**TWITTER SENTIMENT ANALYSIS**

Twitter is a social-networking platform, on which registered users post or interact with other users with messages call as *“tweets”.* It was launched in July 2006. It gained worldwide popularity in 2012, with more that 100 million registered users. In 2018 it has more than 380 million users, with more than 600million tweets a day. These tweets are mostly the reviews, sentiments, views, news, analysis, etc. of millions of users. This project focuses on analyzing the sentiments of these tweets. This means analyzing the sentiments of the tweets to classify if they are either: positive, negative or neutral. Because of the large amounts of data available, we hope to analyze the sentiments with accurate results. These analysis reports can be used in many ways in our day-to-day life. For example, predicting political elections, know if a product of a certain brand is a hit or a flop, stock exchange, reviews of a certain movie, music, and many more. The project aims to build a functional classifier that can give accurate and automatic sentiment classification of any tweet.

We have chosen twitter data because this has better approximation of public reviews compared to the conventional internet reviews or blogs. Also, the response of public in twitter is more prompt compared to the response to blogs. This analysis plays an important role in socioeconomic phenomenon to predict the stock market of a firm. This is done by taking into account all the public sentiments towards that particular firm along with the use of economic tools that associate the relation between public sentiments’ and the firm’s stock market value. Firms can also use this analysis to know how their product is doing in the market, also to know which areas of the market the product is a hit, and which areas have a negative response. Predicting the popularity of a political party and the support a party has among the public is an emerging application of this sentiment analysis.

**NATURAL LANGUAGE PROCESSING:**

The project relies on *“Natural language Processing (NLP)”*. NLP helps in extracting features and patterns of data from large sets of data, for this project, large sets of twitter data. The project also relies on *“Machine Learning (ML)”*. ML is used to identify accurately unlabeled data samples (tweets).

**DEFINITION:**

Natural language processing is theoretically motivated range of computational techniques for analyzing and presenting naturally occurring text or speech at one or more level of linguistic analysis for the purpose of achieving human like language processing for a range of tasks or applications.

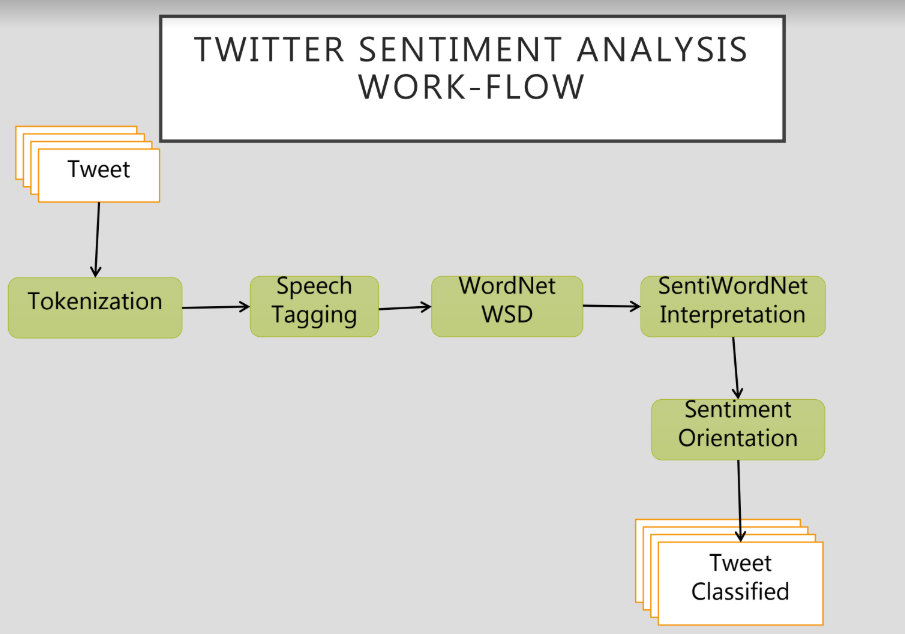
**NLTK:**

The **Natural Language Toolkit**, or more commonly **NLTK**, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics ,cognitive science ,artificial intelligence , information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

The features that the project uses to classification and modelling patterns can be divided into two groups: Bases on Formal Language and Based on Informal blogging. Formal Language features are that deal with formal linguistics and prior sentiment polarity and parts of speech of the sentence. Prior sentiment polarity: some words have the tendency for expressing a particular and specific sentiment. For example, the word *“excellent”* is a strong positive word whereas *“evil”* is a strong negative word. When a strong positive word is used in a sentence, chances are that the entire sentence has a positive sentiment. But twitter data features are more informal. This is because tweets relate with how people express their expressions, also they have to compress their expressions within the 280 characters that twitter provides. Also, these 280 characters may contain hashtags, emoticons, retweets, questions marks, exclamation marks, word lengthening, URL’s, shorthand, word capitalization.

Classification techniques are divided into two categories: Supervised vs. Unsupervised and Non-Adaptive vs Adaptive Techniques. Supervised Technique: When the data sample is pre-labeled, and we use this data to train the classifier. The pre-labeled data is used to extract features, patterns and differences between different classes and then classifying an unlabeled sample. Unsupervised Technique: When there no labeled data for training. Adaptive Classification Technique: Deals with feedback from the environment. In this project feedback from environment is the human telling the classifier if it has done a good or a bad job in classifying a particular tweet, and the classifier learns from this feedback.

Below mentioned figure shows the basic workflow for Twitter Sentiment Analysis:



**Pre-processing:**

Removing non English tweets. Remove URL, Target Mentions, Hash tags, Numbers. Replace seq of repeated characters eg: ‘cooool’ by ‘cool’. Remove Nouns and Prepositions

**Tokenization:**

The word tokenization is the process of splitting statement into words. It is a concept of natural language processing which breaks statements into smaller chunks. Analyzing, classifying and counting these word chunks, one can determine sentiments.

**Stop-word Filtering:**

To make the data machine understandable, it needs to be pre-processed, which includes removing of useless words, which are referred as Stop-Words. Stop-word is a word that a search engine is programmed to ignore both when indexing entries for searching and resulting them in the search enquiry.

**Negation Handling:**

The scope of negation may be limited only to the next word after a negation or may be extended up to other words following negation. Example: The Climate is not good; Machine doesn’t function properly. These examples show that the scope of negation is not fixed and varies based on different linguistics features such as conjunctions, punctuation marks and part of speech (POS) of negation etc. Negation handling is an automatic way of determining the scope of negation and inverting the polarities of opinionated words that are actually affected by a negation. The portion of the sentence that negation affects is called the vicinity or scope of negation.

**Stemming:**

A process of reducing a word to its word stem. Is important in Natural Language Understanding and Natural Language Processing. It actually means going back to the origin of the word.

**Speech Tagging:**

Automatic assignment of descriptors to the given tokens is called Tagging. The descriptor is called tag. The tag may indicate one of the parts-of-speech, semantic information, and so on. So tagging a kind of classification. Parts of speech tagging simply refers to assigning parts of speech to individual words in a sentence, which means that, unlike phrase matching, which is performed at the sentence or multi-word level, parts of speech tagging is performed at the token level.

**Word-Net:**

**WordNet** is a lexical database for the English Language. It groups English words  into sets of synonyms called sysets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license and are freely available for download from the WordNet website. Both the lexicographic data (*lexicographer files*) and the compiler (called *grind*) for producing the distributed database are available.

**Senti Word Net:**

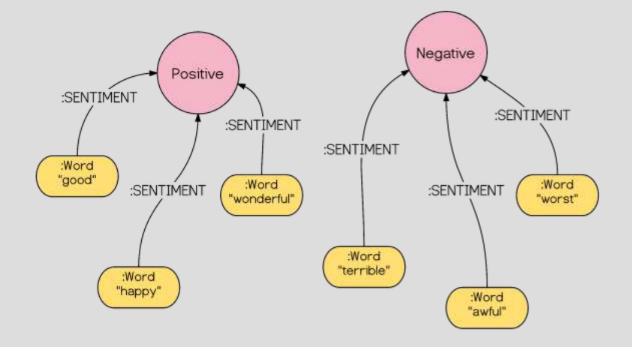
SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. It us an extended version of wordnet which has the sentiment of each word written.

SentiWordNet (SWN) is a global (or general) lexical re-

source [12, 18] developed in 2006. In SWN, each synset

(denoted by sn) of WordNet is associated with three numer-

ical scores, namely Obj(sn),Sub(sn), and Neg(sn). These

****

Examples:

