Naïve Bayes Algorithm for Large Scale Text Classification

Pirunthavi SIVAKUMAR[1], Jayalath EKANAYAKE[2]

(1. *Department of Information and Communication Technology*, *Faculty of Technology*, *Rajarata University of Sri Lanka*, *Mihintale* 50300, *Sri Lanka*; 2. *Department of Computer Science and Informatics*, *Faculty of Applied Sciences*, *Uva Wellassa University of Sri Lanka*, *Badulla* 90000, *Sri Lanka*)

Abstract: This paper proposed an improved Naïve Bayes Classifier for sentimental analysis from a large-scale dataset such as in YouTube. YouTube contains large unstructured and unorganized comments and reactions, which carry important information. Organizing large amounts of data and extracting useful information is a challenging task. The extracted information can be considered as new knowledge and can be used for decision-making. We extract comments from YouTube on videos and categorized them in domain-specific, and then apply the Naïve Bayes classifier with improved techniques. Our method provided a decent 80% accuracy in classifying those comments. This experiment shows that the proposed method provides excellent adaptability for large-scale text classification.

Keywords: Naïve Bayes, Text Classification, YouTube, Sentimental Analysis

1 Introduction

Recently, the complexity of documents and text has increased exponentially, which requires a deeper knowledge of machine learning methods to accurately classify text in many computer applications. Many machine learning methods have reached outstanding output in natural language processing (NLP). The success of these machine learning algorithms relies on their ability to understand the complex patterns and nonlinear relationships between the data. Even though, finding the appropriate text classification technique for solving their research problems is a challenge for many researchers.

A widely used text classification algorithm is the Naive Bayes algorithm due to its easiness and high accuracy[6]. Rather than the other text classification algorithms, the Naïve Bayes algorithm is simple to construct and especially beneficial for extremely huge data sets. In addition to simplicity, the Naive Bayes algorithm is even superior to the most sophisticated classification methods[14].

This research topic is proposing a model which can classify large-scale data sets with more accuracy through sentimental analysis using the Naïve Bayes algorithm. In this research, we selected YouTube as the source of data that can provide a large scale of text data. Thousands of users give positive and negative comments in the comment section for each video. For this reason, we need a way to analyze viewer comments and opinions to rank viewers' opinions about each video through sentimental analysis using the Naive Bayes algorithm. Text preprocessing such as text labeling, tokenization, data stemming, and filtering.

2 Literature Review

Several types of machine learning algorithms have been applied to text classification, including the Naive Bayes algorithm[12][13], KNearest Neighbor (KNN) algorithm[3], Support Vector Machines algorithm[4], and Neural Networks algorithm. Sentimental analysis is a computerized method for analyzing sentiments, opinions, and emotions[5]. Sentimental analysis is used to identify emotional tendencies when an opinion is positive or negative. Sentimental analysis is performed by processing textual data, so it is ready for analysis by text mining methods. Sentimental analysis is one of the natural language processing techniques that can be used to determine the level of sensitivity behind the text, i.e. YouTube comments, Tweets, movie reviews, etc[7].

For an instance, Grammarly is a tool that is utilized to modify the grammar of a document or text, showing the general meaning and pronunciation of the document and creating useful comments such as positive and optimistic. This is also true. It is done through a sentimental analysis of the entire document or text. It is widely used and helps to maintain brand equity as it is used by many large companies to check customer reviews on their products/ services, social networks, and websites.

Some large companies such as Apple, Google, and Trip Advisor use this technique to increase customer loyalty in their service services. Other uses of sentimental analysis include sentiment analysis on Twitter, movie ratings on IMDB, customer reviews on Amazon, and video comments on YouTube[7].

2.1 K-NN Algorithm

K-Nearest Neighbor (K-NN) algorithm is utilized for text classification with the aid of using the K nearest neighbors in the training dataset after which the usage of the label of closest suits to predict[8]. In essence, the K-nearest neighbor algorithm has resulted in a majority out of K-cases that closely resemble a certain "invisible" observation. The similarity between two data points is defined as a measure of the distance. A common method is the Euclidean distance method. Other methods include Minkowski, Manhattan, and the Hamming distance method[7]. The Hamming distance method should be used for the categorical variables. K-nearest neighbor algorithm is a straightforward algorithm that saves all available cases and ranks upcoming cases based on a measure of similarity. It is primarily utilized to classify data points based on the classification of adjacent data points.

2.2 SVM Algorithm

A support vector machine algorithm that determines the fine-tuning of boundaries within the vectors that belong to a particular group and vectors that do not.

A vector is a list of numbers that represent a set of coordinates in a small space. Therefore, while the support vector machine determines the boundaries, the support vector machine decides to draw the best "line" (highest quality hyperplane) that divides the gap into subspaces. Text classification using SVM is simple and straightforward with the use of MonkeyLearn[8].

2.3 Neural Network

A neural network consists of weighted neurons that connect them, learning by processing records in sequence and comparing classifications to actual classifications[11]. A neural network can be described by three components or layers: an input layer, a hidden/intermediate layer, and an output layer. The task of the input layer is to capture the input signal from the external system. The hidden layer is made up of neurons. Neural network research is fully monitored for this reason, as the inputs to the neural network have solutions or outputs. The neural network gets the input values and weights from the input layer, and the function sums the weights and moves the output to a hidden layer that maps the output to the appropriate unit of the output layer. You can include "n" different hidden layers among the input and output layers. Determining the type of hidden layer, the community may be referred to as a single-layer or multi-layered (in the case of multiple hidden layers) neural network[11].

The neural network method had been an extraordinary tool for classifying the text of numerous structures of NN that had been utilized for the utility of classification of text, e.g. Error Back Propagation Algorithm, ADALINE and MADALINE network, etc[10].

Phases concerned in text classification are amassing the dataset, preprocessing, dimensionality reduction, and classifier implementation[10].

2.4 Naïve Bayes

A commonly used text classification algorithm is the Naive Bayes algorithm due to its simplicity and high accuracy[6]. The Naïve Bayes algorithm is simple to construct and especially beneficial for extremely huge data sets. Compared to the other text classification algorithms, the Naïve Bayes algorithm is simple to implement and extremely exhaustive for large-scale data sets. In addition to that, this algorithm is more powerful than the most sophisticated text classification algorithms[14].

Each step of this algorithm for text classification is pretty much straightforward compared to the other text classification techniques.

2.5 Performance Comparison

Compare to the other existing text classification techniques, Naïve Bayes performs best with a prediction accuracy of 87.47%. In addition to that, Naïve Bayes is the fastest classification method so far, training in seconds[15].

Looking at the total run time, including preprocessing, classifier training, and test execution, reveals a significant advance in the Naïve Bayes approach. The "training" of the classifier is purely statistical, so it is less computationally intensive than analyzing the data itself[15].

The Naïve Bayes algorithm came up with some ideal benefits over other techniques. If a person desires to understand precisely why a sentence became categorized to a particular class it’s far quite simple to give the weighted extraordinary phrases withinside the sentence had in the direction of a class. It makes it feasible to locate flaws withinside the classification method and fine-track the set of rules if desired. Another classification techniques are bit “black box” as it takes input and presents the output with a little manner of understanding of why it was given that result[15].

If new training data is added to the dataset, the Naïve Bayes classifier simply recalculates only the affected classes by correcting how many times the words occur within the class. On the other hand, in other classification techniques, these tasks cannot be addressed without retraining completely by wasting a great deal of time and computing power[15].

3 Methodology

3.1 Data Pre-processing

Here, the naïve Bayes algorithm is utilized for text classification of the comments of YouTube videos. The important step even as implementing a machine learning model is the data pre-processing because it will at once affect the result of the model. The greater pre-processing the data, the greater the model performance will be accurate[7]. Fig.1 illustrates the steps of data-preprocessing of the naïve Bayes algorithm.

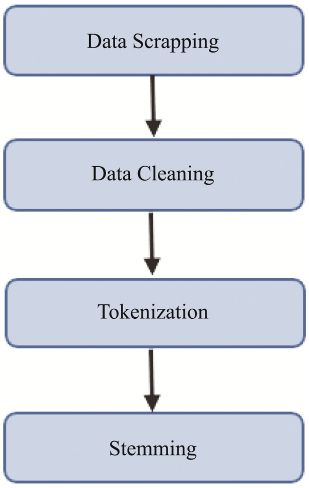

Fig.1 Steps of Data Pre-processing

YouTube remarks contain numerous languages relying on the study of the human population of the commenter. Anyhow, to simplify the sentimental analysis, they modified the data collection scripts to acquire the simplest English comments[7].

Data Scrapping - “YouTube Scraper” was utilized for extracting remarks from YouTube videos. This tool saves the scrapped remarks in a CSV file (Comma Separated Value). But this device isn't that effective. This device extracts all of the remarks from each of the videos. As our PC does now no longer has sufficient processing power and storage, it changed into very much difficult for further processing. Then YouTube Rest API was used to scrape the comments from YouTube. Python language was used to extract the comments of each video using the YouTube APIs. Thousand of comments were extracted per each video since API permits only a thousand comments to be extracted from YouTube[7].

Data Cleansing - Data cleansing performed a large element in constructing a model. It gave the capacity to find out erroneous or incomplete records. Without the right data quality, the very last analysis might go through inaccuracy or incorrect end will be doubtlessly arrived. Extracted comments had been cleaned up to remove the unwanted data and select only the wanted data that is useful in the data analysis process. In this step, punctuation and emoji had been removed from each user review. After this process, the data set will be robust and avoid many common pitfalls from the comments.

Tokenization - Tokenization can be a step that splits longer strings of textual content into smaller portions or tokens. Larger chunks of textual content are frequently tokenized into sentences, sentences are frequently tokenized into words, etc. Further processing changed into done after a bit of textual content has been correctly tokenized. The sequence of strings is divided into parts such as words, keywords, phrases, and different factors referred to as tokens. After some of the text was properly tokenized, further processing was done.

Stemming - Data Stemming is the Text Normalization (or sometimes referred to as Word Normalization) method in the field of Natural Language Processing (NLP) this is used to prepare text, words, and files for further processing. It is the technique of decreasing a phrase to its phrase stalk into suffixes and prefixes. Here, different forms of a word, suffixes, and prefixes were reduced in each sentence.

3.2 Data Transfer

Two types of data set had been obtained after pre-processing.

Training data set has been utilized to train the model and 200 video comments had been used for this process.

The test data set is utilized to evaluate the performance of the model using a measure of performance. The important thing is that there is the data from the training set can not be included in the test suite. If the test set includes the data/ cases from the training data set, it is very difficult to assess whether the algorithm learned to generalization from fixed training or simply remembered it. Hundreds of movies had been selected to check and confirm paradigm. Fig.2 shows the division of collection of data to the training and testing data set.

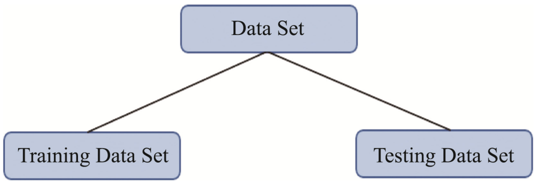

Fig.2 Data Transfer

3.3 Data Analysis

Data Labeling - The words had been categorized into classes which include positive words (+ve), and negative words (-ve). Positive and negative datasets had been obtained from the internet. And by the one’s dataset, the processed information had been categorized into classes (positive and negative). So that, the percentage of positive functions in addition to the negative functions had been calculated accordingly.

Overall Measure - A user’s remarks can also additionally incorporate positive or negative characteristics. It used a mathematical scale to calculate the percentage of positive features in addition to negative features. The average of the datasets collected from YouTube was calculated using the Naive Bayes algorithm. Fig.3 explains the process of the data analysis part of the naïve Bayes algorithm.

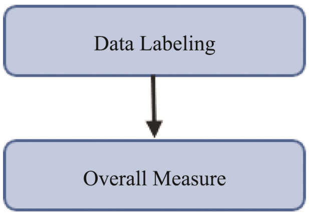

Fig.3 Steps of Data Analysis

4 Results and Discussion

The Naive Bayes algorithm is simple to construct and especially beneficial for extremely huge data sets. Naive Bayes can offer accurate outcomes without the need for much training dataset. It treats the probability of every phrase appearing in a document as though it has been unbiased of the probability of another phrase. This assumption is sort of in no way true of any files we`d desire to classify, which have a tendency to comply with regulations of grammar, syntax, and communication. When we comply with those regulations, a few phrases tend to be correlated with different phrases.

In addition to that, the following equation is useful in the terms of performance which is defeating the other classification algorithms[14].

$$Pr(H|E) = Pr(H) * Pr(E|H) / Pr(E)$$

H - Hypothesis
E - Evidence

Classification tasks involve comparing two (or more) hypothesis,

$$Pr(C_1|W) / Pr(C_2|W)= Pr(C_1) * Pr(W|C_1) / Pr(C_2) * Pr(W|C_2)$$

$$Pr(C_1|W_1, W_2 \ldots Wn) / Pr(C_2|W_1, W_2 \ldots Wn)=Pr(C_1) * (Pr(W_1|C_1) * Pr(W_2|C_1) * \ldots Pr(Wn|C_1)) / Pr(C_2) * (Pr(W_1|C_2) * Pr(W_2|C_2) * \ldots Pr(Wn|C_2))$$

The naive Bayes algorithm utilizes a unique method to obtain the probability that different classes support different attributes. After pre-processing and transferring the data, the algorithm creates a likelihood table after converting the data set into a frequency table by finding the possibilities of classes. Tables-1, 2, and 3 show its working example.

Table 1

Words	Positive/ Negative
Hate	Negative
Fantastic	Positive
Bad	Negative
Super	Positive
Harmful	Negative
Fantastic	Positive
Bad	Negative
Exhausted	Negative
Worst	Negative
Super	Positive
Beautiful	Positive
Admiring	Positive
Exhausted	Negative
Fantastic	Positive
Worst	Negative
Lovely	Positive
Gorgeous	Positive
Awful	Negative
Lovely	Positive
Harmful	Negative

Table 2 Frequency Table

Frequency Table		
Words	Positive	Negative
Hate		1
Fantastic	3	
Bad		2
Super	2	
Harmful		2
Exhausted		2
Worst		2
Beautiful	1	
Admiring	1	
Gorgeous	1	
Lovely	2	
Awful		1
Total	10	10

Table 3 Likelihood Table

Likelihood Table			
Words	Positive	Negative	
Hate		1	1/20=0.05
Fantastic	3		3/20=0.15
Bad		2	2/20=0.1
Super	2		2/20=0.1
Harmful		2	2/20=0.1
Exhausted		2	2/20=0.1
Worst		2	2/20=0.1
Beautiful	1		1/20=0.05
Admiring	1		1/20=0.05
Gorgeous	1		1/20=0.05
Lovely	2		2/20=0.1
Awful		1	1/20=0.05
Total	10	10	
	10/20=0.5	10/20=0.5	

Table 4 Sample Outputs of the Model

Video ID	No of Comments to Extract	Positive Sentiment	Negative Sentiment
yGY484EPe5U	1000	38.9%	61.1%
vbNib_NsVRN	1000	25.2%	74.8%

To the model, the video ID and the number of comments were specified as input and took the percentage as the output of positive and negative sentiment analysis. Table-4 shows some of the sample outputs of naïve Bayes text classification. The YouTube API permits the developers to access video statistics and YouTube channel datasets through REST API calls[7]. There were many comments on sentiment analysis, as a small number of comments can reduce accuracy. Therefore, we collected 1000 comments for each video.

Compared to the other existing methods, this method is straightforward and decent enough and provides high accuracy of 80%. As the YouTube API was used here for extracting the comments from each YouTube video, there was a limitation since it allows only a maximum 1000 number of comments to be extracted. If anyone uses any tool for extracting all the comments from each YouTube video the accuracy might vary. However, to store each scrapped comment, the machine should have more processing power and storage capacity.

5 Limitations

YouTube comments comprise several languages counting on the demography of the commenter. We only considered English movies in our research process.

YouTube APIs were used to extract comments from each video. On YouTube, you can get up to 1000 comments for each video via YouTube's API, so only thousands of comments were extracted for each video.

People generally tend to search for feedback based on their very own opinion. These days one terrible comment may unfold very instantly. People generally tend to reply to those types of feedback very often. So by and large negative comments with more responses can also additionally come to the top withinside the queue. This factor could negatively affect the accuracy of the model.

6 Conclusion

In case we have an issue in classification, we have to solve that issue. If there is a large dataset and variables for training the dataset, the best solution is to utilize the naive Bayes algorithm, which is much faster and more optimistic than the other classification algorithms.

The Naive Bayes algorithm is an efficient and effective way to design a text classification with 80% high accuracy and speed. Naïve Bayes classification algorithm is primarily based totally on the conditional probability of each feature belonging to a class, which the features are decided on with the aid of using feature selection methods. However, the naive Bayes algorithm is based on machine learning, hence the accuracy rate is highly dependent on the training data set. Also, if the test data set contains complicated sentences, the accuracy rate will be lower. Hence, this will continue to be our field of study in the future. YouTube API allows only 1000 user reviews to extract and hence, this factor could negatively affect the accuracy of the model.

7 Recommendations

We couldn't scrape all the comments from YouTube, because we used YouTube APIs to scrape the comments. Thus, we only extracted up to 1000 comments. One can use another technique to scrape all the comments of a particular video from YouTube and do sentimental analysis.

One can use more than one social media (Example: YouTube, Facebook, Twitter, Instagram) and use all the data and do sentimental analysis and output the accurate output.

Here, we have only considered English YouTube videos. In another way, one can use any language videos from any country and follow up with the sentimental analysis. So their expected system can be used in any country.

References

[1] Lo S, Ding L., 2012. Probabilistic reasoning on background net: An application to text categorization, Proc of 2012 *International Conference on Machine Learning and Cybernetics (ICMLC). IEEE Press,2*: 688-694.

[2] Kuang, F., Xu, W. and Zhang, S., 2014. A novel hybrid KPCA and SVM with GA model for intrusion detection. *Applied Soft Computing*, *18*, pp.178-184.

[3] Alpaydın, E., 2010. Introduction to Machine Learning, 2nd Edition Cambridge, MA: The MIT Press.

[4] Indurkhya N. and Damerau J., 2010. *Handbook of natural language processing, second edition.*

[5] Rish, I., 2001. "An empirical study of the naïve Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22,* pp. 41–46.

[6] Sivakumar, P., Rajeswaren, V., Abishankar, K., Ekanayake, J., and Mehendran, Y., 2020. Movie Success and Rating Prediction Using Data Mining Algorithms. *Journal of Information Systems & Information Technology (JISIT)*, 5(2), pp.72-80.

[7] Dao, S., 2018. Text Classification using K Nearest Neighbors. [online] Medium. Available at: <https://towardsdatascience.com/textclassification-using-k-nearest-neighbors-46fa8a77acc5> [Accessed 16 January 2022].

[8] https://monkeylearn.com/. n.d. Text Classification Using Support Vector Machines (SVM). [online] Available at: <https://monkeylearn.com/text-classification-support-vector-machines-svm/> [Accessed 20 January 2022].

[9] Patra, A. and Singh, D., 2013. Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method. *International Journal of Computer Applications, 68(17),* pp. 37-41.

[10] Lakshmi Prasanna, P. and D. Rajeswara Rao, D., 2017. Text classification using artificial neural networks. *International Journal of Engineering & Technology, 7(1.1),* pp.603.

[11] Ekanayake, J., 2021. Bug Severity Prediction using Keywords in Imbalanced Learning Environment. *Int. J. Inf. Technol. Comput. Sci. (IJITCS)*, *13*, pp.53-60

[12] Ekanayake, J.B., 2021. Predicting Bug Priority Using Topic Modelling in Imbalanced Learning Environments. *International Journal of Systems and Service-Oriented Engineering (IJSSOE)*, *11*(1), pp.31-42

[13] upGrad blog. 2021. *Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022 | upGrad blog*. [online] Available at: <https://www.upgrad.com/blog/naive-bayes-explained/#:~:text=Naive%20Bayes%20is%20suitable%20for,input%20variables%20than%20numerical%20variables.> [Accessed 4 February 2022].

[14] analytics Vidhya. 2017. *6 Easy Steps to Learn Naïve Bayes Algorithm with codes in Python and R.* [online] Availableat:<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> [Accessed 4 February 2022].

[15] Edward, E., 2018. *Comparing Methods of Text Categorization*. [online] Diva-portal.org. Available at: <http://www.diva-portal.org/smash/get/diva2:1275337/FULLTEXT01.pdf> [Accessed 1 March 2022].

Author Biographies

Pirunthavi SIVAKUMAR received a B.Sc. degree from Uva Wellassa University of Sri Lanka, Badulla, Sri Lanka, in 2020. She is currently a M.Sc. candidate at Postgraduate Institute of Science, University of Peradeniya, Sri Lanka. She is also a temporary lecturer at the Department of Information & Communication Technology, Faculty of Technology, Rajarata University of Sri Lanka, Mihintale, Sri Lanka. Her main research interests include machine learning and data mining.

E-mail: psivakum@tec.rjt.ac.lk

Jayalath EKANAYAKE received his B.Sc., and M.Sc. degrees from the University of Peradeniya, Sri Lanka. He received his Ph.D. degree from the University of Zurich, Switzerland. He is currently a senior lecturer at the Department of Computer Science and Informatics, Faculty of Applied Sciences, Uva Wellassa University of Sri Lanka, Badulla, Sri Lanka. His main research interests include mining software repositories and pattern recognition.

E-mail: jayalath@uwu.ac.lk