectorization, we obtain a sparse matrix.

4. Lemmatization : This is the process in which we obtain the base or dictionary meaning of the text data fed into the system to group words with the same base meaning together which makes classification easier.

5. Division of data into bigrams and trigrams : This is where the individual text data is put into a set of consecutive words, where bigram refers to groups of 2 consecutive words and trigrams refers to the groups of 3 consecutive words.

C. Training and Testing

1. We use the Support Vector Machine Classifier (SVM) for classification of data.

2. We tokenize the data and vectorize it so that each entry is converted into numerical representation. If the word is present, it is assigned it’s frequency, else it is assigned 0.

3. To create these vectors we use the CountVectorizer from sklearn.

4. We then split the data into a training set and testing set.

5. Even though we have only 2 classes in our data, to generalize it for multiclass classification, we use OneVsRestClassifier with linear kernel where Kernel Coefficient (gamma) = 0.01 and Regularization Parameter (C) = 100.

6. We fit the training data to the classifier and calculate the classification score which comes around 74%, which is an acceptable score.

IV. Results

1. The results are evaluated on the basis of F1 Score and Accuracy.

2. F1 Score is the primary performance measure while Accuracy is the secondary performance measure.

3. F1 Score is calculated using precision (P) and recall (R) values.

$$F1\ Score\ =\ 2 \times \frac{P \times R}{P+R}$$

4. The accuracy score, although acceptable, is slightly less because the tweets contain noisy data which occupies a lot of space.

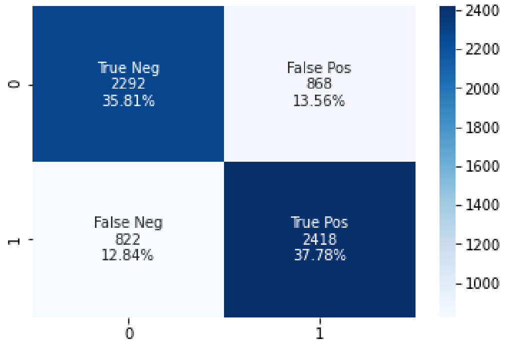

Fig. 1. Confusion Matrix

TABLE I. Result

Algorithm name	**Performance Metrics**			
	Precision	***Recall***	***F1 Score***	***Accuracy***

Support Vector Machine	Positive = 74%
	Negative = 74%
		Positive = 75%
		Negative = 73%
			74%	74%

Conclusion

In this paper, we analysed the results obtained when using Support Vector Machine Algorithm to perform the sentiment analysis on the textual data. To test the performance we used a single dataset which consisted of tweet data. Results are measured in terms of precision, recall and f-score.

The results clearly show the dependency of Support Vector Machine on the input dataset. The performance of these algorithms can be further tested with bigger and varied datasets. Moreover it can also be thought of for classification purposes, which Machine Learning algorithm performs better on which type of dataset and what might be the reasons. This can lead the researchers to the improved versions of machine learning algorithms for classification purposes.

References

[1] Zgheib, W. A., & Barbar, A. M. A Study using Support Vector Machines to Classify the Sentiments of Tweets.

[2] Arora, R. (2012). Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications, 54(13)

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

[4] Shoukry, A., & Rafea, A. (2012, May). Sentence-level Arabic sentiment analysis. In Collaboration Technologies and Systems (CTS), 2012 International Conference on (pp. 546-550). IEEE

[5] Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE.

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