**NATURAL LANGUAGE PROCESSING PROJECT PROPOSAL**

**TITLE: SENTIMENT ANALYSIS ON TWITTER MESSAGES & STOCK MARKET ANALYSIS**

**GitHub Link:** <https://github.com/NitheeshSama/NLP_project.git>

**Team Members:**

**Nitheesh Sama**

**Lokesh Bachu**

**Rajasri Lekkala**

**Sai Krishna Bussa**

**MOTIVATION:**

Twitter has been a stage where people share their views and opinion on a particular product, thing, movie or an issue. The amount of data that is produced from this is so large and an organization that wants to know the public opinion on their products can make use of this data and analyze it to improve their products or services. Sentiment analysis technique can be applied to get the required analysis from twitter dataset. As part of our project, we are using a twitter dataset and perform sentiment analysis to gain insights into the data. This technique can be further modified according to different data to gain insights into them.

**OBJECTIVE:**

In this project we will classify twitter messages into Positive, Negative, Neutral and Irrelevant based on the content of the messages. We will be using python programming with packages like panadas, nltk, ski-kit learn, etc accordingly to perform sentiment analysis on the dataset. After performing the sentiment analysis, we would apply Machine Learning classifiers, evaluate the model and find the accuracy of the model. At the end all the messages will be classified as positive, negative, neutral or irrelevant and the accuracy of the classification can be determined.

After classifying based on sentiment, we use the dataset to analyze the stocks of the companies present in the datasets and provide detailed analysis about the performance of stock in accordance with the sentiment.

**SIGNIFICANCE:**

This type of projects plays very important role in fields of Market Analysis, Brand Monitoring, Business Intelligence Build up and Customer Service. Sentiment analysis is very important to understand and detect customer feelings. It is important to accomplish the project purpose using sentiment analysis because to understand real user opinions, complaints and suggestions, we must again filter the unrelated Tweets (Spam, junk, marketing, news and random) and this method is very helpful in doing so. These types of analysis are also used for strategic purpose by organizations to know their competition.

**FEATURES:**

Some of the important features of our project are

1. **Data Collection**: In this project we are using data collected from twitter and saved in a csv file as a data source. The dataset is obtained from Kaggle.com and contains data collected from twitter on games like **Borderlands**, **Dota-2** and **World of Craft** etc.
2. **Data Processing:** In this project, the data processing is done at different levels like removal of punctuation, removal of stop words, conversion to lower case and Lemmatization etc. These steps are helpful in making the analysis part easy.
3. **Data Analysis:** As part of data analysis, we will be classifying the messages into one of the 4 categories using sentiment analysis.
4. **Data Visualization:**  After the analysis part, we will be visualizing the number of messages in each category, most frequent words in that category and so on (As this is the early stage of the project, we did not come a conclusion on which visualization to use)

After the sentiment analysis part is done, we are planning to apply decision tree classifier on the and then evaluate the resulting model to find the accuracy, important features etc.

**WORKFLOW OR STEPS:**

The steps that will be performed as per the project are

1. **Data Preparation**
2. **Removing punctuation and replacing short words.**
3. **Lower Case Conversion.**
4. **Stop words Removal.**
5. **Lemmatization.**
6. **Analysis**
7. **Machine Learning Model fitting**
8. **Evaluation of model**

**REFERENCES:**

**Data Set reference:** [**https://www.kaggle.com/**](https://www.kaggle.com/)

**DATASET DETAILS:**

**Name: twitter\_training.csv**

**Number of Features: 4**

**Number of rows: 74002**

**Data Source: Twitter**

**Data Set Source: Kaggle.com**

**Project Increment 1**

**Background:**

The reason behind picking the dataset from twitter data is to verify the stock trends of the domains and their product reviews.

**Features design:**

The dataset we have contains three features they are the company’s product name , tweet related to the product and the status of the product.

Here the main feature is the tweet and the status of the product.

**Analysis:**

In this project we are analyzing the data using natural language tool kit and the machine learning algorithms to find the market trends based on the tweet’s status such as positive and negative which reflects the stock market.

Here the data can be on any domain, using this process or project implementation we can apply this method on any type of data for sentiment analysis.

**Implementation:**

1.Reading the data from the dataset.

A screenshot of a computer

Description automatically generated

2.creating data frame from the data object.

3.performing tweet pre-processing by removing stop words , punctuation etc.

4. performing Lemmatization on the dataset .

A screenshot of a computer

Description automatically generated

5.Plotting dataset by visualizing them for top 20 tweets.

A screenshot of a computer

Description automatically generated

6.Applying vectorization the final dataset for perfect data splitting.

7.Splitting data into four datasets such as test and train for both model fitting and model accuracy.

8.Creating Decision tree classifier and performing model fitting.

9.The final step is to do accuracy test of the model.

**Preliminary results:**

A screenshot of a computer

Description automatically generated

**Project management:**

Work completed:

Responsibilities took by Rajasri are Data reading and transformation (Pre-processing)

Lemmatizing and advance text processing is done by Niteesh.

Sai Krishna worked on plotting the data and finding the dependencies.

Lokesh worked on model creation , vectorization of new data for model fitting.

Work to be completed

* Rajasri and Sai Krishna are planning to work on extraction of stock related details of the companies present in the dataset.
* Nitheesh is planning to work on creation of a new dataset for stock market analysis and analysis of the results obtained along with supporting Lokesh in code development.
* Developing the required code for data analysis.

Overall, we are also working on enhancements on the project and using the sentiment of the tweets to analyze stock market trends and various aspects of the stock.

**References:**

Data Set reference**:** [**https://www.kaggle.com/**](https://www.kaggle.com/)

**Project Increment 2**

**Introduction:**

In this increment we focused on analyzing stocks data of a particular company and their behaviors at various stages from the dataset that is created according by fetching information from different sources. Using this dataset we are analyzing the changes in the stock in accordance with the sentiment.

**Feature Design:**

We are having two datasets one is related to the twitter data sentiments and the other about the stock prices,companies that published the games, published year, count of positive and negative tweets and positivity percentage.

In the first data we are analyzing the tweets which has an effect on the stock trends.

The second dataset is all about the stock trends of the companies or brands which we are checking from the first data.

Below is the image explaining the features of the dataset.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Analysis:**

As part of Increment 1, we have performed sentiment analysis on the twitter dataset which has tweets regarding games and some organizations. We also trained the Decision tree classifier model using the dataset and used the test dataset to predict the accuracy of the model. We also determined the important features that are most helpful in finding the sentiment behind the tweet.

In Increment-2, we will be performing stock market analysis of the companies that have been developed the games in the dataset and how the sentiment on the game is affecting the stock performance.

At first, we have created an excel that has the following columns. Game, Company Name, published year, Stock price before game release, Stock price after game release, Number of Positive tweets, Number of negative tweets, Positivity Percent.

Chart, bar chart

Description automatically generated

After that we have plotted a side-by-side bar graph for each game in the dataset showing number of positive and negative tweets for the game. This graph helps us determine the sentiment towards the game on twitter.

Chart, bar chart, histogram

Description automatically generated

Then we plot a side-by-side graph for stock price before, stock price after and positivity percentage. This graph is helpful in determining how the stock performed in accordance with the sentiment on the game.

From the dataset we have loaded, we can see that the stock prices have decreased for those organizations which recorded a low positivity percent after the release of a game from them and some organizations had increase in stock prices after their game is release because the game has a positive sentiment on twitter and positivity percent more than 50.

For further analysis of the stock of an organization, we have taken **Electronic Arts(EA)** company that has built 4 games in our dataset and analyzed the stock movement from 2008 to 2020.

Graphical user interface

Description automatically generated with medium confidence

To perform this, I have imported yfinance package that has the stock market information from a very long time.

I have also highlighted the stock movement 2 months before or after a particular game was released by Electronic Arts. Battlefield game released by EA has a negative sentiment on twitter that can be clearly supported by the downward movement of the stock value of the organization. Similarly, there is positive sentiment towards the other 3 games published by EA. So, there is an increase in the stock value of the company after their release.

Chart, line chart, scatter chart

Description automatically generated

Similarly, from the stock volumes traded graph, we can see that the investors try to hold the stock of this company for few months after a new game is launched in anticipation of changes to the trends in the stock. From this graph it can be determined that longer the low trading volumes of the stock, more extreme (either positive or negative) the sentiment towards the game.

Chart

Description automatically generated

From the market capitalization graph, we can see that there is no impact of the sentiment on the market can as it always remained the same but there was a huge market cap increase at the time of Apex Legends release because the game provided success to the company that was in a downward trend which acted as a booster for increase in market capitalization.

Chart

Description automatically generated

As part of this analysis, we have analyzed only EA stock, but this analysis can also be performed for other stocks in the dataset in the similar way. Further going forward, the project can be extended by process to dynamically load the data into the dataset and predict the sentiment and use it for stock market analysis. Analysis of the twitter data might from long ago might not be so accurate as new data because at present there is a huge increase in the use of twitter than in the past.

**Implementation:**

Input Twitter data

Data Pre-Processing

Negative Data

Positive Data

Mapping data

Finding Frequency in Data

Da

Input is a twitter data from kaggle related to gaming companies or organizations and another dataset manually created to the gaming organization in first twitter data.

In the first step we are going to perform data processing and map them into positive and negative tweets .And we are creating a dataset manually which contains features explaining the stock trends and the negative and positive behaviors of them.

We are analyzing how the tweets effecting the stock prices of a company with taking one as an example.

**Results:**

In the increment 1 we took the data and performed the sentiment analysis and building a ML Model to find the tweet category.

In the second increment we are manually creating a dataset which includes the companies involved in tweet dataset(i.e first dataset). In this we consider one organization as an example and analyze different aspects of the data and find whether the twitter trends are effecting the stocks prices of the organizations.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

In this project our aim is to check how the twitter or some social media transactions effecting the stock prices of an organization. To prove it we took a gaming company (Electronic Arts) as an example.

A screenshot of a computer

Description automatically generated

From above images we can observe the stock trends and the tweets related to the organization which are coinciding with each other at every point of time.

A screenshot of a computer

Description automatically generated

Background And References:

In context with this analysis we referred to course materials and code snippets related to sentiment analysis such as a text processing using lemmatization and other techniques.

Here are the references for the project:

1. <https://pypi.org/project/yfinance/>
2. [**https://www.kaggle.com/**](https://www.kaggle.com/)
3. https://finance.yahoo.com/

1. Abdul-Mageed, M., M.T. Diab, and M. Korayem. Subjectivity and sentiment analysis of modern standard Arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, 2011.

2. Akkaya, C., J. Wiebe, and R. Mihalcea. Subjectivity word sense disambiguation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009), 2009.

3. Alm, C.O. Subjective natural language problems: motivations, applications, characterizations, and implications. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011), 2011.

4. Andreevskaia, A. and S. Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 2006.

5. Archak, N., A. Ghose, and P.G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2007), 2007.

**Future Scope:**

This project can be further extended by making using of dataset that has a continuous inflow of data. Such a dataset will be very much helpful in the stock market prediction. Additional analysis can be made by making use financial terminologies.