Tweet Sentiment Extraction - Extracting support phrases for sentiment labels

Problem Statement

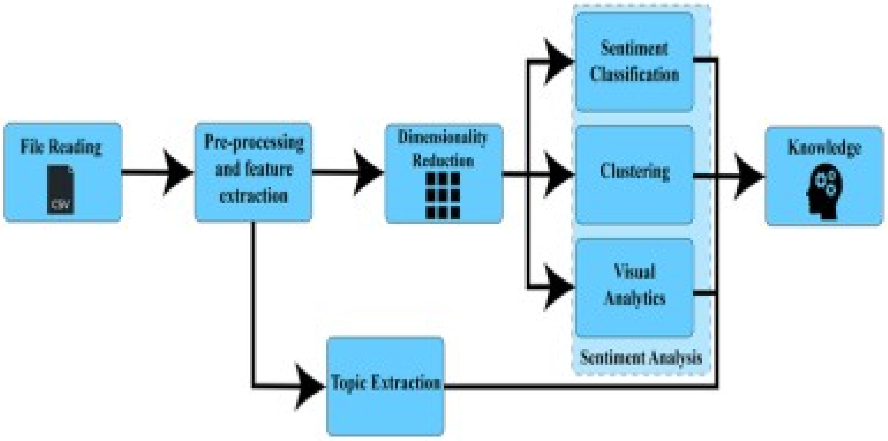
The objective of this project is to look at the labeled sentiment for a given tweet and figure out what word or phrase from the tweet best supports it. With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person's, brand for being viral (positive), or devastate profit because it strikes a negative tone. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds. But, it is difficult to find words that actually lead to the sentiment description.

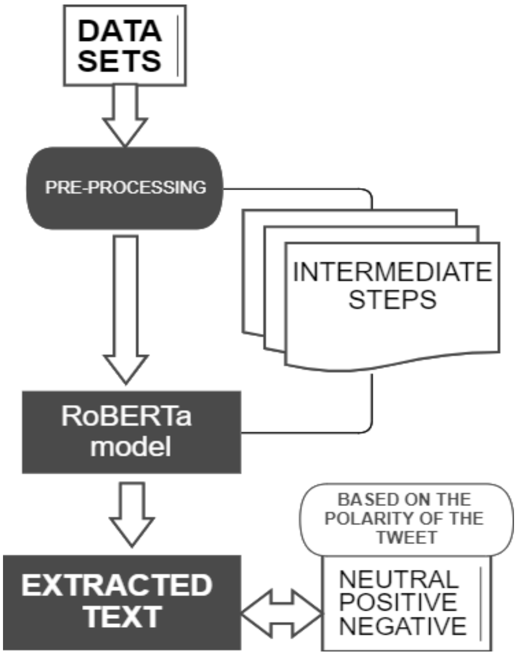
Background

Twitter is a popular microblogging service in which users post status messages, called “tweets”, with no more than 140 characters. Twitter represents one of the largest and most dynamic datasets of user-generated content — approximately 200 million users post 400 million tweets per day. Tweets can express opinions on different topics, opinions concerning brands and products, outbreaks of bullying, events that generate insecurity, predict polarity in political and sports discussions, and acceptance or rejection of politicians, all in an electronic word-of-mouth way. Automatic tools can help decision-makers to ensure efficient solutions to the problems raised. The focus of our work is on the sentiment extraction of tweets.

Sentiment extraction aims at extraction of phrase which decide the emotion or sentiment of the sentences. It also determines emotions, and attitudes reported in source materials like documents, short texts, sentences from reviews, blogs, and news and among other sources. In such application domains, one deals with large text corpora and most often “formal language”. Twitter users post messages on a variety of topics, unlike blogs, news, and other sites, which are tailored to specific topics.

Methodology





Dataset

The datasets used in the Roberta model is retrieved from the Kaggle Tweet Sentiment Extraction Competition. The dataset is in the form of .csv format which can be directly used in the model. There are three sets of data, one for training the model, one for testing the model, and one for submission of the data for the evaluation. The training dataset consists of four columns which are textID, text, selected_text, and the sentiment value and the test dataset consists of three columns which are text_Id, text, and sentiment value and the submission dataset consists of the textId and the selected_text which is obtained as the output. The text Id of the dataset is the unique value of the data and the text is the normal text from which you must extract the tweet sentiment and the selected_text is the desired output that the model must interpret and the sentiment is the polarity of the tweet. The link for the dataset:
https://www.kaggle.com/c/tweet-sentiment-extraction/data

Evaluation Measures

The metric used in the model is the word-level Jaccard score.
Jaccard Similarity over the union is defined as the size of the intersection divided by the size of the union of two sets. we will first perform lemmatization to reduce words to the same root word, to calculate similarity using Jaccard similarity, In our case, “friend” and “friendly” will both become “friend”, “has” and “have” will both become “has”. The formula for calculating the Jaccard score:
score=(1/n)∑i=1 to n (jaccard (gti, dti))
where n is the number of documents, Jaccard is the function provided above, gti is the ith ground truth and dti is the ith prediction

Software and Hardware Requirements

1. Pytorch
2. Fastai framework (optional)
3. Numpy, Pandas Python libraries
4. GPU/TPU support (Google Colab/Kaggle)