Twitter Sentiment Analysis

Final Report

Team Name : Data Pirates
Team Leader: Aman Singhal
Team Details:

Team Name		**Team 5**	
Sr. No	**Participant's Name**	**Email id**	**Phone Number**
1	Aman Singhal	amansinghal1108@gmail.com	9424590441
2	Kshitij Mankar	mankarkshitij@gmail.com	9588415170
3	Anshul Garg	anshul0771@gmail.com	8130735949
4	Pushpum Krishna	pushpum.krishna@gmail.com	8377972108
5	Vinamr Bajaj	lit2019023@iiitl.ac.in	9205022830
6	Vivek Babu Jacob	vivek1jacob@gmail.com	9847517865

Introduction ::

Sentiment analysis :: Sentiment analysis is a set of algorithms and techniques used to detect the sentiment (positive, negative, or neutral) of a given text. This is a very powerful application of NLP. Here are some other examples of use cases of sentiment analysis:

1. A stock investor scanning news about a company to assess overall market sentiment
2. An individual scanning tweets about the launch of a new phone to decide the prevailing sentiment
3. A political party analyzing social media feeds to assess the sentiment regarding their candidate

Problem Statement ::

The proposed Project is Twitter Sentiment Analysis using Natural Language Processing (NLP) . The goal of the project is to classify a given tweet/message to whether the message is of positive, negative, or neutral sentiment.

Approach ::

Using Machine Learning models

The machine learning approach tackles the problem as a text classification task employing classifiers after pre-processing the data.

Models Considered

1.)Naive Bayes

2.)Linear Regression

Various Difficulties ::

Cleaning the tweets-

The tweets given in the data are usually not clean or in a desired structure. So the main difficulty is to clean the data .

Word Ambiguity

Word ambiguity is another pitfall we are facing working on the sentiment analysis problem.The problem of word ambiguity is the impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context.

Processes Involved ::

1. Pre-processing Tweets

This is one of the essential steps in any natural language processing (NLP) task. Following processes were used to clean the tweets.

- **Letter casing**: Converting all letters to either upper case or lower case.
- **Tokenizing**: Turning the tweets into tokens. Tokens are words separated by spaces in a text.
- **Noise removal**: Eliminating unwanted characters, such as HTML tags, punctuation marks, special characters, white spaces etc.
- **Stopword removal**: Some words do not contribute much to the machine learning model, so it's good to remove them. A list of stopwords can be defined by the nltk library, or it can be business-specific.

2. Stemming

It may be defined as the process to remove the inflectional forms of a word and bring them to a base form called the **stem**. The chopped-off pieces are referred to as affixes. The two most common algorithms/methods employed for stemming include the ::

- Porter Stemmer
- Snowball Stemmer

We will be using Porter Stemmer in our process.

3. Lemmatization

It is a process wherein the context is used to convert a word to its meaningful base form. It helps in grouping together words that have a common base form and so can be identified as a single item. The base form is referred to as the lemma of the word and is also sometimes known as the dictionary form.

The most commonly used lemmatizers are the

- WordNet Lemmatizer
- Spacy Lemmatizer
- TextBlob Lemmatizer

We will be using WordNet Lemmatizer in our process.

4. Vectorization

Processing natural language text and extract useful information from the given word or a sentence using machine learning and deep learning techniques requires the string/text needs to be converted into a set of real numbers (a vector) — **Word Embeddings**.

Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.

The process of converting words into numbers are called Vectorization