User Manual

Contents

Introduction

Social media have received more attention in the recent times. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media like Facebook, Twitter, Instagram etc. Social media analysis is the study of people's interactions and communications on different topics and nowadays it has received more attention. Opinion and sentiment mining is an important research area as people spend hours daily on social media and share their opinion, which can help in deriving valuable insight.

Problem Statement

Millions of people share their opinion about different topics on a daily basis on many social media platforms. It is difficult to gather data from different media and aggregate for opinion analysis as each platform has different ways to express opinion. For example, people express in the form of article and pictures mainly in Facebook, Pictures in Instagram, Twitter is mainly focused on text format. Due to these reasons it has increasing become challenging to gain insights and analyze public perspective about a topic.

Problem Solution

In this project, we chose Twitter, as it is one of the widely used social media, in which, people express their opinion in form of text mainly. Twitter is rich source to gather people's opinion on social movements and perform sentiment analysis. We collected tweets related to #MeToo and generated sentiment for each tweet. Finally, we derived the insights like trend of movement in twitter.

Requirements

Functional Requirement

- Retrieve data related to social issues
- Perform rule/pattern mining
- Perform data aggregation from different sources
- Perform classification (building dictionary for custom analyzer)
- Perform comparative analysis (third party tool vs custom analyzer)
- Enrich the data sets with information on
    - Demographics (gender, age)
    - Geospatial classifications (country based)
- Developing solutions for the insights

Non-Functional Requirement

- Find the accuracy for data
- Find the precision for custom model

Techniques Used

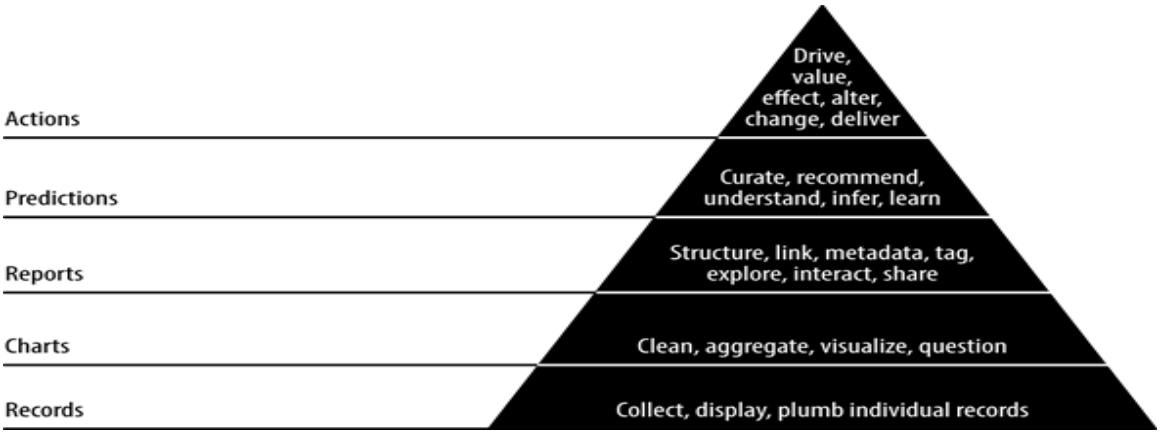
Interviewing:
Interviews are the best solution for data science projects. Everyone has different level of experience in data science concepts. So, we can share with everyone and we can know every individual insight of viewing and analyzing data. In our part, we had bi monthly client meetings to gather outcomes from client based on which we developed our model.

Analyzing Existing Documents:
Analyzing existing documents can prove to be a useful technique in requirement gathering, on its own as well using it to supplement other techniques. Reviewing the current process and documentation can help the analyst understand the business, or system, and its current situation. We found many documents in google which helped us a lot. This will help the analyst formulate questions for interviews or questionnaires to ask of Client, to gain additional requirements. If an analyst is uncertain why certain procedures are in place, this can also help the analyst in asking these questions during interviews. When studying the requirements, the analysts may find problems that they may distinguish on their own. They may also find redundancy, in which steps are unnecessarily repeated. Principles and rules for the organization itself can be discovered by analyzing documents. Analyzing documents can be used as a supplement to information obtained from interviews, questionnaires, and observations. For example, if some of the interview answers are unclear, organizational documents may help in making sense of some of the interviewee's answers. Reviewing existing documents may also assist in understanding why a user performs certain tasks while observing them. A drawback to analyzing documents is that documents may be outdated.

Methodology

Any Data science project needs attentive consideration and result evaluation in the context it is used because the extracted knowledge is significant to assist the decision-making process in an application. Below is the data science methodology, that we as a team have followed, to find the sentiment of the tweet and derive some analysis.

Starting for the bottom up, we followed each step-in order to find the sentiment of text based on a tweet.



Records:
The initial step is data collection. We have collected data for our analysis in two types. The first type is stream data where we retrieved data using python library tweepy. The second type we used to retrieved data is static type where we have retrieved almost 1350k tweet form DataWorld website. After retrieving the data by these two types we stored it in excel and made it ready for performing analysis.

Charts:
After retrieving the data that we need and storing it in excel, we had to perform data cleaning on the retrieved data. To do the preprocessing, we have used python re library, which is a special sequence of characters that helped us match or find other strings or sets of strings, using a specialized syntax held in a pattern. Then we aggregated the cleaned data and visualized it to remove any special characters or symbols and finally stored it in the excel.

Actions:
Our plan to find the sentiment on the text based off on a particular tweet is in two ways. We have calculated the sentiment of a tweet using a python's third-party library, TextBlob. It classifies text into positive, negative and neutral. It gives output in two values, polarity and subjectivity which deals with emotion for a text and opinion respective to topic respectively.

The other way we did sentiment analysis is by designing our own custom analyzer. The process is:

- We started off taking each tweet and remove stop words
- Stem each word
- Add words to the dictonary
- Classify each word as positive, Negative and Neutral.
- Generate overall sentiment of each tweet

Tools and Technology

Python

Python is an interpreted high-level language generally used for programming. It is a multi-paradigm programming language, which supports object-oriented programming and structured programming. Python is easily readable, unlike many other programming languages, it does not use curly brackets to delimit blocks, and semicolons after statements are optional. Python uses whitespace indentation, rather than curly brackets or keywords, to delimit blocks. One of the best things about python is its large standard library, it provides tools suited to many tasks.

Tweepy

Tweepy is open-sourced, hosted on GitHub and enables Python to communicate with Twitter platform and use its API. Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication, so OAuth is now the only way to use the Twitter API.

TextBlob

It is a python library for processing textual data. It provides APIs for some simple tasks like parts-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more. Some other features include Tokenization, spelling correction, and add new models or languages through extensions. We have checked several packages like TextBlob, ParallelDots and found out that Text Blob is the best option based on the accuracy and some other factors like polarity and subjectivity.

Natural Language Toolkit

NLTK is a leading platform for building python programs to work with human language data. It gives easy to use interfaces such as wordnet with some text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Some things that we can do with NLTK are:
Tokenize and tag some text, identify named entities, Display a parse tree.

NLTK Stop words

The main idea of Natural Language Processing is to do some form of analysis, or processing, where the machine can understand, at least to some level, what the text means, says, or implies.
We can recognize ourselves that some words carry more meaning than other words. We can also see that some words are just plain useless and are filler words. Some words mean nothing, unless we are searching for someone who is maybe lacking confidence or hasn't practiced much speaking. For most analysis, these words are useless. We call these words "stop words" which contain no meaning, and we want to remove them. We can do this easily, by storing a list of words that we consider to be stop words. NLTK starts off with a bunch of words that they consider to be stop words, we can access it with the NLTK corpus with: *from nltk.corpus import stopwords*

NLTK Word tokenize

Tokenization is a way to split text into tokens. These tokens could be paragraphs, sentences, or individual words. NLTK provides several tokenizers in the tokenize module.

GitHub

GitHub is a web-based version-control and collaboration platform for software developers. Microsoft, the biggest single contributor to GitHub, initiated an acquisition of GitHub for $7.5 billion in June 2018. GitHub, which is delivered through a software-as-a-service (SaaS) business model, was started in 2008 and was founded on Git, an open source code management system created by Linus Torvalds to make software builds faster.

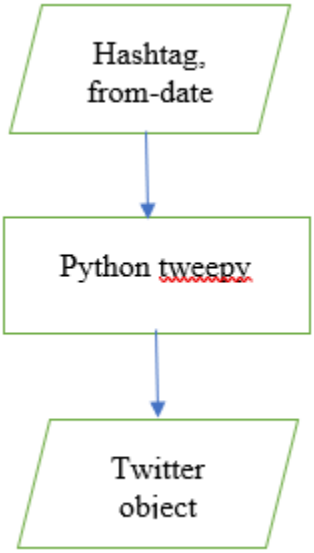
Git is used to store the source code for a project and track the complete history of all changes to that code. It allows developers to collaborate on a project more effectively by providing tools for managing possibly conflicting changes from multiple developers. GitHub allows developers to change, adapt and improve software from its public repositories for free, but it charges for private repositories, offering various paid plans. Each public or private repository contains all of a project's files, as well as each file's revision history. Repositories can have multiple collaborators and can be either public or private.

Data Collection

In this project, we used two types of data related to metoo hashtag. Static data, which is time series metoo dataset we gathered from data.world website. Stream data, which is stream of tweet objects. Metoo data set is collection of around 350,000 tweets related to metoo form October 2017 to December 2018. The data was in the form of Excel file with multiple attributes like tweet id, text, handle etc.

We retrieved stream of tweets using twitter's public python API tweepy. In order to use twitter API for data retrieval, one should have a developers account in twitter, with which we get authentication information: API key, API secret, Access token and Access token secret. Each tweet is a json object with multiple attributes. We retrieved only tweet id and tweet text attributes for our analysis and stored in excel file.

Data collection flow chart: (parellogram- input/output, Rectangle – process)

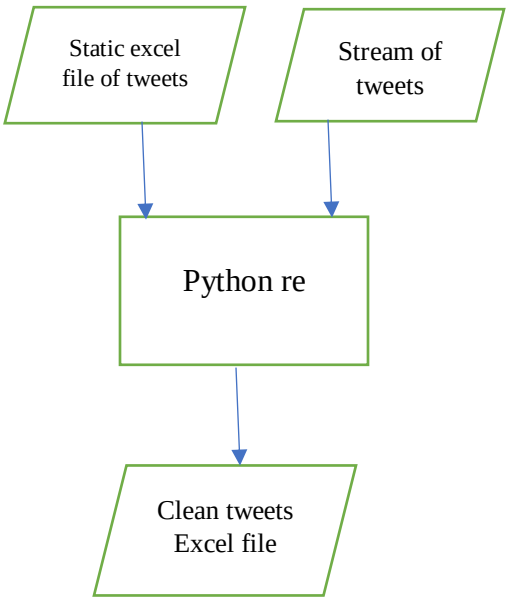


Data Preprocessing

Pre-processing is fundamental to all Natural Language Processing (NLP) Task. Steps needed for pre-processing of text in general depends on the targeted requirement or application. Maximum length of Twitter message is 140 characters, which may include user's actual message, hash tag, URL, emoticons, etc. The embedded URL is usually used to give the source for the detailed description of the content mentioned in the text. In addition, emoticon emphasis the emotion of the user.

In our approach, the description of the web content is not taken to analyze the sentiments of tweets. Hence all such web directions are removed from the tweets by identifying such patterns. To identify such patterns, we used re (regular expression) library in python. Once we remove the unwanted data or noise present in the data we are storing processed tweets in an excel file.

Data preprocessing flow chart: (parellogram- input/output, Rectangle – process)


Static excel
file of tweets
Stream of
tweets
Python re
Clean tweets
Excel file

Sentiment Analysis

Social networks are a rich platform to learn about people's opinion and sentiment regarding different topics as they can communicate and share their opinion actively on social media including Facebook and Twitter. There are different opinion-oriented information gathering systems which aim to extract people's opinion regarding different topics. The sentiment-aware systems these days have many applications from business to social sciences. Since social networks, especially Twitter, contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text.

Methods for Sentiment Analysis

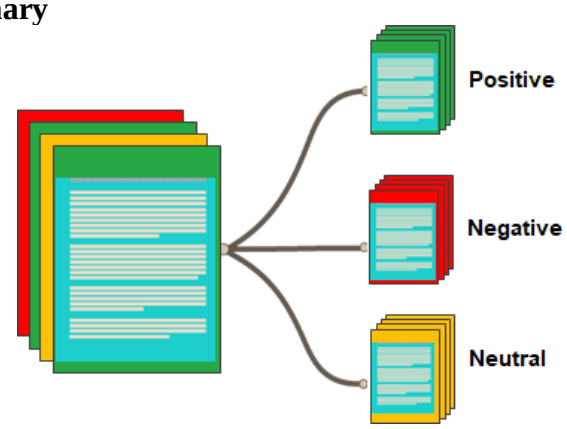
TextBlob:

To perform sentiment analysis, we have chosen two distinct methods. One of them is a third party tool called TextBlob. A detailed description of choosing TextBlob has been outlined in the results section. The second method for analysis is Custom Analyzer developed as per the custom requirements.

Custom Analyzer:

Custom analyzer has been developed after a detailed study of NLTK and its limitation. It has been developed in two stages.

Step 1: Building Dictionary



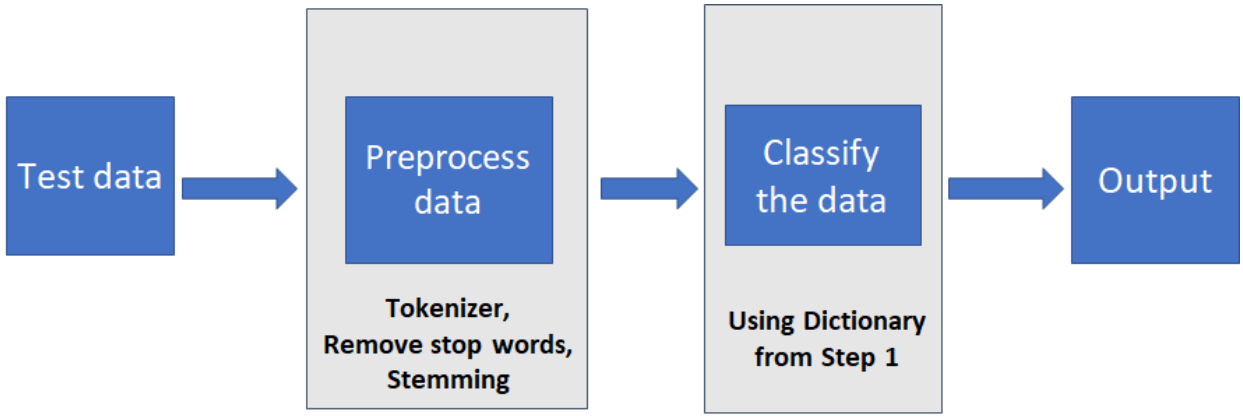
The main goal is to build a dictionary with a number of positive, negative and neutral analyzed both manually and by a third-party library. Natural Language Toolkit is the main backbone for building a dictionary here. It is a python package that can be installed before performing the analysis.

The approach to extract sentiment from tweets is as follows from the dictionary:

1. Start with downloading and caching the sentiment dictionary
2. Download twitter testing data sets, input it into the program.
   Input: It is beautiful day to go for fishing
3. Tokenize each word in the dataset and feed in to the program.
   ['it', 'is', 'beautiful', 'day', 'to', 'go', 'for', 'fishing']
4. Clean the tweets by removing the stop words.
   ['beautiful', 'day', 'go', 'fishing']
5. The multiple forms of each word are counted as one word using stemming.
   ['beautiful', 'day', 'go', 'fishing']
   ['beauti', 'day', 'go', 'fish']
5. For each word, compare it with positive sentiments and negative sentiments word in the dictionary. Then increment positive count or negative count.

Step 2: Text Classification

We have a definite model built from the stage one. We train the custom analyzer through many of the ways with a lot of datasets. The test data is let to preprocessed in various steps as discussed above: tokenizing, removing stop words, stemming.



Result Analysis

Selection criteria for third party tool

With availability of many sentimental analysis tools at hand, there was a need to choose the best suitable for our project. The criteria to select the best tool was to identify a tool that matches closest with a manually analyzed result. To perform this validation, a set of 10 random tweets were considered and first evaluated manually to set the benchmark. The same 10 tweets were then analyzed through each of the sentimental analysis tools. The results are tabulated as follows.

S. No	Text	Manual	Textblob	ParallelDots	Aylien
1	I love this sandwich	Positive	**Positive**	**Positive**	**Positive**
2	This is an amazing place!	Positive	**Positive**	**Positive**	**Positive**
3	I feel very good about these beers	Positive	**Positive**	**Positive**	**Positive**
4	This is my best work	Positive	**Positive**	**Positive**	**Positive**
5	What an awesome view	Positive	**Positive**	**Positive**	**Positive**
6	Tomorrow is Wednesday	Neutral	**Neutral**	**Neutral**	**Neutral**
7	I am tired of this stuff	Negative	**Negative**	**Negative**	**Negative**

8	I can't deal with this	Negative	**Negative**	**Negative**	**Negative**
9	He is my sworn enemy!	Neutral	**Neutral**	**Negative**	**Positive**
10	I am here	Neutral	**Neutral**	**Neutral**	**Neutral**

The closest match to the benchmarked output was that of **TextBlob**. Moreover, we need a model that can predict the nature of the tweet in simple metrics with the best possible match to the manually obtained result. TextBlob can be customized based on the classification requirement. The additional features that TextBlob has are language translation and detection. Hence, the choice for a sentimental analysis tool was TextBlob.

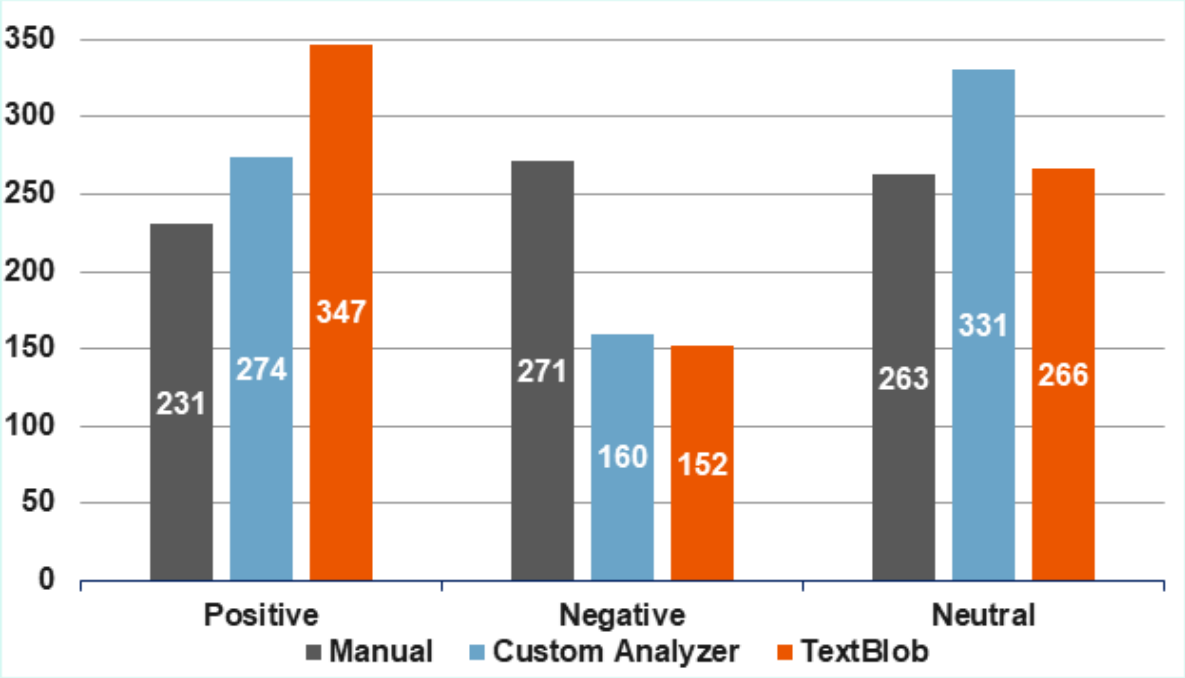
Analysis

The project has been divided into two phases for the better understanding and drawing clearer conclusions from the study.

Phase I:

To perform sentiment analysis, the major component of our project, we had validated the results of a dataset of 765 records in weather domain. The collection of 765 tweets has been used as benchmark for our project. This dataset contains 11 attributes out of which 2 attributes are most important for our project. They are tweet data and tweet description. These tweets have been manually validated to find out the sentiment of the tweet. The results of the manual validation are noted.
The same set of tweets have been analyzed through TextBlob. TextbBlob classified the number of tweets as positive, negative and neutral.

Through the developed custom analyzer, this dataset has been analyzed and the number of positive, negative and neutral tweets are noted. Below is a graph of the detailed count of the number of tweets analyzed positively, negatively and neutral.



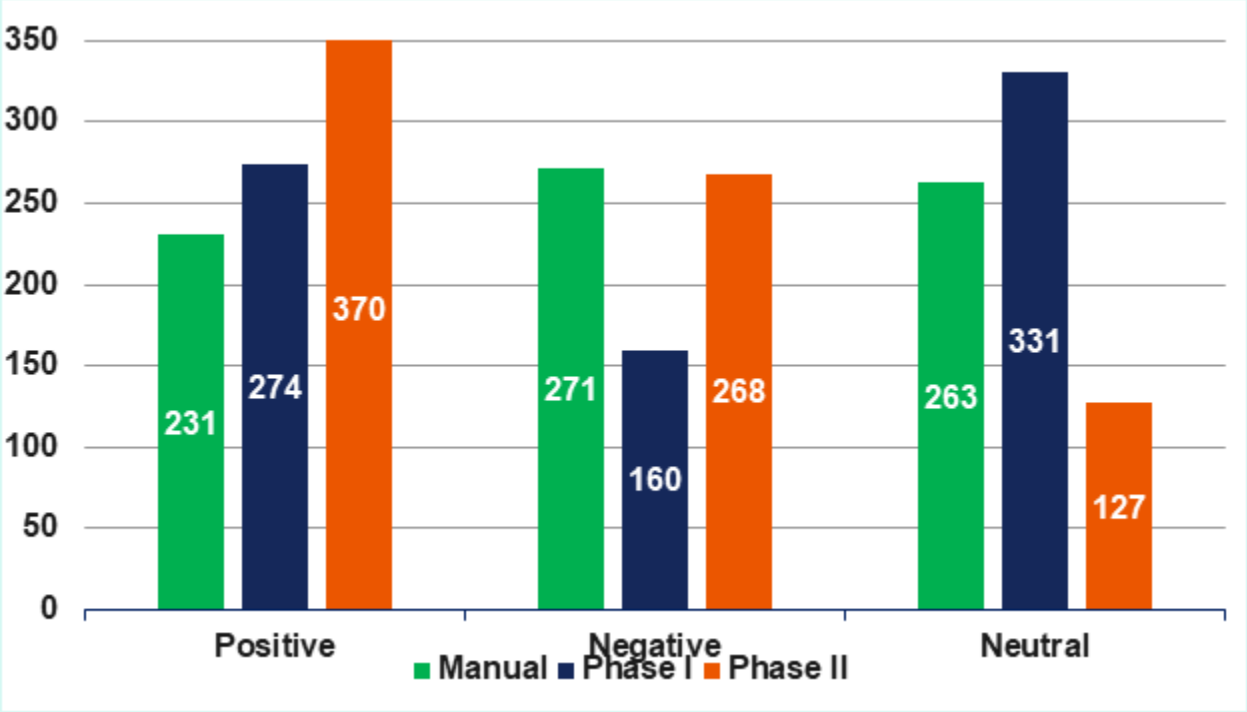
We observe the difference in number of tweets analyzes sentiment wise. The closest to the manual analysis is the number of words matched through custom analyzer. The least difference observed for positive tweets is for manual and custom analyzer. Custom analyzer and TextBlob have analyzed negative number of tweets equally.

Accuracy of the custom analyzer is necessary to predict the further use of the analyzer for other purposes. The accuracy is calculated after matching the number of pass and fail cases. To calculate the accuracy of custom analyzer, we consider pass as a case where the manual result matches with Custom Analyzer and TextBlob. We consider fail as a case where the manual result doesn't match with Custom Analyzer and TextBlob. Thus, the accuracy of the custom analyzer is 52%. The accuracy of TextBlob is 48%.

Phase II:
The main objective of the project is to improve the accuracy of the custom analyzer. To do this, we add more words to the existing collection of words. Refer to the bibliography section to know the source of the words added. Below is a tabulated version of the number of words added to the dictionary for the project.

	Positive Words	**Negative Words**	**Neutral Words**	**Total Words**
Phase I	398	517	38644	39,559
Phase II	792	1081	39709	41,582

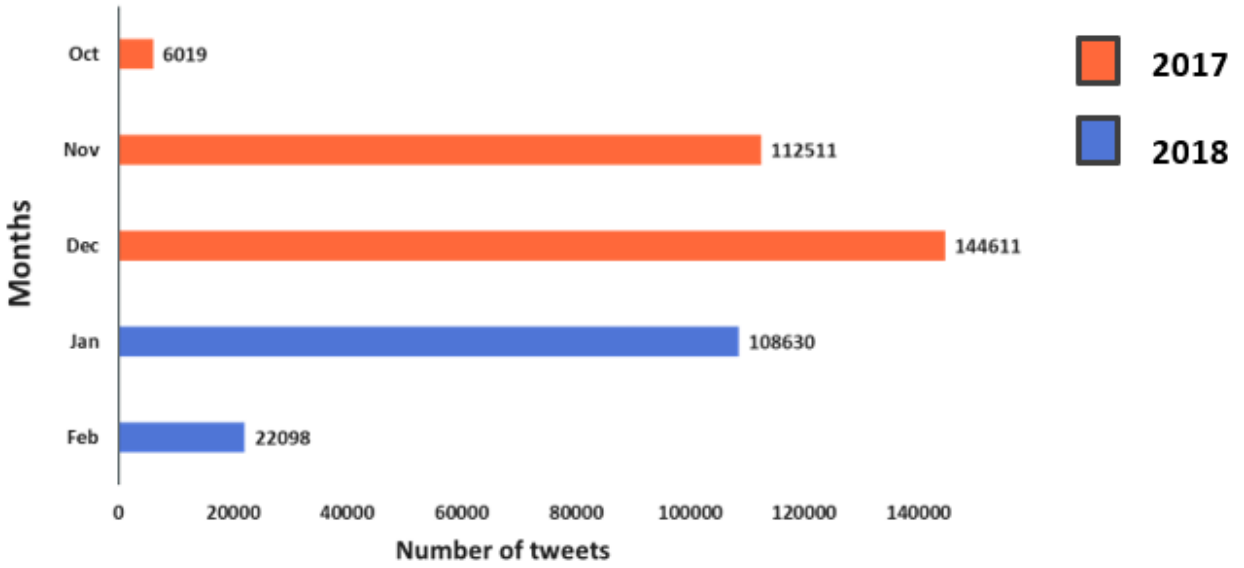
After adding words to dictionary, the analysis performed again manually, using TextBlob and Custom Analyzer on the weather dataset. The results do not differ for manual analysis and for the analysis done by the custom analyzer.
Plotted is a graph with the change in the sentiment analysis observed with respect to Phase I and Phase II when analysis is performed by Custom Analyzer.



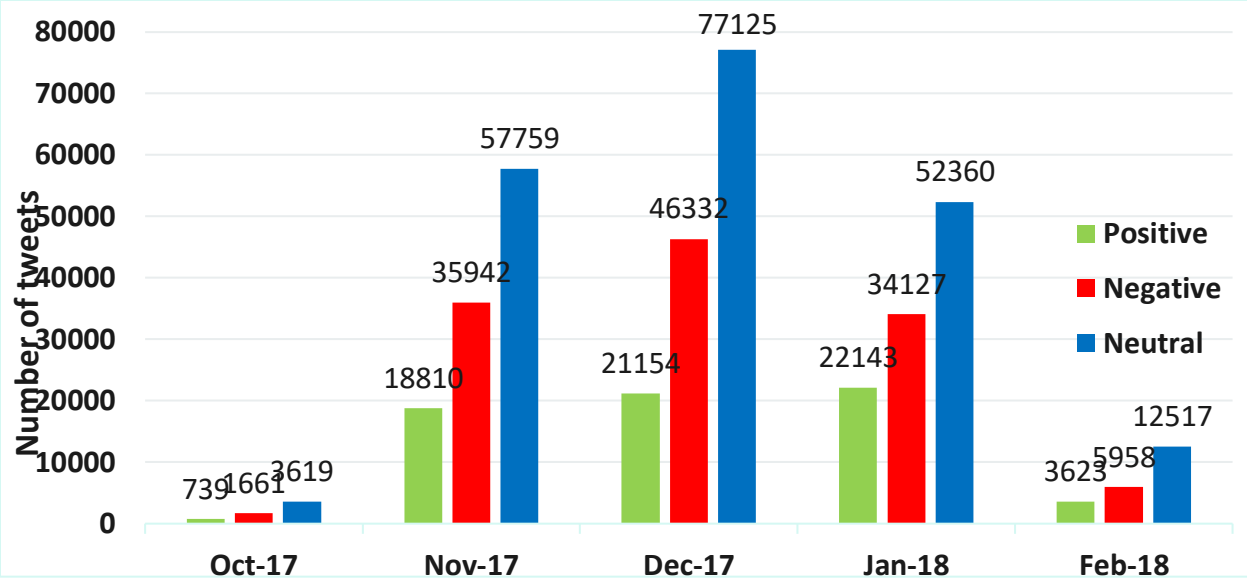
After improvising the dictionary, the accuracy of the custom analyzer is calculated. We see that the change in number of words, changes the count of the positive, negative and neutral tweets. The pass cases and fail cases also change.
Here, the number of pass cases is 53%. It means the 53% of the manual results match to the results analyzed by the Custom Analyzer. Hence, we can conclude that the accuracy of the custom classifier has been increased.

MeToo Analysis

There is an important reason to study the social movements and how people react towards them on social media. One such study is on #MeToo, the most active social movement in today's times. The Me Too movement (or #MeToo movement), with many local and international alternative names, is a movement against sexual harassment and sexual assault. We have a collection of 393,869 tweets on this issue. These tweets have been collected over a period of 5 months from October 2017 to February 2018. The outcomes of have been plotted as below.



This data set was analyzed by the custom analyzer to find out how many tweets have been classified as positive, negative and neutral. The results can be seen below.



Validation

Precision

While precision is a measure of the variation among survey estimates, over repeated application of the same sampling procedures, accuracy is a measure of the difference between the survey estimate and the true value of a sentiment parameter.

		Predicted	
		Positive	***Negative***
Actual	***Positive***	**True Positive**	**False Negative**
	Negative	**False Positive**	**True Negative**

Precision = True Positive / (True Positive + False Positive)

While precision is a measure of the variation among survey estimates, over repeated application of the same sampling procedures, accuracy is a measure of the difference between the survey estimate and the true value of a sentiment parameter.

The results observed have three different parameters: positive, negative, neutral.

Since precision deals with only two parameters, we have the following considerations to calculate precision of our model. The dataset contains 765 unique records on tweets with the results analyzed as positive, negative and neutral.

Case 1: Neutral as Positive
In this case, we have considered the count of neutral tweets as positive. In doing so, we evaluate the precision parameters as follows:

	TP	FP	Precision
Custom classifier	428	177	0.707438
TextBlob	447	166	0.729201

- Precision of Custom Analyzer is 70.74% when we consider neutral tweets as positive tweets.

- Precision of TextBlob is 72.92% when we consider neutral tweets as positive tweets

Case 2: Neutral as Negative

In this case, we have considered the count of neutral tweets as negative. In doing so, we evaluate the precision parameters as follows:

	TP	FP	Precision
Custom classifier	151	68	0.689498
TextBlob	183	95	0.658273

- Precision of **Custom Analyzer** is **68.94%** when we consider **neutral tweets** as **negative** tweets.
- Precision of **TextBlob** is **65.82%** when we consider **neutral tweets** as **negative** tweets.

Limitations

1. Data retrieval process is only possible in twitter as other social media have security restrictions.
2. Twitter allows to retrieve only past 7 days of data from the current date.
3. No access to demographic data like age of the person, location, gender
4. Unable to find proper hashtag related dataset for analysis.
5. To retrieve historic data is very expensive.
6. Even storage of retrieved data is an issue.
7. Huge amount of storage capacity is required to store historic data.
8. Texts containing redundant data which is not required for doing any analysis.
9. Emotion for a tweet cannot be determined with numerical data.
10. Existing packages for doing sentiment analysis are not accurate.
11. As many packages and API's are available, Choosing the proper package is difficult.
12. Availability of data.
13. Unable to retrieve data from tweetID attribute.
14. More neutral words
15. Unable to retrieve gender or place for a tweet.
16. Improving accuracy of our custom classifier.
17. As neutral words are High, unable to get accurate results.

Conclusion

Twitter is a major social media platform that has experienced tremendous growth in communication globally. However, the clarity of these insights and the effectiveness of the derived sentiment information when applied are critically dependent upon the underlying approach and its ability to accurately evaluate the opinions expressed by users. In this project, we presented results for sentiment analysis on Twitter. We use two classification tasks: One is existing third-party API and other is custom analyzer. We presented a comprehensive set of experiments for both these tasks on manually validated data, static dataset of tweets and stream of tweets. Our goal was to develop custom analyzer to process tweets and improve model accuracy. We tentatively conclude that our model performs better as per our work for sentiment analysis for #MeToo twitter data. In future work, to improve the model accuracy, it will be good to explore even richer linguistic analysis, for example, topic modeling, adding n-grams and using better model for classification.

References

1. Zimbra, David & Abbasi, Ahmed & Zeng, Daniel & Chen, Hsinchun. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. ACM Transactions on Management Information Systems. xx, No. x. 10.1145/3185045.
2. Shravan I.V. (2016). Sentiment Analysis in Python using NLTK http://opensourceforu.com/2016/12/analysing-sentiments-nltk/
3. Weather dataset from data world website- https://data.world/crowdflower/weather-sentiment
4. #MeToo dataset - https://data.world/rdeeds/350k-metoo-tweets