VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: M. Tech integrated CSE Specialization with Business Analytics

Course: Natural language processing

Slot: A2

Faculty: Premalatha

Title: Twitter Sentiment Analysis

Team Members:

Roshan Kumar S -20MIA1156
Linga Harish kumar-20MIA1127
Harsha Vardhan Reddy -20MIA1083

ABSTRACT

Twitter is a micro-blogging website that allows people to share and express their views about topics, or post messages.

There has been a lot of work in the Sentiment Analysis of twitter data. This project involves classification of tweets into three main sentiments: positive, neutral and negative. In this project, the use of features such as removing stop words, regular expression, tokenization, and stemming is observed. We are using ml classification algorithm like Logistic Regression, Support Vector Machines(SVM) and Decision Tree.

KEYWORDS

NumPy, Pandas, Sklearn, Seaborn, nltk, ML, spacy

INTRODUCTION

Sentiment analysis is the task of finding the opinions and affinity of people towards specific topics of interest. Be it a product or a movie, opinions of people matter, and it affects the decision-making process of people. The first thing a person does when he or she wants to buy a product online, is to see the kind of reviews and opinions that people have written. Social media such as Facebook, blogs, twitter have become a place where people post their opinions on certain topics. The sentiment of the tweets of a particular subject has multiple usage, including stock market analysis of a company, movie reviews, in psychology to analyse the mood of people that has a variety of applications, and so on. Sentiments of tweets can be categorized into many categories like positive, negative, and neutral. The two types of sentiments considered in this classification experiment are positive and negative sentiments. The data, being labelled by humans, has a lot of noise, and it's hard to achieve good accuracy.

Currently, the best results are obtained by Decision Tree with an accuracy of 90%. The other algorithms used in this project are SVM and Logistic Regression and we would be comparing these in the upcoming sections.

Literature Review

The proposed method by Nigam et al. (2000) provides a novel approach to text classification that combines the EM algorithm and the Naive Bayes classifier. The method is able to effectively leverage both labelled and unlabelled data to improve the accuracy of text classification. The experimental results showed that the method outperforms other popular text classification methods, particularly when the labelled data is limited. This study provides a valuable contribution to the field of text classification and highlights the importance of using both labelled and unlabelled data for improved performance.

In conclusion, the paper by Otter et al. (2020) provides a comprehensive overview of the use of deep learning in NLP. The authors have discussed the various applications and models used in NLP and have highlighted the recent advancements and challenges in this field. This paper provides valuable insights into the current state-of-the-art techniques used in NLP and deep learning and serves as a useful reference for researchers in this field.

In conclusion, the paper by Young et al. (2018) provides a comprehensive overview of recent trends in deep learning-based NLP. The authors have discussed the various applications and models used in NLP and have highlighted the recent advancements and challenges in this field. This paper serves as a useful reference for researchers in this field and provides valuable insights into the current state-of-the-art techniques used in NLP and deep learning.

Data Set Description

We have 11,020 rows and 16 columns.

We did some pre-processing like

Converting tweets to lower case, removed hashtags / URL / @ Using re (regular expression)

Dataset url: https://www.kaggle.com/datasets/gpreda/pfizer-vaccine-tweets

Implementation

There 3 sentiments namely 'Positive', 'Negative' and 'Neutral' in the sentiment column. We are going to follow the steps written below

1. Data Pre-processing. (in data pre-processing we have did regular expression , tokenization)

2. Stemming

3. Polarity

4. Model Building

5. Model Accuracy

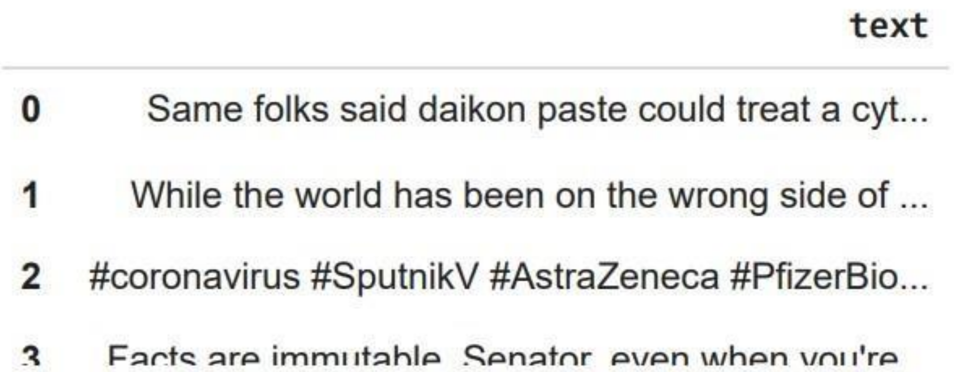
We have used the **nltk / spacy / re** for processing the text**.** So my first step was to import libraries like nltk , spacy , re , seaborn / matplotlib , sklearn , pandas.

1)Data Pre-processing: Machine Learning models cannot cope with messy data. Therefore, data pre-processing is an extremely important step as it affects the ability of our model to learn.

The training and the testing dataset consist of the text(tweets) & polarity columns. I have dropped the unnecessary columns that are not useful for what we are trying to predict. There is not much difference between 'Extreme Positive' and 'Positive' and 'Extremely Negative' and 'Negative', therefore I have replaced extremely positive with positive sentiment and extremely negative as negative. This will also help in fast processing. Data pre-processing includes dropping null values, removing punctuations, URLs, hashtags, stop words, and converting all words in the tweets to lowercase. I have also used stemming in the pre-processing the text process.

STEMMING:

Stemming is the process of reducing the word to the root word by removing its suffix. I have used Porter Stemmer for this code.



	text
0	Same folks said daikon paste could treat a cyt...
1	While the world has been on the wrong side of ...
2	#coronavirus #SputnikV #AstraZeneca #PfizerBio...
3	Facts are immutable, Senator, even when you're

Machine Learning:

- **SVM**
- **Logistic regression**
- **Decision tree**

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Logistic Regression

Logistic Regression is a “Supervised machine learning” algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature.

That means Logistic regression is usually used for Binary classification problems.

Binary Classification refers to predicting the output variable that is discrete in **two** classes.

A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Noncancerous, etc.

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Results And Discussion

The highest accuracy among these models is achieved using Decision tree with 90% where Logistic Regression with 84% and SVM with 76%.

Conclution

So, we were able to did the sentiment analysis on the tweets like classifying them whether they are positive / neutral / negative tweets.

Future Work

So, in future if u want to read a person’s sentiment, u can use this model

Which we have designed.

Reference

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Recent+trends+in+deep+learning+based+natural+language+processing&btnG=

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Otter%2C+D.+W.%2C+Medina%2C+J.+R.%2C+%26+Kalita%2C+J.+K.+%282020%29.+A+survey+of+the+usages+of+deep+learning+for+natural+language+processing.+IEEE+transactions+on+neural+networks+and+learning+systems%2C+32%282%29%2C+604-624.&btnG=

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Nigam%2C+K.%2C+McCallum%2C+A.+K.%2C+Thrun%2C+S.%2C+%26+Mitchell%2C+T.+%282000%29.+Text+classification+from+labeled+and+unlabeled+documents+using+EM.+Machine+learning%2C+39%2C+103-134.&btnG=