# Twitter Topic Based Sentiment Analysis Using Deep Learning

**1.1 Introduction**

Artificial Intelligence, sometimes calledmachine Intelligence, it is demonstrated by machines in contrast to the natural intelligence displayed by humans and animals. Colloquially, the term "artificial intelligence" is used to describe machines that mimic "cognitive" functions that humans associate with other human minds such as "learning" and "problem solving. Artificial Intelligence can be classified into three different types of systems: analytical, human-inspired, and humanized artificial intelligence. The traditional problems (or goals) of AI research include reasoning, [knowledge representation](https://en.wikipedia.org/wiki/Knowledge_representation), [planning](https://en.wikipedia.org/wiki/Automated_planning_and_scheduling), [learning](https://en.wikipedia.org/wiki/Machine_learning), [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [perception](https://en.wikipedia.org/wiki/Machine_perception) and the ability to move and manipulate objects.

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python's design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with its notable use of [significant whitespace](https://en.wikipedia.org/wiki/Off-side_rule). Its language constructs and [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aims to help programmers write clear, logical code for small and large-scale projects. Python is [dynamically typed](https://en.wikipedia.org/wiki/Dynamic_programming_language) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigm), including [procedural](https://en.wikipedia.org/wiki/Procedural_programming), object-oriented, and [functional programming](https://en.wikipedia.org/wiki/Functional_programming).

**1.2 Objectives of Research**

* To implement an algorithm for automatic classification of text into positive, negative or neutral.
* Sentiment Analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest.
* To provide a method for analyzing sentiment score in noisy twitter streams.
* To develop a program for customers review on a product which allows an organization or individual to sentiment and analyze a vast amount of tweets into a useful format.

**1.3 Problem Statement**

Twitter is a popular social networking website where users posts and interact with messages known as “tweet”. This serves as a mean for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on such tweets to gather insights into products or to conduct market analysis

Using social media, models are built for classifying “tweets” into positive, negative, and neutral classes. “Sentiment Analysis” is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as **opinion mining**, deriving the opinion or attitude of a speaker. Sentiment Analysis helps us to automatically transform unstructured information into structured data of public opinions about products, services, brands, politics or any other topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, product reviews, product feedback, customer service etc….

Deep learning is also known as deep structured learning or hierarchical learningis part of a broader family of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) methods based on artificial neural networks. Learning can be [supervised](https://en.wikipedia.org/wiki/Supervised_learning), [semi-supervised](https://en.wikipedia.org/wiki/Semi-supervised_learning) or [unsupervised](https://en.wikipedia.org/wiki/Unsupervised_learning) .Deep Learning is nothing is nothing but a part of machine learning. Deep Learning helps us to classify a review as positive, negative or neutral. The model will take whole review as an input and provide percentage ratings for checking whether the review conveys a positive or negative sentiment.

# 2. Review of Literature

**A. Opining Mining:**

Opinion mining refers to the broad area of natural language processing, text mining, computational linguistics, which involves the computational study of sentiments, opinions and emotions expressed in text .Although, view or attitude based on emotion instead of reason is often colloquially referred to as a sentiment. Opinion mining has many application domains including accounting, law, research, entertainment, education, technology, politics, and marketing. In earlier days many social media have given a web users avenue for opening up to express and share their thoughts and opinions.

**B. Twitter Sentiment Analysis:**

The sentiment can be found in the comments or tweet to provide useful indicators for many different purposes .Sentiment analysis is a natural language processing techniques to quantify an expressed opinion or sentiment within a selection of tweets. Sentiment analysis refers to the general method to extract polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases .There has two main approaches for extracting sentiment automatically which are the lexicon-based approach and machine-learning-based approach.

**C. Techniques of Sentiment Analysis:**

The semantic concepts of entities extracted from tweets can be used to measure the overall correlation of a group of entities with a given sentiment polarity .Polarity refers to the most basic form, which is if a text or sentence is positive or negative. However, sentiment analysis has techniques in assigning polarity such as:

1. Natural Language Processing (NLP**)**

NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules .Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level.NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world

1. **Case-Based Reasoning (CBR)**

Case-Based Reasoning (CBR) is one of the techniques available to implement sentiment analysis. CBR is known by recalling the past successfully solved problems and use the same solutions to solve the current closely related problems CBR does not require an explicit domain. This and the application of database techniques make the maintenance of large columns of information easier.

1. **Artificial Neural Network (ANN)**

Artificial Neural Network (ANN) or known as neural network is a mathematical technique that interconnects group of artificial neurons. It will process information using the connections approach to computation. ANN is used in finding the relationship between input and output or to find patterns in data.

**D. Application Programming Interface (API)**

Alchemy API performs better than the others in terms of the quality and the quantity of the extracted entities. As time passed the Python Twitter Application Programming Interface (API) is created by collected tweets [30]. Python can automatically calculated frequency of messages being re-tweeted every 100 seconds, sorted the top 200 messages based on there-tweeting frequency, and stored them in the designated database [12]. As the Python Twitter API only included Twitter messages for the most recent six days, collected the data needed to be stored in a different database

# 3. Data Collection:

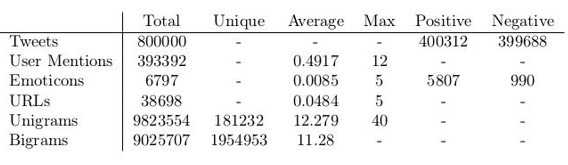
## REQUIREMENTS

* **Software:** Anaconda – Spyder.
* **Language:** Python
* **Modules Used:**
  + - Pickle
    - Nltk
    - Collections
    - Keras model import Sequential,load\_model
    - keras layers import Dense
    - numpy
    - sklearn.tree import DecisionTree Classifier
    - scipy.sparse import lil\_matrix
    - sklearn.naive\_bayes, import Multinomial NB
    - sklear ensemble import Random Forest Classifier
    - sklearn.feature\_extraction text import text Transformer

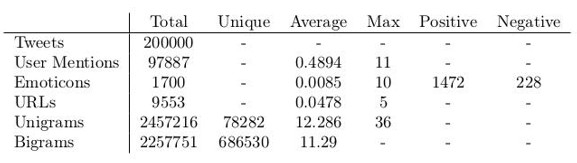
## DATASET DESCRIPTION

The data given is in the form of a comma separated values files with tweets and their corresponding sentiments. The training dataset is a csv file of type tweet id, sentiment, tweet where the tweet id is a unique integer identifying the tweet, sentiment is either 1 (positive) or 0 (negative), and tweet is the tweet enclosed in " ". Similarly, the test dataset is a csv file of type tweet id, tweet.

**Table1:** **Statistics of pre-processed train dataset**

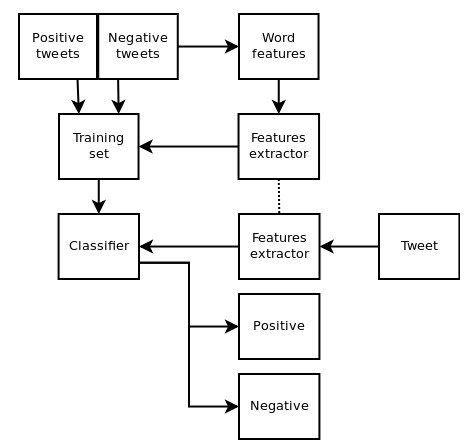
****

**Table 2: Statistics of pre-processed test dataset**

****

The dataset is a mixture of words, emoticons, symbols, URLs and references to people. Words and emoticons contribute to predicting the sentiment, but URLs and references to people don’t. Therefore, URLs and references can be ignored. The words are also a mixture of misspelled words, extra punctuations, and words with many repeated letters. The tweets, therefore, have to be pre-processed to standardize the dataset.

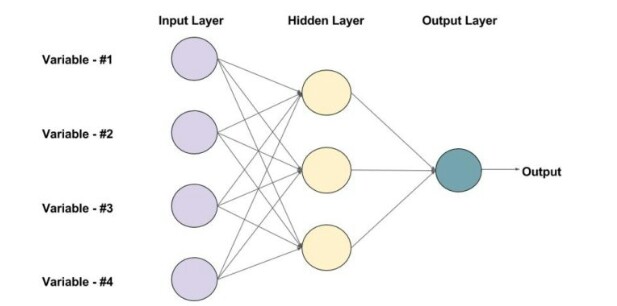
## PROPOSED METHOD AND IMPLEMENTATIONS

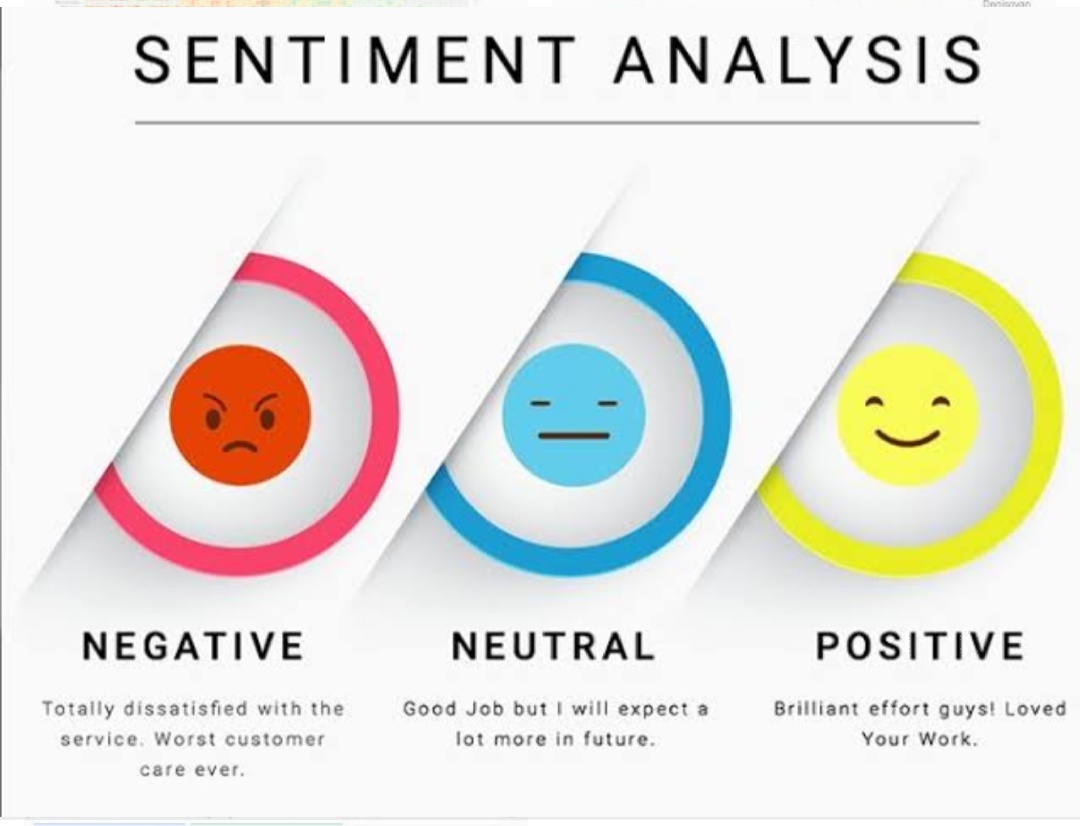


# 4. Methodology

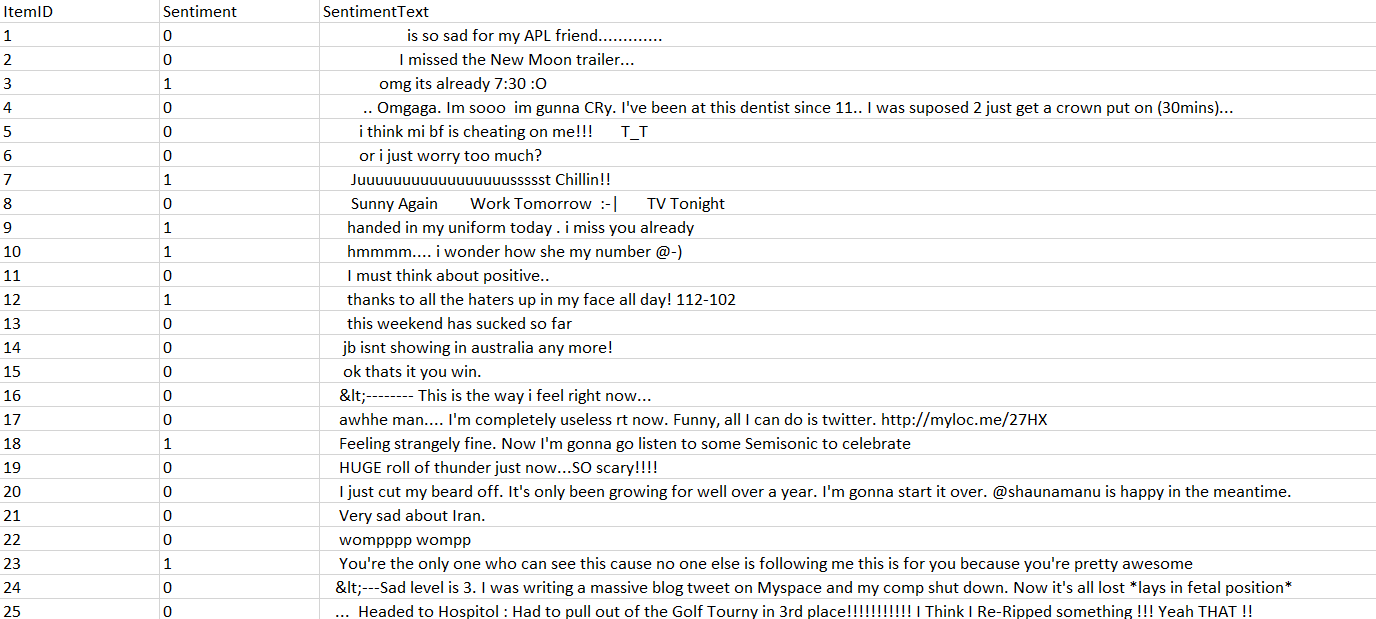
**4.1. Exploratory Data Analysis**

**4.1.1. Figures and Tables**

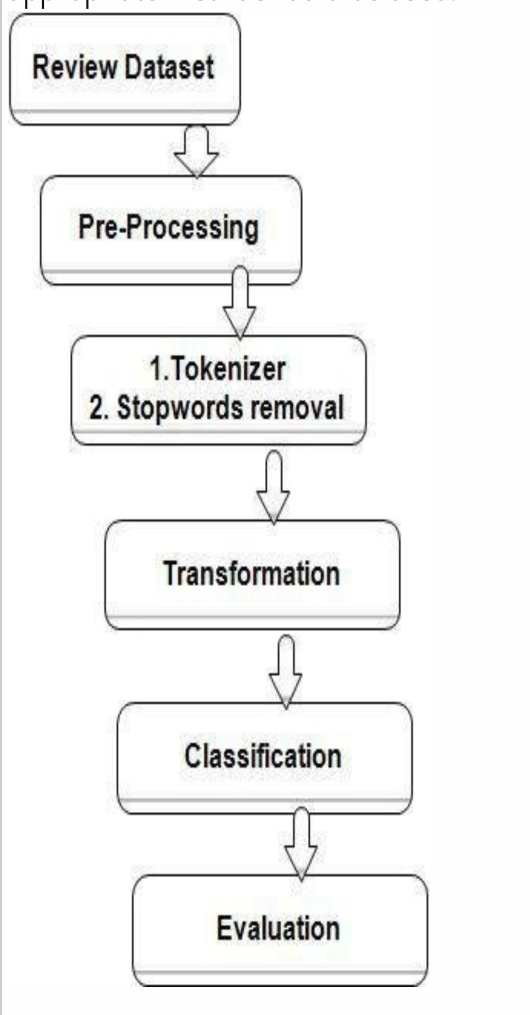
****

****

In the above figure we can see three expressions. “Negative” refers to sad expression, “Neutral” refers to neither positive nor negative expression and “Positive” refers to satisfactory expression.

****

In the above table we observe three columns. The 1st column indicates item ID, 2nd column indicates sentiment and the 3rd column indicates sentiment text. Based upon the sentiment text the sentiment will be either 1 or 0.

****

The above diagram shows the steps to evaluate Sentiment Analysis

**4.2. Data Modeling**

Basically, there are 4 parts to this:

1. Getting Twitter Data based upon hash tags
2. Training and saving the models (Word2Vec, TF-IDF and SVM model)
3. Using the model for Sentiment Classification
4. Using individual sentiments to do Topic modeling.

There are a few activities which need to be done once:

1 .under train \_models - execute train\_word2vec.py to train the model and save it in pickle format

2 .under train \_models - execute train\_tfidf.py to train the model and save it in pickle format

3 .under train \_models - execute train\_classifier.py to train the model and save it in pickle format

The Classifier accuracy is around 78% in test dataset.

Once, the above is completed, the models are ready to predict.

Keep running the twitter\_data.py in order to collect more samples of data.

Once, everything is done, run all\_together.py to classify the tweets into positive and negative sentiments and do a Topic modeling on each dataset separately.

# Findings and Suggestions

Our output will predict whether a given text is positive, neutral or negative sentiment.

* Some of the positive texts are:-

I am happy, I am a good girl, I like dancing etc….

* Some of the negative texts are:-

He is a bad boy, everything wrong etc….

If it is positive text then the output will be “True” and if it is a negative text then the output will be “False”

# Conclusion

**Twitter sentiment analysis** comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy, so that we can use this model for future use according to the results. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%. But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don’t perform well when the number of classes increased. Also it’s still not tested that how accurate the model will be for topics other than the one in consideration. Hence sentiment analysis has a very bright scope of development in future.