Schmaltz Surveyor

Sentimental Analysis of Twitter

Minor Project - IS6C06
Under the Guidance of Mrs. Nandini BM
By
Nithyashree Arunachalam
Pradyoth P
Tejasvini SJ
Shashank BU

Introduction

Social media has gained immense popularity and has become a major global platform to stay connected as well as express opinions.
A huge amount of content is created on various topics and comments are posted on these platforms daily.
The feedback received on a certain piece of content can be either negative or positive.
Other than this, receiving negative feedback, on various occasions might affect the mental health of the content creators and, in some cases, might also lead to cyberbullying.
Social media is a necessity in today's time to stay connected, informed, and relevant and therefore such issues must be tackled.
Schmaltz refers to excessive sentimentality. Therefore this project has been named Schmaltz Surveyor, which aims to analyze the sentiments expressed by Twitter users.

- Sentimental Analysis reads people's sentiments or emotions towards particular things or topics.
- Sentiment analysis is a machine learning tool that will help analyze and categorize the texts as positive or negative.

Literature Survey

- Xing Fang, and Justin Zhan[1], worked on Sentiment Analysis on Amazon Online Products Review by collecting data from amazon.com. They figured out the issues of categorical sentiment polarity using Machine Learning Algorithms. They used many libraries that hold Naive Bayesian, SVM, and Random Forest.
- Faizan[2] built a model for the analysis of feeling using the KNN algorithm with unigram, bigram, and n-gram features. They then performed training and testing of their model on the US Airlines data set, for which they attained an accuracy of 65.33 %.
- Chirag Kariya and Priti Khodke's[3] paper explains various steps involved in the analysis of Twitter sentiments along with the various tools that are used to perform Twitter sentiment analysis. Amongst the various algorithms available, the KNN algorithm is used to increase the efficiency of sentiment analysis whereas Naive Bayes for simple and efficient sentiment analysis by classifying the tweets as either positive, negative, or zero.
- Akshay Amolik et al.[4] proposed sentiment analysis, and they accurately classified tweets by using Feature Vector and classifiers like Naïve-Bayesian and SVM. Exception of lower recall and accuracy, Naïve Bayesian has better precision as a comparison to SVM. SVM gives better results when it comes to accuracy. With the increase of training data, the accuracy of classification will also increase.

Proposed System

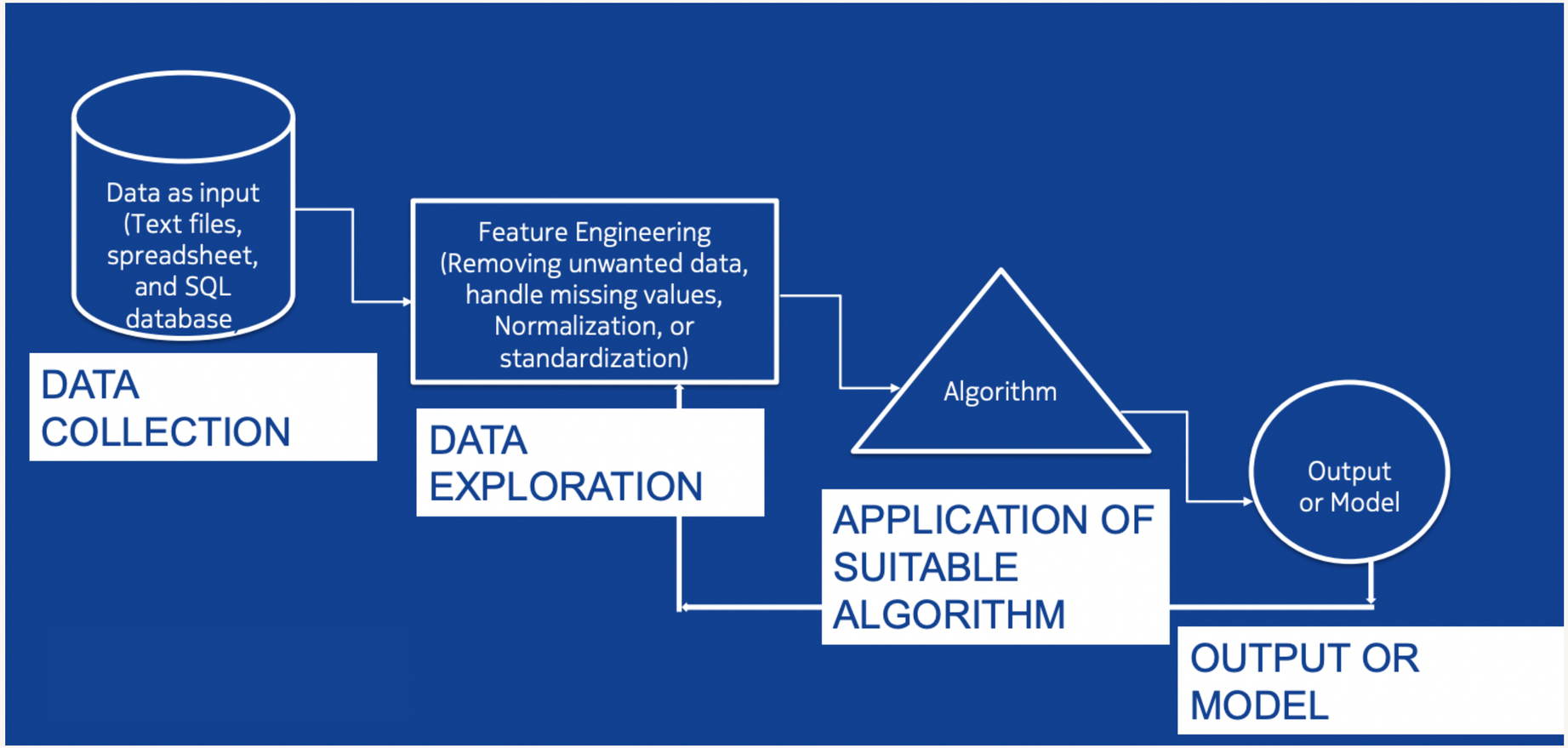
The objective of this whole project is to classify the social media content into positive or negative using machine learning algorithms.
It will help us analyze and see the extent of the positive or negative extent of these comments and content on social media.

A dataset taken consists of comments and texts extracted from various social media platforms.
The dataset is then converted into the numerical form using vectorization methods in NLP.
These numbers are used to train ML models to make predictions.
The ML models or classifiers as the name suggests classifying the text into positive and negative. These classifiers are nothing but machine learning algorithms that automatically order or categorize data.

Scikit learn library which is a python library will be used extensively throughout the whole project. Scikit-learn is one of the most useful libraries for machine learning in Python.
It contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction

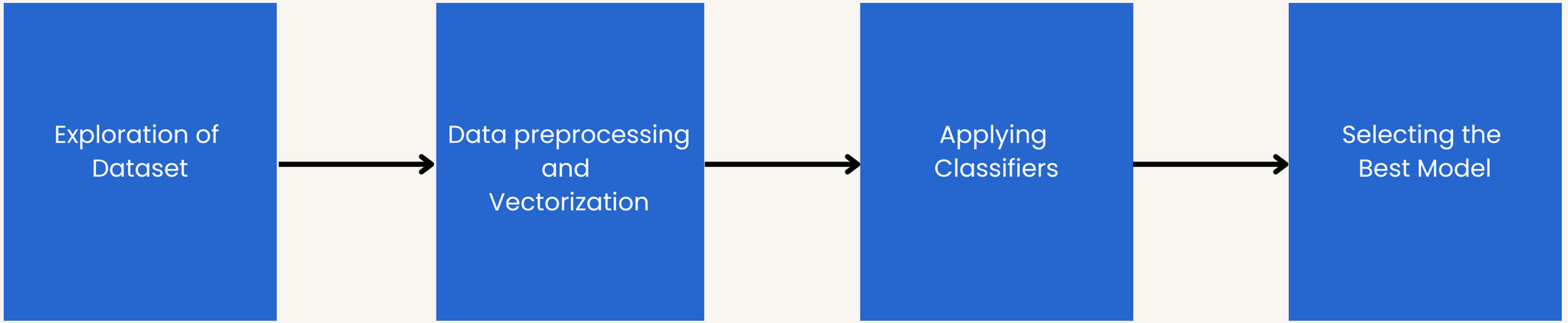
Flow of the Project

Based on the typical practises, the project will be covered in 4 phases.



Project Design

The project is partitioned into 4 Phases.



Exploration of Dataset

- Creating a database that has a list tweets
- This database can be divided into training and testing data.

Data preprocessing and Vectorization

Step 1: Checking for missing values.

Step 2: Text Normalization.
Normalize the text data as texts from such online platforms usually contain inconsistent language and the use of special characters in place of letters. To tackle such inconsistencies in data, Regex.

Step 3: Lemmatization
Lemmatization is the process of grouping together the different forms of a word so they can be analyzed as a single item. For example, we do not want the Machine Learning algorithm to treat eating, eats, and eat as three separate words because they convey the same message. Lemmatization helps reduce the words to their root form.

Data preprocessing and Vectorization

Step 4: Removal of stop words

Stop words are a set of commonly used words in the language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.

Step 5: Converting to Numerical Form

The dataset needs to be converted into the numeric form so that it can be put through classifiers.

- Count Vectorization: Count Vectors can be helpful in understanding the type of text by the frequency of words in it.
- TF IDF: TF-IDF means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word but it also provides a numerical representation of how important a word is for statistical analysis.

Choose the one with the more optimum result.

Applying Classifiers

- Logistic Regression
- Linear Support Vector Classifier or Support Vector Machine
- Random Forest
- Naive Bayes

Selecting the Best Model

- Drawing comparison among the 4 to choose the model to be used

Conclusion

The result of the whole project is to

- Classifying the tweet into positive and negative would help us understand and analyze the extent of positivity and negativity on Twitter.
- Sentiment Analysis is a great way to analyze the response to a particular tweet.
- The next step is to deploy the model as a backend to a web application that determines the toxicity of a comment which is provided as an input by the user.

References

- [1]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1.
- [2] Faizan. "Twitter Sentiment Analysis" International Journal of Innovative Science and Research Technology (2019)
- [3] Chirag Kariya and Priti Khodke. "Twitter Sentiment Analysis" 2020 International Conference for Emerging Technology (INCET) Belgaum.
- [4]Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6.