Study of Sentiment Analysis Through a Machine Learning Perspective

Sherlin Angel Narayanan and Md L. Ali, Ph.D.
Department of Computer Science and Physics, Rider University

ABSTRACT

When a person is scrolling through their social media feed, YouTube feed, or simply on Google, they are shown a variety of websites, posts, ads, and so on. If the information that is displayed is randomly presented, it does not benefit the company presenting the data or the user viewing the data. However if data is analyzed respectively to the users to see what data users seem to like and what data users show a disliking to, then both parties can benefit. The emotions of users will be analyzed with respect to the posts or the information being presented on the internet in an analysis known as Sentiment analysis to display products or info that users like and will likely purchase. Therefore, if companies rely on such analysis, appropriate methods must be used to conduct such analyses. In this research, different machine learning algorithms were analyzed and compared to highlight the best ways to benefit internet users through sentiment analysis. Multinomial Naive Bayes, Complement Naive Bayes, Passive aggressive Classifier, Logistics Regression Classifier, Support Vector Machine, and Decision Trees were the algorithms that were analyzed in this study through three different ngrams. Accuracy was the primary metric used for comparison, however precision, F1, and recall were also used as comparison metrics. At the culmination of the analysis, logistics regression with unigrams was found to have the highest accuracy of 63.76% in sentiment analysis.

INTRODUCTION

Sentiment analysis is the analysis of the emotional information within textual data. The analysis takes data to extract, interpret, and classify the emotions as positive or negative within the data usually in reviews, and social media posts to help businesses understand how to better target consumers emotionally through their products [1]. However, it is crucial to find appropriate technology to conduct such sentiment analysis in order to obtain accurate results. A branch of computer science, specifically Artificial Intelligence (AI), that studies data and attempts to create algorithms for the purpose of mimicking human behavior and learning by collecting past data to use for future outcomes is known as machine learning (ML). ML slowly learns from patterns and improves the accuracy of its algorithms for better performance [2]. It is crucial to realize the accuracy or general performance of specific machine algorithms in focus in order to adequately utilize the appropriate technology that provides the best outcome of results. In this study 6 different algorithms were compared and analyzed to identify that were tested in this study.

LITERATURE SURVEY

Before proceeding with completing research pertaining to this current study, it is imperative to gather together previous research that was conducted on topics that may prove useful for the study at present. Therefore, several papers were obtained from background research on sentiment analysis through machine learning. The summary of most of the papers that were studied for background research was organized and categorized into papers that focused on hybrid algorithms, and general supervised learning algorithms. The summarized information can be visualized in Tables 1 and 2 below. **Tables 1** and **2** depict the details of the paper associated with the information, the models that were used and produced the highest accuracy values, datasets, as well as performance metrics shown through accuracy for the different categories of algorithms.

Table 1: Hybrid and Deep Learning Algorithms Tested

Paper	ML Models	Dataset	Accuracy (%)
Naresh et al.	SMO + DT	Twitter	89.47
	KNN + SVM		76
Sadhasiva et al.	NB+SVM	Official Product Site	78.69
Rajput et al.	MNB+SentiWordNet	Twitter	86
Suneera et al.	CNN	News	80.27
	MLP		78.74
	LSTM		70.41
	CNN+LSTM		78.74

SMO - Sequential Minimal Optimization , DT- Decision Trees, KNN - K-Nearest Neighbors, SVM - Support Vector Machine, NB - Naive Bayes, MNB - Multinomial Naive Bayes, SWN - SentiWordNet, CNN- Convolutional Neural Network, MLP - Multilayer Perceptron, LSTM - Long Short-Term Memory

Table 2: Supervised Learning Algorithms Tested

Paper	ML Models	Dataset	Accuracy (%)
Elmurngi et al.	SVM	Movie Reviews	81.35
Narendra et al.	AH	Movie Reviews	98.02
Rahman et al.	MNB	Movie Reviews	88.50
Suneera et al.	LR	News	82.74
Alam et al.	NB	Twitter	91.81
Alshamsi et al.	NB	Twitter-Airline Tweets	97.65
Poornima et al.	LR	Twitter	86.23

SVM - Support Vector Machine, AH - Apache Hadoop, MNB - Multinomial Naive Bayes, LR - Logistics Regression, NB - Naive Bayes

EXPERIMENTAL SETUP

The dataset that was used for this study was retrieved from Kaggle and was movie reviews from Rotten Tomatoes [3]. 3 datasets were available within the module in which the data was retrieved. 1 dataset was a sample file showing the general setup of what the data offers. The second dataset was a file with 156,060 entries, that served as the training data with the ID of the phrase and sentence, the actual phrase, followed by the sentiment of the phrase in order to train the algorithm to accurately determine the sentiment of the phrase within the dataset. The third dataset was a smaller file with 66, 292 entries, that served as the testing data with the ID of the phrase and sentence, and the actual phrase but not the sentiment of the phrase so the algorithm could be tested. The dataset was visualized with regard to positive, negative, and neutral sentiments through means of graphical data.

Figure 1 shows the spread of those sentiments, categorized into five classes as neutral, somewhat positive, somewhat negative, positive, and negative reviews, along with the approximate values of those sentiments present in the dataset. **Figure 2** shows the simplified sentiment polarity of negative, positive and neutral emotions through a pie chart. The characterized data was then preprocessed in terms of modifying the data to eliminate errors and redundant, or unnecessary data. As term frequency-inverse document frequency (TF-IDF), word embedding, vectorizers, and tokenization are all measures that are normally applied to datasets to provide accurate results, as indicated by background research, similar approaches were applied to the dataset [1]. In addition to vectorizer, and tokenization for preprocessing, an important tool present in this study was the application of the n-grams model to the dataset that was split 75:25 as training and testing data. The six models that were selected were individually applied with unigrams, bigrams, and trigrams n-grams model following the preprocessing. The models were then evaluated and recorded with a set of metrics of accuracy, precision, recall, F1, and confusion matrices.

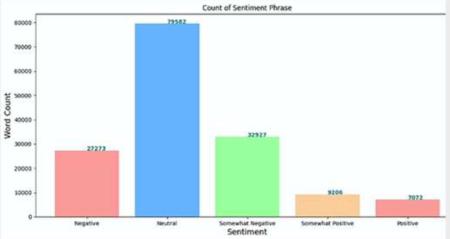

Figure 1: Bar Chart for Sentiment of Reviews

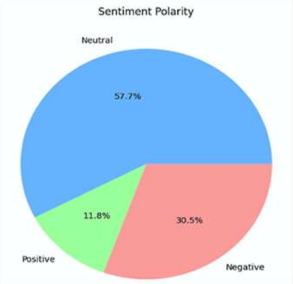

Figure 2: Pie Chart for Sentiment Polarity

EXPERIMENTAL RESULTS

Figure 3 depicts the relationship between the three types of ngrams that were applied to each algorithm to the accuracy of the algorithm. The exact accuracy, precision, recall and F1 values that were obtained for the unigrams of the six algorithms can be observed in **Table 3**. Of the six algorithms that were used, the logistics regression classifier presented the greatest accuracy, without factoring in other features, as indicated below. Figures 4 and 5 depict the confusion matrix for the highest accuracy algorithm and the lowest accuracy algorithm, respectively.

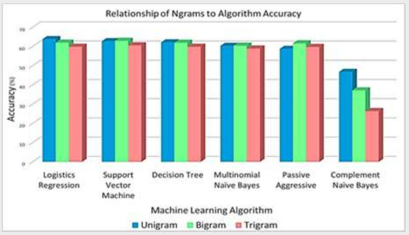

Figure 3: Relationship of Ngrams to Accuracy of Algorithm

Table 3: Actual Metric Values of Each Algorithm with Unigrams

ML Models	Accuracy	Precision	Recall	F1
Multinomial Naïve Bayes	60.25	51.22	44.87	47.18
Complement Naïve Bayes	46.74	39.39	45.24	40.28
Decision Tree	62.08	52.91	50.84	51.61
Support Vector Machine	62.68	54.72	47.08	49.9
Logistics Regression	63.76	57.03	45.96	49.51
Passive Aggressive	58.69	48.49	44.59	46.1

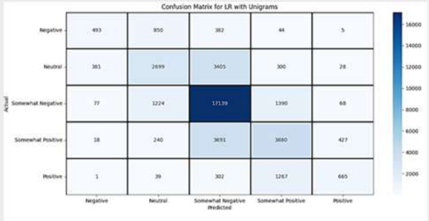

Figure 4: Confusion Matrix for LR with Unigrams

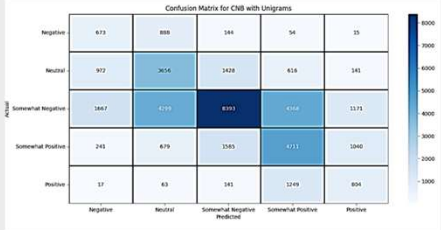

Figure 5: Confusion Matrix for CNB with Unigrams

CONCLUSION

Sentiment analysis is a very important tool that can be used to benefit businesses, social media users, and customers as it allows for all parties to express their emotion and be understood. Therefore, having functional and efficient algorithms for the purpose of conducting sentiment analysis is imperative, which was the goal of this study. The maximum accuracy attained through analysis of six different algorithms in this study was 63.76%. Unigrams was also found to be the n-grams model that presents the best performance results. The algorithm presented with the greatest accuracy was supported by previous studies as well, since the paper by Suneera and others, as well as that of Poornima and others presented logistics regression as a high-accuracy algorithm when compared with the performance of algorithms such as support vector machines, decision tree, and Naive Bayes [4 & 5]. While the greatest accuracy and precision value obtained through the unigrams model for this study through logistics regression is still a low number when attempting to analyze an algorithm, it can potentially be enhanced through certain features. Identification of a central algorithm, as well as additional features, that can provide a relatively high accuracy serves as the basis for further research. Future implementations such as multiple datasets from different sources, additional preprocessing, feature scaling, and k-fold cross validation can be integrated to attain higher performance rates and perform efficient sentiment analysis.

REFERENCES

[1] Dang NC, Moreno-García MN, De la Prieta F. "Sentiment Analysis Based on Deep Learning: A Comparative Study." Electronics. 2020. 9(3):483.
[2] K. Reyes, "What is deep learning and how does it work [updated]," Simplilearn.com, 08-Dec-2022. [Online]. Available: htt ps://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning. [Accessed: 10-Dec-2022].
[3]Adam, Rizqy. "Tugas Sentiment analysis_rizqy Adam." Kaggle, Kaggle, 4 Dec. 2022, https://www.kaggle.com/code /juko911/tugas-sentiment-analysis-rizqy-adam/comments.
[4] Suneera C, Prakash J. "Performance Analysis of Machine Learning and Deep Learning Models for Text Classification." India Council International Conference. 2020. 1-6.
[5] Poornima A, Priya K. "A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques." 6th International Conference on Advanced Computing and Communication Systems. 2020. 493-496.