International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Communication Sentiment Analyzer using Machine Learning with Naive Bayes Bernoullinb

M. Prabu, Mayank Singh Aithani, Niroj Deb, Pratyush Joshi

Abstract: ***In this never-ending social media era it is estimated that over 5 billion people use smartphones. Out of these, there are over 1.5 billion active users in the world. In which we all are a major part and before opening our messages we all are curious about what message we have received. No doubt, we all always hope for a good message to be received. So Sentiment analysis on social media data has been seen by many as an effective tool to monitor user preferences and inclination. Finally, we propose a scalable machine learning model to analyze the polarity of a communicative text using Naive Bayes' Bernoulli classifier. This paper works on only two polarities that is whether the sentence is positive or negative. Bernoulli classifier is used in this paper because it is best suited for binary inputs which in turn enhances the accuracy of up to 97%.***

Keywords: Sentiment analysis, machine learning, polarity, Naive Bayes' Bernoulli.

I. INTRODUCTION

Sentiment Analysis is a process of recognizing and examining the data based upon the sentiments, feelings, reviews, and thoughts of a human being. In this paper, we have used only one machine learning classifier which is also called the Naive-Bayes' Classifier [1]. Here we have integrated the classifier with a commonly used machine learning algorithm logistic regressions. The sentiments and emotions conveyed through the messages provide an indication of multiple aspects of human behavior [2]. For example, sadness, stress, disappointment, anger can indicate that the message conveys a negative meaning. On the other hand, it can be deduced that a person has a positive mood according to the nature of his message. Generally, users possess [6] varying behaviors or emotions. If its intensity rate of shown sentences is varying, then it indicates that the user is suffering from chronic depression. Most people feel sad or depressed at times. It's a normal reaction to loss of life's struggles. But when intense sadness -- including feeling helpless, hopeless, and worthless -- lasts for many days to weeks and keeps you from living your life, it may be something more than sadness. You could have clinical depression is a treatable medical condition.

According to the Centers for Disease Control and Prevention (CDC), 7.6 percent of people over the age of 12 have depression in any 2week period. This is substantial and shows the scale of the issue. According to the World Health Organization (WHO), depression [11] is the most common illness worldwide and the leading cause of disability. They estimate that 350 million people are affected by depression, globally. To detect mental illness or disorders, various studies about health systems use sensor devices. Stress can be detected using electroencephalogram signals, with an average [4] accuracy of 80.45%. There are a lot of studies that make use of textual information to detect physiological illness. There are different machine learning classifiers that can be used to perform emotion classification approaching an average accuracy of 80%. This research is done in order to find the best algorithm to identify the polarity of messages received by the people. We have tried various machine learning algorithms namely Random Forest, Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes' CNN [13] BernoulliNB to classify the sentiment of the message one receives in two polarities.

II. RELATED WORK

Leah and Bruce (1996) [16] have previously differentiated the multinomial model to the (BIM) binary independence model which was traditionally used with the PRP, the communication sentiment wordage for this model. They have done an analysis (theoretical) that stated that multinomial does not model does not efficiently call document length assumptions. They have used a single and small dataset with acutely small vocabularies. Their event model no longer remains a multinomial after normalizing document length.

III. PROPOSED SYSTEM

We have downloaded the data set online from some GitHub account. Our research was started by classifying the data set through learning the polarity of phrases and sentences in order to classify a message as positive or negative on different polarity levels. We have used the Naive-Bayes' Classifier to classify the messages [3] for a detailed analysis of the sentiment of a message.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Mr. Prabu, Asst. Professor Srm Institute of Science and Technology, Ramapuram, Chennai. India.
Mayank Singh Aithani, B. Tech Student at SRM Institute of Science and Technology, Ramapuram Chennai, India.
Niroj Deb, B. Tech Student at SRM Institute of Science and Technology, Ramapuram Chennai, India.
Pratyush Joshi, B. Tech Student at SRM Institute of Science and Technology, Ramapuram Chennai, India.

This method can also be applied for tweets which can give us the polarity of a tweet while other papers have used other classifiers like Logistic Regression, SVM, Bagging, Random Forest but none of them took binary inputs as their polarity measure. Our classifier which is NB Bernoulli works on binary inputs which have helped us to attain a high level of accuracy. We have used the online dataset. Then we have divided the dataset into two parts. One is used to train the model by using the classifiers and the other part is used to test the dataset and see the outcome of our model. We have used a slightly different approach which is we have classified the dataset into only two polarities which is positive and negative. This helped us to attain a very high accuracy. However, people use very diverse grammatical structures that too in different languages. This makes the classification of data rather challenging.

IV. METHODOLOGY

This section presents the methodology behind the proposed Communication Sentiment Analyser. The first step in this research was to manage the subjective tests to analyze the model of the proposed system.

A. Data Source

a) Test Data:

For test data, we can use the NLTK [12] library of Python. NLTK consists of corpora. Corpora consist of a structured set of text files that are used to perform analysis. In corpora, there are various types of text files like quotes, chats, reviews, history, etc. From these corpora, we have to use files of movie reviews. We should make sure the text file we have downloaded

should be converted to vectors. Rename it to test_vectors.txt.

b) Train Data:

To make use of train data download any reviews data. It can be any movie review or product reviews. It should be large enough to train the model. Convert the .txt file into vectors and rename the file as train_vectors.txt as vectorization is our major approach in this proposed system. Train data is used to train the classifier which we have to build.

B. Preprocessing Data

The data we have obtained as training data and test data are not fit for drawing out features. Mostly the data has usernames, empty spaces, special characters, stopwords, emojis, short forms, #tags (hashtags), timestamps, uniform resource locators (URLs), etc. Thus to make this data fit for classifying a message into the two known polarities we need to extract these above-mentioned constraints. We then have to replace all the above emojis and short forms with their respective meanings as in:-), =D, =), LOL, ROFL, etc. are replaced with a positive polarity (i.e. 0 or 1 whichever we have considered). Once we have finished this we are ready to process our data [7] using our classifier for the required output.

Table- I: Sample Review and Processed Review

SAMPLE REVIEW	@leonardodicaprio acted great. He is awesome in the movie #wolfofwallstreet
PROCESSED REVIEW	great, awesome

Table- II: Removed and modified content

CONTENT	ACTION
Punctuations	Remove
Words not starting with alphabets	Remove
@username	Remove
Uppercase characters	Convert to lowercase
URLs	Remove
Numbers	Remove
#words	Remove
Emojis	Replace
Whitespaces	remove

C. Data Classification

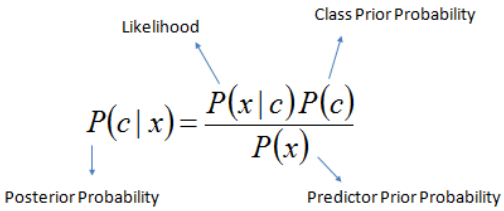
a) Building Classifier

To classify any message using machine learning [5] we received into two polarities (negative or positive), we constructed a classifier which has several other machine learning classifier. To construct a classifier we have used a library of python Scikit-learn. It is a very efficient and the most powerful library in python which helps us to integrate different classifiers into one. Scikitlearn also includes tools for classification, regression, clustering, and visualization. To install Scikit-learn we have to just type a one-line command in your command prompt (if windows) or terminal (if Linux) that is 'pip install sci-kit-learn. So to construct our classifier, we have used three different inbuilt classifiers [8] which already come in the Scikit-learn library which are :

• *Naive Bayes Classifier:* This classifier is a machine learning technique that is used for probabilistic classification of data. This is based on Bayes' Theorem with naive acceptance of freedom between each pair of characteristics.

The reference code for training the dataset through NB is given below:

"# Init the Gaussian Classifier
model = GaussianNB()
# Train the model
model.fit(xtrain, ytrain)
# Predict Output
pred = model.predict(xtest)"



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 1.Posterior Probability

• *BernoulliNB Classifier*: It uses an NB classification algorithm to classify the data. It uses NB for multivariate distribution of data that is there are many characteristics but we have to assign them to have a binary value or boolean variable (true or false). Hence every class has to be divided into two different polarities. BernoulliNB takes in binaries as input. The reference code for training the data through the Bernoulli classifier is given below.

"#Init the BernoulliNB classifier
nb_model = BernoulliNB(alpha=1.0, binarize=None, class_prior=None, fit_prior=True)
#Train the model
nb_model.fit(train_vec, ytrain)
#Predict Output
pred = model.predict(test_vec)"

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

Fig. 2. Likelihood of a document given a class C_k

• *Logistic Regression Classifier*: In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. It is also called Maximum Entropy Classifier.

b) Training Classifier

A script written in python is used to classify the training sets. Once, we have trained the classifier, we can also check the accuracy of the testing set.

Table- III: Generated Polarities of Reviews

MOVIE REVIEWS	POLARITY
Boring, Youngish, Idiotic, Worst	Negative
Best Movie ever	Positive
Overrated	Negative
Time Pass, Moderately funny	Negative

D. Classifying Messages

Once we have trained all of our classifiers we need to classify our messages [14] from different persons into their respective polarities which can be either positive or negative.

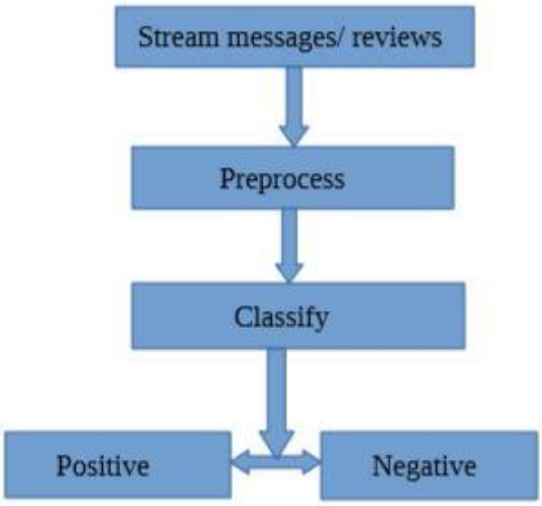


Fig. 2. Flowchart of Message Classifier.

V. RESULTS AND DISCUSSION

Here we have reviewed all the accuracies for the models used in the past papers and have represented their accuracies in the below bar-graph.

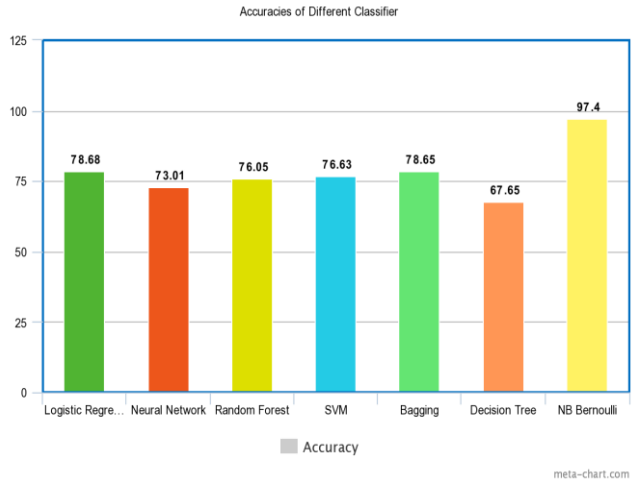


Fig. 3.Comparision of Classifiers (source:meta-chart.com)

It is quite evident from the above data that our model has the highest accuracy. Hence, we can say that the proposed model is more efficient comparatively.

VI. CONCLUSION

Sentiment analysis is used to identify people's attitudes, opinions [15] and their mental state. The opinions of people can either be positive or negative. To do sentiment analysis of the messages we need to first feed the message into our model. Then we need to apply our classifiers using Scikit-learn. Our model has an accuracy of approximately 95% - 97% depending upon the size of the dataset given to train the classifier. So this method can also come in very handy to differentiate between good and bad movie reviews so that they can easily filter out the movies they need to watch.

This model can also be used in other aspects like marketing (product reviews), etc.

VII. FUTURE WORK

In the future, we can make a web-based application or a mobile application using this model. We can improve our system by classifying complex grammatical structures like sarcasm, and deal with sentences with double or maybe multiple meanings. We can also make this for other languages like Hindi, French, German, Japanese, Russian, etc. to provide more local sentiment analysis.

REFERENCES

1. H. Zang, 3The optimality of Naïve-Bayes´, Proc. FLAIRS, 2004.
2. C.D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, pp. 234-265, 2008.
3. A. McCallum and K. Nigam, 3A comparison of event models for Naive Bayes text classification´, Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998.
4. M. Schmidt, N. L. Roux, and F. Bach, 3Minimizing finite Sums with the Stochastic Average Gradient´, vol. 162, pp.83-112, 2017.
5. T. Mitchell, Machine Learning, McGraw Hill, 1997.
6. G.Vinodhini*, RM.Chandrasekaran Sentiment Analysis and Opinion Mining: A Survey International Journal of Advanced Research in Computer Science and Software Engineering.Volume 2, Issue 6, June 2012 ISSN: 2277 128X. Page-286.
7. Svetlana Kiritchenko, Xiaodan Zhu, Saif M. Mohammad Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research 50 (2014) 723-762. Page no-727.
8. Walaa Medhat a,* Ahmed Hassan B, Hoda Korashy b Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. Page no-1098.
9. D. Davidov, O. Tsur, and A. Rappoport, ‘‘Semi-supervised recognition of sarcastic sentences in Twitter and Amazon,’’ in Proc. 14th Conf. Comput.Natural Lang. Learn., Jul. 2010, pp. 107–116.
10. C. C. Liebrecht, F. A. Kunneman, and A. P. J. van den Bosh, ‘‘The perfect solution for detecting sarcasm in tweets #not,’’ in Proc. WASSA, Jun. 2013, pp. 29–37.
11. S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression vulnerable college students,” Cognition & Emotion, vol. 18, no. 8, pp. 1121–1133, 2004.
12. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” J. Mach. Learn. Res., vol. 12, pp. 2493–2537, Nov. 2011.
13. S. Poria, E. Cambria, and A. F. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in EMNLP, 2015.
14. J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. C. M. Fong, “Quantitative study of individual emotional states in social networks,” IEEE Transactions on Affective Computing, vol. 3, no. 2, pp. 132–144, April 2012.
15. C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu, ‘‘Identifying breakpoints in public opinion,’’ in Proc. 1st Workshop Soc. Media Anal., Jul. 2010, pp. 62–66. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

AUTHORS PROFILE

M. Prabu is Asst. Professor at Srm Institute of Science and Technology, Ramapuram, Chennai. He has years of experience as a professor in Computer Science..

Mayank Singh Aithani is a third year student at SRM Institute of Science and Technology, Ramapuram Chennai, pursuing his BTech in Computer Science. He has done research in machine learning and deep learning. He is currently working on his projects in Artificial Intelligence and Computer Vision. He is a permanent Artificial Intelligence programmer at Alphabt Chennai. This is his first paper on machine learning and its algorithms.

Niroj Deb is a third year student at SRM Institute of Science and Technology, Ramapuram Chennai, pursuing his BTech in Computer Science. He has done research in machine learning and deep learning. He is currently working on his projects in machine learning and web development. . This is his first paper on machine learning and its algorithms.

Pratyush Joshi is a third year student at SRM Institute of Science and Technology, Ramapuram Chennai, pursuing his BTech in Computer Science. He has done research in machine learning and deep learning. He is currently working on his projects in machine learning and cloud computing. This is his first paper on machine learning and its algorithms.