Sage Plagarism Report

Powered by schoolworkspro.com

Hari Sharan Shrestha
150212

28 Results Found

18.67 % Match Percentage

Submitted to: Softwarica	**1.81%**
Submitted to: Softwarica	**1.56%**
Submitted to: Softwarica	**1.39%**
Source: What is Artificial Intelligence (AI) for Documents? 14 Truths Link: https://blog.bisok.com/general-technology/what-is-artificial-intelligence-in-computer	**1.08%**
Submitted to: Softwarica	**0.98%**
Submitted to: Softwarica	**0.87%**
Submitted to: Softwarica	**0.8%**
Source: (PDF) Sentiment Analysis in Twitter | PRASAD NAR - Academia.edu Link: https://www.academia.edu/75847207/Sentiment_Analysis_in_Twitter	**0.76%**
Submitted to: Sunway	**0.73%**
Source: Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique Link: https://doi.org/10.1155/2022/2500772	**0.73%**
Submitted to: Softwarica	**0.72%**
Source: Sentiment Analysis | Comprehensive Beginners Guide | Thematic | Thematic Link: https://getthematic.com/sentiment-analysis/	**0.69%**
Submitted to: Softwarica	**0.69%**

Submitted to: Softwarica	**0.69%**
Submitted to: Softwarica	**0.69%**
Source: XGBoost Algorithm: Long May She Reign! | by Vishal Morde | Towards Data Science Link: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d	**0.62%**
Source: Twitter Sentiment Analysis by Rajani Shree Manjappa, Aditya Kumar :: SSRN Link: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3633673	**0.59%**
Source: Novel email spam detection method using sentiment analysis and personality recognition | Logic Journal of the IGPL | Oxford Academic Link: https://doi.org/10.1093/jigpal/jzz073	**0.56%**
Submitted to: Softwarica	**0.42%**
Source: Imbalanced Classification Problems in R Link: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/	**0.35%**
Source: Social Media Sentiment Analysis Using Twitter Datasets - DataScienceCentral.com Link: https://www.datasciencecentral.com/social-media-sentiment-analysis-using-twitter-datasets/	**0.28%**
Submitted to: Softwarica	**0.28%**
Source: Lexicon-Based Sentiment Analysis: A Tutorial | KNIME Link: https://www.knime.com/blog/lexicon-based-sentiment-analysis	**0.28%**
Source: ERD With Attributes Data Types Link: https://en.wikipedia.org/wiki/Hui_people	**0.24%**
Source: Cloud-based email phishing attack using machine and deep learning algorithm | SpringerLink Link: https://link.springer.com/article/10.1007/s40747-022-00760-3?	**0.24%**
Submitted to: Softwarica	**0.24%**
Submitted to: Softwarica	**0.21%**

REFLECTIVE REPORT
Sentiment Analysis using
Logistic regression, Naïve Bayes
ST7072CEM – Machine Learning
Name: Hari sharan shrestha College ID: 150212 GitHub Repository Link:
https://github.com/softwarica-github/Hari_sharan_shrestha_150212_stw7072cem.git
Contents

Sentiment analysis is a field of natural language processing that aims to identify and extract subjective information, such as opinions and emotions, from text data.

It is a challenging problem due to the complexity and variability of human language.

Machine learning methods have shown great potential in solving this problem, as they can learn patterns and relationships from data and use them to make predictions on new, unseen data.

In this report, we will discuss three machine learning algorithms that we have used for sentiment analysis of the IMDB movie review dataset: Logistic Regression, Linear SVM, and Multinomial Naive Bayes.

In today's world social media is everywhere, people can express their opinions, feedback and attitude towards various things on the social media platforms and on the internet.

Many people refer to those opinions to make their decisions.

Many companies and organizations use those reviews to improve their services.

This is where sentiment analysis comes into play.

Sentiment analysis is a process of analyzing a text or a document to extract the polarity of the text i.e.

negative or positive.

As mentioned before, a lot of reviews and opinions are available on various social media platforms, ecommerce sites, and microblogging platforms, with this various companies, politicians and brands use sentiment analysis to improve their marketing strategies or provide better customer services.

Sentiment analysis can be used to extract people's opinions or sentiment about a single or multiple domain.

Another way for defining sentiment analysis, is the implementation of various tools and techniques that can be used to extract subjective information from a text, particular language, sentence or a document.

Sentiment analysis is an application of Natural language processing (NLP) and is one of the fastest growing research topics in that area.

According to the paper Sentiment analysis: Machine Learning Approach sentiment analysis also known as opinion mining can be used for getting opinions about any events or products and it helps people to make decisions about a specific service or item by looking at how much it is preferred or how good or bad it is.

The diagram below shows the processes involved in sentiment analysis:

Sentiment analysis can be performed using techniques like machine learning, neural networks and lexicon-based approaches.

Researchers have performed sentiment analysis using those techniques.

Problem and Data sets

According to estimates, 80% of the world's data is unstructured and not arranged in a certain way.

As with reviews, emails, chats, social media, surveys, and articles, the majority of this information is text-based.

These works typically require extensive research and a lot of time to comprehend.

By automating business processes, cutting down on hours of human processing, and obtaining actionable insights, the sentiment analysis system enables the corporation to make sense of this enormous amount of unstructured text.

Sentiment analysis on the IMDB dataset is important for a number of reasons.

First and foremost, it allows us to gain a better understanding of how audiences feel about particular movies or genres.

This can be valuable information for filmmakers and movie studios, as it can help them to tailor their marketing and distribution strategies to better appeal to their target audiences.

For example, if sentiment analysis shows that audiences have a strong negative reaction to a particular movie, studios can adjust their promotional efforts or distribution plans in order to minimize losses.

Sentiment analysis on the IMDB dataset can also be useful for consumers who are looking information about movies before they make a decision to watch them.

By analyzing the sentiment of reviews, consumers can get a better sense of what to expect from a particular movie, which can help them make more informed decisions about whether or not to see it.

This can be particularly useful for people who are looking to save time and money by avoiding movies that are unlikely to meet their expectations.

For this task I have collected 50k review data set from kaggle with positive and negative sentiment.

There are altogether 25k positive sentiment and 25k negative sentiment.

And separated nearly 40k for training dataset and 10k for test dataset.

Now need to explore sentiment analysis on the IMDB movie review dataset using machine learning algorithms such as Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes

Literature Review:

AUTHOR TOPIC MEASUREMENT DESCRIPTION

BoPang, Lillian Lie A Sentimental Education: Sentiment Analysis using subjectivity summarization n based on Minimum Cut.

Accuracy (Performance)

This paper proposed that SVM and NB are better technique for improving the performance of a model up to 86.4 %.

Erik.Boiy, Pieter Automatic Sentiment Accuracy (speed This paper shows

Hens, Koen Dschacht, Marie Francine Moens

Analysis in Online Text

and size) the varying level of accuracy when symbolic and machine learning methods were applied to different social network dataset.

Doreen Hii Using Meaning specificity to aid negation handling in sentiment analysis

Accuracy (Performance)

This paper compared the accuracy of 1-,2-,3-,4-, gram and suggested 4- gram is the best performing window size in Lexicon method

Dr. Sefer Kurnaz, and Mustafa Ahmed Mahmood.

Sentiment Analysis in data of Twitter using Machine learning algorithm

Accuracy (Time) This paper proposes a new technique which offers an accuracy of 98 % when compared with Deep learning method, SVM and Maximum Entropy method.

Xiaomie Zou, Jing Yang, and Jianpei Zhang.

Microblog sentiment analysis using social and topic context.

Accuracy (Performance)

This paper proposes a new method to identify the polarity of the sentiment and shows the structure similarity has a better accuracy than user direct relations.

Logistic Regression

Logistic Regression is a classification algorithm that models the probability of the binary target variable (positive or negative sentiment) as a function of the predictor variables (words or features) using a logistic function.

In other words, it tries to find the best decision boundary that separates the positive and negative examples in the feature space.

The logistic function maps the input to a probability value between 0 and 1, which can be interpreted as the confidence level of the prediction.

The algorithm optimizes the parameters of the logistic function using gradient descent or other optimization techniques.

To implement Logistic Regression for sentiment analysis, we first need to preprocess the text data by removing stop words, stemming or lemmatizing the words, and converting them into numerical features using methods such as bag-of-words or TF-IDF.

Then, we can split the data into training and testing sets, and fit a Logistic Regression model on the training data.

We can tune the hyper parameters such as the regularization strength and the learning rate using cross-validation on the training set.

Finally, we can evaluate the performance of the model on the testing set using metrics such as accuracy, precision, recall, and F1 score.

Figure 1: Logistic regression

Support Vector Machine

Linear SVM (Support Vector Machine) is another classification algorithm that tries to find the best hyper plane that separates the positive and negative examples in the feature space with the maximum margin.

The margin is defined as the distance between the hyper plane and the closest data points from each class.

The algorithm tries to maximize the margin while minimizing the classification error.

Linear SVM can also handle non-linearly separable data by using kernel functions that map the input to a higher-dimensional space where the data becomes separable.

To implement Linear SVM for sentiment analysis, we can follow a similar preprocessing and feature extraction procedure as Logistic Regression.

Then, we can split the data into training and testing sets, and fit a Linear SVM model on the training data.

We can tune the hyper parameters such as the regularization strength and the kernel function using cross-validation on the training set.

Finally, we can evaluate the performance of the model on the testing set using the same metrics as Logistic Regression.

Figure 2: Linear SVM

Naive Bayes

Multinomial Naive Bayes is a probabilistic classification algorithm that models the conditional probability of the target variable given the predictor variables using Bayes' theorem and the assumption of independence between the predictor variables.

In other words, it tries to find the class that has the highest posterior probability given the observed evidence.

Multinomial Naive Bayes is commonly used for text classification tasks such as sentiment analysis and spam filtering, as it can handle high-dimensional and sparse feature spaces efficiently.

To implement Multinomial Naive Bayes for sentiment analysis, we can again follow the same preprocessing and feature extraction procedure as the previous algorithms.

Then, we can split the data into training and testing sets, and fit a Multinomial Naive Bayes model on the training data.

We can tune the hyper parameters such as the smoothing parameter using cross-validation on the training set.

Finally, we can evaluate the performance of the model on the testing set using the same metrics as before.

Figure 3: Naive Bayes

Experimental Setup

In this report, we have applied machine learning methods to perform sentiment analysis on the IMDB movie review dataset.

Specifically, we have used three algorithms: Logistic Regression, Linear SVM, and Multinomial Naive Bayes.

Pre-processing

Before applying these algorithms, we have preprocessed the text data by removing stop words, HTML tags and noise, special characters, and applying stemming.

The stopwords and special characters were removed by using NLTK library.

For the stemming of the tokens Porter stemming was implemented.

Porter stemming is a stemming algorithm developed by Martin Porter in 1980.
Stemming is the process of reducing words to their root or base form, called the stem, by removing suffixes and prefixes.
The goal of stemming is to reduce the dimensionality of the feature space and group similar words together, which can improve the performance of natural language processing tasks like text classification, information retrieval, and sentiment analysis.
The Porter stemming algorithm consists of a set of rules for removing common suffixes from English words.
These suffixes include -s, -es, -ed, -ing, -ly, and -ment, among others.
The algorithm applies these rules in a specific order, starting with the longest suffix and proceeding to the shortest.
For example, the word "jumping" would be reduced to "jump" by removing the -ing suffix according to the algorithm's rules.
Alos, the HTML tags were removed using Beautiful soup library and other noise were removed using
This step is essential to reduce the dimensionality of the feature space and remove irrelevant information that could negatively affect the performance of the models.
Figure 1: Pre processing
Feature selection
Then, we have used two methods for word embedding: Bag of Words and TF-IDF.
Bag of Words represents the text as a vector of word counts, where each word corresponds to a feature.
The frequency of each word in the document is used as the value of the corresponding feature.
This method does not consider the semantic relationships between words, but it is simple and efficient for large datasets.
On the other hand, TF-IDF stands for Term Frequency-Inverse Document Frequency, which is a statistical measure that reflects the importance of a word in a document relative to the entire corpus.
It assigns a weight to each word based on its frequency in the document and the inverse frequency of its occurrence in the corpus.
This method gives more weight to rare words that are more informative for the classification task.
The CountVectorizer function from the sklearn library was used for the Bag of Words (BOW) approach.
The CountVectorizer was configured with a minimum document frequency of 0, a maximum document frequency of 1, and an n-gram range of 1-3.
The minimum document frequency of 0 means that any terms with a frequency lower than 0 will be ignored during vocabulary creation.
Similarly, the maximum document frequency of 1 means that any terms with a frequency greater than 1 will be ignored.
The n-gram range of 1-3 indicates that the CountVectorizer will accept unigram, bigram, and trigram tokens from the pre-processed data.
The same parameters were used for the TFIDF approach.
Figure 2: feature selection
After feature engineering, the data was split into training and test datasets.
Several machine learning algorithms were implemented to compare their performance.

The first algorithm used was Logistic Regression, which employed L2 regularization by default to prevent overfitting of the model.
The hyper parameter maxiter was set to 500, which specifies the maximum number of iterations for the algorithm to converge to obtain the accuracy.
A random state value of 42 was also used to ensure the consistency of the algorithm's output across different runs and machines.
The default parameters were used for Linear Support Vector Machine and Multinomial Naive Bayes algorithms, similar to Logistic Regression.
The accuracy of each algorithm was calculated and evaluated using metrics such as confusion matrix, precision, recall, f1-score, and support.
Logistic regression
Figure 1: Accuracy metrics of Logistic Regression
The upper metrics is obtained by using feature selection of BOW and the metrics below is obtained using TFIDF.
The weighted accuracy for BOW and TFIDF is 74%.
And the overall metrics such as precision, recall, f1-score and support are aslo 74%.
Figure 2: Confusion matrix of logistic regression
The above confusion metrics is obtained using BOW and TFIDF with Logistic Regression.
The true positive values are 4027 and 3998, the true negative value predicted are 3996 and 4029.
While False Positive and False Negative values are (1364 and 1393) and (1417 and 1384) respectively.
Linear SVM (Support Vector Machine)
Figure 3: Accuracy metrics of linear SVM
The upper metrics is obtained by using feature selection of BOW and the metrics below is obtained using TFIDF.
The weighted accuracy for BOW is 75% and TFIDF is 25%.
And the overall metrics such as precision, recall, f1-score and support are not same as compare to logistic regression method.
Figure 4: Confusion matrix of linear SVM
The above confusion metrics is obtained using BOW and TFIDF with linear SVM.
The true positive values are 5390 and 5391, the true negative value predicted are 126 and 0.
While False Positive and False Negative values are (1 and 0) and (5287 and 5413) respectively.
Multinomial Naive Bayes
Figure 5: Accuracy metrics of Naïve Bayes
The upper metrics is obtained by using feature selection of BOW and the metrics below is obtained using TFIDF.
The weighted accuracy for BOW and TFIDF is 74%.
And the overall metrics such as precision, recall, f1-score and support are aslo 74%.
Figure 6: Confusion matrix of Naïve Bayes
The above confusion metrics is obtained using BOW and TFIDF with lnaive bayes.
The true positive values are 4004 and 4020, the true negative value predicted are 4015 and 4002.
While False Positive and False Negative values are (1387 and 1371) and (1398 and 1411) respectively.
Comparison Table of Accuracy of model
Methods Bow Score TFIDF Score

Logistic regression 0.7512 0.75
SVM 0.5829 0.5112
Naïve Bayes 0.751 0.7509

Discussion and Conclusion

In conclusion, this report aimed to explore sentiment analysis on the IMDB movie review dataset using machine learning algorithms such as Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes.

The dataset was preprocessed using various techniques such as stop word removal, HTML stripping, noise removal, special character removal, and stemming.

Two methods of word embedding, Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TFIDF), were employed to convert the textual data into a numerical representation.

After splitting the dataset into training and test data, three machine learning algorithms were implemented, and their performances were compared.

Among these algorithms, Logistic Regression and Multinomial Naive Bayes achieved the highest accuracy of 75%, outperforming Linear Support Vector Machine.

The Logistic Regression algorithm used L2 regularization by default, which helped to prevent over fitting of the model.

In contrast, the Multinomial Naive Bayes algorithm was based on the Bayes theorem and was relatively simple and computationally efficient.

While the results obtained from the Logistic Regression and Multinomial Naive Bayes algorithms were promising, it is essential to note that these algorithms are not perfect and may not generalize well to other datasets.

Additionally, there are several factors that could affect the accuracy of these algorithms, such as the quality and size of the dataset, the pre-processing techniques employed, and the hyper-parameters used in the algorithms.

Despite these limitations, sentiment analysis using machine learning algorithms can be a useful tool in various applications such as customer feedback analysis, social media monitoring, and product reviews.

With the increasing availability of large datasets and improvements in machine learning algorithms, sentiment analysis is likely to become even more prevalent in the future.

mastery, M., 2020.
Naive Bayes for Machine Learning.
[Online] Available at: https://machinelearningmastery.com/naive-bayes-for-machine-learning/ [Accessed 3 5 2023].

Narkhede, S., 2018.
Understanding Confusion Matrix.
[Online] Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 [Accessed 1 4 2023].

SAI, 2017.
Sentiment Analysis Using Deep Learning Techniques: A Review.
[Online] Available at: https://thesai.org/Publications/ViewPaper?Volume=8&Issue=6&Code=ijacsa&SerialNo=57 [Accessed 1 5 2023].

statology, 2020.
Introduction to Logistic Regression.
[Online] Available at: https://www.statology.org/logistic-regression/ [Accessed 2 5 2023].

VIDYA, A., 2021.
Support Vector Machine(SVM): A Complete guide for beginners.

[Online] Available at: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a- complete-guide-for-beginners/ [Accessed 1 5 2023].