**Detailed project proposal**

**Student Name: Patnam Nithyanand Reddy**

**Student Number: 19025019**

**Module Supervisor: Justin Mifsud**

**Module Code: 7COM1039**

**Title: US AIRLINE SENTIMENTAL ANALYSIS ON TWITTER**

**SUMMARY OF THE PROJECT:**

* In conjunction with the brands Twitter and other microblogging sites, this is a useful tool for practically real-time marketing, public opinion, and consumer knowledge mining. As a result, automated sentiment analysis research focuses on compiling and analysing natural language information generated by people. Mastering machine learning, which encompasses data cleansing and transformation, feature development, model selection, and parameter selection, is the most successful strategy. Documents have been thoroughly examined in recent years, and it is now recognised that very simple techniques like as textual conversion and Naive Bayes models can provide acceptable results, but that appropriate tuning can be difficult, resulting in somewhat limited outcomes (75 percent to 85 percent F1 scores for the average dataset). However, even with a medium-sized dataset, a high proportion of success can result in thousands of better classified products, as well as thousands of lost or angry consumers in any industry. There are tweets in the existing data sets of six US airlines, and we must forecast whether the tweets are good, negative, or neutral. It's a typical supervised task in which we're given a problem statement and asked to categorise it into one of several categories. Experiments demonstrate Naive Bayes, logistic regression, decision tree classification, and random forest with domain specific terminals, as well as checking for data imbalance and binary groups. The Naive Bayes classification systems are shown in the results. Filtering stopwords is critical for improving predictive efficiency, and the experiment reveals that a stopword collection should be domain specific. The conclusion is that there is no optimal technique to represent training and stopword collection in sentiment analysis. As a result, this study proposes that a comparison framework may be used to fine-tune prediction trends for a particular problem: a comparison framework can compare multiple training settings on the same data set in order to identify the best trained models for a specific real-world situation.

**AREA OF STUDY:**

Because Twitter is an open source of important information, its data is thoroughly studied. Around 2009, the first articles on Twitter data sentiment analysis were released. Twitter offers a variety of public APIs with various data harvesting capabilities (but the public API is also becoming more and more limited because it also sees value). However, there are a plethora of publicly available datasets for sentiment analysis. The following is a common pattern in sentiment analysis documents, particularly those involving Twitter data.

The authors define the issue's field and the study's purpose. The product review is the most popular topic, and the research seeks to identify elements that might improve its classification. Several authors focus on 'creating a list of product attributes' in order to make "microblogging websites rich data sources for the research of opinion." Authors looking for "automatic techniques of separating positive from negative feedback" (Dave, Lawrence, and Pennock, 2003) and "automatic methods of separating positive from negative feedback" (Pak and Paroubek, 2010). The authors choose a dataset to work with. Some people use automatically acquired datasets and try to create training datasets using various ways. Several commentators point to the problematic nature of Twitter data collecting in the dataset section, with 'labelling consistency' being one of the most significant difficulties.

In the "Comparison of Predefined Classes Section" of these publications, the subject is discussed (Barbosa and Junlan, 2010). Some people employ unusual training methods to construct a training data collection, such as integrating twitter functionality with remote monitoring to collect training data in the form of tweets. Making use of emoticons (Go, Alec and al., 2009). Others have pre-existing datasets, such as IMDB (Pange, Lee, & Vaithyanathan 2002), or crowdsourced solutions, such as IMDB (Pange, Lee, & Vaithyanathan 2002). The Sentiment Intensity Twitter Dataset (Saif et al, 2003) and Stanford Twitter Gold Sentiment, among others, are available 'normal' datasets for monitoring outcomes. The most important task is usually to explain the feature engineering process. This includes preprocessing approaches that effectively convert text into data that is then processed by machine learning algorithms. The techniques can be divided into the following broad categories.

Methods of linguistic and replacement (Dave, Kushal and Pennock, 2003) To change the N-gram. 'Word N-grams functions are the most basic feature for Twitter sentiment analysis' (Jianqiang et al, 2017). There are numerous viewpoints on the optimum n-gram dimensions; however, this thesis contends that the data set and domain may be used for the environment.

Conversion to a POS system (Speech Part). '[These] methods have demonstrated the accuracy of POS tags. Such POS tags, in my opinion, are important markers of feelings (Wiebe and Riloff, 2005).

Text pattern recognition (POS groups, syntactic trees..) Semantics and syntax were the emphasis of the methods. 'Contextual semantic approaches,' 'Conceptual semanthetic approaches,' and 'Entity-Level Sentimental Analysis Approaches,' for example, have traditionally been utilised in language studies (Saif, Francis and Alani, 2016). The tailored implementation is the major focus of these efforts. The authors demonstrate in the experiment that the proposed characteristics are effective and can improve dataset prediction. In the evaluation portion, a base score is always supplied, and the outcome almost always exceeds the base score. A basic score, uncertainty matrix, F1 score, or other metrics like accuracy and reminder may be used to assess the situation (Jianqiang and Xiaolin, 2017). Machine learning approaches are at the cutting edge of the field, yet the majority of articles focus on the function output aspect. (Dave, Kushal, and Pennock, 2003) Most studies identify the algorithms employed and compare the outcomes of different techniques (Jianqiang and Xiaolin, 2017). However, most evaluations do not go into detail about the algorithms (parameters, Deployments, etc..).

**KNOWLEDGE THAT IS REQUIRED:**

Sentiment Analysis is a system that learns to interpret polarity statements, from positive to negative. By using instances of emotions in text to train systems to learn how to recognise feelings without human feedback.

Machine learning, to put it simply, allows computers to learn new tasks without having to be explicitly programmed. Sentiment analysis models can be taught to interpret things like meaning, sarcasm, and misapplied phrases in addition to their literal meanings.

**FRAMEWORK FOR SIMULATION:**

Step 1: Download the Kaggle data and upload it to the Google Collaboratory.

Step 2: Import all Python packages that are required.

Step 3: Data preparation in order to fully comprehend the data before to sentimental analysis.

Step 4: Create a graphical representation of the data.

Step5: Create Machine Learning models like Decision Trees, Logistic Regressions, Random Forests, and KNNs, and analyse their categorization results.

Step 6: Compare the results to your prior researcher's Literature work.

Step 7: Assign future work in order to improve the model's construction.

**SOFTWARE NEEDS ARE AS FOLLOWS:**

Google Research's Colaboratory, or simply "Colab," is a product. Colab is a web-based tool that allows anyone to write and run arbitrary Python code, making it ideal for machine learning, data analytics, and training.

Python is an open source object-oriented programming language that is versatile and simple to learn. It has a large number of libraries and tools to help data scientists with their work. Python also has a big user base, which allows developers and data scientists to ask each other questions. Python has been used as a service by data scientists for many years and will continue to be the preferred language for data scientists and developers.

**PYTHON'S IMPORTANCE:**

Python is a computer language that data science consultancies allow its development team and data scientists to utilise. Python learned the most significant programming language in an extraordinarily short amount of time.

Data scientists must manage a large amount of data known as big data. Python has been a popular choice for dealing with large amounts of data due to its ease of use and extensive range of python libraries.

**PLAN FOR THE PROJECT:**

|  |  |  |  |
| --- | --- | --- | --- |
| TASK NAME | START | END | DURATION  (days) |
| Topic selection | 1/06/2021 | 22/06/2021 | 5 |
| Dataset selection | 23/06/2021 | 28/06/2021 | 5 |
| Initial draft preparation | 29/06/2021 | 08/07/2021 | 10 |
| Proposal acceptance | 9/07/2021 | 14/07/2021 | 5 |
| Literature review | 11/08/2021 | 04/09/2021 | 25 |
| Load data into google colab | 5/09/2021 | 09/09/2021 | 5 |
| Preprocessing the data | 10/09/2021 | 9/10/2021 | 30 |
| Building sentimental analysis | 10/10/2021 | 29/10/2021 | 20 |
| Test the results | 30/10/2021 | 08/11/2021 | 10 |
| Final draft preparation | 09/11/2021 | 23/11/2021 | 15 |
| Final report submission | 24/11/2021 | 03/12/2021 | 10 |

**DESCRIPTION OF THE DATASET**:

A nostalgic look at the challenges of each major US airline. Since February 2015, Twitter data has been scraped, and contributors have been asked to identify positive, negative, and neutral reasons for categorisation, followed by negative reasons (such as "late flight" or "rude service"). There are about 14000 tweets in this dataset, each with a different property. This information was obtained via Kaggle.

**REFERENCES:**

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R.J. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R.J. Twitter data sentiment review. The proceedings of the Workshop on Language in Social Media (LSM 2011) were released.

Open Source Datasets, Appen.com https://appen.com/resources/datasets/,

https://docs.aws.amazon.com/comprehend/latest/dg/how-it-works.html, Amazon Web Services, 2020.05.25

L. Barbosa and J. Feng, 2010. From skewed and noisy results, robust sentiment detection on Twitter. Posters from the 23rd international conference on computational linguistics are included in the proceedings (pp. 36-44). The Association for Computational Linguistics (ACL) is a non-profit organisation dedicated to the study of language

A. Bifet and E. Frank, 2010. Twitter streaming data is used to explore sentiment information. In a forum on discovery research held around the world (pp. 1-15). Heidelberg and Berlin: Springer.

K. Dave, S. Lawrence, and D.M. Pennock, 2003. Opinion analysis and semantic classification of advertising feedback from the peanut gallery. The 12th international conference on the World Wide Web is published in the proceedings (pp. 519-528).

D. Davidov, O. Tsur, and A. Rappoport, 2010. Twitter hashtags and smileys were used to improve emotion learning. Posters from the 23rd international conference on computational linguistics are included in the proceedings (pp. 241-249). The Association for Computational Linguistics (ACL) is a non-profit organisation dedicated to the study of language

A. Go, R. Bhayani, and L. Huang, 2009. Distant supervision for Twitter emotion grouping. Stanford, 1(12), p.2009, CS224N project study.

A.F. Hidayatullah, 2015. Stemming's Impact on Indonesian Tweet Sentiment Analysis 2(1), pp.127-132, Proceeding of the Electrical Engineering Computer Science and Informatics.

Z. Jianqiang and G. Xiaolin, 2017. On Twitter sentiment analysis, a comparison study of text pre-processing methods was conducted. IEEE Access, vol. 5, no. 5, pp. 2870-2879

Kaggle.com US Airline Sentiment

2020.05.25, https://www.kaggle.com/crowdflower/twitter-airline-sentiment