**Detailed project proposal**

**Student Name: Patnam Nithyanand Reddy**

**Student Number: 19025019**

**Module Supervisor: Justin Mifsud**

**Module Code: 7COM1039**

**Title: TWITTER US AIRLINE SENTIMENTAL ANALYSIS**

**Project Summary:**

A valuable tool for almost real-time marketing, public opinion, and customer knowledge mining in conjunction with the brands Twitter and other microblogging services. The research on automated sentiment analysis focuses therefore upon the compilation and analysis of natural language content created by users. The most successful approach is mastered machine learning, which includes data cleaning and transformation, the generation of features and the choice of model and the selection of parameters. Documents have been extensively reviewed in recent years and it is understood that relatively simple techniques such as textual conversion and Naive bayes models can produce acceptable results and good tuning can be difficult, producing relatively limited results (75 percent to 85 percent F1 scores for the average dataset). But even with a mid-sized dataset, many percentage levels of success can mean thousands of better classified materials, meaning thousands of lost or unhappy customers in any business sector. In the existing data sets of 6 US airlines we have tweets and we have to predict whether the tweets are positive, negative or neutral. It is a standard supervised job, where a problem statement is given to us and categorized into a predetermined category. Experiments show Naive Bayes, logistic regression, decision tree classification, random forest with domain specific terminals, and check whether the data has not been imbalanced or whether the groups are not binary. The results show the Naive Bayes classification systems. To improve the predictive efficiency, filtering stopwords is crucial; and the experiment shows that a collection of stopwords should be domain specific. The inference is that in sentiment analysis, there is no optimal way to model training and stopword collection. This paper therefore suggests that a comparison framework can be used to fine tune prediction trends for a given problem: a comparison framework can compare various training settings on the same data set so that the best trained models for a given real-life problem can be identified.

**Research Area:**

Twitter data is extensively investigated, as Twitter is an open source of useful information. The early papers on Twitter data sentiment analysis were published around 2009. Twitter provides public APIs with different data collection capabilities (but the public API is also becoming more and more limited because it also sees value). However, for sentiment analysis, there are a wide number of publicly accessible datasets. In documents on sentiment analysis, in particular about Twitter data, the following is a popular pattern.

Authors identify the field of issue and the study goal. The most popular subject is the review of the product and the research aims to recognize features that can enhance their classification. A number of authors concentrate on 'generating a list of product attributes' so that 'microblogging websites are rich data sources for the study of opinion. 'Authors seeking 'automatic methods of separating positive from negative feedback' (Dave, Lawrence and Pennock, 2003), (Pak and Paroubek, (2010).). Authors identify their dataset. Some use datasets that are automatically collected and attempt to build training datasets by using various techniques. It is the dataset segment where several writers highlight the problematic nature of Twitter data collection, some key issues are 'labeling consistency.'

The problem is discussed in the "Comparison of Predefined Classes Section" section of these works (Barbosa and Junlan, 2010). Some people use unexpected training methods to develop a training data collection or to use twitter functionality using remote monitoring, which consists of tweets for our training data. Using emoticons (Go, Alec and al., 2009). Such others have existing datasets, such as IMDB (Pange, Lee & Vaithyanathan 2002), which are compiled by others or crowdsourced solutions. There are available 'normal' datasets used for monitoring outcomes, such as the Sentiment Intensity Twitter Dataset (Saif et al, 2003) and Stanford Twitter Gold Sentiment and others. The key task typically is to explain the process of feature engineering. This includes the techniques of preprocessing which effectively transform the text into data which is processed by algorithms of machine learning. The techniques can be categorized into the next large categories.

§ Linguistic and replacement methods (Dave, Kushal and Pennock, 2003) To transform N-gram. 'The simplest feature for Twitter sentiment analysis is Word N-grams functions' (Jianqiang et al, 2017). There's a variety Opinions on the ideal n-gram dimensions; this thesis argues that Possibly the data set and the domain are used for the environment.

* Conversion to POS (Speech Part). '[These] methods have shown that POS tags are accurate. Intuition is that such POS tags are strong indicators of feelings (Wiebe and Riloff, 2005).
* Recognition of text patterns (POS groups, syntactic trees..) The approaches focused on semantics and syntax. Those strategies are traditionally focused on language studies such as 'Contextual semantic approaches', 'Conceptual semanthetic approaches' and 'Entity-Level Sentimental Analysis Approaches' (Saif, Francis and Alani, 2016). The main emphasis of these works is the personalized implementation. In the experiment, the authors show that the features proposed work effectively and can produce an improved prediction on the dataset. A base score is always provided in the assessment section, and almost always the outcome exceeds the base score. The assessment may be a simple score, uncertainty matrix, F1 score, or other metrics like accuracy and reminder (Jianqiang and Xiaolin, 2017). Machine learning techniques are the state of the art in the area, but most of the papers concentrate on the function output portion. Most papers identify the algorithms used, and compare the results between different algorithms (Dave, Kushal and Pennock, 2003), (Jianqiang and Xiaolin, 2017). But most reviews do not detail the algorithms (parameters, Deployments, etc..).

**Required Knowledge:**

Sentiment Analysis is a learning machine that analyzes polarity messages, from good to bad. By training tools to learn how to detect feelings without human feedback, by examples of emotions in text.

Simply put, machine learning lets computers learn new tasks without explicit programming. Models of sentiment analysis can be educated in reading stuff like, meaning, sarcasm and misapplied terms beyond mere meanings.

**Simulation Framework:**

Step 1: Download the data from Kaggle and load the Data on Google colaboratory

Step2: Import all required Python packages

Step3: Data preprocessing for understanding the data in depth most before the sentimental analysis

Step4: Visualise the data graphically

Step5: Build the Machine learning models such as Decision tree, Logistic Regression, Random Forest, KNN and get their classification report.

Step6: Compare the results with your Literature work made by previous researcher.

Step7: Give the future work for better building the model

**Software requirements:**

Colaboratory, or shortly "Colab," is a Google Research product. Colab enables anyone to write and apply arbitrary python code through the browser and is suitable for machine learning, data analytics and training.

Python is an open source object-oriented language, adaptable and easy to understand. It has a rich collection of libraries and tools to simplify tasks for data researchers. Python also has a large user base in which developers and data scientists can ask questions from others. Python has been used by data science as services for many years and will continue to be the top option for data scientists and developers.

Importance of Python:

Data science consultancies enable their development team and data scientists to use Python as the language of their programming. In an incredibly brief period of time, Python became familiar with the most important programming language.

A great deal of data known as big data needs to be managed by data scientists. Python has become a common option to deal with big data with easy use and a wide variety of python libraries.

**Project plan:**

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
|  |  |  |  |

**Dataset Description:**

A nostalgic study of each big U.S. airline's problems. Since February 2015, Twitter data were scrapped and contributors were requested to first identify positive, negative and neutrals, followed by negative reasons for categorization (such as "late flight" or "rude service"). This dataset contains more than 14000 tweets with 15 different attributes. This data has been taken from Kaggle.

**References:**

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R.J. Twitter data sentiment review. The Workshop on Language in Social Media (LSM 2011) published its proceedings (pp. 30-38).

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