Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Sentiment analysis of Social Media Text-Emoticon Post with Machine learning Models Contribution Title

To cite this article: V Jagadishwari *et al* 2021 *J. Phys.: Conf. Ser.* **2070** 012079

View the article online for updates and enhancements.

Sentiment analysis of Social Media Text-Emoticon Post with Machine learning Models Contribution Title

Jagadishwari V, Indulekha A, Kiran Raghu, Harshini P

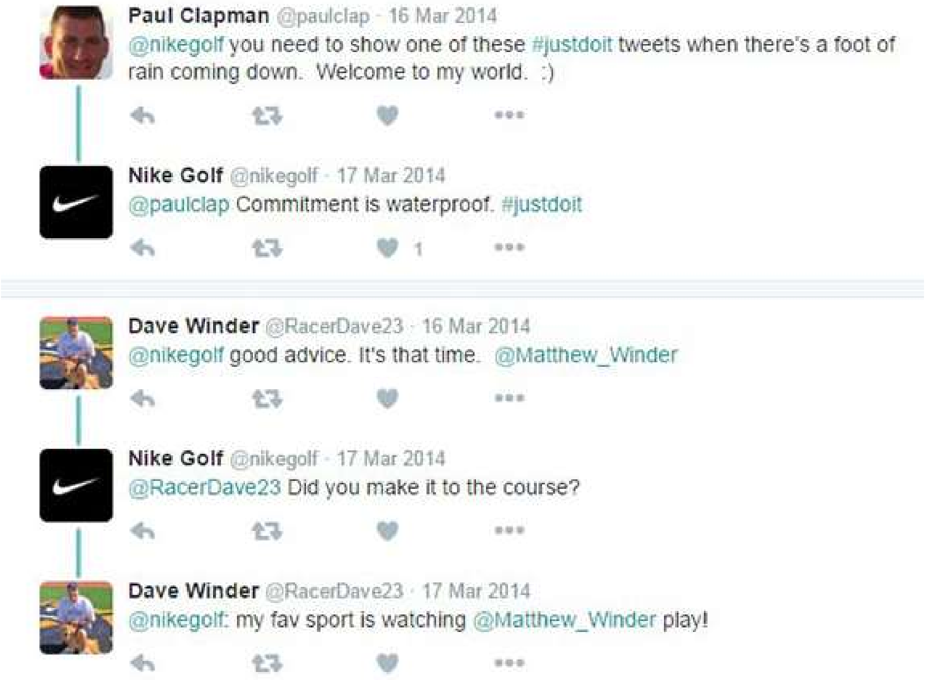
CMR Institute of Technology, Bengaluru
jiswariv@yahoo.com

Abstract. Social Media is an arena in recent times for people to share their perspectives on a variety of topics. Most of the social interactions are through the Social Media. Though all the Online Social Networks allow users to express their views and opinions in many forms like audio, video, text etc, the most popular form of expression is text, Emoticons and Emojis. The work presented in this paper aims at detecting the sentiments expressed in the Social Media posts. The Machine Learning Models namely Bernoulli Bayes, Multinomial Bayes, Regression and SVM were implemented. All these models were trained and tested with Twitter Data sets. Users on Twitter express their opinions in the form of tweets with limited characters. Tweets also contain Emoticons and Emojis therefore Twitter data sets are best suited for the sentiment analysis. The effect of emoticons present in the tweet is also analyzed. The models are first trained only with the text and then they are trained with text and emoticon in the tweet. The performance of all the four models in both cases are tested and the results are presented in the paper.

Keywords: Terms—Sentiment Analysis, Emoticon, SVM, Naive bayes

1. Introduction

The advancements in technology has changed the way people socialize and communicate with each other. Most of the communications have become virtual owing to the rapid growth of Social Media like Twitter, Facebook Instagram etc. People are allowed to express their opinion on various social and political issues. These Online Social Networks are also used to discuss about films, sports music etc. In brief anything under the Sun can become a topic of discussion and people interested in them can voice their opinion. Every opinion expressed has an associated feeling and emotion which is termed as the Sentiment. Analyzing the Sentiments from the social media posts has become important research area and has gained prominence in recent days. Researchers from the domains of Business, psychology, Computer Science etc are working on it. Most of the Social Media platforms allow the use of audio, video and text for expression the views. The most popular form of expression is text owing to its simplicity. Twitter is a front-runner for textual opinions. Twitter has limited the number of characters to be used in a Tweet, hence opinions are expressed in very short phrases in a crisp fashion. Hence Twitter Network data is extensively used by researchers in their analysis. An example Text Tweet is shown in the Fig. 1. Along with the text , the other high impact form of expression is Emoticon and Emojis. These are also simple but powerful means of communication. Every tweet or a text post in Social media is accompanied by a emoticon or an Emoji. Research has gone into analyzing the sentiments from Emojis as well. The Fig 2 shows a sample of tweets with Emojis. A lot of work has gone into analyzing the sentiments of Social media posts mainly the Twitter network. The state of art work done in this direction is discussed in the subsequent section.



Paul Clapman @paulclap · 16 Mar 2014
@nikegolf you need to show one of these #justdoit tweets when there's a foot of rain coming down. Welcome to my world. :)

Nike Golf @nikegolf · 17 Mar 2014
@paulclap Commitment is waterproof. #justdoit

Dave Winder @RacerDave23 · 16 Mar 2014
@nikegolf good advice. It's that time. @Matthew_Winder

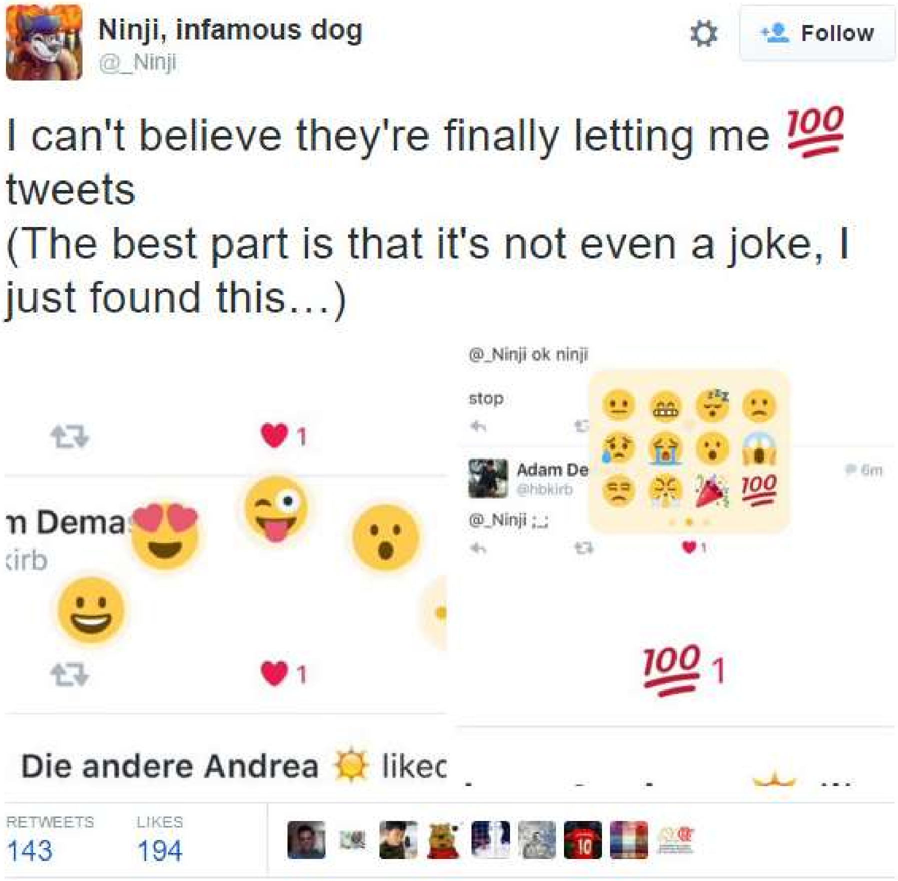
Nike Golf @nikegolf · 17 Mar 2014
@RacerDave23 Did you make it to the course?

Dave Winder @RacerDave23 · 17 Mar 2014
@nikegolf: my fav sport is watching @Matthew_Winder play!

Fig. 1. Sample Text Tweet

2. Related works

Human race is the most evolved species with an intellectual behaviour. People working in the fields of philosophy and religion [1] have discussed the contribution of emotions and Sentiments to People's behaviour. Charles Darwin has documented [2] about the relation between Evolution and Emotions way back in 1872. Today Sentiment analysis is continuing in a different flavor. With the Technical advances, Sentiment analysis has become an important research area in Information Technology. With advent in Machine Learning and Artificial intelligence work is progressing in the direction of making the machines understand human Emotions. An approach to detect emotion from a particular text by building an emotion embedding model is described in [3]. They first built an emotion embedding word model using the collected Tweet data annotated with hashtags. Next, they extract the representative emotional word in each sentence of the ROC story data.



Ninji, infamous dog
@_Ninji

I can't believe they're finally letting me 💯 tweets
(The best part is that it's not even a joke, I just found this…)

@_Ninji ok ninji

stop

@_Ninji ;_;

Fig. 2. Sample Tweet with Emoji

The representative emotional word is then used to classify the emotion of the sentence leveraging the cosine similarity. A corpus based approach was proposed by Faisal Muhammad Shah et al [3] involved classifying the text into categorical and dimensional model. There are 3 stages in establishing the sentiment in this model. They form a automatic emotional corpus by merging two computational model. It called CorpusBased of Emotion (CBE). A lot of research has gone into analyzing the sentiment of tweets , the authors of [4] have proposed an approach that uses four kinds of text based techniques. keyword spotting method, lexical affinity method, learning based method and hybrid method. For detecting emotions, lexical affinity and learning based method is combined to classify multi class emotions. Another similar work on tweets was done by Dilesh Tanna et al [5] where a method to analyse the sentiments of the emoji is discussed. The emoji unicode is converted to textual sequences and then its labelled. Emoji data inclusion improves the accuracy of the analysis when compared to analysing just the textual information. Artificial Intelligence and machine learning models are also implemented for sentiment analysis, a framework to detect depression using AI was built by Mandar Deshpande et al [6]. Analysis of facial expressions to detect the emotions is also done using Machine learning and CNN models. The authors of [7] have used Haar feature classifier for face feature extraction and it is further preprocessed for facial landmarks. These facial landmarks with the dataset is trained using SVM algorithm and then classify them on the basis of eight universal emotions. Convolutional Neural Networks are also used in sentiment anlaysis from facial expressions [8] [9].

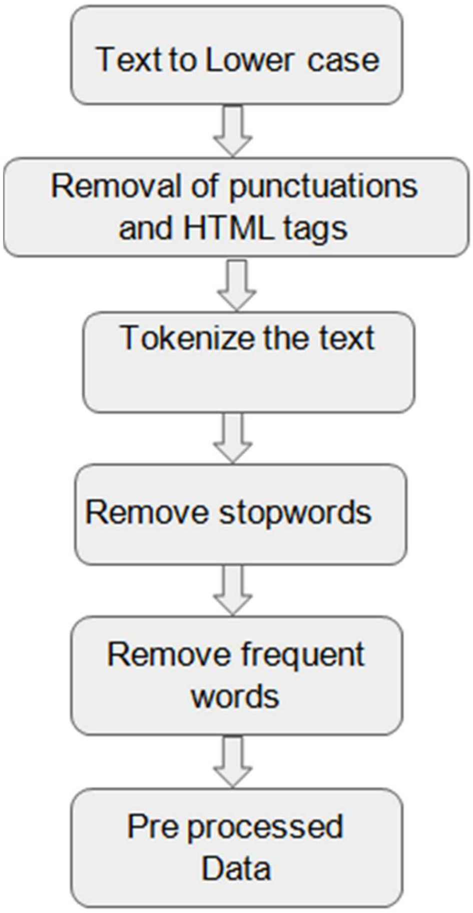


Fig. 3. Data Preprocessing

3. Methodology

3.1 Dataset

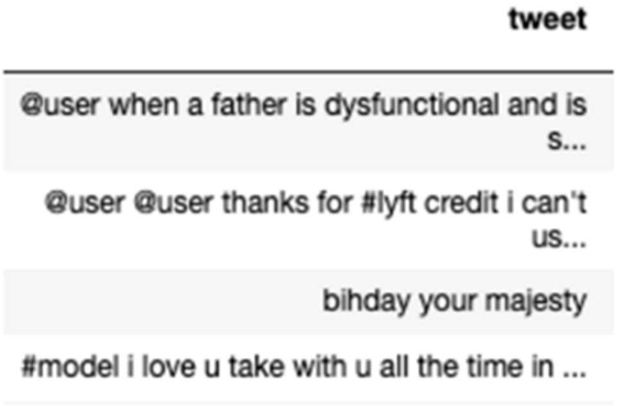
The dataset used is a Twitter dataset which was downloaded from 'Kaggle'. It can be downloaded in following link https://www.kaggle.com/youben/twitter-sentiment analysis. The tweets include the user's name, the text tweet and emoticons. Emojis are not included in the tweet. The dataset has 300 hundred thousand tweets containing three columns. First column contains the tweet ID, second contains the sentiment score and third column contains the actual tweet. The sentiment score takes two values- either one (1) or zero (0). A score of one (1) indicates a positive sentiment and a zero (0) score indicates a negative score. The output is a floating point number which lies in the range or zero to one. The 70% of the dataset was used in training the models and the remaining 30% was used to test the models.

3.2 Data Preprocessing

Pre-processing is the set of transformations applied to the data before doing the actual analytics. Data collected from various sources exists in raw format which needs to be processed before it is analysed. It helps in transforming the raw data into a clean data set which excludes noise and thus gives better results. Preprocessing the text involves various steps which is depicted in the Figure and explained below:

- Conversion of text into uniform format: Tweets will contain letters both in upper and lower case, all letters are converted into lowercase to maintain uniformity which will speed up model training.

- Removal of punctuation and HTML tags: Punctuation marks and other special characters do not impact the sentiment analysis hence they are removed.
- Tokenization: Tokenization is the process of splitting text into smaller units called tokens. In the tweets sentences or paragraphs are broken down to individual words.
- Stopword removal: Most common words appearing in the sentences like “a”, “this” etc are called Stopwords. The presence or absence of these words do not play a role in analyzing the emotion of the sentence, hence these stop words are removed and their removal helps in reducing the computation time and also helps in processing large amount of data.
  The steps included in sentiment analysis with emoticons are as follows:
- Convert emoticon to unicode.
- Check the unicode sequence in the dataset and match it with the respective textual sequence.
- Replace the Emoticon unicode sequence with the textual sequence.



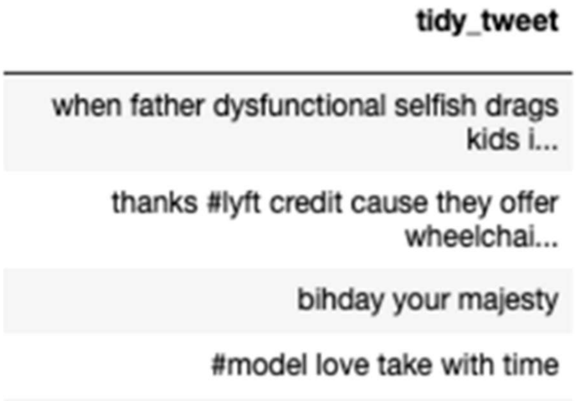
tweet
@user when a father is dysfunctional and is s...
@user @user thanks for #lyft credit i can't us...
bihday your majesty
#model i love u take with u all the time in ...

Fig. 4. Tweet before Preprocessing

The Fig 4 shows a tweet before applying the text preprocessing steps and the Fig 5 shows the same tweet after preprocessing. It can be seen that unwanted charcters and words those that are useful in the sentiment analysis are removed and the tweet shown in Fig 5 can be used for effective sentiment analysis. The tweet after preprocessing is also called as a tidy tweet.

3.3 Sentiment Analysis using Machine Learning Models

Sentiment analysis is done on the preprocessed Twitter Data sets was implemented using four different Machine Learning Models. The four models that were built were Bernoulli Bayes, Multinomila Bayes, Regression and SVM. Naive Bayes classifiers are a family of simple “probabilistic classifiers” based on applying Bayes’ theorem with strong independence assumptions between the features. Naive Bayes approach classifies text based on posterior probabilities. The Naive Bayes classifier is a classic classification algorithm that classifies text based on the probabilities of events. It requires less training data and time taken to train these models are less which in turn reduces the memory consumption and the CPU time. Two models from the Bayes family implemented are Bernoulli Naive Bayes and Mutinomial Naive Bayes.



tidy_tweet
when father dysfunctional selfish drags kids i...
thanks #lyft credit cause they offer wheelchai...
bihday your majesty
#model love take with time

Fig. 5. Tweet after Preprocessing

Support-Vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.SVM is more preferred when the dataset is large. It considers certain textual properties like high dimensional feature spaces, irrelevant features (dense concept vector), and sparse instance vectors.

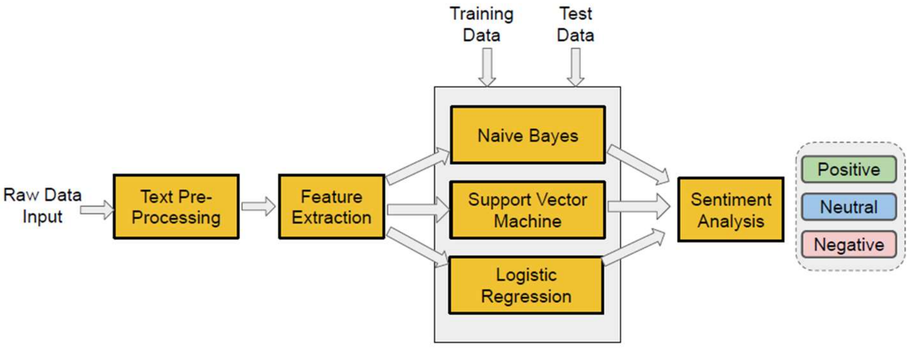
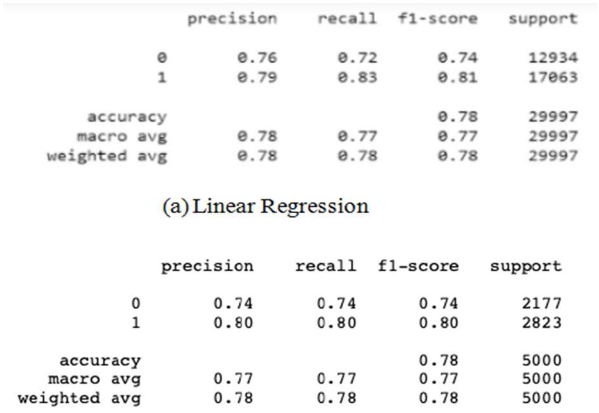


Fig. 6. Framework of the Sentiment Analysis

The Fig 6, shows the workflow of the sentiment analysis using the four models. The cleaned Twitter data set is used in building and testing the model. One Twitter dataset which contains 300 thousand tweets was used to train and test the models. 70% was used to train the model and 30% was used for testing the models. In most of the related work on sentiment analysis, the emoticons were not used and only text was used in training the models. In this work the Machine learning models were trained with text data and the model was tested and the same models were also trained using both the text and emoticon present in the tweet. This was done to check if the accuracy has an effect with the use of emoticons. The performance of the models is analyzed using the classification reports. The classification reports of the model shows the accuracy, precision, recall and F1 score. The Precision, Recall and F1 score are computed for each class in the classification problem. In the Twitter dataset which is used for the sentiment analysis in the paper consists of two classes positive sentiment indicated as class 0 and negative sentiment indicated as class 1 in the diagram. The weighted average and macro average of these scores of all the classes are also displayed in the report. This report for Linear Regression and SVM is shown in Fig 7 and the classification report of Bernoulli Bayes and Multinomial bayes is shown in Fig 8. The inference drawn from this report is that the precision, recall and F1 score is almost the same for Linear Regression, SVM models nad Multinomial bayes. The Bernouli Bayes has a lower value as compared to the other models. The accuracy of all the models are shown in the Fig 9 for comparison of all the models. It is clear that the use of emoticons has a negligible effect on the performance of the model. The Bernoulli Naive Bayes model has shown the best performance with an accuracy of 89 percent followed by Mutinomial Bayes model which makes it clear that the Naive Bayes family of classifiers performs well in sentiment analysis.

4. Conclusion

The work presented in this paper aimed at implementing machine learning models for sentiment analysis of Social Media posts. The Twitter data set was used to train and test the models. Tweets consist of short text messages, Emoticons and Emojis. The machine learning models tht were implemented were Bernoulli Bayes, Multinomial Bayes, Regression and SVM. The models were trained only with the text message removing the Emoticons and Emojis and then tested and their performance was evaluated. To study the effect of Emoticons the models were also trained with text and Emoticons and their performance were analysed.



```
              precision    recall  f1-score   support

           0       0.76      0.72      0.74     12934
           1       0.79      0.83      0.81     17063

    accuracy                           0.78     29997
   macro avg       0.78      0.77      0.77     29997
weighted avg       0.78      0.78      0.78     29997
```

(a) Linear Regression

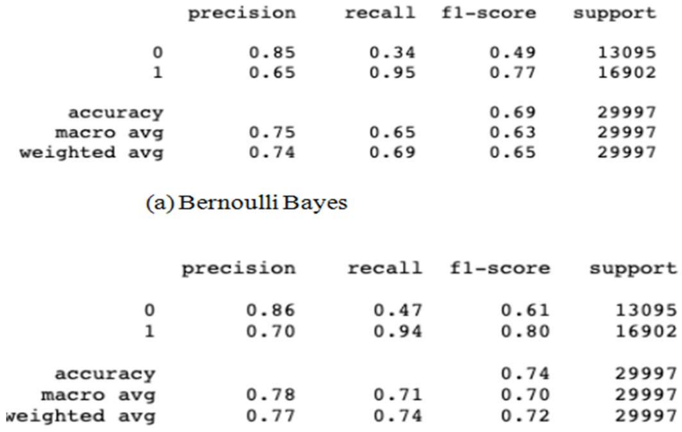
```
              precision    recall  f1-score   support

           0       0.74      0.74      0.74      2177
           1       0.80      0.80      0.80      2823

    accuracy                           0.78      5000
   macro avg       0.77      0.77      0.77      5000
weighted avg       0.78      0.78      0.78      5000
```

(b) SVM

Fig. 7. Classification report of Linear Regression Model and SVM



```
              precision    recall  f1-score   support

           0       0.85      0.34      0.49     13095
           1       0.65      0.95      0.77     16902

    accuracy                           0.69     29997
   macro avg       0.75      0.65      0.63     29997
weighted avg       0.74      0.69      0.65     29997
```

(a) Bernoulli Bayes

```
              precision    recall  f1-score   support

           0       0.86      0.47      0.61     13095
           1       0.70      0.94      0.80     16902

    accuracy                           0.74     29997
   macro avg       0.78      0.71      0.70     29997
weighted avg       0.77      0.74      0.72     29997
```

(b) Multinomial Bayes

Fig. 8. Classification report of Bernoulli Bayes and Multinomial Bayes

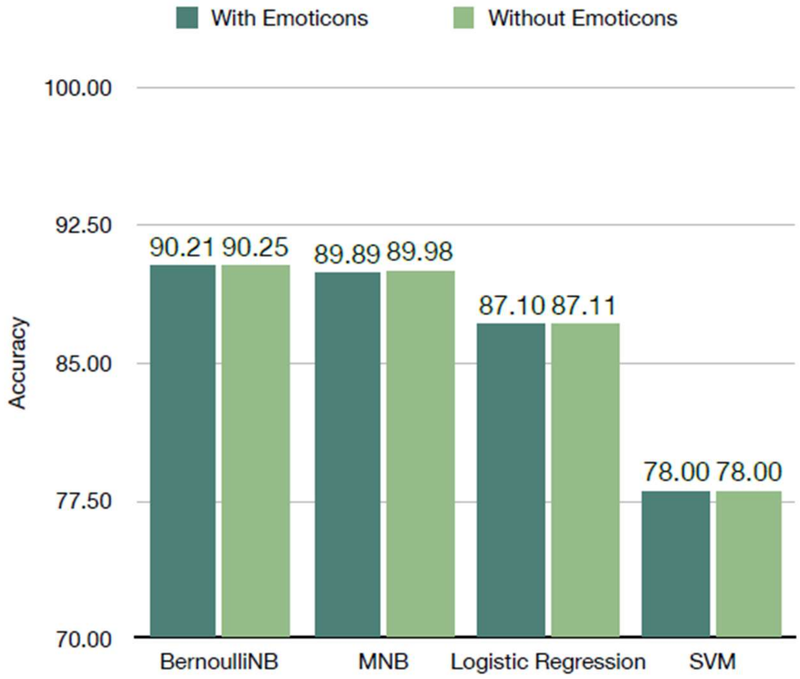


Fig. 9. Accuracy of the Machine Learning Models

The experimental results proved that the Bayes family of classifiers perform well in Sentiment analysis giving a very high accuracy and also it was evident that Emoticons have a negligible effect on accuracy of Sentiment analysis.

References

[1]. C. Bell, *Essays on the Anatomy and Philosophy of Expression*. J. Murray, 1824.
[2]. C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
[3]. F. H. Rachman, R. Sarno, and C. Fatichah, “Cbe: Corpus-based of emotion for emotion detection in text document,” in 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE*). IEEE*, 2016, pp. 331–335.
[4]. D. Tanna, M. Dudhane, A. Sardar, K. Deshpande, and N. Deshmukh, “Sentiment analysis on social media for emotion classification,” in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). *IEEE*, 2020, pp. 911–915.
[5]. F. M. Shah, A. S. Reyadh, A. I. Shaafi, S. Ahmed, and F. T. Sithil, “Emotion detection from tweets using ait-2018 dataset,” in 2019 5th International Conference on Advances in Electrical Engineering (ICAEE). *IEEE*, 2019, pp. 575–580.
[6]. M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS). *IEEE*, 2017, pp. 858–862.
[7]. H. Tuli, M. Singh, and N. Singh, “*Facial emotion recognition system using machine learning*.”
[8]. A. Bag, “*Real time facial expression recognition using convolution neural network algorithm*.”
[9]. D. Y. Liliana and T. Basaruddin, “Review of automatic emotion recognition through facial expression analysis,” in 2018 International Conference on Electrical Engineering and Computer Science (ICECOS). *IEEE*, 2018, pp. 231–236