Sentiment Analysis on Movie "Kashmir Files" Using Machine Learning

Anjali, Pinki, Varun Sharma

Department of Computer Science & Engineering, Lyallpur Khalsa College Technical Campus, Jalandhar

Corresponding author: Varun Sharma , Email: varunsharma@lkcengg.edu.in

Sentiment Analysis which is also known as "Opinion Mining" is a process of analyzing, emotions, sentiments, attitudes and expressions expressed in written language. Now, these opinion can be mined from any social media platform like Twitter, Facebook as people are more likely to share their opinions on these platforms. Twitter is one such platform which is very popular most people are using this to express their opinion. The research addresses the sentiment analysis of movie "Kashmir files" by using trending hash tages on Twitter. Sentiment can be of three types- 1) Positive 2) Negative 3) Neutral Sentiment Analysis using Machine Learning makes use of various libraries such as numpy, pandas, sklearn, TextBlob etc. and Supervised Machine Learning algorithms like- Naïve Bayes, Support Vector Machine (SVM) etc. can be used for classification of tweets i.e. to identify the tweet whether it is positive, negative or neutral. In this work project, we will use Supervised Machine Learning algorithms like Naïve Bayes, Support Vector Machine (SVM) K-nearest neighbour (KNN) for classification of tweets and a comparison will be made based on their accuracy score.

Keywords: Sentiment analysis, Machine learning, Naïve bayes, SVM.

2023. In Vikram Dhiman & Pooja Dhand (eds.), *Emerging Trends in Engineering and Management*, 7–12. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.56155/978-81-955020-3-5-02

1. Introduction

In the current century, a total of 5 billion people around the world uses the internet which is equivalent to 63 percent of the world's total population and among those there are 4 billion active social media users globally. Social media applications like Twitter, Facebook, YouTube are being used as a platform for expressing opinions, expressions, attitudes etc. Thus, public opinions are the best source of feedback for business stake holders about their products & services.

Those important opinion or feedback can be analyzed by the process of "Sentiment analysis". Sentiment analysis is a mathematical procedural study of people's thoughts and opinions which can be negative or positive. Sentiment analysis is correlated with "text mining" or "data mining". The basic purpose of Sentiment Analysis is to assure the polarity of natural language by performing Supervised or Unsupervised classification.

The purpose of our research is to apply the classification techniques on the tweets retrieved by using the Twitter API on movie "Kashmir Files" and to classify the tweets based on the polarity value. The polarity values lies between -1 and +1.

- If polarity value > 0 , then the tweet would be considered as positive..
- If polarity value < 0, then the tweet would be considered as negative.
- And if the polarity value = 0, then the tweet would be considered as neutral.

In our work study, data pre-processing steps has been taken to achieve the better analysis results. Supervised Machine Learning algorithms like Naïve Bayes, Support Vector Ma- chine (SVM) & K-Nearest Neighbours (KNN) has been used for tweets classification & their performance has been measured by their accuracy score.

The main contribution of this paper :-

- To apply supervised machine learning algorithm on tweets for text classifications.
- To conduct a comparative analysis between algorithm- Naïve bayes, SVM, & KNN.

2. Methodology

Data Pre-Processing: Tweepy library has been widely used to retrieve the data/tweets from the twitter API by executing a python- based script. We've retrieved 5000 tweets by using the "Kashmir Files" name as a keyword. Below figure represents the total number of tweets being used in our work study.

In order such that our machine learning algorithms per- forms better we need to do from data preprocessing steps. Data pre-processing includes the following steps in Sentiments Analysis process:- We will use natural language processing toolkit (NLTK) for pre-processing techniques. We will first remove punctuations, hyperlinks & emoticons before starting data pre- processing steps.

2.1 Tokenization of Tweets

Tokenization is breaking the row data into small clunks in order to interpret the meaning of test by analyzing the sequence of the words. We will use word tokenize () class of NLTK library for doing tokenization of tweets.

Example:-

Sentence- "My name is Anjali. I am a student."

Tokenized sentence will be- ["My name is Anjali", "I am a student"]

1. **Removing stop words**-Stop words are unnecessary words which does not add much weighted in analyzing polarity of the sentence.

Example:-

"a", "the", "are", "is" etc. are STOPWORDS which do not add much value in analyzing polarity of the sentence.

2. **Lemmatization:**-Lemmatization is the process of reducing infected words to their WORDSTEM. We will us wordNetLemmatizor() class of NLTK library for doing lemmatization.

Example:-

1) Word doing will become do.
2) Words like finally, finalized will become final.
3) Words like going, goes, gone will become go.
4) Bag of Words:- Bag of words creates a vector table which holds count of word occurences in the given text disregarding grammar and word order but keeping multiplicity.

Example:-
Sentence1: He is a good boy.
Sentence2: She is a good girl.
Sentence3: Boys & girls are good.
The vector table of the above sentences will looks like: For creating vector array, we will use Count_Vectorizer()class of sklearn.feature_extraction.

Fig. 1: Vector Table

	good	boy	girl
Sentence 1	1	1	0
Sentence 2	1	0	1
Sentence 3	1	1	1

3. Implementation

Naïve Bayes-
Naïve Bayes is a Supervised Machine Learning algorithm which is widely used for text classification problems. Naïve bayes is a probablitistic model which use Bayes theorem to solve the classification problems. It assumes that all the features or data attributes are independent of each other.
Mathematically, Bayes theorem is as:
P(a/b) = P(b/a) * P(a)/ P(b)
Where a is class label b is the attribute set while P(a) and P(b) is the prior probability.
By applying this particular model, we got the following results :-

Accuracy is: 0.87
Classification Report is:

	precision	recall	f1-score	support
Negative	0.66	0.87	0.75	245
Neutral	0.90	0.86	0.88	533
Positive	0.95	0.88	0.91	722
accuracy			0.87	1500
macro avg	0.84	0.87	0.85	1500
weighted avg	0.88	0.87	0.87	1500

Fig. 2: Accuracy score and Classification Report of Naïve Bayes.

Support Vector Machine (SVM)-
Support Vector Machine is another popular Supervised Machine Learning algorithm used for text classification. It uses a hyper-plane in order to divide the classes. The position & orientation of hyper-plane is entirely depends upon the position of Support vectors. It aims on maximizing the margin between the hyper-plane and the marginal planes.
By applying this particular model, we got the following results :-

```
Accuracy is: 0.9053333333333333
Classification report is:              precision    recall  f1-score   support

    Negative       0.98      0.69      0.81       245
     Neutral       0.87      0.95      0.91       533
    Positive       0.92      0.95      0.93       722

    accuracy                           0.91      1500
   macro avg       0.92      0.86      0.88      1500
weighted avg       0.91      0.91      0.90      1500
```

Fig. 3: Accuracy score and Classification Report of SVM.

K-Nearest Neighbour (KNN)-
KNN is a non-parametric Supervised Machine L earning algorithm used for classification as well as regression prob- lems. It finds out the similarity between the new data point and the available cases or the data points & and put the new case or data point into the category that is the most similar in the available categories. It uses Eucledian's distance (can be Manhattan distance) for finding similarity between the data points.
By applying this particular model, we got the following results :-

```
Accuracy is: 0.8853333333333333
Classification Report is:              precision    recall  f1-score   support

    Negative       0.99      0.73      0.84       245
     Neutral       0.76      0.99      0.86       533
    Positive       0.99      0.86      0.92       722

    accuracy                           0.89      1500
   macro avg       0.91      0.86      0.87      1500
weighted avg       0.91      0.89      0.89      1500
```

Fig. 4: Accuracy score and Classification Report of KNN.

Following figure depicts the flow diagram used in carrying out the process:

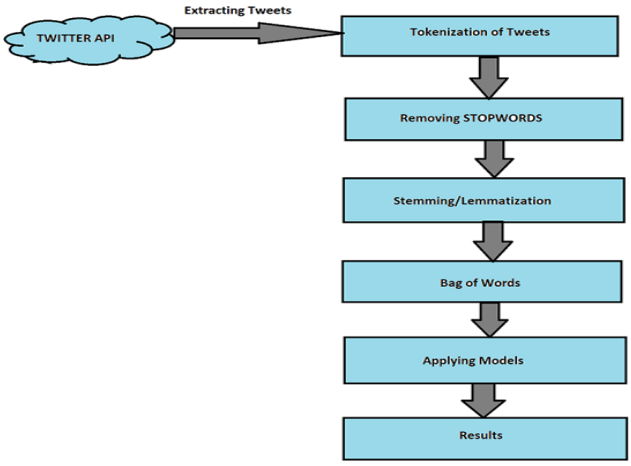

Fig. 5: Flow Chart of Sentiment Analysis Process.

4. Figures and Tables

The below fig. 6 depicts the total number of positive, negative and neutral tweets predicted by the models and fig. 7 depicts the Word Cloud formed by using the tweets.

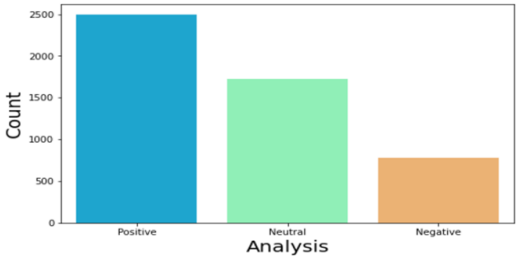

Fig. 6: Total number of Positive, Negative and Neutral Tweets.

Fig. 7: WordCloud depicting the most frequent words in Tweets.

5. Conclusion

Sentiment Analysis have always been an interested and fascinated topic in Machine Learning. It aims to identify people's opinions, attitudes and expressions.

In our work study, we first extracted the tweets from Twitter by using twitter API and we tried to show the Positive, Negative or Neutral category using Textblob library. After that we have applied Naïve Bayes, SVM and KNN model and compared their performance based on accuracy score.
From their comparison we have found that SVM model has performed well as compared to Naïve Bayes and KNN.

6. Acknowledgement

We thank the IKG-PTU and Lyallpur Khalsa College Technical Campus for giving us this opportunity. We are sincerely indebted for their efforts and time.

References

[1] "Redy Brahmananda A., & Vasundhra D.N., & Subhash P. (2019): Sentiment Research on Twitter Data"
[2] "Rasool Abdul (2019): Twitter Sentiment Analysis:A Case Study for Apparel Brands"
[3] https://datareportal.com/global-digital-overview
[4] https://www.smartinsights.com/social-media- marketing/social-media-strategy/new-global-social- media-research/
[5] https://www.analyticsvidhya.com/blog/2021/08/a-friendly-guide-to-nlp-bag-of-words-with-python- example/