# Exploring different approaches for sentiment analysis.

Sridhar Ashwin
*School of Computing*
*Dublin City University*
Dublin, Ireland
ashwin.sridhar3@mail.dcu.ie

Nair Rohit
*School of Computing*
*Dublin City University*
Dublin, Ireland
rohit.nair2@mail.dcu.ie

Foster Jennifer
*School of Computing*
*Dublin City University*
Dublin, Ireland
jennifer.foster@dcu.ie

***Abstract*— Social media lets users share their stories, experiences or opinions with everyone on the internet. It has become a place where you can make your voice heard. With the ever-increasing use of such platforms, the amount of textual data is also growing. This huge data can be used to analyse the general sentiment of users and is known as sentiment analysis. Opinion mining or sentiment analysis helps us interpret the emotions from any given data. Using NLP, we can extract high quality information from text and use it to improve customer experience or gather information of the masses on a particular issue or topic. This project explores different approaches to classify datasets obtained from Twitter, Amazon and IMDB. The 3 different datasets are used train and validate 5 models and their accuracy is used to determine which model performs better for each dataset.**

**Keywords— sentiment analysis, NLP, XGBoost, SVM, Naïve Bayes classifier**

## I. Introduction

The digital era has given users access to the internet and this generates a huge amount of data each second. Every click on the internet generates data which can be captured by the website or the application. In our digital world, data is the new oil. Twitter, Facebook, LinkedIn and other such social media platforms give users a stage to voice their opinion.

People nowadays are keen to know others' opinion about a product/service before they are buying and make their decisions based on the reviews of others. Big E-commerce companies like Amazon or eBay receive extremely high number of reviews on their products. Several reviews for a specific product, brand, individual personality etc. are very helpful in deciding the perception of people towards the product`. Hence it becomes necessary to create algorithms to automate the classification of distinctive reviews based on their polarities particularly: Positive, Negative and Neutral. This automated classification mechanism is referred to as Sentiment Analysis.

Sentiment analysis involves analysing the data and extracting the sentiment attached to it. Natural Language Processing (NLP) enables us to extract emotions from human language. It involves tokenization of sentences, i.e. breaking down sentences into words and deriving meaning from it based on the sentiment value associated with each word.

Each word has a sentiment value attached to it based on its intensity and level of emotions that it conveys. This project explores lexicon based and machine learning based algorithms to compares their accuracy and performance on different data sources.

Three datasets from different domains have been used in this project. Traditional machine learning algorithms like Naive bayes and Support Vector Machine (SVM) have been performing well in classification tasks and our attempt in this work is to compare the performances of SVM, Naive Bayes, XGBoost and the formula based method for predicting the sentiment of the sentence. We pre-processed the data by removing punctuations and stop words, followed by lemmatization, tokenization and tf-idf for feature extraction.

## II. Literature Review

There are systems that have currently implemented both, lexicon and machine learning based algorithms. The most common approach is a mathematical method to derive sentiment from data collected.

Reference [7] extracted tweets from twitter to determine the effects of an average person's tweets over fluctuation of stock prices of a multinational company. Work in this paper focuses on unsupervised learning using sentiment-lexicon based approach. Sentence level approach is being used here. Here, extraction of tweets is done from twitter4j library. The extracted tweets are then pre-processed and cleaned by removing unwanted words in the sentences. The tweets are then split into tokens which are compared with a dictionary of words and scores are assigned to words where a positive word is 1, negative word is -1 and neutral word is 0. The final sentiment of the tweet is calculated by adding all the scores of the individual tokens.

Xing Fang & Justin Zhan [6] use product reviews from E-commerce websites like Amazon.com to perform both sentence-level and review level categorization. This paper tackles a fundamental problem of sentiment analysis, namely sentiment polarity categorization. They compared Naïve Bayesian, Random Forest and SVM classification. They also implemented parts of speech (POS) tagging to remove parts of speech that do not contribute to the sentiment. The dataset consists of 5.1 million reviews from amazon belonging to 4 major categories: beauty, books, electronics and home. Feature vector formation method is finally used to compute the sentiment of the sentence.

Yige Wang [2] combined POS tagging with Naïve Bayes classifier to train the model which is used to predict the sentiment of movie reviews. With POS tagging, the complexity was reduced by a huge margin as the number of words that had to be trained decreased, but the overall accuracy of the model decreased by 7%. The experiment shows that the sentiment of a sentence or a document is mainly dependent on adjectives and verbs and hence only certain types of words are selected to train the model using POS tagging thereby reducing computational complexity.

Reference [3] compares Naïve Bayes approach and SVM to predict sentiment using review dataset. The dataset they considered of airline reviews which was collected from twitter. They found the accuracy of SVM to be higher than Naïve Bayes, the accuracy of SVM was found to be 83% and

Naïve Bayes provided an accuracy of 77%. The result shows that in the case of airline reviews, SVM algorithm gives better results than Naive Bayes algorithm.

H. Parveen and S. Pandey [4] use Naïve Bayes algorithm to reduce the overhead. The current systems lack the capability to handle huge datasets and are very time consuming. The study proposes the use of Hadoop and MapReduce architecture to handle big data. They also used emoticons to extract sentiment from the data. The whole emoticon was converted into its equivalent word and the sentiment associated with the word was used.

U. Kumari, A. Sharma and D. Soni [5] use SVM classification technique to find the sentiment from a product review dataset. The accuracy of the model was tested with datasets of 4 different products. The highest and the lowest accuracy the model achieved were 90.99% and 88.03% respectively. Though the accuracy of the model is high, the datasets used for testing were considerably small.

M. Wongkar and A. Angdresey [1] also use tweets to assess the political sentiment of Indonesian people during the 2019 presidential elections. A comparison was carried out using Naive bayes method, Support-vector networks and K-Nearest neighbour methods. The method used to collect data from twitter is data crawler in this study. Probability of each word being positive or negative is calculated and multiplied. The result is used to classify the sentence as positive or negative. Twitter is a platform where people mostly use text of less than 140 words to publish their opinion. Hence tweets are one of the most effective datasets for sentiment analysis. In this study, it was found that the Naive Bayes method is the most accurate with 80.1% accuracy.

Bo Pang and Lillian Lee [8] have used IMDB movie review dataset. They have used three machine learning methods namely Naïve Bayes, Maximum Entropy classification and support vector machines. The paper classifies movie reviews as either Positive or Negative. This was one of the first attempts to take this approach. Bag of words method was used and Unigrams were taken as features for this classification problem and the end result showed that the system performed well with either of the approaches.

Sida Wang and Christopher D. Manning [9] conclude variants of Naïve Bayes and Support Vector Machines (SVM) behave differently depending on the model variant, feature used and task/dataset. They have identified simple Naïve Bayes and SVM variants that provide great results when compared to most published results on sentiment analysis. A new variant NBSVM that was identified in this paper works well on snippets and longer documents, for sentiment, topic and subject classification and is found to be better than most published results. The work introduced different variants like MNB (Multinomial Naïve Bayes), Multivariate Bernoulli NB (BNB) and NBSVM (Naïve Bayes Support Vector Machines). MNB is found to perform better than BNB in most cases.

## III. METHODOLOGY

### A. Domain Understanding

Huge amount of textual data is generated every second online and offline. This data can be used by organisations to detect the sentiments of the users or customers regarding their products or company. It can also be used to analyse sentiments towards other topics in general. As different social media platforms help users voice their opinions publicly, it becomes all the more important to tap on this valuable resource and put it to good use. Analysing customer opinion is vital for organisations as it helps them work on their brand image and cater to consumer needs actively.

### B. Dataset

Three standard datasets will be used to train and test the models. Stanford Artificial Intelligence Laboratory contains labelled datasets of IMDB movie reviews and amazon user reviews. There are 12000 positive and negative movie reviews for training as well as testing.

The Amazon dataset has approximately 120000 records and the ratings are on a scale of 5 which was acquired from Stanford Network Analysis Platform. This will have to be converted to positive or negative to train the model. The twitter dataset consists of 1.5 million tweets labelled with 1 and 0. 1 being positive and 0 being negative. This dataset was acquired from Thinknook.com.

### C. Data Cleaning, Pre-Processing And Transformation

Every document, irrespective of which data source it is from, is converted to lower case, tokenised, stripped of stop words and punctuations, and lemmatised. Additionally, emoticons are also replaced with the sentiment word that they are convey. To see the frequency of words in the documents, we graph out a word cloud. The dataset is then split into training and validation datasets, in an 80:20 ratio of 40000 and 10000 documents respectively.

A new count vector of all documents is used to keep a track of occurrences of words which occurs in at least 0.005% of the documents and not more than 95%. This data frame is used to for training and prediction in supervised learning. It should be taken care that all the words which occur in the training dataset should be present in testing dataset with their counts.

The IMDB dataset is segregated into 4 files of positive training data, negative training data, positive testing data and negative testing data with 12500 reviews each. The testing set, 5000 random records from the positive and negative sets are extracted and create a data frame. Similarly, the remaining 7500 positive and negative records are concatenated with the 12500 positive and negative training datasets. Effectively, our training dataset now consists of positive and negative records, with 20000 records each. Due to the data being segregated into positive and negative records, we concatenated them in a manner that the positive records appear in the top half of the said dataset and similarly the negative dataset appear in the bottom half. Using this dataset in this state would induce bias in the model and to avoid that, we reshuffle the training and testing datasets such that the positive and negative records are organised in a random manner.

### D. Modelling

*a. Lexicon Based*: The lexicon-based approach is not supervised learning. From every document we extract positive and negative words. A dictionary is being used to identify positive and negative words and sentiment scores are calculated in the following manner [7]:

$$\frac{Positve\ Words - Negative\ Words}{Total\ Number\ Of\ Words}$$

*b. Naive Bayes:* Naive Bayes classifier is a probabilistic approach based on Bayes rule and assumes that the attributes are conditionally independent. Multivariate Bernoulli model and the multinomial model are generally used in text mining. The Multinomial Naive Bayes model uses information about the number of times a word appears in a document. It treats each occurrence of a word in a document as a separate event. These events are assumed independent of each other. Hence, the probability of a document, given a class, is the product of the probabilities of each word event, given the class.

*c. Support Vector Machine:* Support Vector Machine finds a hyperplane in an N dimensional space (N- number of features) which distinctly classifies the data points. SVM can be used to solve both regression and classification problems. However, it is primarily used for classification. SVM can be of two types, namely linear and nonlinear.

*d. XGBoost:* Extreme Gradient Boost, also known as XGBoost, is a class ensemble machine learning algorithm. It can be used for classification or regression problems. It is good with handling outliers. It can handle large datasets and is not prone to overfitting. But tuning it becomes a bit difficult due to the amount of hyperparameters involved.

Each model is trained and validated using the three datasets mentioned above. The IMDB dataset is a dataset with balanced classes. The twitter and Amazon datasets had to be balanced and under sampled to ensure similar nature of data.

## IV. Results and Evaluation

The performance of each model on the 3 data sets are as follows:

Table I Model Performance

| **Dataset** | **Accuracy** | | | |
|---|---|---|---|---|
| | **SVM** | **Naïve Bayes** | **XGBoost** | **Lexicon Based** |
| **Twitter** | 73.9 | 74.8 | 72.5 | 65.3 |
| **IMDB** | 87.0 | 82.4 | 85.7 | |
| **Amazon** | 86.0 | 86.0 | 84.1 | 71.8 |

Table 1 shows the performance scores of each model. SVM performs the best for the IMDB dataset. For the amazon product reviews, both SVM and Naïve Bayes had an accuracy score of 86%. Naïve Bayes classifier outperformed others for the twitter data set.

## V. Conclusion

Data is being generated by everyone these days. Users post their views and form opinions based on what others share online. The goal of this project is capitalising on this huge source of data, find which approach is better for each dataset and explore the reason behind an approach outperforming the other. Businesses need to know what their customers think about their products and it is humanly impossible to go through each and every review. The goal is to automate the process and give the businesses an overview of the sentiment of the customers towards their product. A decision, if not backed by data, will not yield good results. The results will be evaluated by calculating the accuracy percentage of the model based on the difference in predicted sentiment and actual sentiment. This will help us determine which approach would be better for extracting user sentiment for each subdomain.

## References


[1] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, 2019.

[2] Y. Wang, "Advanced Naïve Bayes Algorithm Design with Part-of-Speech Tagger on Sentiment Analysis," in *International Conference on Computer Systems, Electronics and Control (ICCSEC)*, Dalian, 2017.

[3] A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, 2019.

[4] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," in *2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, 2016.

[5] U. Kumari, A. Sharma and D. Soni, "Sentiment analysis of smart phone product review using SVM classification technique," in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, 2017.

[6] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, 2015.

[7] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," in *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016.

[8] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning," in *Empirical Methods in Natural Language Processing (EMNLP)* , Philadelphia, 2002.

[9] S. Wang and C. D. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification".

[10] B. Pang and L. Lee, Opinion mining and sentiment analysis, now Publishers Inc., 2008.