Social Media Sentiment Analysis Using Twitter Dataset

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

Computer Science & Engineering

(BIG DATA ANALYTICS)

Submitted by:

Tushar Sharma(20BCS3837)

Sachin Pareek(20BCS3817)

Amit Kumar(20BCS3767)

Under the Supervision of

Ms. Neeru Bala

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
APEX INSTITUE OF TECHNOLOGY

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,
PUNJAB

March, 2023;

Table of Contents

TIMELINE

1. 27-02-2023

- Data research
- Data Gathering
- Data filtering

2. 27-03-2023

- Project Preliminary Design
- Continuous evaluation and enhancements

3. 27-04-2023

- Final Testing and Evaluation

1 INTRODUCTION

1.1 Problem Definition:

The problem definition for social media sentiment analysis using Twitter dataset would be to analyses a large collection of tweets related to a particular topic or event, and to determine the overall sentiment of the tweets. This can be done by classifying each tweet as positive, negative, or neutral based on the language used, the context of the tweet, and other factors.

The goal of the sentiment analysis could be to understand the public perception of a product, brand, or political campaign, or to identify trends and patterns in social media conversations. The results of the sentiment analysis can be used by businesses, marketers, and policymakers to make informed decisions and to improve their communication strategies.

Some of the challenges in social media sentiment analysis include dealing with noise and sarcasm in social media data, handling different languages and dialects, and ensuring the accuracy and consistency of the sentiment classification. However, with the right tools and techniques, social media sentiment analysis can provide valuable insights into the attitudes and opinions of social media users.

1.2 Project Overview/Specifications:

In order to complete this project, few goals must be met:

- Research machine learning approaches for sentiment analysis.
- Changing the machine learning algorithm to improve accuracies of model.
- Making use of customized machine learning methods with NLP in different scenarios.
- To put the machine learning algorithm to the test, use real data from the machine learning data repository.
- Project will tend to analyze the sentiments of the tweets by weighing them under the labels neutral, positive and negative. This will help us to analyze the trends and sentiments of the twitter users for a particular scenario or issue concerned. This will leverage NLP, some classifying ML algos and deep learning methodologies.
- It will help us to draw insights from a dataset of tweets and classify them into categories according to similarity of topics using the techniques mentioned above.

1.3 Hardware Requirements:

Name of component	Specification
Processor	Pentium-IV
RAM	Atleast 4 GB
Hard Disk	2 GB or more

1.4 Software Requirements:

- Chrome
- Visual studio code
- Windows operating system
- Python Environment
- Required libraries installed

1.4.1 Libraries Used:

- **Pickle:** It is used for serializing and de-serializing python object structures
- **Streamlit:** It is used for easy to create and share beautiful, custom web apps for machine learning and data science.
- **Requests:** Requests library is one of the integral part of Python for making HTTP requests to a specified URL.
- **NumPy:** It is a library consisting of multidimensional array objects and a collection of routines for processing those arrays.
- **Pandas:** It is a Python library for data analysis.
- **Keras**: An open-source software library that provides a Python interface for artificial neural networks.It acts as an interface for the TensorFlow library.

2 LITERATURE REVIEW

2.1 Existing System:

There have been numerous studies on social media sentiment analysis using Twitter datasets. Here are some of the key findings and approaches:

1. **Machine Learning and Deep Learning Techniques**: Several studies have used machine learning and deep learning techniques to classify tweets into positive, negative, or neutral categories. These techniques include Support Vector Machines (SVM), Naive Bayes, Random Forest, and Convolutional Neural Networks (CNN). These models use features such as text, hashtags, emoticons, and user information to classify tweets.
2. **Lexicon-Based Approaches**: Lexicon-based approaches use pre-built sentiment lexicons, such as SentiWordNet and AFINN, to determine the sentiment of a tweet. These approaches assign a score to each word in the tweet based on its sentiment value in the lexicon, and then calculate the overall sentiment score for the tweet. However, these approaches can be limited by the size and accuracy of the lexicon.
3. **Hybrid Approaches**: Some studies have used hybrid approaches that combine machine learning and lexicon-based techniques to improve the accuracy of sentiment classification. For example, a study by Zhou et al. (2015) used a hybrid approach that combined a sentiment lexicon with a Support Vector Machine classifier to classify tweets.
4. **Emotion Analysis**: Emotion analysis is a subset of sentiment analysis that focuses on identifying specific emotions in social media data. Several studies have used machine learning techniques to identify emotions such as joy, anger, sadness, and fear in tweets. These approaches can provide more nuanced insights into the emotional responses of social media users.
5. **Multilingual Sentiment Analysis**: With the increasing use of social media around the world, multilingual sentiment analysis has become an important area of research. Studies have explored techniques for sentiment analysis in languages such as Arabic, Chinese, and Spanish, using machine learning and lexicon-based approaches.

2.2 Proposed System:

- A dataset from the "Kaggle" website is used as a training dataset in this suggested method. To improve machine performance, the inserted dataset is first verified for duplicates and null values. The dataset is then divided into two sub-datasets, say "train dataset" and "test dataset", with a 70:30 split. The "train" and "test" datasets are then given as text-processing parameters.

- Text-processing removes punctuation symbols and words from the stop words list and returns them as clean words. These clean words are then used to transmit to "Feature Transform". The clean words obtained from text processing are then utilized for 'fit' and 'transform' to create a vocabulary for the machine in feature transform. The dataset is also passed for "hyperparameter tuning," which is used to discover the best values for the classifier to utilize based on the dataset.

- After obtaining the settings from the "hyperparameter tuning," the machine is fitted with a random state using those values. The trained model's state and characteristics are retained for future testing with unseen data. Using classifiers from module sklearn in python, the machines are trained using the values obtained from above.

2.3 ADVANTAGES OF PROPOSED SYSTEM:

Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of tweets are sent and received and it is difficult as our project is only able to test tweets using a limited amount of corpus. Our project, thus sentiment analysis is proficient of filtering tweets giving to the content of the tweet and not according to the domain names or any other criteria.

- Good Efficiency
- Greater accuracy

3. PROBLEM FORMULATION:

Approach:
The sentiment analysis problem can be tackled using various machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines (SVMs).

The following are the steps involved in developing a machine learning model for sentiment analysis:

1. Data collection: Collect the Twitter dataset related to the topic of interest using Twitter APIs or third-party data providers.

2. Data pre-processing: Clean and preprocess the text data by removing stop words, stemming or lemmatization, handling special characters, emojis, and URLs.

3. Feature extraction: Convert the preprocessed text data into numerical features using techniques such as bag-of-words, term frequency-inverse document frequency (TF-IDF), or word embeddings.

4. Model selection: Select a suitable machine learning algorithm and tune its hyperparameters using cross-validation techniques to obtain the best model performance.

5. Model evaluation: Evaluate the performance of the model using various metrics such as accuracy, precision, recall, and F1-score.

6. Model deployment: Deploy the trained machine learning model as a web service or application to classify new tweets in real-time.

4. RESEARCH OBJECTIVES:

- Research machine learning approaches for tweet sentiment analysis.
- Changing the machine learning algorithm to improve accuracies of model.
- Making use of customized machine learning methods with NLP in different scenarios.
- To put the machine learning algorithm to the test, use real data from the machine learning data repository.
- Project will tend to classify tweets as neutral, positive and negative, addresses using Natural Language Processing (NLP) and also set priorities and segregate them using perceptron models and other deep learning methodologies.
- It will help us to draw insights from a dataset of tweets and classify them into categories according to similarity of topics using the techniques mentioned above.

5. METHODOLOGY:

5.1 INTRODUCTION

This chapter will explain the specific details on the methodology being used in order to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering. So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself.

5.2 IMPLEMENTATION AND CODING PHASE

This project is developed by using Python Language and combining with the many machines learning and a neural network algorithm. It contains important function for preprocessing the dataset. Then, the dataset is going to be used to train and test either the model of the machine learning achieves the objectives of the project

6. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

CHAPTER 1 (i): PROBLEM DEFINITION

This chapter will tell the problem about which we are going to discuss through our project.

CHAPTER 1 (ii): PROJECT OVERVIEW/ SPECIFICATION

This chapter will cover the overview of **Social Media Sentiment analysis using Twitter Dataset.**

CHAPTER 1 (iii): HARDWARE

This section will talk about all the hardware that we are going to use in our sentiment analysis system.

CHAPTER 1 (iii): SOFTWARE

This section will talk about all the software and libraries that we are going to use in our sentiment analysis system.

CHAPTER 2: LITERATURE REVIEW

This chapter include the literature available for **Social Media Sentiment analysis using Twitter Dataset.** The findings of the research will be highlighted which will become basis of current implementation.

CHAPTER 2 (i): EXISTING METHOD

This chapter will provide basic knowledge about the Sentiment Analysis which are continuously used since many years.

CHAPTER 2 (ii): PROPOSED METHOD

This chapter will describe all those new features which we will include in sentiment analysis system.

CHAPTER 3: PROBLEM FORMULATION

This chapter will cover all the details about those problems that we are going to solve through our sentiment analysis system.

CHAPTER 4: RESEARCH OBJECTIVE

This chapter will cover all the purposes of this sentiment analysis that will going to be fulfilled after creation of this system.

CHAPTER 5: METHODOLOGY

This chapter will cover the technical details of the proposed approach.

CHAPTER 6: CHAPTER PLAN

This chapter will provide information about the chapter- by - chapter topics and tools usedfor evaluation of proposed method.

CHAPTER 7: REFERENCES

This section will show all the resources from which we have taken help.

7. REFERENCES:

- www.google.com
- https://www.wikipedia.org
- Social Media Sentiment Analysis using Machine Learning : Part — I | by Deepak Das | Towards Data Science
- Twitter Sentiment Analysis With Python | Introduction & Techniques (analyticsvidhya.com)
- Twitter Sentiment Analysis using Python - GeeksforGeeks
- Social Media Sentiment Analysis On Twitter Datasets | IEEE Conference Publication | IEEE Xplore
- Twitter sentiment analysis using deep learning methods | IEEE Conference Publication | IEEE Xplore
- Advances in Distributed Computing and Artificial Intelligence Journal : 9, Regular Issue 3, 2020 - Ediciones Universidad de Salamanca - Torrossa
- https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
- https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
- https://www.tandfonline.com/doi/abs/10.1515/jisys-2015-0032
- https://j-scdss.com/index.php/files/article/view/30/38
- https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441/1852
- https://www.aclweb.org/anthology/W11-0804.pdf