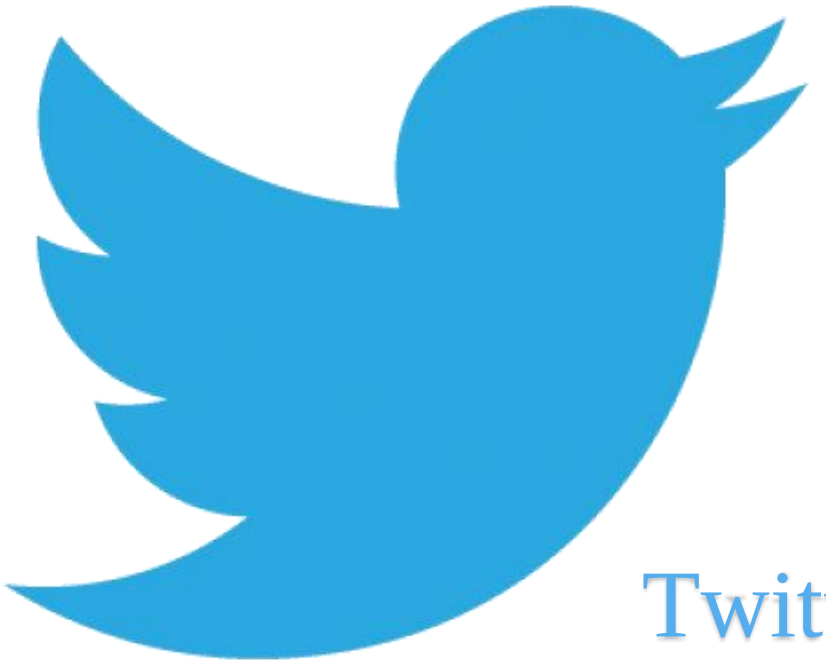
Kingdom Of Saudi Arabia
Ministry Of Education
Shaqra University
College Of Computer Science and Information Technology



Twitter sentiment analysis

By: Raghad Al-moqhim and Lina Al-fawzan.

Supervisor: Dr. Samia Amor Dardouri.

Abstract

In this report, we address the problem of sentiment classification on twitter dataset. We use a number of machine learning and deep learning methods to perform sentiment analysis. We use a hugging face Twitter dataset with a 124m tweets from January 2018 to December 2021, the accuracy of the dataset is approximately 85%.

Keywords: Twitter sentiment analysis, Classification, Machine learning, Prediction, Pie chart, Analyze, Visualization, Probability.

Contents:

Introduction

Nowadays, the age of Internet has changed the way people express their views and opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc.

Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising.

People mostly depend upon user generated content over online to a great extent for decision making. For e.g., if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision.

The amount of content generated by users is too vast for a normal user to analyze. So, there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but, there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis. It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks.

For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of sentiment analysis.

Problem

Sentiment analysis in the domain of micro-blogging is a relatively new research topic so there is still a lot of room for further research in this area. A decent amount of related prior work has been done on sentiment analysis of user reviews, documents, web blogs/articles, and general phrase-level sentiment analysis. These differ from twitter

mainly because of the limit of 280 characters per tweet which forces the user to express opinions compressed in a very short text. The best results were reached in sentiment classification using supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labeling required for the supervised approach is very expensive. Some work has been done on unsupervised and semi-supervised approaches, and there is a lot of room for improvement. Various researchers testing new features and classification techniques often just compare their results to baseline performance. There is a need for proper and formal comparisons between these results arrived through different features and classification techniques to select the best features and most efficient classification techniques for particular applications.

Objectives

Being extremely interested in everything having a relation with the Machine Learning, the independent project was a great occasion to give me the time to learn and confirm my interest for this field.

The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities.

We can use Machine Learning in Finance, Medicine, almost everywhere. That's why we decided to conduct our project around the Machine Learning, regardless of it being our course.

Methodologies

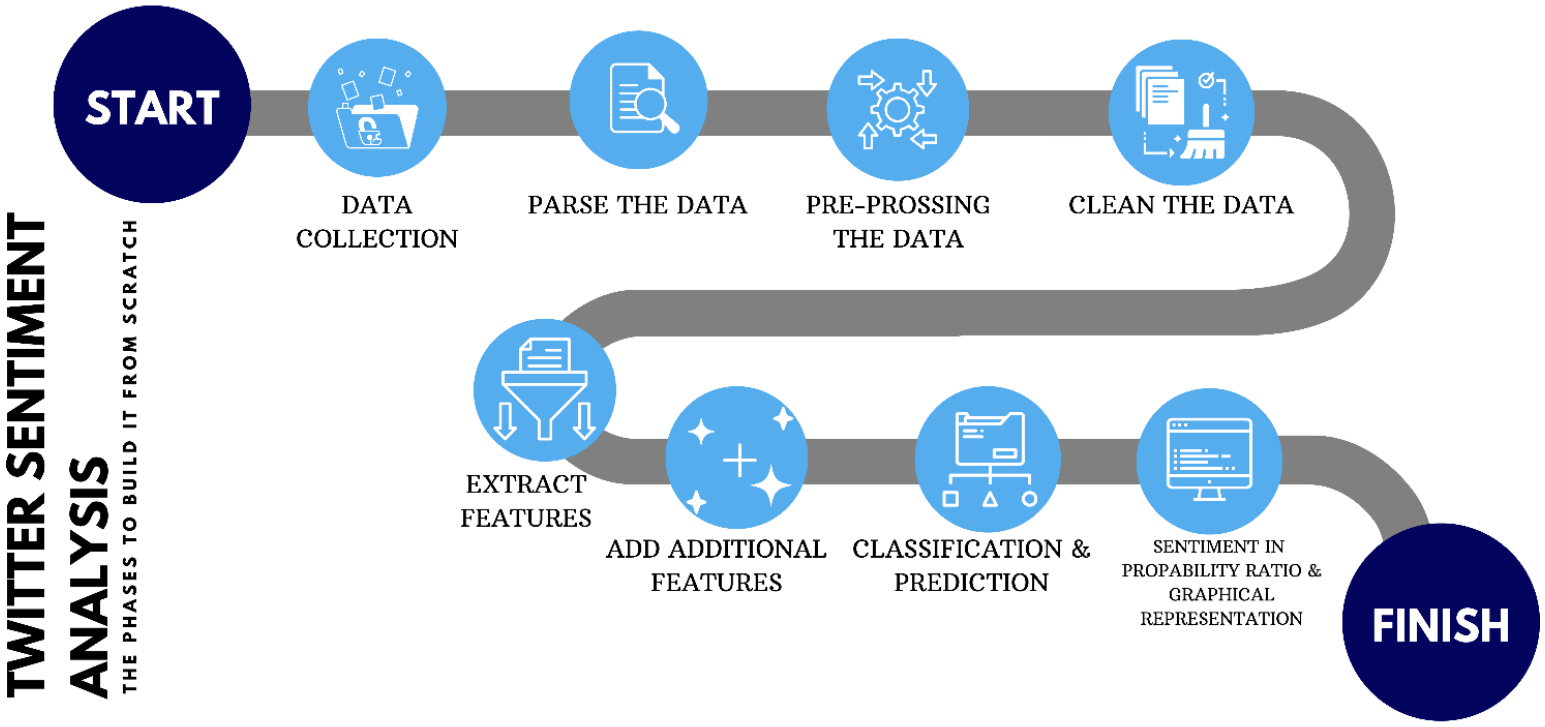


Fig 1: The methodology we used to build a sentiment analysis system using Twitter dataset

- **Data (Tweet) Collection**

The dataset collected is imperative for the efficiency of the model. The division of dataset into training and testing sets is also a deciding factor for the efficiency of the model. The training set is the main aspect upon which the results depend. The dataset accuracy approximately 85% as shown in the *fig2*.

Model	F1	⬆ Accuracy
cardiffnlp/twitter-roberta-base-dec2021-emotion		0.845

Fig 2 The accuracy of the dataset.

- **Parse the data**

Tokenizers are one of the core components of the NLP pipeline. They serve one purpose: to translate text into data that can be processed by the model. Models can only process numbers, so tokenizers need to convert our text inputs to numerical data.

This process done by a Python library called Transformers that has an AutoTokenizer package which was very helpful in this phase. The use of AutoTokenizer modules, parsing with everyday expressions, parsing with string techniques together with the split() and strip() techniques.

- **Data cleaning**

The data collected was not in the correct format. Information cleaning is the way toward guaranteeing that information is right, predictable, and usable. Usually, datasets need to cleanse because they consist of a lot of noisy or unwanted data called outliers. The existence of such outliers may lead to inappropriate results. Data cleaning ensures the removal and improvisation of such data and results in a much more reliable and stable dataset.

Data cleaning can be done in given ways:

- Monitors errors: The entry point or source of errors should be tracked and monitored constantly. This will help in correcting the corrupted data.
- Process standardization: The point of entry should be standardized. By standardizing the data process, the risk of duplication reduces.
- Accuracy validation: Data should be validated once the existing database is cleaned. Studying and using various data tools that can help in cleaning the datasets is very important.
- Avoid data duplication: The identification of duplicate data is a very mandatory process. Several AI tools help in identifying duplicates in large corpora of data.

The above-mentioned steps are a few of the many ways to clean datasets. Making use of these methods will end up giving good, usable, and reliable datasets.

- **Pre-processing of the datasets**

The preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the tweets as desired. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction.

Below are a few steps used for pre-processing of tweets:

- Removal of re-tweets.

- Converting upper case to lower case: In case we are using case sensitive analysis, we might take two occurrence of same words as different due to their sentence case. It important for an effective analysis not to provide such misgivings to the model.

- Stop word removal: Stop words that don't affect the meaning of the tweet are removed (for example and, or, still etc.). Uses WEKA machine learning package for this purpose, which checks each word from the text against a dictionary.

- Twitter feature removal: User names and URLs are not important from the perspective of future processing, hence their presence is futile. All usernames and URLs are converted to generic tags or removed.

- Stemming: Replacing words with their roots, reducing different types of words with similar meanings. This helps in reducing the dimensionality of the feature set. In other word, it is an element procedure of delivering morphological variations of a base word. The words like “chocolatey,” “chocolates” are converted to their root word “chocolate.”

- Special character and digit removal: Digits and special characters don't convey any sentiment. Sometimes they are mixed with words, hence their removal can help in associating two words that were otherwise considered different.

- Creating a dictionary to remove unwanted words and punctuation marks from the text.

- Expansion of slangs and abbreviations.

- Spelling correction.

- Generating a dictionary for words that are important or for emoticons.

- Part of speech (POS) tagging: It assigns tag to each word in text and classifies a word to a specific category like noun, verb, adjective etc. POS taggers are efficient for explicit feature extraction.

- **Feature extraction**

The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect is used to

compute the positive, negative and nature polarity in a sentence which is useful for determining the opinion of the individuals' using models like unigram, bigram.

Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification task.

Some examples feature that have been reported in literature are:

- Words and Their Frequencies: Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. showed better results by using presence instead of frequencies.

- Parts of Speech Tags: Like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

- Opinion Words and Phrases: Apart from specific words, some phrases and idioms which convey sentiments can be used as features.
  e.g., cost someone an arm and leg.

- Position of Terms: The position of a term with in a text can affect on how much the term makes difference in overall sentiment of the text.

- Negation: Is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

- Syntax: Syntactic patterns like Collections are used as features to learn subjectivity patterns by many of the researchers.

- **Classification and prediction**
  - The features extracted are passed to the classifier.

  - The model we built is predicting the sentiment of the tweet.

- **Sentiment in probability ratio and graphical representation**

We use a matplotlib library to visualize the resulting graph. A graph indicating the ratio of each of the following probabilities: positive in green, negative in red, and neutral in orange.

Probability ratios are values ranging from 0 to 1. Probability ratios represented in our program as decimals. If a probability for a category of the three possibilities (negative, positive, neutral) equal to 0, then it is impossible. If a probability equal to 1, then it is certain to be.

Results and Discussion

We have successfully built a Twitter Sentiment Analyser Program by Python. The results were great and accurate. We will explain it from several aspects as follows:

A. Program Graphical User Interface

We build a GUI using the customtkinter library, so users can interact with a computer program. This GUI consists of an **entry widget** that provides us with the Tweet that the user wrote. And **three buttons** first one "predict" which gives the user the output in decimal form, the second one "visualization" which gives the user the output in percentage form and a pie chart, and the third one "clear" which is reset all form values. And a **big label** for the program name. And the last one is a **Twitter icon**.

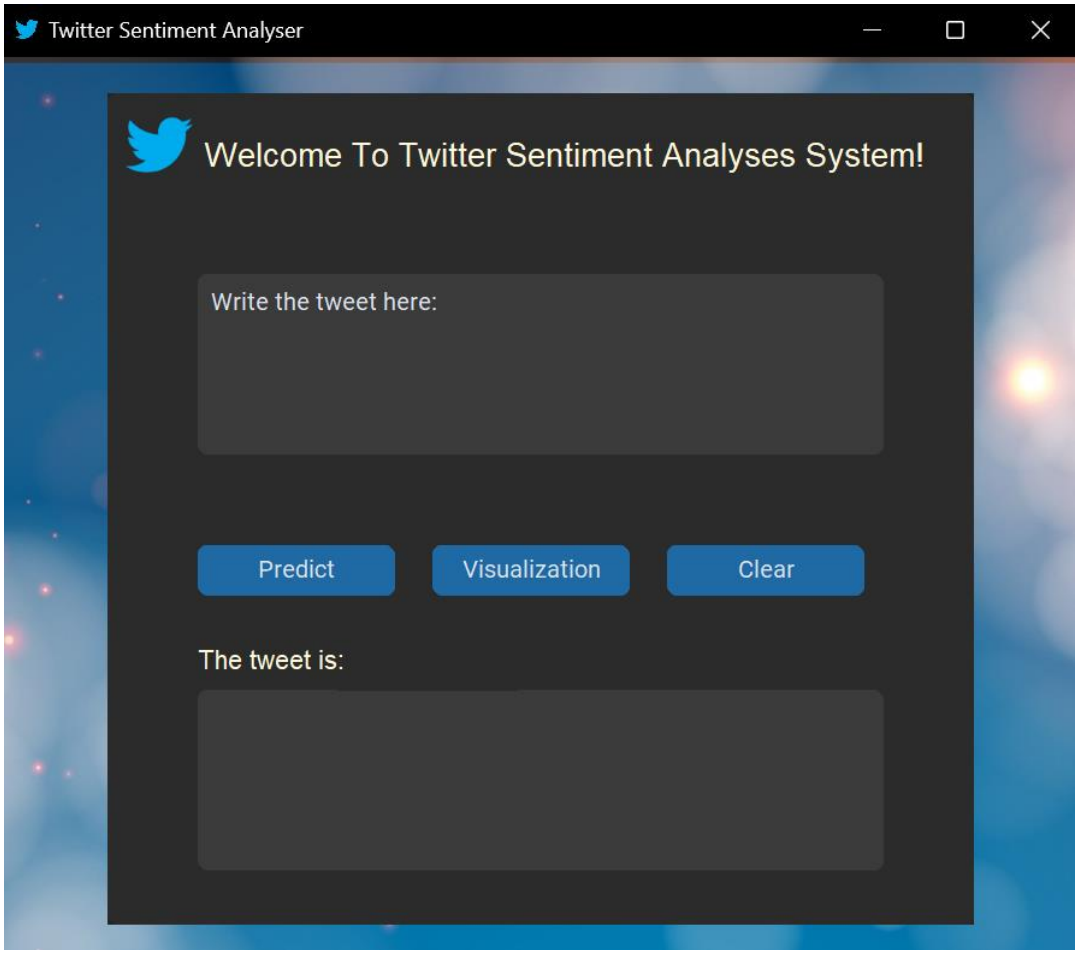


Fig 3: The Program GUI

B. Sentiment result by probability decimal ratio.

In this part, the output will represent as a decimal ranging from 0 to 1, and the probability that has the highest value will be adopted as a result of the tweet sentiment analysis. The tweet in which we will do the sentiment analysis by the program illustrated in *fig3*. The predict result illustrated in *fig4*.

Fig 4

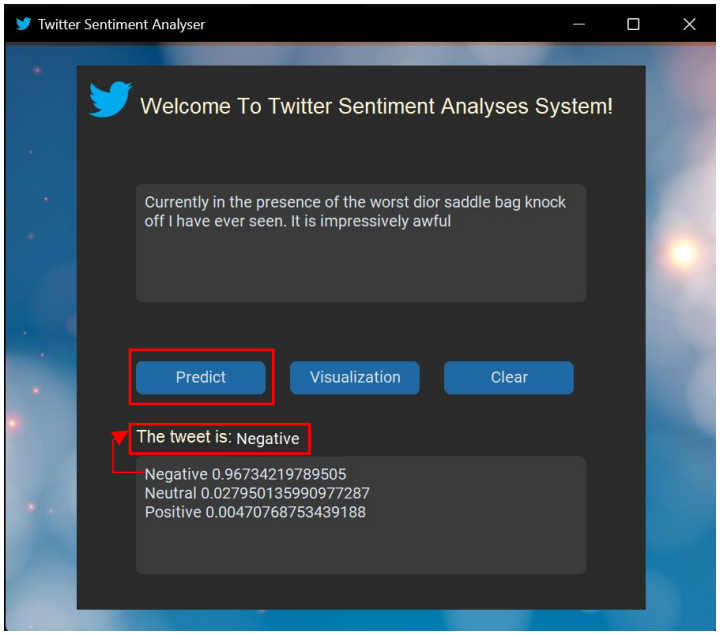


Fig 5

C. Sentiment result by a pie chart and probability percentage ratio.

In this part, the output will represent as a percentage ranging from 0% to 100%, within a pie chart. It will be a great benefit because it facilitates the user to read the long accurate numbers.

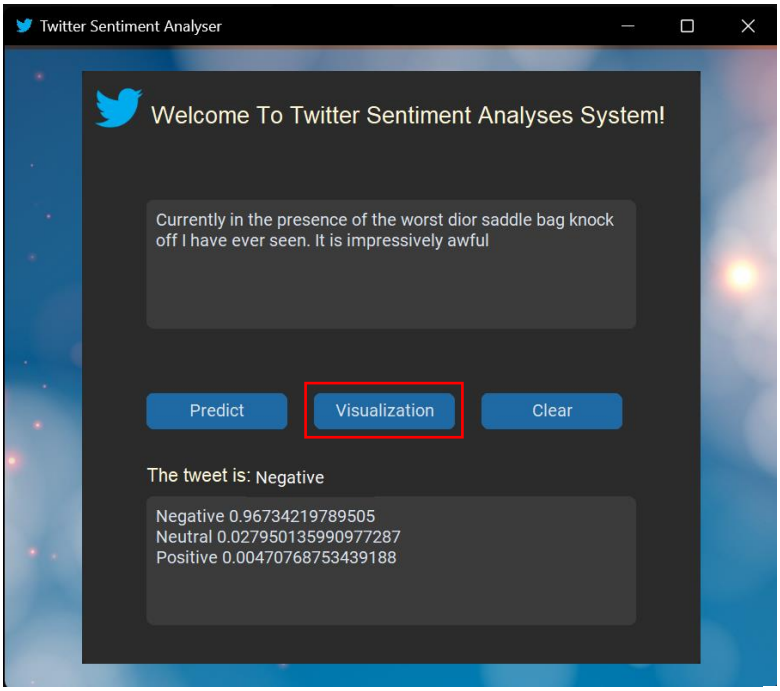


Fig 6

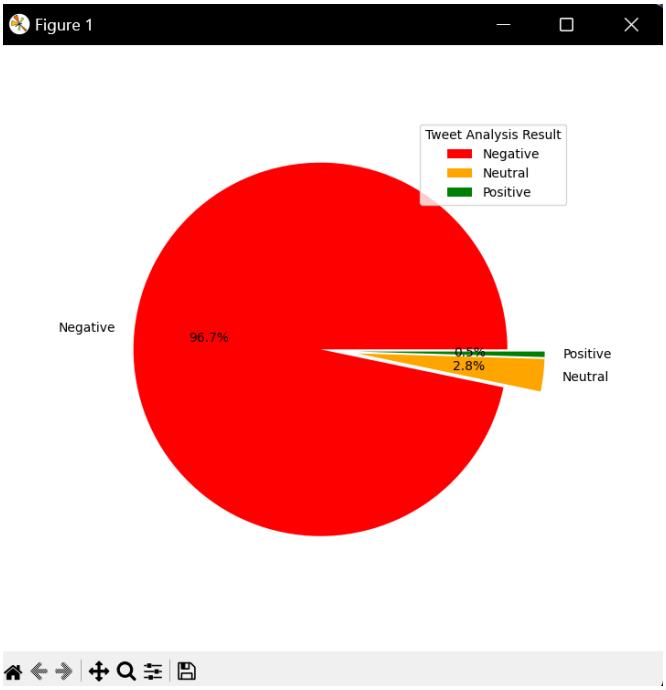


Fig 7

Conclusion and Recommendations

The work in this paper is done to classify a relatively huge corpus of Twitter data into three groups of sentiments, positive, negative, and neutral respectively. Higher accuracy is achieved by using sentiment features instead of conventional text classification. This feature can be used by various establishments, business organizations, entrepreneurs, etc., to evaluate their products and get a deeper insight into what people say about their products and services. The recommendation includes working not only in the English language but in other regional languages too. Also, it will include analysis of complex emotions like sarcasm and generate a hybrid classifier to get the best accuracy.

References

- ✓ Yener, Y. (2021, December 16). *Step by Step: Twitter Sentiment Analysis in Python - Towards Data Science*. Medium. https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d
- ✓ Silaparasetty, N. (2022, September 23). *Twitter Sentiment Analysis for Data Science Using Python in 2022*. Medium. https://medium.com/@nikitasilaparasetty/twitter-sentiment-analysis-for-data-science-using-python-in-2022-6d5e43f6fa6e
- ✓ *Getting Started with Sentiment Analysis on Twitter*. (n.d.). https://huggingface.co/blog/sentiment-analysis-twitter
- ✓ Wicaksana, P. Y. (2022, September 22). *Twitter Sentiment Analysis with Transformers Hugging Face (RoBERTa)*. Medium. https://medium.com/mlearning-ai/elon-musks-twitter-sentiment-analysis-with-transformers-hugging-face-roberta-49b9e61b1433
- ✓ Duong, B. T. (2022, December 22). *Twitter Sentiment Analysis with Deep Learning using BERT and Hugging Face*. Medium. https://medium.com/mlearning-ai/twitter-sentiment-analysis-with-deep-learning-using-bert-and-hugging-face-830005bcdbbf