2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT) | 978-1-6654-7647-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMPACT55510.2022.10029067

Comparison of SVM and Naïve Bayes for Sentiment Classification using BERT data

Shivani Rana[1], Rakesh Kanji[2], Shruti Jain[3]

[1, 2] Department of Computer Science and Engineering, Jaypee University of Information and Technology, Solan, HP, India
[3]Department of Electronics and Communication Engineering, Jaypee University of Information and Technology, Solan, HP, India

Abstract: Expressing views related to any common interest of area is becoming very general due to the excessive use of internet on the handheld devices like mobiles. Everyone is free to share his opinions within his social network. Having some useful knowledge from those opinions is the challenge for text classification techniques. In this work, the movies related opinions in the form of long and more than one sentence are analyzed to categorize them as positive or negative sentiments towards the movie. In this paper, the machine learning approach using the supervised learning methods is used for the implementation of the model. The data is prepared using the BERT model where plain text is directly converted into a numerical data form. The experiments are conducted using the Support Vector Machine and Naïve Bayes algorithms. Accuracy, F1 score, and other metrics have been evaluated for the comparison of the results of both algorithms. The results show that the predictions by SVM have high accuracy than Naïve Bayes.

Keywords: Machine Learning, Sentiment Analysis, BERT, Naïve Bayes, SVM.

I. Introduction

With the advancement in the technology and internet age, a huge volume of data is generated by internet users on the web. The data may be from online businesses, social networking sites, blog spots, social media, etc. People have changed the way of exchanging their ideas among themselves. The large volume of data is generated by expressing one's own opinion about the products, by comments, reviews, tweets, etc. All the data generated so far by this type of information exchange is in unstructured form, so to get any type of information or patterns from the available data it is needed to be converted into the structured form. This led to the quick rise of Data Analytics. Data Analytics is the process to investigate data to draw inferences about the information it contains. The process of manual extraction is merely not possible as billions of data are generated every day. So, there is a need for an automated system for this extraction of information. For automation of the label or tags are to be assigned to the textual data such as emails filtering, retrieval of information, language detection, customer feedback analysis, etc. is known as Text Classification. The two common applications of text classification are e-mail spam filtering and Sentiment Analysis.

Sentiment Analysis is to determine the polarity of the opinions (comment is expressing the positive, negative or neutral) shared publically. The tagging of the text is done either manually or by automated labelling. The tremendous size of data leads to an increase in classification memory requirements and computation time [1]. The available data is redundant and insignificant; to address these problems feature extraction from the text came into existence to reduce the size of the text by removing unrelated and inappropriate words. Preparing a pre-defined set of rules for text labelling requires lots of domain knowledge. So, machine learning-based approaches are getting lots of attention for classifying the text based on observation of the data. These models run in two steps (a) Hand-crafted features are extracted from the available text and (b) Extracted features are fed into the classifier for prediction. For feature extraction using a machine learning NLP classifier, each text is transformed into a numerical representation in the form of a vector representation. A commonly used approach is Bag of Words (BOW) and its variants. Various machine learning classification algorithms used for text classification are Naïve Bayes (NB), Support Vector Machine (SVM), Hidden Markov Model (HMM), Gradient Boosting trees, and random forest. The limitations of the above-mentioned approach arise as dependency on hand-crafted features requires lots of feature engineering and analysis to have good performance [2]. So, the need for embedding models arises. Some of the models with their advancement sequence as LSA, word2vec, contextual embedding model (3-layer bidirectional Long Short-Term Memory (LSTM), embedding models using transformers, Generative Pre-trained Transformer (GPT), and Bidirectional Encoder Representations from Transformers (BERT)) [4]. BERT is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. BERT mainly focuses on local consecutive word sequences, which provide local context information.

Literature Review

This section presents the findings of the review of literature conducted to do this research work.

S.no	Author/Year	Contribution	Results
1	Vishal A Kharde *et al.* 2016 [1]	-Presents basics steps in sentiment analysis -provides an overview of various approaches in	i. SVM and naïve Bayes method have the highest accuracy and are regarded as baseline methods

978-1-6654-7647-8/22/$31.00 ©2022 IEEE

		sentiment analysis i.e. SVM, etc. -presents different challenges in sentiment analysis	ii. Different challenges of sentiment Analysis are presented- entity recognition is taken as a base for further work
2.	Rohitash Chandra *et al.* 2021 [3]	Presents a framework for employing deep learning-based language models via LSTM recurrent Neural Networks for sentiment analysis	The region-based study is conducted and to be extended to other regions
3.	Md. Shoaib Ahmed *et al.* 2021[4]	-User thoughts regarding COVID-19-related issues of health, and vaccination are studied to contribute to new interventions. - Clustering of sentiments based on topics	-active users are analyzed based on their participation -Detected variation in users' involvement as well as in their sentiments after a specific time interval. -in a future study to be continued in labeling them as +ve and -ve

Literature Survey: Text Classification using BERT, Naïve Bayes, SVM

S.no	Author/Year	Contribution	Results
1.	M.Thangaraj, *et al* 2018. [2]	Paper presents the analysis of various text classification techniques, their strengths, and weaknesses	Semi-supervised learning approach improves classification efficiency and is found suitable for labeling problems while handling a greater number of instances of an entity.
2	Shervin Minaee 2021 [5]	Presents a qualitative analysis of the performance of different DL models -presents a summary of 40 datasets for TC	-TC has seen great progress in recent years. -neural embedding, attention mechanism, self-attention, transformer, BERT, and XLNet has contributed to the fast progress of DL models
3.	Amirsina Torfi Feb 2021 [6]	Categories and addresses the different aspects and applications of NLP benefitting DL	The survey proves to be a guide for students and researchers for a basic understanding of the integration of NLP with deep learning
4.	Qian Li *et al.* 2021[7]	Presents the details of TC methods according to the text involved and method used for feature extraction and classification and their pros and cons	-Feature extraction and classifier design is improved by the shallow model to great extent. -metrics of single and multi-label tasks are evaluated
5.	Abdul Mohaimin Rahat *et al.* [8]	-Naïve Bayes and SVM are compared on the Airline review dataset.	- SVM provides better results than Naïve Bayes.
6.	Himanshu Batra *et al.* [9]	-Three variants of BERT for Sentiment Analysis are analyzed	- The ensemble model attains 6-12% improvement over others.
7.	Habeebullah shah *et al.* 2020[10]	-Naïve Bayes's performance is observed and analyzed for sentiment classification on user reviews online with proper selection of NB variants.	- Naïve Bayes achieved less accuracy than SVM and more than Random Forest Algorithm. - Selection of proper variant of NB depends upon the specific problem

From the literature, it has been found that the sentiments in the form of long sequences of words comprises of more than one sentence are not an easy task. Getting the opinion by remembering the long sequences is the main challenge addressed in this work. The dataset used in this work comprises this type of data to have the overall analysis regarding each sentiment huge number of words have to be taken into consideration. In this work, authors have used different approaches such as text or sentiment classification. As previously the work has been done to pre-process the text and then predictions are made based on machine learning algorithms [11]. In this paper, the authors used the BERT algorithm. This work differs as the text preparation phase also generates negative values (Output of BERT), which can't be used in this particular method. To overcome this problem Gaussian Naïve Bayes and SVM methods are used. We generated a classification report based on various metrics for both the Naïve Bayes and SVM methods. For comparative purposes, various performance metrics are evaluated.

II. Methodology

BERT is fine-tuned for classification tasks and the other approach followed for text classifications task is preparing the data by removing the punctuations, stopwords hashtags, etc. then the formation of vector-based input matrix of the text to be fed into the model. This work tries to follow a different approach of using BERT capabilities with a machine learning model for predictions. To get the sentiment of a particular text shared by a user publically, supervised learning approaches are used where training is done on labeled data for generating predictions on unknown data. A step-by-step approach is followed for sentiment analysis. SA is the automated process of reading a text for opinion mining. The main motive of this work is to analyze our data with the Naïve Bayes and SVM algorithms. Fig 1 presents the working of our proposed work.

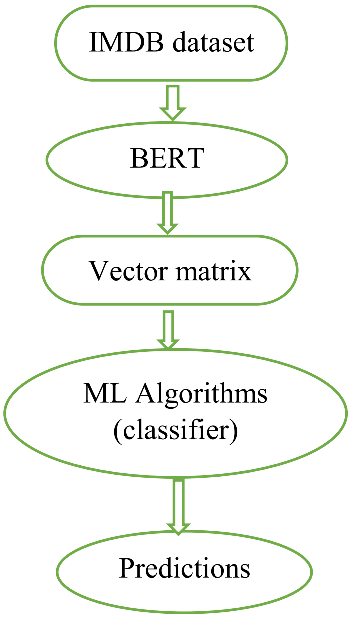

Fig1. Proposed Work

The authors have used the IMDB dataset for our work. This dataset is based on movie reviews, basically binary sentiments classification. All the simulations were carried out using Python programming. It contains around 5K labeled reviews. In our work, the data is directly fed into the BERT model to have the vector matrix of each review. Then supervised learning approaches are used for making predictions. BERT is not a word embedding, but it is a method to generate word embeddings. To have better knowledge of text BERT works bidirectionally to read the text from left to right and from right to left. It simultaneously takes previous and next tokens into an account for language modeling. The input of BERT is trained as shown in Fig 2 and Fig 3 respectively.

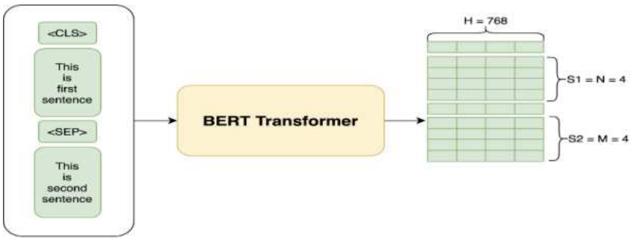

Fig 2: A very High view of BERT [12]

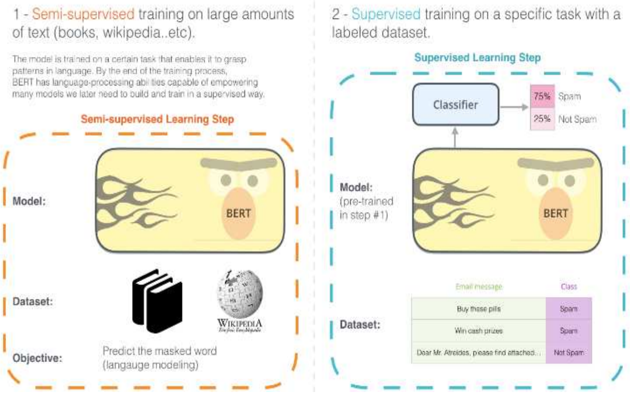
1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.
Semi-supervised Learning Step
Model:
BERT
Dataset:
WIKIPEDIA
Objective:
Predict the masked word (langauge modeling)
2 - Supervised training on a specific task with a labeled dataset.
Supervised Learning Step
Classifier
75% Spam
25% Not Spam
Model:
(pre-trained in step #1)
BERT
Email message
Class
Buy these pills
Spam
Win cash prizes
Spam
Dear Mr. Atreides, please find attached...
Not Spam
Dataset:

Fig 3: Two ways of using BERT [13]

BERT is used to avoid the problems of hand-crafted feature generation as stated earlier [14]. The model is trained in a semi-supervised way [15, 16] for generating language modeling as word embedding for a large amount of text in a single review or opinion. NB and SVM [17, 18, 19] are used for the classification. *Naïve Bayes* is based on applying Bayes theorem with the "naïve" conditional independence assumption of feature matrix as expressed by Eq. (1).

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y), \qquad (1)$$

for all *i,* this relationship can be simplified by Eq. (2).

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} \quad P(x_i \mid y)}{P(x_1, \dots, x_n)} \qquad (2)$$

It is an extremely fast and most popular method for classification-based applications. In our work, we are using Gaussian NB for classification (its likelihood features are Gaussian) as expressed by Eq. (3).

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \qquad (3)$$

SVM approach suits the applications with high-dimensional spaces. Kernel-level functions can be varied for varying decisions. The basic principle behind SVM is creating a hyper-plane that separates the dataset into classes. The support vectors (points lying close to both classes) were detected first. The proximity is calculated between our dividing plane and the support vectors. SVM aims to maximize the margin (distance between the points and the dividing line) called optimal hyper-plane. In our work RBF kernel is used as expressed by Eq. (4)

$$\exp\left(-\gamma\|x-x'\|^2\right) \tag{4}$$

where γ is specified by parameter *gamma*, must be greater than 0?

III. Results and Discussion

In the various studies, different methods of machine learning are used for classification purposes. For experimental purposes, the authors have used the 80% of the data for the training purpose and the random state is 40. The Naïve Bayes and SVM techniques are used for the predictions. Authors have experienced in our work that the result of the SVM model is providing good results with a high F1 score. The confusion matrix for NB and SVM is shown in Fig 4 and Fig 5 respectively.

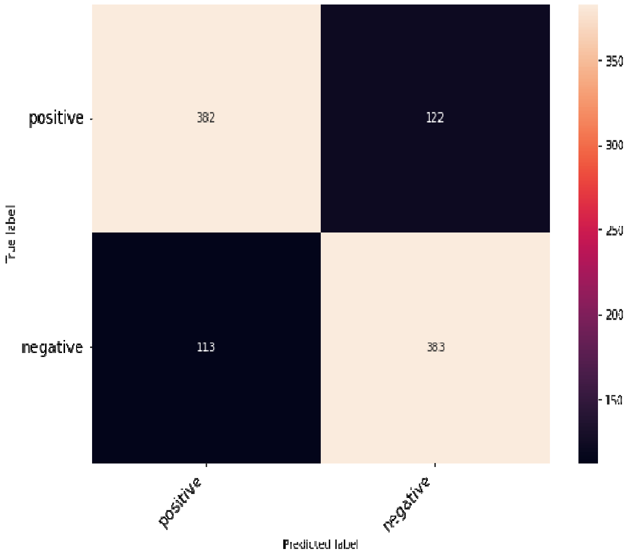

Fig. 4 : Confusion Matrix for Naïve Bayes

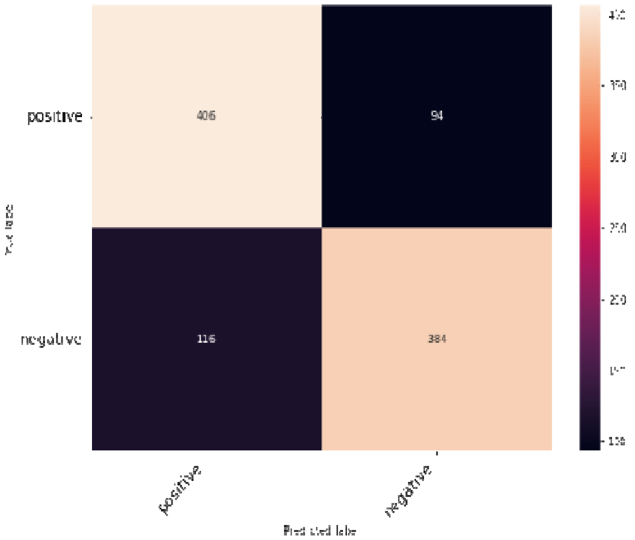

Fig 5: Confusion Matrix for SVM

Based on the confusion matrix, different parameters like Accuracy, Precision, Recall, and F1 score are evaluated and the results of the experiments are shown in Table 1.

Table I . Experimental Results of proposed Model

	SVM (%)	Naïve Bayes (%)
Accuracy	79	77
Precision	80	77
Recall	77	76
F1score	79	76

The authors have used the SVM Radial Basis Function kernel (RBF) with *gamma* variable and Gaussian function for the Naïve Bayes has been used. In most of the previous studies, multinomial NB is used for text classification as vector matrices are generated from the text. In this work, the BERT output matrix is used instead of using tf-idf for creating vector matrices. Some of the data have negative values so multinomial NB can't work in our model. Based on the Confusion matrix the Area Under Curve (AUC) for NB and SVM is shown in Fig 6 and Fig 7 respectively.

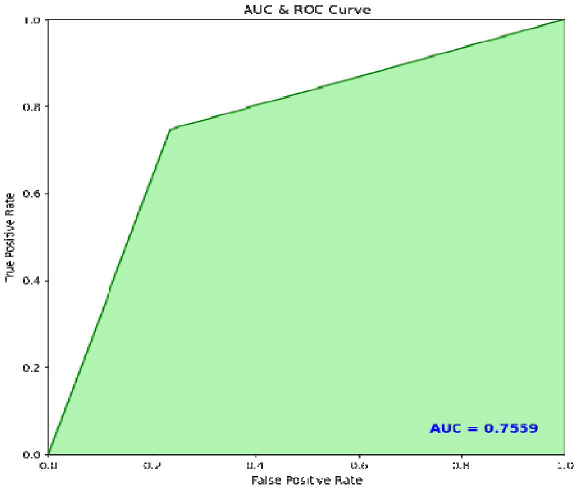

Fig 6 : AUC curve for Naïve Bayes

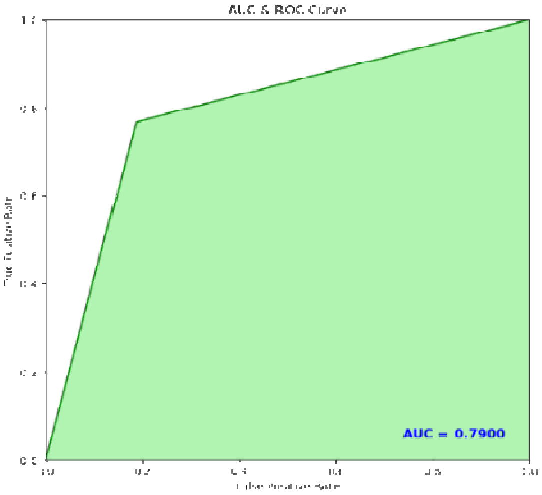

Fig 7: AUC curve for SVM

The plotting of the graph with the true positive rate and the false positive rate at different classifications threshold values forms the ROC curve and the AUC presents the performance of our model. In this work, the 0.79 AUC value for SVM shows that 79% of predictions are true. Whereas 75% of predictions of Naïve Bayes are correct. The work depicts that the SVM classifiers-based prediction outperforms the Naïve Bayes approach in either text cleaning and preparation phases or directly feeding plain text to BERT for vector-matrix generation is used. The data prepared with different methods are unable to enhance the Naïve Bayes performance as compared to SVM

IV. Conclusion and Future work

In this paper machine learning-based algorithms namely SVM and Naïve Bayes are studied. The SVM RBF kernel with gamma variable works well than the c variable. SVM outperforms then Naïve Bayes as its F1 score measure is higher but not up to the expected values. 2.5% accuracy improvement has been observed when using SVM over NB. The text data is directly fed as plain text into the BERT model, reducing the text preparation time but the result of the predictions is not achieved to a great extent. More machine learning models may be compared with the same approach to achieve good accuracy. Also, deep learning algorithms can be exploited to have better or expected performance for text classification so will continue to work with deep learning methods in future work.

References

[1] Vishal A kharde. SS Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Application*, Vol. 139, 2016.

[2] M Thangaraj, M Sivakami, "Text Classification Techniques: A Literature Review*,"* *Interdisciplinary Journal of Information, Knowledge, and Management*, 2018.

[3] Rohitash Chandra, Aswin Krishna, "Covid-19 sentiment analysis via deep learning during the rise of novel cases," *Journal Pone*, 2021.

[4] Md Shoaib Ahmed, Tanjim Taharat Aurpa, Md Musfique Anwar, "Detecting sentiment dynamics and clusters of twitter users for trending topics in Covid-19 pandemic*," Journal PLOS ONE*, August 2021.

[5] Shervin Minaee, "Deep Learning Based Text Classification: A Comprehensive Review," *ACM*, Vol 1, January 2020.

[6] Amirsina Torfi, Rouzbeh a Shirvani, Yaser Keneshloo, Nader Tavaf, Edward A Fox, "Natural Language Processing Advancements by Deep Learning: A Survey," *arXiv: 2003.01200v4 [cs.CL*], Feb 2021.

[7] Qian Li, Hao Peng, Jianxin Li, "A Survey on Text Classification: From Shallow to deep learning," *IEEE Transactions on neural networks and learning systems*, vol31, October2021.

[8] Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum, "Comparison of Naïve Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," *Proceedings of the SMART-2019, IEEE ConferenceID:46866., 8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019 College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India.*

[9] Himanshu Batra, Narinder Singh Punn Sanjay Kumar, Sonali Agarwal, "BERT- Based Sentiment Analysis: A Software Engineering Perspective," axXiv:2106.02581v3[cs.cv] 2 July 2021.

[10] Habeebullah Shah Quadri, RK Selvakumar, "Performance of Naïve Bayes in Sentiment Analysis of User Reviews Online," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* Vol 10 issue 2 December 2020.

[11] Ankita Yekurke, Gayatri Sonawane, Harshada Chavan, Pradhyumn Thote, Anuja Phapale, "Sentiment Analysis using Naïve Bayes, CNN, SVM," *International Research Journal of Engineering and Technology,* Vol 9 issue 04, April 2022.

[12] J Devlin, M W Chang, K Lee, K Toutanova, "Bert Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805* (2019).

[13] Jay Alammar. "The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)", Github.io. https://jalammar.github.io/illustrated-bert/ (accessed August 06,2022).

[14] Touseef Iqbal, Shaima Qureshi, "The Survey: Text generation models in deep learning," *Journal of King Saud University-Computer and Information Sciences*, April 2020.

[15] AO Salau, Shruti Jain, "Adaptive Diagnostic Machine Learning Technique for Classification of Cell Decisions for AKT Protein," *Informatics in Medicine Unlocked*, 7 January 2021, 100511.

[16] Shruti Jain, DS Chauhan, "Computer Aided Design and Synthesis for Marker Proteins of HT Carcinoma Cells: A Study," *In: Paul S. (eds) Biomedical Engineering and its Applications in Healthcare. Springer, Singapore*, pp 399-419, 2019.

[17] Shruti Jain and Meenakshi Sood, "SVM Classification of Cell Survival/ Apoptotic Death for Color Texture Images of Survival Receptor Proteins," *International Journal on Emerging Technologies* 10(2): 23-28(2019).

[18] Shruti Jain, Ayodeji Olalekan Salau, "An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins," *Cogent Engineering*, 6(1): 1599537, 1-14, 2019.

[19] Navdeep Prashar, Meenakshi Sood, Shruti Jain, "A Novel Cardiac Arrhythmia Processing using Machine Learning Techniques," *International Journal of Image and Graphics*, 20(3), 2050023(17 pages), 2020.