Exploring different approaches for sentiment analysis

Ashwin Sridhar
Student ID: 20210149
School of Computing
Dublin City University
Dublin, Ireland
ashwin.sridhar3@mail.dcu.ie

Rohit Nair
Student ID: 20210378
School of Computing
Dublin City University
Dublin, Ireland
rohit.nair2@mail.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 to train and validate 5 models and their accuracy is used to determine which model performs better for each dataset.**

Keywords*— sentiment analysis, NLP, Extreme Gradient Boosting, Support Vector Machine, Naïve Bayes classifier, Logistic regression

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 an 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 compare 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, Extreme Gradient Boosting (XGBoost) and the lexicon-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 generating count vectors for feature extraction.

The results of our research suggest that SVM performs the best overall and the lexicon based approach had the lowest accuracy score for each dataset. Naive Bayes outperforms SVM for the Amazon dataset, but only by a small margin while SVM had the highest accuracy score for the other 2 data sets.

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.

In their work [7], Bakshi et al. 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.

Fang & Zhan [6] use product reviews from E-commerce websites like Amazon.com to perform both sentence-level and review level 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.

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 suggests 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.

Rahat et al. [3] compare Naïve Bayes approach and SVM to predict sentiment using a 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 suggests that in the case of airline reviews, SVM algorithm gives better results than Naive Bayes algorithm.

Parveen & 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.

Kumari et al. [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.

Wongkar & 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, SVM 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 texts 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.

Pang & Lee [8] have used IMDB movie review dataset. They have used three machine learning methods namely Naïve Bayes, Maximum Entropy classification and SVM. 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.

In their work [9] Wang & Manning conclude variants of Naïve Bayes and 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 are being used to train and validate the models. The twitter dataset consists of 1.5 million tweets labelled as 1 and 0. 1 being positive and 0 being negative. This is based on data from 2 different sources, one provided by the University of Michigan and the other by Niek Sanders. This dataset was acquired from Thinknook.com.

Stanford Artificial Intelligence Laboratory contains labelled datasets of IMDB movie reviews. There are 12500 positive and negative movie reviews for training as well as testing, making a total of 50000 reviews.

The Amazon dataset, released in 2018, has approximately 233.1 million reviews for 29 categories and 8.1 million for the category “Movies and TV” selected for this research. It contains reviews from May 1996 to October 2018. The ratings are on a scale of 5 which are converted to either positive or negative to train the model.

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 a similar nature of data.

C. Data Cleaning, Pre-Processing And Transformation

Every document, irrespective of which data source it is from, is converted to lowercase, tokenized, stripped of stop words and punctuations, and lemmatised. Additionally, emoticons are also replaced with the sentiment word that they 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 occur in at least 0.005% of the documents and not more than 95%. This data frame is used for training and prediction in supervised learning.

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 is a simple, yet fast, accurate and a reliable machine learning algorithm. Naive Bayes classifier is a probabilistic approach used in a variety of classification tasks based on Bayes rule and assumes that the attributes are conditionally independent and contribute equally to the outcome. It is found to work particularly well with Natural Language processing problems. 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: SVM is a supervised machine learning algorithm that can be used for both regression and classification problems. Each data item is plotted in a n-dimensional space (where n is the number of features) with the value of each feature being a value of a particular coordinate. A hyperplane that separates the data points is then found to perform classification. In SVM two parallel lines on either side of the hyperplane are created called margins which are tangents to the nearest positive and negative points respectively. These positive/negative data points that lie on the marginal plane are what are called as support vectors. The distance between the hyperplane and the margins is called the marginal distance and this distance is determined with the help of support vectors. The best hyperplane for a classification problem in SVM is simply the one where the marginal distance is the greatest. SVM can be of two types, namely linear and nonlinear. Linearly separable models are the models where the data points can be easily separated using a hyperplane which is a straight line. Non linear separable models cannot be classified with only a straight line as a hyperplane. SVM kernels are used to separate non linear models.

d. Logistic Regression: Logistic regression model is a machine learning algorithm used for classification problems. Logistic regression predicts whether something is true or false instead of predicting something continuous. Instead of fitting a line to the data, logistic regression fits an “S” shaped logistic function which is known as a sigmoid function. The curve goes between 0 and 1 because the sigmoid function always takes these two values as maximum and minimum. Logistic regression models have a certain number of fixed parameters that depend on the number of input features and they output categorical prediction. Optimization algorithm like gradient descent or probabilistic methods like maximum likelihood can be used to train the logistic regression model. It is one of the simplest machine learning models.

e. XGBoost: Extreme gradient boost, also known as XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. In boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and boosts the attributes that led to misclassification in the previous tree. The tree that comes next in the sequence will learn from an updated version of the residuals. XGBoost is different from other boosting methods in its use of regularised boosting which helps avoid the overfitting problem. One of the best features of XGBoost is that it can handle missing values automatically. Performance evaluation can be performed at each iteration and it enables early stopping by finding the optimal number of iterations. It can handle large datasets but tuning it becomes a bit difficult due to the amount of hyperparameters involved. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage.. It is one of the simplest machine learning models.

IV. Results And Evaluation

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

TABLE I. ACCURACY SCORES IN PERCENTAGE

	Twitter	**IMDB**	**Amazon**
Lexicon	63.2	65.81	72.15
Logistic Regression	71.63	82.58	80.24
Naïve Bayes	74.27	82.8	85.73
SVM	75.27	86.95	84.98
XGBoost	73.95	86.34	84.87

Table 1 shows the accuracy scores for each model in the 3 different data sets that were used for the experiments. The lexicon-based method consistently performs subpar compared to the machine learning models. It is an initiation-based method where we assume that the sentences having a greater number of positive or negative words belong to that respective class. This is mainly because of the base assumption that the number of positive or negative words determine the overall sentiment of the sentence and each word is given equal importance. Its high dependency on a dictionary also contributes to

A significant improvement in performance is seen with machine learning algorithms compared to the lexicon-based method. This might be because the machine learning algorithms, unlike the lexicon-based method, associates different weights with different words which helps in deciding the class of the sentence effectively.

In our experiments Logistic Regression had the lowest accuracy score out of the 4 machine learning algorithms. Naïve Bayes performed the best for the Amazon dataset with an accuracy score of 85.73%. It performs competitively for the twitter dataset too. SVM has the highest accuracy score for the Twitter (72.27%) and IMDB (86.95%) dataset and performs well with the Amazon (84.98%) dataset too. XGBoost also performs well but not better than SVM with accuracy scores of 73.95%, 86.34% and 83.46% for Twitter, IMDB and Amazon data sets respectively.

Our results suggest that SVM performed better overall while Naive Bayes had a higher accuracy score, by a small margin, for the Twitter dataset. The lexicon-based method did not perform well and had the lowest accuracy score across all 3 datasets.

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.

All models were run on a local system which limited the processing capabilities for tuning the models. An instance of higher computing power can be used to tune the model and possibly yield better results. For future work, a weighted average ensemble approach can also be used.

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.

[11] A. Mass, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts,“Learning Word Vectors for Sentiment Analysis”. *The 49th Annual Meeting of the Association for Computational Linguistics,* 2011.

[12] J. Ni, J. Li, J. McAuley,*“*Empirical Methods in Natural Language Processing (EMNLP)*”*, 2019.