A Model to Analyse Subjective Sentiment Information from Twitter Corpus Using Machine Learning Approach

Dolly Khandelwal
PG Scholar, CSE, SSTC (SSGI), CSVTU, Bhilai, Chhattisgarh, India.

Dr. Megha Mishra
Sr. Assitant professor, CSE, SSTC (SSGI), CSVTU, Bhilai, Chattisgarh, India.

Abstract – Sentiment analysis is the process of extracting and analyzing the sentiments or opinions from the reviews expressed by the people towards the specific topics of interest using machine learning. In recent years, the Internet has become an important source of information, because of the increase in social network contacts, discussion forums and blogs. Twitter is the famous microblogging site to post their opinions on various topics. As opinions are important for efficient decision making, sentiment analysis is used to determine the information that is relevant for users. In this paper, an approach is used that classifies the opinions into categories as positive or negative or neutral. The classification is done on the three multidimensional fields that are politics, entertainment, and companies which gives the overview of what is going around the world. The dataset is created by using the Twitter API. The classifiers used in this paper are Naive Bayes, Maximum Entropy, and Baseline. Finally, unigram feature is used for feature extraction and the performance of different machine learning classifiers are compared.

Index Terms – Sentiment Analysis, Twitter, Machine Learning, Polarity Classification, Classification Algorithms.

1. INTRODUCTION

Microblogging sites are gaining much popularity and attention as millions of people are sharing their views, opinions and suggestions on the daily basis. As people used to express their thoughts and feelings, there has been an enormous growth in the social media that has become a valuable resource for pattern and trend analysis. Among the various microblogging sites twitter is widely used because of the prevalence by the famous personalities. Twitter allows the users to post their messages called tweets up to the 140 characters.

With the rise of social media to extract the useful information from the large amount of opinionated text is a formidable task. So, Sentiment analysis comes into the play, which extract, organize and analyze the information which can be used by the people for better decision making.

Sentiment analysis is the process of finding the opinions and affinity of people towards a specific topic of interest. Sentiment analysis is mainly concerned with the polarity classification of the opinions of each tweet. However, various machine learning techniques are employed to analyze the sentiment from the tweets.

Following is the figure which explains the process to perform sentiment analysis. The input is the text and output is the polarity of the given text.

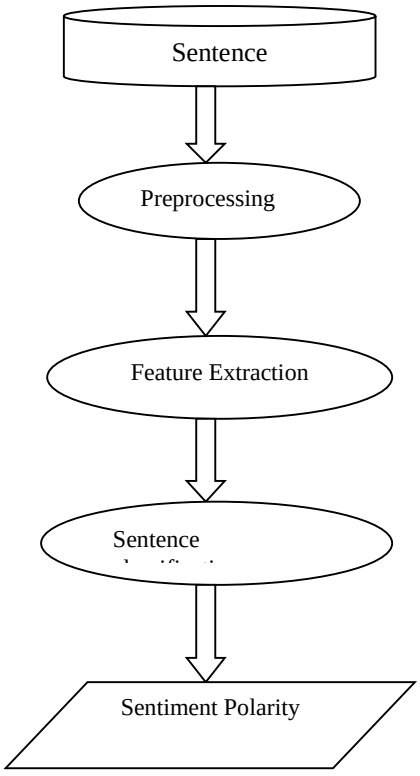

Fig 1: Process of Sentiment Analysis

As twitter data are a challenge to the sentiment analysis because of the short posts, but it is also an advantage as users write straight to the point post. Twitter is a tricky language

because users use hashtags, spelling mistakes, URLs as tweets are confined to 140 words. Sentiment analysis is a system that takes the sentences as input, analyze it and determine the polarity of tweets.

2. RELATED WORK

In the paper for sentiment analysis in various algorithms and methods used by various researchers are investigated.

In paper [1], the author described the various applications and challenges faced by the sentiment analysis. The challenges can be solved by the more innovative technology.

In [2], the author explored the approaches, techniques, challenges and applications can be used for sentiment analysis. Business organizations use sentiment analysis for their growth. Many researchers have found that SVM provide high accuracy.

The author in [6] has presented an approach for classifying the user opinions towards the political candidates suing different classifiers. They arrived at a conclusion that SVM gives good performance on text categorization and for performance evaluation precision and recall are used.

In paper [7], the author implemented a model for sentiment analysis on movie reviews using NLP and machine learning techniques. In their research, the preprocessing schemes are applied to the data. They used different feature selection schemes to determine the behavior of two algorithms – Naive Bayes and SVM. The model has been extended for higher ngrams. The results of this research show that linear SVM gives more accuracy.

The author has proposed a solution to the language barrier in the sentiment analysis by using the Google translator. The dataset collected from twitter is used for polarity classification. The google translator can be used with any other language. They build models using Naive Bayes and Maximum entropy algorithms. Here, Naive Bayes performance was the best in classification {8]

In paper [9], the author performed the sentiment analysis on the latest Hollywood and Bollywood movies. With the classification of tweets into positive, negative or neutral the accuracy achieved by SVM is 75% and 65% in Naive Bayes. The accuracy can be increased by increasing the training data

In paper [10], the author presented the comparative study of machine learning and lexicon based approach along with some evaluation metrics. Using machine learning methods their research shows that SVM performs better than Maximum Entropy. The accuracy can be improved by cleaning the data and using the higher n- gram feature for sentiment analysis.

3. PROPOSED METHOD

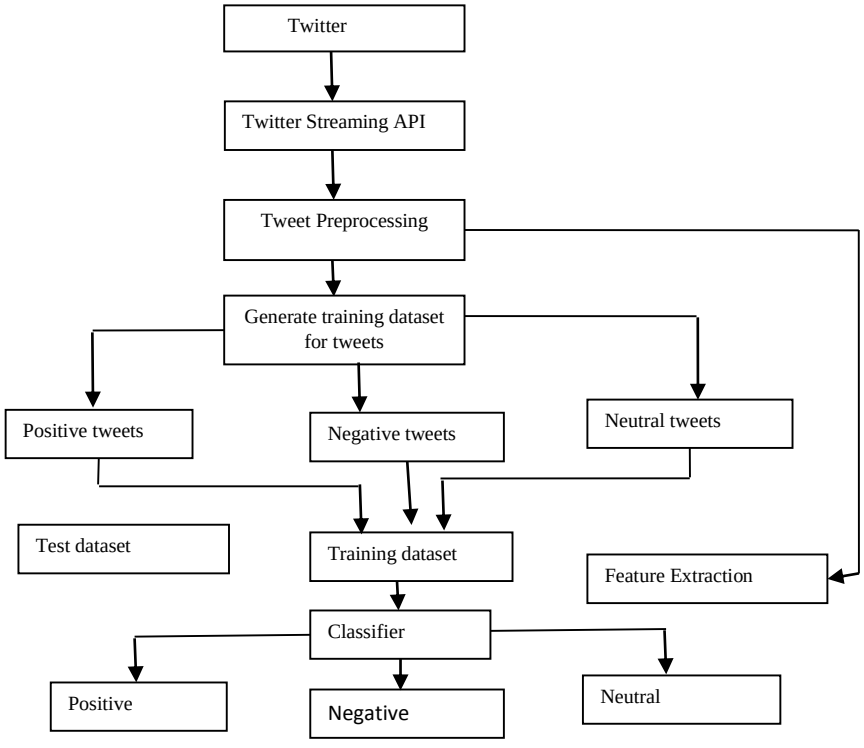

Figure 2: System Architecture

This section presents the techniques and algorithms that are used to analyze sentiment from tweets. The approach used in this paper for polarity classification are data pre-processing, feature extraction and the machine learning classification algorithms. Pre-processing of data is done to optimize the dataset. The machine learning algorithms used are Naive Bayes, Baseline and Maximum entropy for determining the performance of feature selection schemes.

A. Creation of dataset

Dataset is extracted by collecting the tweets using the twitter API. As twitter offers a user friendly API it is easy to create data. The data are collected using the tweepy. Tweepy is an easy-to-use python library that enables us to access the Twitter API. The tweets containing information such as date, time, author and more are returned in JSON format.

B. Preprocessing

Pre-processing is the important step in sentiment analysis. The tweets obtained are raw and unstructured data. For better performance the dataset is to be cleaned. So, pre-processing involves the following steps:

- Filtering – Filtering is nothing but simply cleaning the raw data. In filtering process, following are removed.

a) URL- With the short length of tweets the users are sharing information using the URL. Basically, the URLs are pointing to websites for sharing images. As URL does not contribute to determine the sentiment of the tweet. Hence, it is replaced by a word URL.

b) Username – To refer other users @symbol is used in Twitter. Usernames are not required for the sentiment classification it is replaced by the word USERNAME.

c) Hashtag – Hashtags refers to the topic or some useful information. So, it is replaced by the exact same word without.

d) Lower case – The capital letters are converted to lower case. As it will be easy to process the tweets by machine

- Tokenization- Tokenization is the process that splits the text into the sequence of tokens.
- Removal of stop words- The stopwords such as ‘a’, ‘an’, ‘the’,’to’, ‘of’,’are’,’this’ are removed in this step.

C. Feature Extraction – Feature extraction is the process where relevant features are extracted from the data used in the classification process. It makes classifiers more efficient by reducing the amount of data. The proposed feature is unigram approach.

- Unigram- Unigrams is the simplest model for the feature word selection. It consists of the individual words present in the document. In this the data are segmented into the tokens and every single word is treated like a feature.

D. Classification – Classification is an instance of supervised learning methods that assigns a class label .After creating the feature vector, classification is done using the supervised learning methods. In this work, Naive Bayes and Maximum Entropy are used and the performance is compared.

- Naive Bayes- Naive Bsayes is the basic text classification machine learning algorithm. It is a probabilistic classifier based on Bayes theorem that is easy to understand and implement. For text classification the class labels are known and the goal is to create the model. The main advantage of naive bayes is that it is independent of other features.
- Baseline – Baseline classifier counts the negative and positive words and classify it.

Maximum Entropy

Maximum entropy is a discriminative classifier that allows classification with more than two discrete classes. The principle is to model all features that are known and no assumptions is to be made regarding other unknown features. This classifier choose the one which has maximum entropy.

4. RESULT

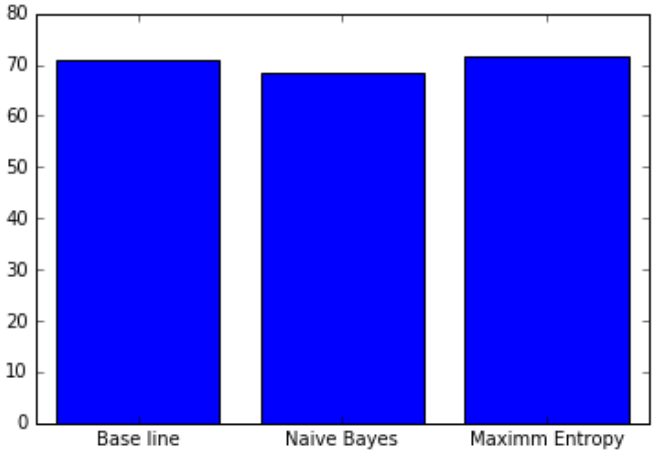

Fig 4.1: Comparison of Accuracy of Three Classifiers

The above figure shows the accuracy of the three different algorithms.

5. CONCLUSION

This paper presents an approach which automatically classifies the opinion into positive, negative and neutral. In the proposed system, the sentiment analysis is used to determine the opinions on the topics entertainment, companies and politics. Sentiment analysis is done by using feature selection schemes in combination with classification algorithms. Twitter is the

popular microblogging site that analyses the public mood and predict the future trend. In this paper, the tweets are collected and analyzed using supervised machine learning techniques. To classify tweets the different machine learning classifiers used are Naive bayes, baseline and Maximum entropy. The features are extracted using the unigram approach. The results show that Maximum Entropy gives more accuracy that other two algorithms. We conclude that sentiment analysis helps the organization and individuals to derive the insight from social media.

In future work, other feature selection methods can be used for increasing the accuracy. The tweets can be classified into the extremely positive, positive, neutral, negative and extremely negative. The negation handling and other classification algorithms can be used to improve the accuracy.

REFERENCES

[1] Hemalatha in "Preprocessing the Informal Text for efficient Sentiment Analysis" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 2, July – August 2012.
[2] Mr. Saifee Vohra, 2 Prof. Jay Teraiya, "Applications and Challenges for Sentiment Analysis : A Survey, "in International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, February- 2013
[3] Kowcika, Aditi Gupta, Karthik Sondhi, Nishit Shivhre and Raunaq Kumar, in "Sentiment Analysis for Social Media, "*International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 7, July 2013.*
[4] K. S. Tai, "Sentiment Analysis of Tweets: Baselines and Neural Network Models, "in *CS229 Final Project December 13, 2013.*
[5] Walaa Medhat, " Sentiment analysis algorithms and applications: A survey, " *in Ain Shams Engineering Journal(2014) 5, 1093-1113.*
[6] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan and Ms. Kalpana Dange, "Sentiment Analysis Using Support Vector Machine, "in International Journal of Innovative Research in Computerand Communication Engineering Vol. 2, Issue 1, January 2014
[7] Haseena Rahmath P in "Opinion Mining and Sentiment Analysis - Challenges and Applications, "in International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 3, Issue 5, May 2014.
[8] Gautami Tripathi1 and Naganna S., Feature selection and classification approach for sentiment analysis, "in Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015.
[9] Neelima, Dr. Ela Kumar, IndiSent Analysis in Twitter using Machine Learning Methods, "in International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2015.
[10] Akshay Amolik et al. in "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, "in International Journal of Engineering and Technology (IJET) Vol 7 No 6 Dec 2015-Jan 2016.
[11] ishal A. Kharde ,S.S. Sonawane , Sentiment Analysis of Twitter Data: A Survey of Techniques, "in International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

Author

Dolly Khandelwal, a research scholar currently pursuing Master of Technology from Shri Shankaracharya Technical Campus since 2014. She received the Bachelor of Engineering degree in Computer Science and Engineering from Shri Shankaracharya Technical Campus in 2014.