#MeToo Analysis.
How can we spread awareness?

Big Data Lab 2 Project Report

CMPT 733

Project Members:

Ipsita Dey (301301082)

Martha Garcia (301361939)

Saumya Dwivedi (301367213)

Project Supervisor

Prof. Jiannan Wang

Prof. Steven Bergner

April 14, 2019

Department of Computing Science

Simon Fraser University

Motivation and Background

Social Network Sites has become a huge platform today for people to share their opinions,experiences and interests. It has thus also become a source for unlimited amount of data. For this very reason Social media analytics has become a trending research area. Most of these projects look for patterns and insights to build solutions which would help promote a product,service or increase the value of a business in general.
We were motivated by the immense amount of scope present in this area but we also wanted to do something impactful, therefore we decided to analyse the #MeToo dataset

The term MeToo was coined by Tarana Burke for the first time in 2006. She could not respond to a 13-year-old survivor who had confided in her about a difficult situation. She said she wished that she had simply told the girl, "Me too" and henceforth she used this term for the campaign to promote "empowerment through empathy" for sexual harassment survivors of underprivileged communities and women of color in America.

The hashtag #MeToo went viral when Alyssa Milano's tweeted it in 2017.She posted her story and encouraged fellow victims to tweet their own story under this hashtag so that people could be made aware of the magnitude of the problem and the people who had suffered could feel that they were not alone.The hashtag #MeToo went viral overnight. It was tweeted more than 500,000 times in 24 hours. The noteworthy volume and virality of this dataset was another reason we decided to on this topic.

MeToo has today become a huge movement against sexual harassment faced by women.
Leading news articles and surveys quote that more than 80% of women in the world have experienced workplace sexual harassment but it is still one of the most tabooed subject in our society thus incidents of this nature remain to be some of the most unreported crimes.
Thus we wanted to contribute to this movement in our small way by exploring the data and using hybrid text-based and community-based method find trends of how these incidents are impacting our society; have they increased or decreased over time.
With this project we hope to help police or government agencies who monitor and work on helping victims of sexual abuse.

Problem Statement

MeToo dataset has a lot of unstructured data. By analysing the data and finding insights we wanted to spread awareness and pass helpful information to police and government agencies who were working or reducing such crimes.
Using text mining techniques we wanted to answer questions like:

- Can we identify fake vs real accounts? So as to analyse actual data only.
- Can we identify who is more involved in the movement? The number of users who were involved in this movement and how they changed over time.
- What were the trending topics?
- Can we find any similarity in beliefs and emotions of different users and group them together ?
- Which language was used the most by users involved in this movement?
- The locations from where maximum traction was gained by this movement?

Challenges faced:

- The twitter API we used did not have the location coordinates because of which we were not able to grab the actual location of the users. We had to rely on the location field which was filled by the user. Due to this reason the locations filled in were not accurate or in a uniform format. It was difficult to analyse this unstructured location data.
- Gathering historical data from twitter can be quite challenging without getting the premium API.
- Many preprocessing and topic modelling techniques could not be used for analyzing the corpus due to the huge volume of data. We have used sampling techniques to mitigate the problem.

Data Science Pipeline and Methodology

Data collection

We collected data from twitter and instagram. Our main goal during this data collection was to gathered data from 2017, 2018 and 2019 where #METOO was used.

Tools Used:Tweepy and Scrappy

➔ Why these tools?
- ◆ Tweepy is a very robust tool that helped us to extract the data from twitter.
- ◆ Scrapy was used to scrape as many features from instagram as possible since there is no API which we could use to extract that data

The collection of the twitter data was done using Tweepy. We registered our app on Twitter Developers to collect the keys necessary to perform the extraction. Doing the extraction using tweepy got us the data from 2019, now for historical data from 2018 and 2017 we used this dataset(https://data.world/rdeeds/350k-metoo-tweets). We took the ID attached to this dataset and query the tweepy API by the tweet ID to retrieve the necessary fields to match our 2019 dataset. In both cases the output is a csv file, one for 2019 and one for 2018 and 2017 with the following features:

- ❖ Created date
- ❖ Is retweet
- ❖ Text
- ❖ Text retweet
- ❖ Language
- ❖ Hashtags
- ❖ Mentions
- ❖ Location
- ❖ Geo location
- ❖ Geo country
- ❖ Geo coordinates
- ❖ User created at
- ❖ Screen name
- ❖ Name
- ❖ Profile_description
- ❖ Total_number_of_tweets
- ❖ Is_verified
- ❖ Followers_count
- ❖ Friends_count
- ❖ Retweets_count
- ❖ Favorite_count
- ❖ Url
- ❖ Listed_count
- ❖ Default_profile
- ❖ Default_profile_image
- ❖ has_extended_profile

For the instagram dataset we used scrapy. We build the scrapper to gather data from instagram where the #METOO was used and were able to go back until 2017. The scrapper receives the hashtag to be look at as a parameter and returns a json file with the following features:

- ❖ Caption
- ❖ Comment_count
- ❖ Display url
- ❖ ID

- Is_verified
- Likes count
- Location id
- Location latitude
- Location longitude
- Location name
- Owner full name
- Owner ID
- Owner username
- Shortcode
- Tagged users
- Taken at timestamp

Data Integration

Tools Used: Pandas

- ➔ Why these tools?
  - ◆ Pandas Dataframes are useful to manage the schemas better by adding, removing and updating columns in different schemas

Once we finished collecting the data we ended up with 3 different files.

1. Metoo.json with all instagram data
2. Metoo_2019.csv with twitter data from 2019
3. Metoo_2018_2017.csv with twitter data from 2018 and 2017.

The goal from this section is to make the instagram schema match the twitter schema. In order to do so we took the Metoo.json(1) file and map each of the features extracted to a corresponding match from the twitter schema and discard useless information.

The mapping was done as follows

At the end of this phase we combined both datasets and added 2 columns to the existing schema:

- ID: unique identifier
- Source: column to identify if this was data coming from twitter or instagram

Data Cleansing

The main purpose of the data cleansing phase is to remove stop words, remove unwanted characters, break words , tag parts of speech, lemmentation and produce an extended dataset with the results from the data cleansing phase.

Tools Used: NLTK library

- ➔ Why this tool?
  - ◆ NLTK is a rich library which has many NLP related functions.
  - ◆ The library is easy to use with Python and gives efficient results

Instagram	Map Rule	Twitter
Caption	IF Caption does NOTcontains "repost" or "regram" THEN Text=Caption	Text
Comment_count	Maps 1-1	Followers_count
Display url	Ignore	
ID	Ignore	
Is_verified	Maps 1-1	Is_verified
Likes count	Maps 1-1	Favorite_count
Location id	Ignore	
Location latitude	Maps 1-1	Geo coordinates
Location longitude	Maps 1-1	Geo coordinates
Location name	Maps 1-1	Geo location
Owner full name	Maps 1-1	Name
Owner ID	Ignore	
Owner username	Maps 1-1	Screen name
Shortcode	Ignore	
Tagged users	Maps 1-1	Mentions
Taken at timestamp	Maps 1-1	Created date
Caption	IF caption contains "repost" or "regram" THEN TRUE ELSE FALSE	Is retweet
Caption	IF Caption does contains "repost" or "regram" THEN Text retweet=Caption	Text retweet
	Default to "en"	Language
Caption	IF Caption contains "#" THEN extract all words associated with every "#"	Hashtags
	NULL	Location
	NULL	Geo country
	NULL	User created at
	NULL	Profile_description
	Default to 0	Total_number_of_tweets
	Default to 0	Friends_count
	Default to 0	Retweets_count
	NULL	Url
	Default to 0	Listed_count
	Default to FALSE	Default_profile
	Default to FALSE	Default_profile_image
	Default to FALSE	has_extended_profile

First we created a data dictionary with mappings for all the languages to have a proper description of each of them instead of just initials. Then we created a tweet_full column that

had a combination of tweet and retweets columns. Next we remove non supported languages from the dataset. The reason behind this is that we are using NLKT library to remove stop words and it only support certain languages, meaning that everything that did not satisfied the library scope was removed from the dataset. We then cleaned the hashtags columns to remove the "#" character and the mentions column to remove the "@" character and other unwanted characters from the tweets itself. Then we break some words such as "i'm" into "i am", etc. After cleaning the dataset we then moved into removing the stop words, tag parts speech, stem filter and finally lemmentation using the NLKT library.

At the end of this phase we added the following columns to the existing schema:

- Language desc: full description of language.
- Tweet full: Column that consolidated tweets and retweets.
- Pre proc tweet: Removed unwanted characters and break words from tweet
- Pos tags: Applied tag parts speech with NLKT library.
- Stem Filter: Applied stem with NLKT library
- Lemma Filter: Applied lemmentation with NLKT library.

Fake account detection

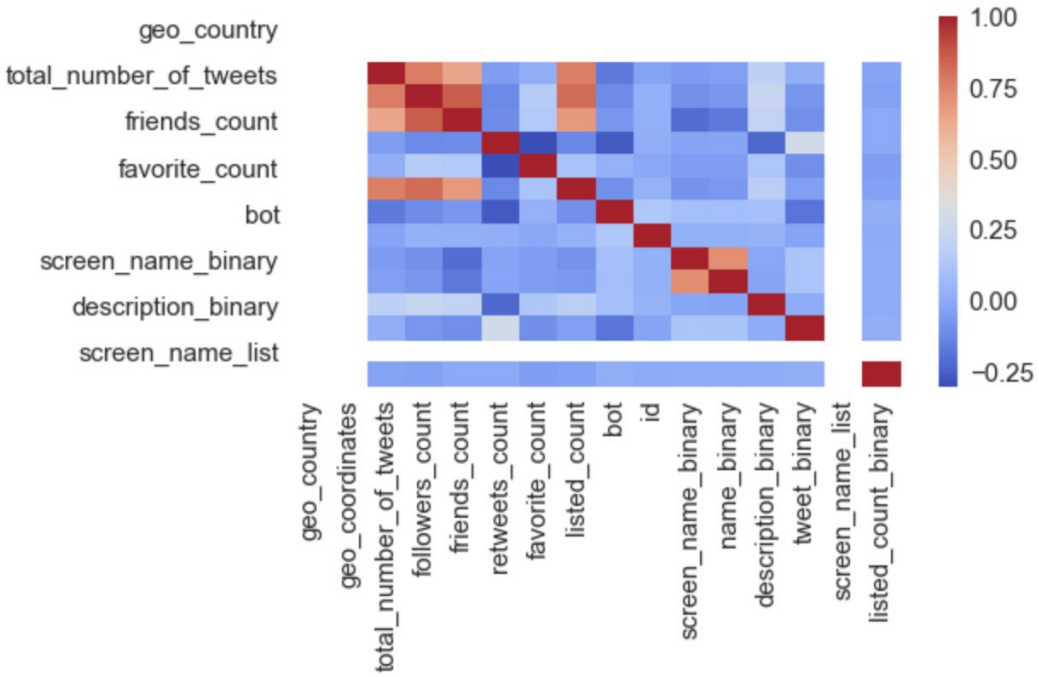
During this phase the main goal is to detect if the account who is tweeting about #METOO is an actual person or a bot. For the last years there has been an increasing number of bots that help increase the popularity of certain topics like politics, propaganda, etc. The main goal for this section is to identify and remove all those tweets that are coming from bots so we can get more valuable insights from the data.

Tools Used: BOTOMETER API,Spearman Correlation, Decision Tree and Random Forest, Text mining NLP techniques

- Why these tools?
  - Botometer api is an open source api to detect bots and humans
  - Spearman correlation gives us the importante between each feature and how strong their relationship is from one another
  - Random Forest and Decision Trees algorithms have been used in the past in attempts to correctly classify bots and humans

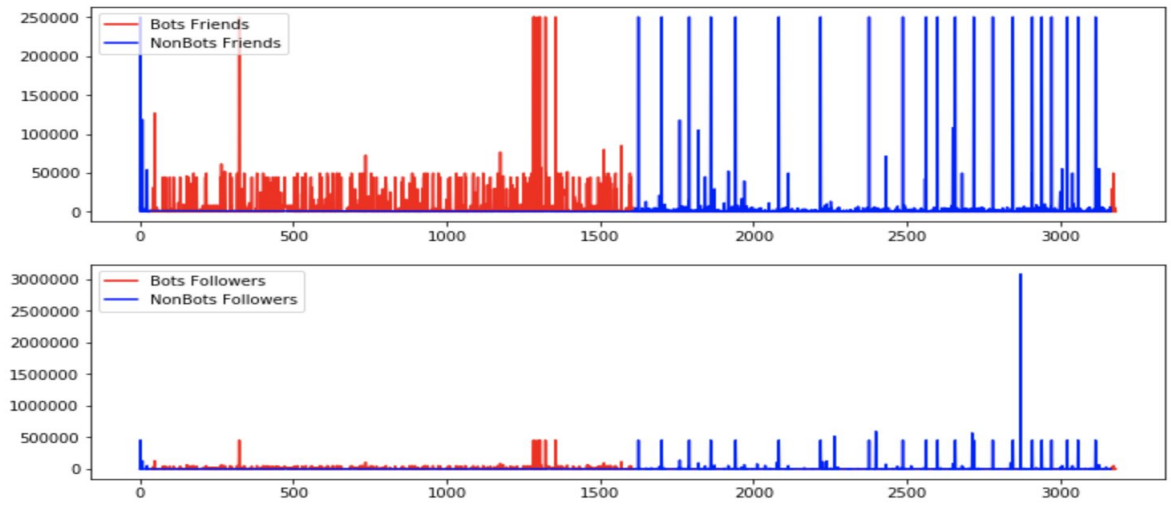
The first step was to gather a labeled dataset to train our model. We chose a subset from our main dataset and use the BOTOMETER API in order to label our data as bot or human. Once we had all of our data labeled, we make sure that it was a balanced dataset so our model was as accurate as possible.

Then, we applied the feature Spearman correlation of each of the features and analyzed the dataset in detail.

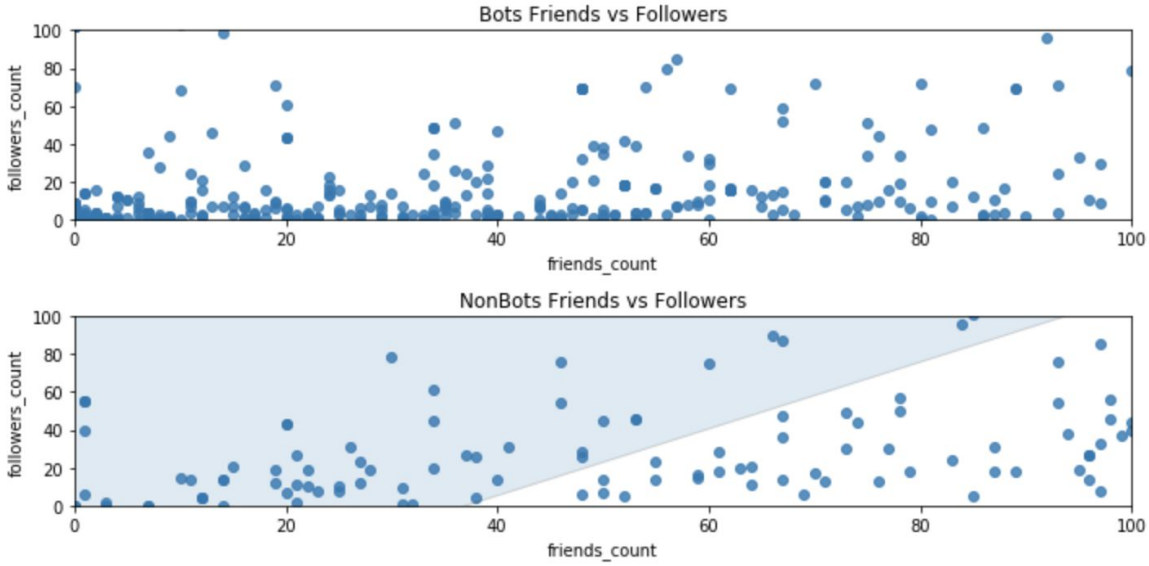


As a result from the Spearman correlation we saw that there was a strong correlation between the following features:

- Listed count
- Friends count
- Total number of tweets



- Followers count



Based on this correlation we performed feature engineering and feature extraction. The feature engineering consisted of identifying if the screen name, profile description, tweet status or name contained any of the words that are regularly part of a bot or if it was in a list of bot names. Next, feature extraction was done on those features who were transformed during feature engineering.

These features were passed on to the decision tree and random forest algorithms. Out of these two the Random Forest Algorithm performed better with an accuracy of 73%.

We then worked on creating our own custom algorithm following the same NLP techniques with has a bag of words, a list of known username bots, but this time we followed our EDA results in order to make specific conditions on the amount of followers, if an account is verified or not, friends count, listed count to classify the user accordingly.

We trained our algorithm and test it, which achieved an 87.234% accuracy. Based on the analysis and percentage confidence that our model is working as expected we created 2 different files.

- Prediction_bot_humans.csv: File that has all the test data classified as bot or human
- Humans.csv: File which contains ONLY the tweets that were classified as non-bots.

As a result from this analysis we could see that around 10% from the whole dataset was marked as bots, which is a very good data insight because that means that bots are not interfering with the impact of the movement nor making fake stories to raise the popularity of the movement.

Topic Modelling

The main purpose of this module is to find the main topics associated with the corpus of the tweets collected so that we can understand the opinion of people and get insights about the similarly in their beliefs. Additionally we use these topic similarity to detect communities and calculate the sentiments of the communities which is described in the preceding sections.In order to perform LDA we first preprocess the data according to the following criteria and form the corpus.

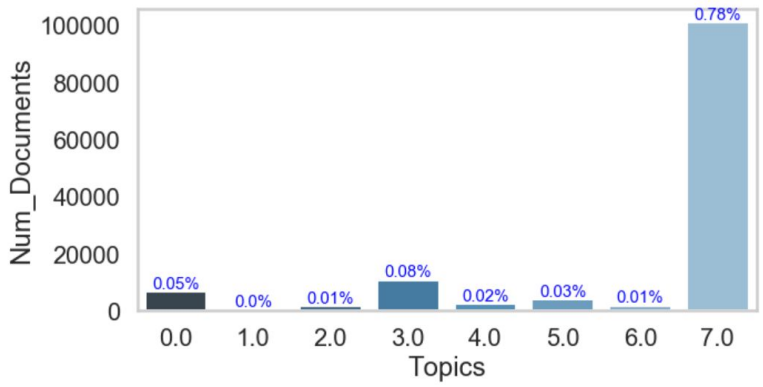
- Drop any column that has missing values
  - Filter out the retweets since they are textually same as the original tweets.
  - Filter out records if they are NAN
  - Merge records by users.
  - Remove # from hashtags and convert to lower case
  - Remove punctuations and tokenize the tweets

- ignore words that appear in less than 20 documents or more than 10% documents

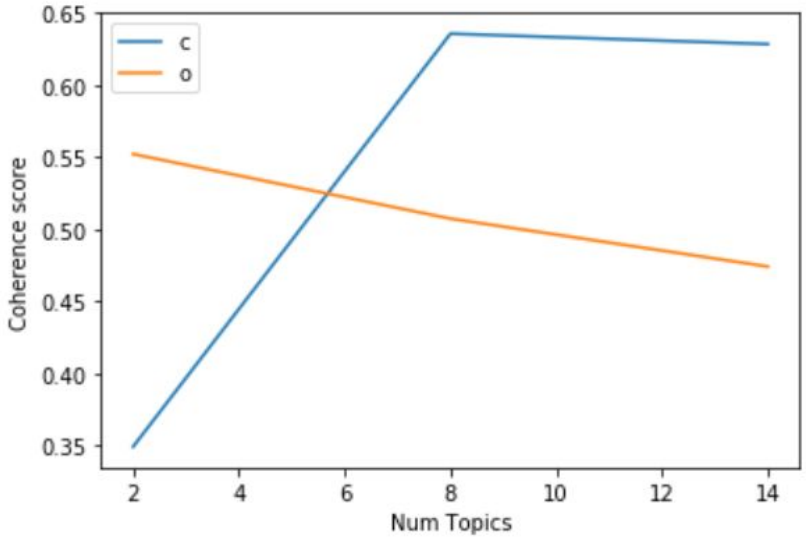
Algorithms Used: LDA and MalletLDA.

Why these algorithms?

- Topic modelling takes a single text or corpus and looks for patterns in the use of words, it is an attempt to inject semantic meaning to vocabulary. It assumes that any piece of text is composed by selecting words from possible baskets of words where each basket corresponds to a topic. The process iterates multiple times till it has the most likely distribution of words into baskets which are named as topics.The challenge, however, is how to extract good quality of topics that are clear, segregated and meaningful. We have tried to implement 2 variants of Latent Dirichlet Allocation(LDA) as it is a popular algorithm for topic modeling for small documents like tweets.
- Topic modelling because they allow sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
- We analysed the accuracy and performance of both the basic version of LDA which is probabilistic topic modelling and the MalletLDA which is implementation of Gibbs sampling In order to get the optimal number of topics we build many LDA models with different number of topics and choose the one with the highest accuracy. Accuracy is measured by coherence score. Higher the score, better the model.
- Mallet model gave high coherence score of .63 as compared to LDA.Though Mallet LDA gave better results, it was quite time consuming. So we performed most of the analysis with basic LDA
- After training the models we could use it to perform various data analysis. For instance, the document distribution, topic distribution, calculating similarity, detecting communities and finding sentiments which will be described in the subsequent sections.



Number of documents in each topic and their percentage contribution



Accuracy measure of both the models

Network Clusters and influencers

Tools Used: Gephi, Python, Pyspark GraphFrames

- ➔ Why these tools?
  - ◆ Gephi is an interactive visualization tool used to visualize the community clusters.
  - ◆ Python and Pyspark are used to process the graphs and perform data analysis on these graphical models

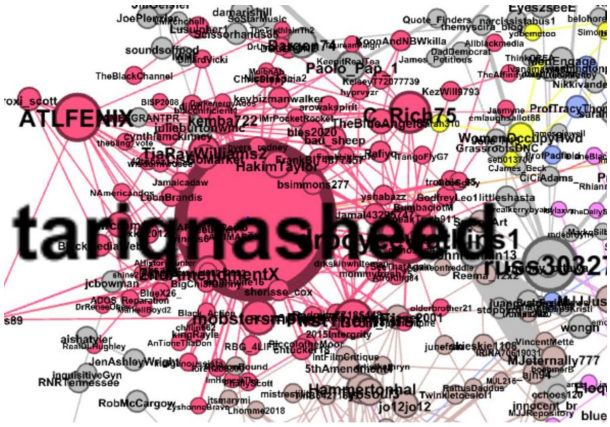
Algorithms:

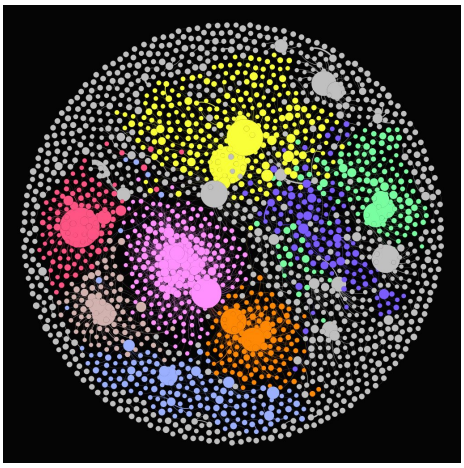
- Label propagation assigns labels to previously unlabeled data points. At the start of the algorithm, subset of the data points have labels (or classifications). These labels are propagated to the unlabeled points throughout the course of the algorithm.
- Louvain Modularity and InfoMap are popular algorithms to extract communities.

The main purpose of this module is to detect communities that are formed based on implicit and explicit relationships through social interactions. To this end, we create a re-tweet network, and a Similarity based topic network with the tweets containing the #Metoo tags and detect meaningful communities using Label Propagation Algorithm, Louvain Modularity and InfoMap community detection algorithms.

It is inferred that the communities detected in the the Re-tweet network represents much stronger connection as re-tweeting a user would mean direct agreement or disagreement to the content. The communities formed as a result will be direct relationships. To create a Similarity based network, we first represent the tweets into a vector-space model. The method we use is called Latent Dirichlet Allocation(LDA) which is a generative probabilistic model. We can view LDA as a dimensionality reduction technique. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. We used the Louvain Modularity algorithm and InfoMap to extract communities.

The Louvain Modularity algorithm is a very popular method of community detection. Modularity is a scale value between -1 and 1 that measures the density of edges inside communities to edges outside communities. Optimizing this value results in the best possible grouping of the nodes of a given network. In the Louvain Method of community detection, first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated. InfoMap uses the probability flow of random walks on a network as a proxy for information flows in the real system and decompose the network into modules by compressing a description of the probability flow. Taking this approach, InfoMap uses an efficient code to describe a random walk on a network and thus finding community structure in networks is equivalent to solving a coding problem.





Once we detect the communities using the most suitable algorithm discussed above, we find the influencers of each of these communities. The influencers can be found by centrality measures. We have used pagerank for this purpose. The red cluster is the enlarged portion of the entire network and it shows the main influencer as tariqnasheed.

Sentiment Analysis

Sentiment Analysis helps us understand the emotion and opinion of people. We have used a weighted average of the sectiment scores of tweets as well as hashtags because the sentiments of the tweets are often calculated by rule based methods which is Fast, easy to understand, good performance on a certain domain but not on all domains because of hard to construct rules, rules may have ambiguity and different domains require different kinds of rules which is not very accurate always, however hashtags always convey either topics or emotions. So instead of relying on just one score we try to analyze both the features. Given a set of hashtags where each hashtag is associated with a set of tweets, our main goal is to infer the sentiment polarities. The hashtag-level sentiment classification inherently bases upon the tweet-level sentiment analysis results. The idea is to assign a sentiment value to each hashtag based on the sentiment value of the neighbouring/co-occurring hashtags. i.e. a hashtag used frequently with other hashtags with negative sentiment value will most likely to be having a negative sentiment value itself. Hence we use message propagation algorithm to determine the scores as depicted in the paper "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach"

Algorithms Used: Textblob library to calculate lexicon sentiments and Loopy Belief Algorithm to calculate scores for hashtags.

- ➔ Why these tools?
  - ◆ TextBlob library provides the sentiments scores i.e subjectivity and polarity scores based on lexicon analysis which has high accuracy though rule based.
  - ◆ Loopy Belief Algorithm classifies each node in a graph through belief message passing and has proven a very good performance in practice

Our ultimate goal is to assign each hashtag with a proper sentiment label. We make the Markov assumption that the determination of sentiment polarity can only be influenced by either the content of corresponding tweets or sentiment assignments of neighbor hashtag. In order to do this we applied the Loopy Belief Algorithm. The algorithm classifies each node in a graph through belief message passing. When applying the algorithm we end up with the sentiment polarities and define a threshold that will be a turning point into each of the sentiments. Negative values are negative comments, zero means neutral and everything above zero is positive. Finally we get two separate scores for tweets and hashtags and then calculate the probabilities of each after taking an average and define a threshold for labelling the tweets as positive, negative or neutral based on their combined effect to proceed with the rest of the data analysis.

Visualization

The purpose of this module is to create a web application that shows our EDA results as well as some valuable insights gathered from the dataset using D3.

Tools Used: D3, mpld3 library

- ➔ Why these tools?
  - ◆ D3 is a very powerful visualization tool with an extensive set of capabilities to show attractive visualizations
  - ◆ MPLD3 library converts matplotlib code into D3 code.

In order to create our web application we brought different components together. First the EDA performed after the Data Cleansing phase was translated into D3 code using the mpld3 library. Next we developed a native D3 code to create a bubble chart and a word cloud, and finally we integrated all of this into a web application that combines all this different components.

Evaluation

From our analysis we were able to determine the following:

- The ratio of fake/bot accounts as compared to the real human account was very low, which proved that the majority of data and stories were shared by humans and were real. To gather only the authentic data and generate accurate results we removed the bot accounts that we found and performed further analysis only on the real accounts.
- The trend of users who contributed towards this movement. The number of users who tweeted under the hashtag #MeToo increased in October, 2017 , the number decreased a little bit and in 2019 the number increased again which showed that the movement gained traction again in 2019 and is still active.
- The influencers of this movement were identified and clubbed together to form a network of retweets and topics. In this manner we were able to group together users with similar sentiments,interests and similar opinions towards this movement.
- We were able to analyse the sentiments of people towards this movement over time.
- We also realised that most of the data gathered was from the retweets instead of tweets
- The most widely used language by users to tweet under this movement was English followed by Spanish and French
- The country where this movement gained most traction was USA. The most prevalent areas in the USA were Atlanta and Austin.

Data Product

To showcase our findings and to help spread awareness about this movement we created a web application which could be used by police and government agencies to be aware of the trend of such incidents occurring in the society.

In the future with the help of premium Twitter API we hope to add pin exact locations from where these tweets are posted by users. We would like to locate NGO's and safe houses close to these locations with maximum traction in major cities.

The web application has the following components:

- EDA: We performed exploratory data analysis on the MeToo dataset and visualized the following:
    - The number of tweets related to #MeToo posted over time: MeToo became viral in October 2017 and we counted the number of tweets posted under this hashtag over time.We can clearly see the peak reached in October 2017 when Alyssa Milano's tweet went viral after which there is a dip in the number. We also see a boost again in 2018 after which the number of tweets stayed pretty high even in 2019. This a clear indication that movement is still very active and people still find support and help by sharing their stories. We can also see that such incidents of abuse continue to be on the rise even after so much traction.
    - The number of tweets versus retweets: We also compared the number of retweets with the tweets posted by users and could clearly see that most of the data was collected from the retweets. This is an encouraging finding in the sense that people sympathised with the victims stories and supported them by retweeting their stories.
    - Identifying Fake accounts: It was important to identify the fake/bot accounts so that we could ensure that the data we collected was from authentic sources and our analysis was authentic. We were able identify fake accounts and we discovered that the number of fake accounts as compared to human accounts is very low which meant that most of the data was taken from real sources. We removed the data taken from the fake accounts and performed further analysis on data taken from real accounts.
    - Predominant languages used by users involved in this movement- We identified the languages people used to tweet under the MeToo hashtag and English was found to be used predominantly, followed by Spanish and French. The other languages that were also used were German,Swedish,Hindi and Italian.
    - Locations where this movement was most prevalent- To get a sense of the locations where this movement gained maximum traction we identified places from where most of the tweets were posted.Atlanta and Austin were found to be locations from where people tweeted the most.
- Bubble Chart depicting the users involved with #MeToo over time:We wanted to see the trend of users involved in the MeToo movement for the past three years, that is 2017,2018 and 2019.We tracked users who tweeted under this hashtag and plotted them as per the number of tweets and the number of followers these users they have to understand how viral these tweets were. The most interesting fact to notice in this

chart is the sudden rise in the number of users in 2019. The movement gained traction again in 2019 which shows that the fight is far from over.

- Word Cloud displaying most popular words used by users while tweeting under #MeToo:We wanted to identify the words used the most by users who tweeted under the #MeToo hashtag. We used a WorldCloud to represent the most popular words used. Visible patterns and repeating words emerged.
- Retweet Network : We were able to form community clusters based on common retweet networks. Each community is represented by a different node and color.The size of the node showcases how influential that node is in that community.

Please visit our website on this URL: http://sfu-metoo.us-east-2.elasticbeanstalk.com/

Lessons Learnt

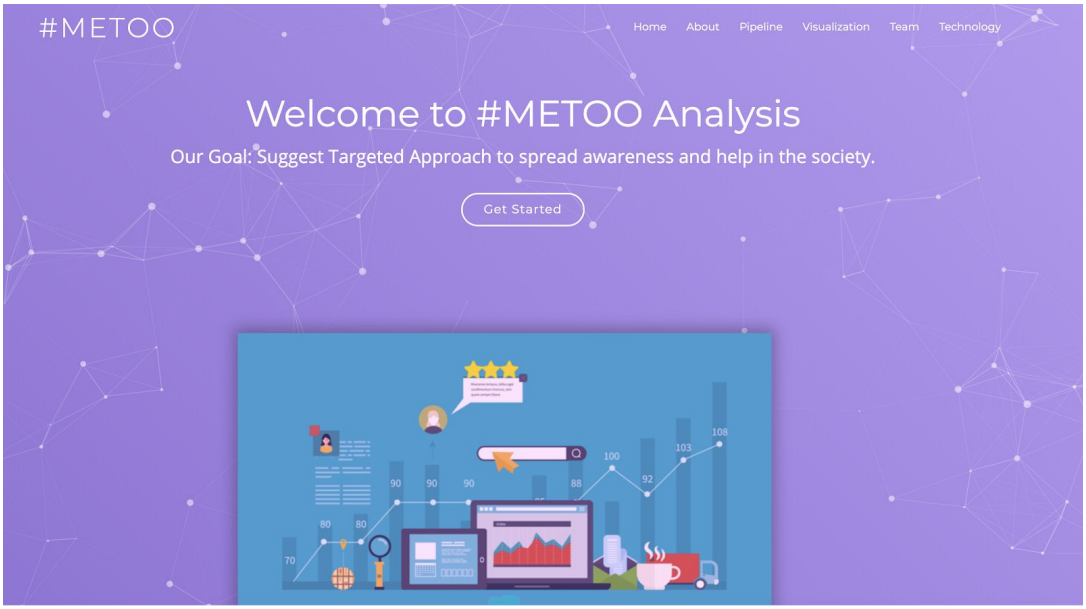
- Social network analysis metric like Modularity can be used to detect sub-communities in networks. We also realized that, the smaller a sub-communities is. the slower viral content spreads.
- Message propagation algorithm can be used to determine sentiments from hashtags to gave additional insights
- Techniques to apply filters so as to ignore words judiciously to get better topic clusters.
- Text mining gave better results than ML algorithms when it came to identifying Fake accounts.

Summary

#MeToo Analysis aims to bring awareness to the community through government of NGO's who are looking for a targeted community to reveal some legal information or spread the movement among different communities. We tried to achieve it by implementing community clustering using graphical analysis which would group people of similar beliefs and emotions and also find the primary influencers. So the concerned bodies can educate society and bring social reforms by the aid of the influencers to influence like minded people. The main goal of our project is to provide an informative platform that can be useful for this entities to see how the movement is behaving since its peak moment in 2017 to this days and identifying the main people involved in the movement.

The volume of data that we collected from twitter and instagram helped us analyze different algorithms and its applications in greater depth. We tried to build hybrid text mining algorithms to classify fake vs real accounts and sentiment analysis which outperformed the traditional techniques.

We consolidated all our findings into one single website that shows the results from our analysis that is available online for everyone who finds it useful to take to the next level.



References

Jalli, Bhavika Reddy , Siyan Chen, Cixing Li and Vidya Mansur. "Community Detection and Trend Analysis of #MeToo Event on Twitter." *Community Detection and Trend Analysis of #MeToo Event on Twitter.* (2018): 11. Web.

Wang, Xiaolong, Furu Wei and Ming Zhou. "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach." *Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach.*(2018): 11. Web.

Matapalli Revanth, Elango Varun and Ramesh Vignesh. "Twitter Bot or Not." *Twitter Bot or Not.*(2017): 8. Web.