Twitter Profile Recommendation System

Mehul Kohli, Ritika Singhal

I. PROBLEM STATEMENT

Implement a section for recommending other twitter profiles to the user, based on the content the users post on twitter about the currently trending topics.

II. SIGNIFICANCE

The aim of this project is to make an addition to the existing "Who to Follow" section on Twitter. The current system makes recommendations based on the user's phone and email contacts, the twitter profiles he follows, and the posts the user has liked.
Our system will recommend twitter profiles and their tweet content to the user based on how similar both users' tweet content is, taking into account what all topics they share content about. This is the approach that has not been implemented by Twitter yet which makes our project new and innovative.

III. DATA DESCRIPTION

Data Source	Tweepy: Twitter API (www.twitter.com)
Data Size	730K Tweets (within the time frame of late September 2020 to early October 2020)
Data Description	Tweets containing the keywords: ["trump", "biden", "covid", "coronavirus", "mask", "distancing", "vote", "voting", "election", "vaccine", "virus", "work from home", "wfh", "lockdown", "capitol", "white house", "black lives matter", "blm", "racist", "racism", "white supremacy", "white supremacist"]. All these keywords were included in the search query keeping in mind the following trending topics/people: Donald Trump, Joe Biden, US elections 2020, Covid - 19, Racism The retweets were not treated separately and were treated just like any other tweet.
Relevant data characteristics used	Each tweet object contained several attributes such as hashtags, retweeted status etc., but we were only concerned with the tweet text and the usernames.

Table 1: Data Description

IV. LITERATURE REVIEW

There have been quite a few research papers that have been published on similar topics. We intend to take inspiration from some of these papers that we feel would help us in our implementation.

The research paper by Kozo Chikai and Yuki Arase from Osaka University had presented a paper on analysis of similarity measures between short texts. In their study, they have compared the state-of-the-art methods for estimating text similarities to investigate their performance in handling short text, specially, under the scenario of short text conversation. They've also implemented a conversation system using a million tweets crawled from Twitter, but the scope of this project only relates with the text similarity component.

A paper on content-based similarity of twitter users by Stefano Mizzaro, Marco Pavan and Ivan Scagnetto from University of Udine via delle Scienze, Italy, proposes a method for computing user similarity based on a network representing the semantic relationships between the words occurring in the same tweet and the related topics. They have used a specially crafted network to define several user profiles to be compared with cosine similarity and also an initial experimental activity to study the effectiveness on a limited dataset.

Another interesting paper by Hind AlMahmoud and Shurug AlKhalifa presents a framework for discovering similar users on Twitter, with the intention of finding applications in profiling users for social, recruitment and security reasons. The framework contains a novel formula that calculates the similarity between users on Twitter by using seven different signals (features). The signals are followings and followers, mention, retweet, favorite, common hashtag, common interests, and profile similarity. The proposed framework is scalable and can handle big data because it is implemented using the MapReduce paradigm. It is also adjustable since the weight and contribution of each signal in calculating the final similarity score is determined by the user based on their needs. The accuracy of the system was evaluated through human judges and by comparing the system's results against Twitter's Who To Follow service.

Work has also been done on analysis of user sentiments for recommending possible friends to follow. In the paper by Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti, a user recommendation technique is proposed based on a novel weighting function, named sentiment-volume-objectivity (SVO) function, which takes into account not only user interests, but also his sentiments. Such a function allowed them to build richer user profiles to employ in the recommendation process than other content-based approaches.

A paper on document clustering using Latent Dirichlet Allocation(LDA) and K - Means Clustering by Peng Guan, Yuefen Wang, Bikun Chen and Zhu Fu presented a method based on LDA and K-means (LDA_K-means). In order to improve document clustering effect with K-means, they discovered the initial clustering centers by finding the typical latent topics extracted by LDA. The dataset used was the 20 Newsgroups dataset. They have made use of a document-topic matrix and distributed each vector to the nearest cluster using the Jensen-Shannon Divergence.

V. APPROACH

A. Approach 1: Applying K - Means Clustering on Tf-IDF features and Entity-Based Sentiment

Our initial approach to this problem was to combine entity-based sentiment present in the user's tweets with the Tf-IDF features; however, we realized that this approach had a drawback of users not being able to know about the opinions opposite to those of their own. To overcome this issue, we discarded this approach and tried another approach.

B. Approach 2: Applying K - Means Clustering on Tf-IDF Features and Text Subjectivity

By using text subjectivity as a feature instead of sentiment, users who expressed an opinion on a certain topic (i.e. more subjective tweets) would have similar subjectivity scores. Adding it as a feature could enable them to look at diversified content, but this approach failed to give us good clusters as there were many dissimilar tweet documents that ended up in the same cluster.

C. Approach 3: Latent Dirichlet Allocation (LDA) and K - Means Clustering

This approach enabled us to extract a number of topics from the tweet documents using LDA to distinguish users based on the topics contained in their tweets This approach gave us an initial soft-clustering of tweet documents, after which the final K - Means algorithm was performed on the probability distribution of documents. This approach gave us the best results.

VI. IMPLEMENTATION

Step - 1: Applying LDA and extracting topics from the Tweet Documents

The idea behind using LDA before applying K-Means is to ensure that each final cluster we form contains users who tweet about the same topic (represented by that cluster) and are thus related.

Data Preparation for LDA Steps

- Concatenating together each user's collection of tweets and treating them as one tweet document.
- Cleaning and pre-processing the tweets' text.
- Building a vocabulary for a bag of words model.
- Forming bigrams and trigrams from the text and adding to the vocabulary.

Implementation of the LDA Model

The LDA model takes as input the bag of words model, and defines each topic as a distribution of words, using which we can obtain a probability distribution of the topics in each document. Once probability distributions for each document are obtained, they can be used as a feature matrix for the K - Means algorithm which would cluster tweet documents having similar distributions together.

The obtained probability distribution matrix *M* can be formulated as follows:

M[i, j] = Probability of document i belonging to a topic j.

Hence, each row of this matrix is stochastic and is treated as a data point for the K-Means Algorithm in the next step.
The optimal number of topics to be extracted was determined by applying the elbow method on the coherence score values, which turned out to be **10** in our case.

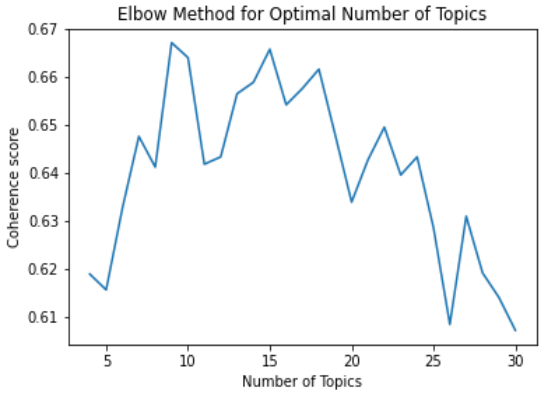

Fig. 1: Coherence scores for upto 30 topics

Step - 2: Applying K-Means Clustering on the output of Step - 1

By applying K-Means clustering algorithm on the probability distribution matrix, we clustered documents that have similar probability distributions. The optimal number of clusters was obtained by applying the elbow method, which turned out to be **10** in our case.

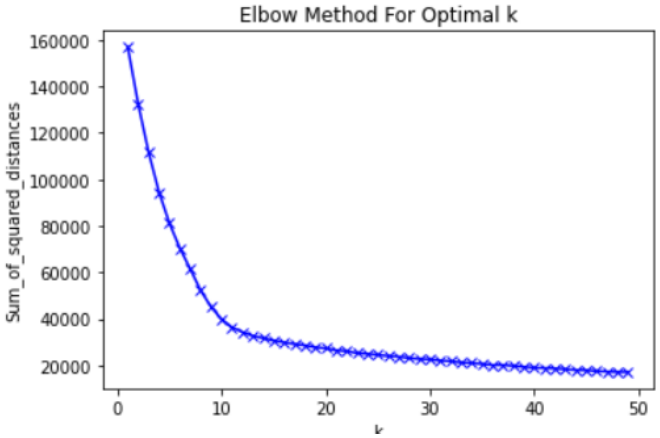

Fig. 2: Elbow method for optimal number of clusters

Step - 3: Applying KNN Algorithm on the output of Step - 2

For a given test tweet document (test-case), the cluster label will be identified from step-2. However, each cluster has thousands of tweet documents, that is why there is a need to prioritize which samples should be shown as recommendations.

Hence, we applied euclidean distance to calculate a test case's distance from all the data points within the same identified cluster and return the usernames along with their tweets that correspond to the closest *k* data points.

VII. RESULTS & EVALUATIONS

The results for this project are the final recommendations of our model to the twitter users. Some of the examples of recommendations are shown below:

S. No.	User's Tweet Document	Recommendations
1.	Please wear masks, observe social distancing, and wash hands regularly to keep yourself and your loved ones safe from the deadly covid19. Working from home this summer! Trying to be productive this quarantine season.	1. **@ravavyr** tweeted: @itylergarrett What do you mean? Can you elaborate on "political usage of the virus"? I figure the reason we are paying is because the government failed to pass the bills to provide covid relief for the people and it's all been because of partisan bullshit that doesn't belong in the relief bill 2. **@TexasCorn** tweeted: @USDA made a new announcement that there is a second round of coronavirus aid to producers for as much as $14 billion. The program provides potential aid to a broader group of farmers who were not eligible for the earlier program. 3. **@mej_joe** tweeted: Trump has a plan for coronavirus: herd immunity. Say goodbye to your parents, your neighbors, and your friends. Because the Trump plan will kill millions in the next four years. Being sick with a virus he called a hoax doesn't make Trump any less of a white supremacist, fascist mass murderer. 4. **@zezezebe** tweeted: Whatever your political perspective, let's remember that Donald Trump is a human being. An American. A father. A husband. A brother. A cousin. A second cousin. A bankrupt. A lactose-tolerator. An omnivore. A biped. A hominid. A primate. A symmetrical animal. A vertebrate. 5. **@Better4Hughes** tweeted: Meanwhile, over on facebook. A certain Federal LNP MP is promoting (bad) legal advice to people who decide to break Victoria's Covid restrictions. And still @ScottMorrisonMP stays silent. Enabling yet again his dangerous behaviour
2.	I'm a proud democrat, supporting Joe Biden for this presidential election. We will win by a landslide!!	1. **@mojoandjasper** tweeted: This is NOT F**KING OKAY. THIS is why we need to show up this November. We can't give up. For Breonna. Demand change and vote out those who won't allow it. https://t.co/XEInmFBwQV Michelle and I hope that the President, First Lady, and all those affected by the coronavirus around the

	#Biden2020 The democrats are winning! Biden is ahead on the national polls. Joe Biden is the kind of president we need to recover from this crisis. We cannot recover until there is a change of power in this country!	country are getting the care they need and are on the path to a speedy recovery. 2. **@Adequate_Scott** tweeted: There is no way out of ~all this~ without truly grappling with the malevolence and rot of the leaders in both sides. That's not handwaving Trump. This laser like focus on Trump, and the accompanying "we'll have time to argue about Reparations after Biden wins" is the handwave. 3. **@KayS57** tweeted: Biden responds to Trump blaming him for the lack of a national mask mandate by correctly pointing out that Trump is in fact the president https://t.co/0aLTPGHEUK 04 #VoteEarly Anti-Abortion Zealots Demand a Vote to Fill Ginsburg's Seat | By: Ed Kilgore https://t.co/hRV96jzsuy I think the Bush brothers, George & Jeb, should endorse Joe Biden ASAP. Trump steals from kids-cancer charities. You don't think he'll try to steal this election? BREAKING: Barack Obama will host two Biden campaign fundraisers with Kamala Harris next Friday. 4. **@KatneaB** tweeted: No Wisconsin absentee ballots were found in mail discovered in a ditch in the Fox Valley last week, the state's top election official said. The White House has been pushing conspiracies about voter fraud based on this discovery. https://t.co/Z2wuXdzSeJ 5. **@pmramani** tweeted: No Wisconsin absentee ballots were found in mail discovered in a ditch in the Fox Valley last week, the state's top election official said. The White House has been pushing conspiracies about voter fraud based on this discovery. https://t.co/Z2wuXdzSeJ

Table 2: Recommendation Results

The text in the table above is the concatenation of the respective users' latest tweets. In the first user tweet, the tweet doc is about masks and covid19. We can clearly see that all the recommended users to the first user also talk about covid19 and coronavirus. Similarly, the second user talks about Democrats, Biden, elections and President. Our model makes recommendations to this user and each recommendation discusses the same topics in their tweets.

In a similar way, to verify our final recommendations, we used human judgement to read our test users' tweet documents along with the tweet documents of those users who were recommended to these test users by our model. Upon evaluating the recommendations with this method, we found that the test users and their recommendations were talking about the same keywords and trending topics. Before the recommendations, we had extracted the results from LDA (Topic Modeling) which gave us insights about the topic analysis of the given tweet documents. The extracted topics are as shown in the visualization below:

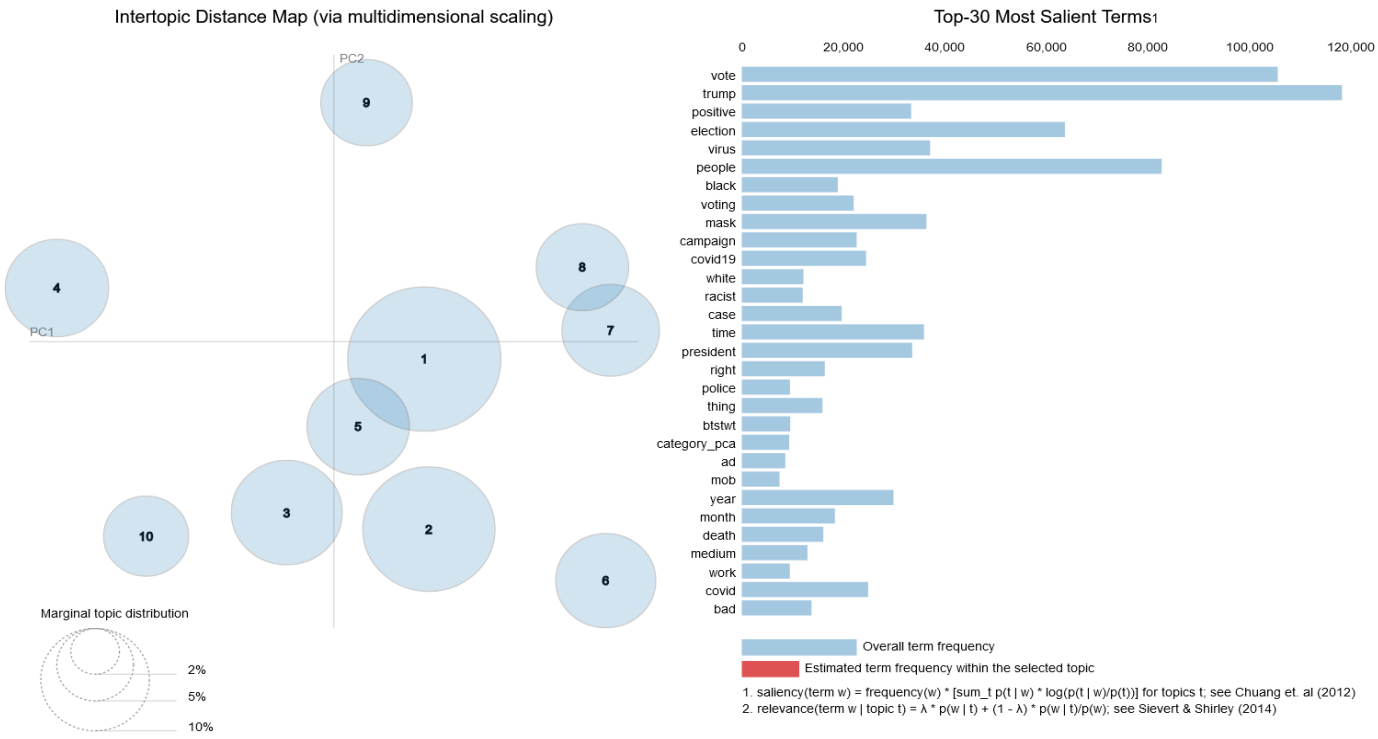

Fig. 3: LDA Results (Topics)

Here is the list of top 5 keywords in the extracted topics from LDA Model. These keywords tell us about the tweets which are categorized in that topic. This is a form of descriptive analysis of each of the topics.

Topic Number	Top 5 keywords
0	virus, case, people, mask, covid
1	people, white, racist, thing, month
2	election, trump, power, vote, president
3	mob, week, story, vaccine, twitter
4	peoples_choice*, bts, vote, year, glow_squid*
5	positive, mask, trump, doctor, covid19
6	trump, president, campaign, ad, year
7	time, medium, work, family, good
8	black, people, police, protect, trump
9	right, leader, election, state, political

Table. 3: Top 5 keywords of LDA Topics * these words are bigrams, joined by a '_'

A. Evaluation strategy for LDA

To evaluate our LDA model, we used Coherence Score metric. This score was used to evaluate our model with different numbers of topics and then choose the best one. We got our best coherence score of 0.67 with the number of topics being 10.

B. Evaluation strategy for the K-Means Clustering

To evaluate the final clusters formed, we used Silhouette Score and Davies-Bouldin Index as the metrics. Our clusters performed well in terms of these metrics with 0.40 being the Silhouette Score and 1.03 being the Davies-Bouldin Index. Through these evaluation strategies, we were able to verify our results.

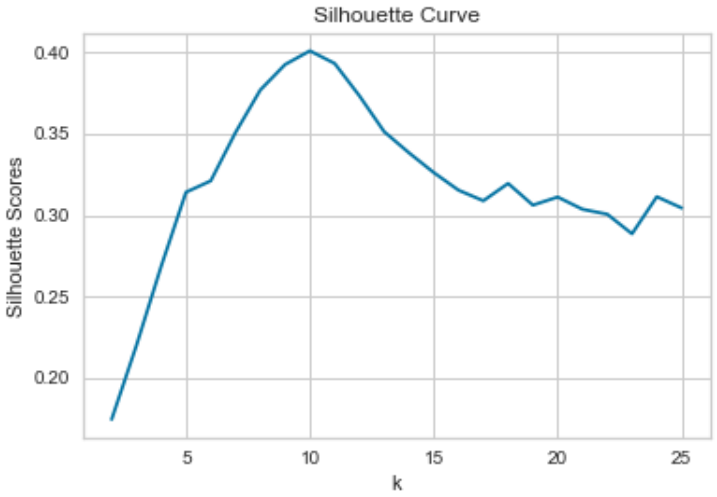

Fig. 4: Silhouette Score vs Number of Clusters

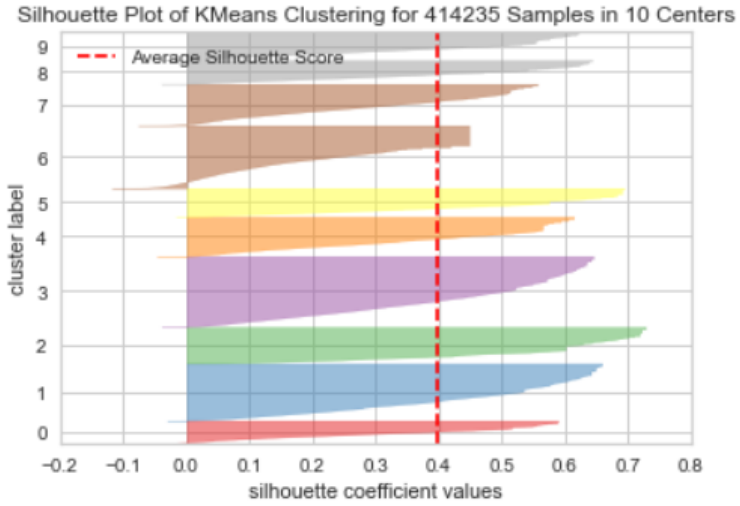

Fig. 5: Silhouette visual plot of the 10 Clusters

VIII. CONCLUSION, LESSONS LEARNED & FUTURE WORK

With the success of our evaluation, we concluded our project which successfully gives good recommendations to twitter users based on the tweet content. Our approach of using Topic Modeling along with K-Means Clustering and KNN algorithms worked for our project and gave us best results.

Through this project, we learned that a good evaluation strategy is extremely important for any project's completion. We need to evaluate our model to ensure we are getting the desired results. Another thing to remember is implementing different approaches to make sure that we select the best one for our project. This project taught us that extracting latent topics could prove useful in clustering text based on their semantic meaning and dimensionality reduction.

Although this project is a success, there are some limitations which give rise to the scope of future work on this project. This includes the inability of our model to deal with real time streaming data on a large scale. It also does not take into account the graphic content in the tweets. So, as a part of the future work, we can extend this project to incorporate real time data through Apache Spark and analysing graphics like images and videos. This will make our project more diverse and more powerful for recommendations on Twitter.

IX. REFERENCES

[1] Kozo Chikai and Yuki Arase (2016). *Analysis of Similarity Measures between Short Text for the NTCIR-12 Short Text Conversation Task*.

[2] Stefano Mizzaro, Marco Pavan and Ivan Scagnetto (2015). *Content-Based Similarity of Twitter Users*.

[3] Muhammad Moeen Uddin, Muhammad Imran and Hassan Sajjad. *Understanding Types of Users on Twitter*.

[4] Carolina Fócil-Arias, Jorge Zúñiga, Grigori Sidorov, Ildar Batyrshin, and Alexander Gelbukh. *A tweets classifier based on cosine similarity*.

[5] Hind AlMahmoud and Shurug AlKhalifa (2018). *TSim: a system for discovering similar users on Twitter*.

[6] Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti. *A Sentiment-Based Approach to Twitter User Recommendation.*

[7] Peng Guan, Yuefen Wang, Bikun Chen and Zhu Fu. *K-means Document Clustering Based on Latent Dirichlet Allocation.*