 doi:10.20944/preprints202106.0196.v1

Article

Similarity approximation of Twitter Profiles

Niloufar Shoeibi [1]*, Nastaran Shoeibi[2], Pablo Chamoso[1,3], Zakieh AlizadehSani [1], and Juan M. Corchado [1,3]

[1] BISITE Research Group, Universidad de Salamanca, Salamanca, Spain; Niloufar.shoeibi@usal.es
[2] Babol Noshirvani University of Technology, Babol, Mazandaran, Iran.
[3] Air Institute, IoT Digital Innovation Hub, Salamanca, Spain
* Correspondence: Niloufar.shoeibi@usal.es; Tel.: +34-617-939-365

Abstract: Social media platforms are entirely an undeniable part of the lifestyle from the past decade. Analyzing the information being shared is a crucial step to understand humans behavior. Social media analysis is aiming to guarantee a better experience for the user and risen user satisfaction. But first, it is necessary to know how and from which aspects to compare users with each other. In this paper, an intelligent system has been proposed to measure the similarity of Twitter profiles. For this, firstly, the timeline of each profile has been extracted using the official Twitter API. Then, all information is given to the proposed system. Next, in parallel, three aspects of a profile are derived. Behavioral ratios are time-series-related information showing the consistency and habits of the user. Dynamic time warping has been utilized for comparison of the behavioral ratios of two profiles. Next, Graph Network Analysis is used for monitoring the interactions of the user and its audience; for estimating the similarity of graphs, Jaccard similarity is used. Finally, for the Content similarity measurement, natural language processing techniques for preprocessing and TF-IDF for feature extraction are employed and then compared using the cosine similarity method. Results have presented the similarity level of different profiles. As the case study, people with the same interest show higher similarity. This way of comparison is helpful in many other areas. Also, it enables to find duplicate profiles; those are profiles with almost the same behavior and content.

Keywords: Twitter; Social Media; Social Networking; Social Network Analytic; Graph Analytic; Text Similarity; Natural Language Processing; User Engagement.

1. Introduction

Social media platforms are now a part of the lifestyle of human beings of any age. This popularity has its advantages and disadvantages. The great benefit is the faster and easier communication and overcoming physical limitations [1].

Social networks are defined as the many overlapping networks that link and move friendships, information, money, power, etc. By analyzing social networks, we can gain new insights into culture, politics, history, and many other things. In other words, user's connections are a significant factor in what they know and how they think [2]. Social network analysis allows us to quantify the connections between individual points. It helps to find patterns in the connections that sustain the society [3]. How the individual is connected or disconnected from people, groups, or populations; in other words, how the individual distributes his or her energy across different social groups over time or explores how an idea, belief, or disease passes through the individual's network [4].

Social media analysis aims to understand people's behavior, provide more safety and achieve higher user satisfaction. There are different malicious types of activities; rumor control [5], detecting fake news and stopping its propagation [6], fake and bot detection [7], and detecting duplicate profiles.

As much it's essential to have a safe society, it is crucial to guarantee the safety of the virtual community users. One step to make this society more secure is to detect

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

and remove these duplicate profiles. These profiles can induce unethical thinking or activities, such as sexist ideologies [8]. Sometimes, the aim is to detect specific topics in real-time[9]. Other times, assuring cybersecurity is a challenge because the safety of social networks is a tremendous concern due to mass user engagement, so the aim is to detect criminal activities and eliminate them [10,11]. Considering the possibilities that social media platforms give for illegal activity, it is crucial to analyze the behavior of users.

For detecting duplicate profiles, it is mandatory to know how to compare two profiles [12]. Thinking about how to detect duplicate profiles has leaded to implementing a methodology for comparing Twitter users, considering three aspects: behavioral similarity, audience similarity, and context similarity. Having the information derived from the users' comparison enables to detection of duplicate profiles, eliminates these frauds, and protects users' policies and privacy, leading to a secure virtual society to make life more comfortable, secure, and updated[13].

The research's questions:

- How to define the similarity of the profiles?
- From which aspects are two profiles similar?
- Which similarity measurements can calculate the distance of the two profiles?
- Which features define the selected aspects of a profile to calculate the similarity?

As the behavior of users on social media is all related to human beings, countless features are affecting their behavior, and there may be a reason for certain behaviors, while for others, no. Also, the chain of transformation of each act flows through all the users worldwide, as in the butterfly effect, creating a stochastic-dynamic environment, which makes it more challenging to analyze and find behavioral patterns. But the biggest challenge in this kind of research is to define the aspects that profiles can be similar to each other.

This article focuses on Twitter, enabling two-way communication and allows any user to interact with other users quickly and easily. Twitter allows users to generate content by posting "tweets" and sharing other users' content by "retweeting." The proposed architecture considers three aspects of similarity measurement, calculates the similarity, and decides that they are replicated or not. These aspects are *the audience similarity*, who are interacting with the profile and its contents, which is calculated using graph network analytic. *the behavioral similarity* that are ratios of activities of account, for example; for two accounts are tending to constantly post the same amount of tweets in the morning between 9:00 to 12:00 am. *the content similarity*, there are two possibilities for checking how similar the contents are. One is to check how many tweets/retweets are the same or to measure the text similarities using the text classification models to calculate the similarity of the concept of the tweets.

For doing it, the user's timeline is extracted as a list of Tweet objects, which are the entities containing all the information of each tweet. Then in parallel, the audience, the list of the users interacting with the primary user are obtained; besides, the behavioral ratios, the time-series-related features are calculated; moreover, the content, the tweets, and retweets the user has posted are collected. For comparing the audience, the inter-communications network of the primary profile is created and later compared to the other profiles' audience. Also, it is possible to measure the overlap of a user's audience with another.

For the next aspect, the frequency of the user's activities is calculated during the time and being compared to another user's features using dynamic time warping (DTW)[14].

For checking the context similarity, two ways have been taken into account; How many tweets are precisely the same? How much are the concepts of the posts on different users' timelines similar? They can be in the same language or not. Natural language processing techniques have been utilized to preprocess, consisting of the tokenization, translation, dictation correction, lemmatization, and extract features by employing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

TF-IDF. Also, for context similarity measurement and similarity measurement, cosine distance is used

This paper has been organized as follows: In Section 2, the related work is presented, Then, in Section 3, the architecture of the proposed method is described. In section 4, a successful case study is outlined, and its results are over-viewed. Finally, in Section 6, conclusions are drawn, and future lines of research are discussed.

2. Review of the state of the art

Many works have been done on data analysis [15], but the focus is on data extracted from social media platforms [16–18] in this paper. In Social Media Analysis, there are many rooms left for investigation and improvement of the existing tools and algorithms. In the literature, several pieces of research have been done on data extraction. However, it is necessary to consider that each platform has its policies for data extraction and publishing. For instance, Twitter allows researchers to extract public information via Twitter official APIs and conduct academic research to make improvements. However, in general, most of the research in this area is related to taking advantage of social network data and apply Artificial Intelligence algorithms, such as machine learning methods (supervised and unsupervised), deep learning, graph theory, etc. In this paper, the focus is on the calculation of the similarity of the profiles on Twitter.

A social network can be interpreted as a complex network graph consisting of nodes connected by edges. The nodes represent the users in the network, and the edges define the connections between these users. Social network analysis requires specific analysis tools; Akhtar et al. in [19], conducted a comparative study of these tools in general graph analysis and social network analysis. They conducted a comparative study of four social network analysis tools (NetworkX, Gephi, Pajek, and IGraph) based on platform, runtime, graph type, algorithm complexity, input file format, and graph features.

Semantic analysis is major technology in Natural Language Processing (NLP) applications. Such as text similarity estimation, text classification, speed recognition, etc. Chen et al. introduced A framework for semantic similarity detection that is deep reinforcement learning for the Siamese attention structure model (DRSASM). It automatically detects the word segmentation and word distillation features and proposes a new recognition mechanism model to improve semantics [20].

There are many strategies for Similarity detection depending on the final goal, such as Euclidean distance, Pearson correlation coefficient, Spearman's rank correlation coefficient, and others. Chicco et al. review applying The Siamese neural network architecture for complicated data samples that have different dimensions and types of features [21].

Semantic similarity detection in text data is one of the challenging obstacles of Natural Language Processing (NLP). Due to the versatility of Natural language, it is challenging to represent rule-based methods for detecting semantic similarity patterns. Chandrasekaran et al. determines the evolution of several available semantic similarity methods and reviews their pros and cons. Classifies by the underlying policies as corpus-based, hybrid approaches, knowledge-based, and deep neural network-based methods [22].

In [23] they introduced a Siamese model of the Long Short-Term Memory (LSTM) network to assess the semantic similarity between texts that works. They add word-embedding vectors enhanced by synonymic data to the LSTMs, based on a fixed size vector to encode the underlying meaning implied in a sentence. They constrain the sentence representations detected by the proposed model to create a structured space whose geometry reveals complex semantic similarities. It reduces subsequent procedures for relying on a simple Manhattan metric.

Siamese Neural network is a method for computing similarity with demanding less training data. An architecture with language-independent features for finding short text similarity detection in multiple languages and domains proposed in [24]. They used

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

these corpora from shared tasks: ASSIN 1 and ASSIN 2 with Portuguese journalistic texts and N2C2 (English clinical texts). Then implemented the proposed SNN by Mueller et al. in two forms. The evaluation is done by calculating the Pearson correlation (PC) and the Mean Squared Error (MSE) among the models' predicted values and corpora's gold standard. This method held better results in both languages and domains.

BERT is a method to merge topics by pre-trained contextual representations. For pairwise semantic similarity detection, Peinelt et al. proposed a unique topic-informed BERT-based structure. This advanced architecture performance over strong neural baselines beyond different classes of English language datasets. Adding topics to BERT helps in determining domain-specific problems [25].

There are many approaches for improving text classification performance, such as centroid-based classifier, multinomial naïve bayesian (MNB), support vector machines (SVM), convolutional neural network (CNN). However, Park et al. presented a cosine similarity-based methodology to enhance the performance. For increasing the precision of classifiers, This methodology merges cosine similarity and conventional classifiers, And then the Conventional classifiers with cosine similarity are named enhanced classifiers. Enhanced classifiers are applied to famous datasets such as 20NG, R8, R52, Cade12, and WebKB, And they show notable improvements in accuracy. Also, word count and term frequency-inverse document frequency (TFIDF) is more suitable in terms of the performance of the classifier [26].

A variety of users and content in online social networking sites (OSN) will cause a fear of identity theft attacks (profile cloning), malware attacks, or structural attacks of cybercriminals. Profile cloning is stealing existing users' identities and creating duplicate accounts with the existing users' credentials. Chatterjee et al. proposed a way to supervise the threat of profile cloning in social networks. Users can use it to prevent cloned, and fake profiles and identity theft [27].

In [28] a detection technique is proposed for discovering fake and cloned profiles on Twitter. For detecting profile cloning, they used two methods: similarity measures and the C4.5 decision tree algorithm. In Similarity measures, Similarity of characteristics and Similarity of network relations are analyzed. C4.5 applies a decision tree by considering information gain. These two methods help in detecting clone profiles and prevent them.

A framework for finding cloned profiles in social networks is stated in [29]. It will analyze user profiles, friends and follower networks, and posting habits. This framework has three parts: Twitter Crawler, Attribute Extractor, and Cloning Detector. The best classification performance is with the decision tree, and the average accuracy of classifying the real or fake posts was 80%.

In the next section, the proposed method, combining the information extracted from the three aspects of profile behaviors, is presented. A couple of examples of knowledge inquiring of each of these aspects are reviewed in state-of-art, and the ideas have helped design and implement the proposed method.

3. Proposed Model

Twitter which is the most news-friendly social media platform, is the primary focus for applying this tool. However, the proposed methodology can be used on different social media platforms with a few modifications. Twitter has a unique feature among all social media: On Twitter, users can respond to each other's tweets and "like" each other's tweets, or leave a comment, and leave comments to share their opinions and viewpoints. Tweets can include text, photos, videos, links, etc. Users can also share the status of other users' tweets by retweeting them. The data relating to the tweets and some information about the profiles are provided as Tweet Object in JSON format[30]. In this section, the architecture of the platform for measuring the similarity of profiles is presented. The platform extracts data from a user's Twitter timeline, analyzes them, and transforms it into meaningful information.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

Algorithm 1 Proposed Method's Algorithm.

Inputs: Screen_names of the twitter profiles.
Output: Similarity of the profiles
Step 2, 3, 6, 10 are executing in parallel **Begin**

1: for **each** Twitter User
2: **Extract the timeline**
3: **Feature Engineering**
4: *Extracting Primary Features*
5: *Extracting Advanced Features*
6: **The Graph of Audience**
7: *Extracting the Audience*
8: *Building the directed network graph of interactions*
9: *Calculating the graph measurements*
10: **Content Preprocessing**
11: *Text Preprocessing*
12: *Tokenization*
13: *Translation*
14: *Dictation Checking*
15: *Stopwords Removal*
16: *Lemmatization*
17: *TF-IDF Vectorizer*
18: **end fore**
19: **Similarity checking**
20: for **each** Twitter User
21: ***Advanced-Features Similarity Approximation***
22: Dynamic Time Warping (DTW)
23: ***Graph Network Similarity Detection***
24: The number of Same Audience
25: Graph Similarity Detection
26: Jaccard Similarity
27: ***Content Similarity Measuring***
28: The number of same tweets/retweets
29: Cosine Similarity
30: **end fore**

End

The proposed architecture is presented in Fig 1. The designed system aims to calculate the similarity of the profiles based on their network of the audience, behavioral traits, and context similarity. This architecture consists of five main components; *Timeline Extraction*, *Advanced Feature Extraction*, *The Graph of Audience*, and *Content Processing*. These aspects are discussed in their respective subsections.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

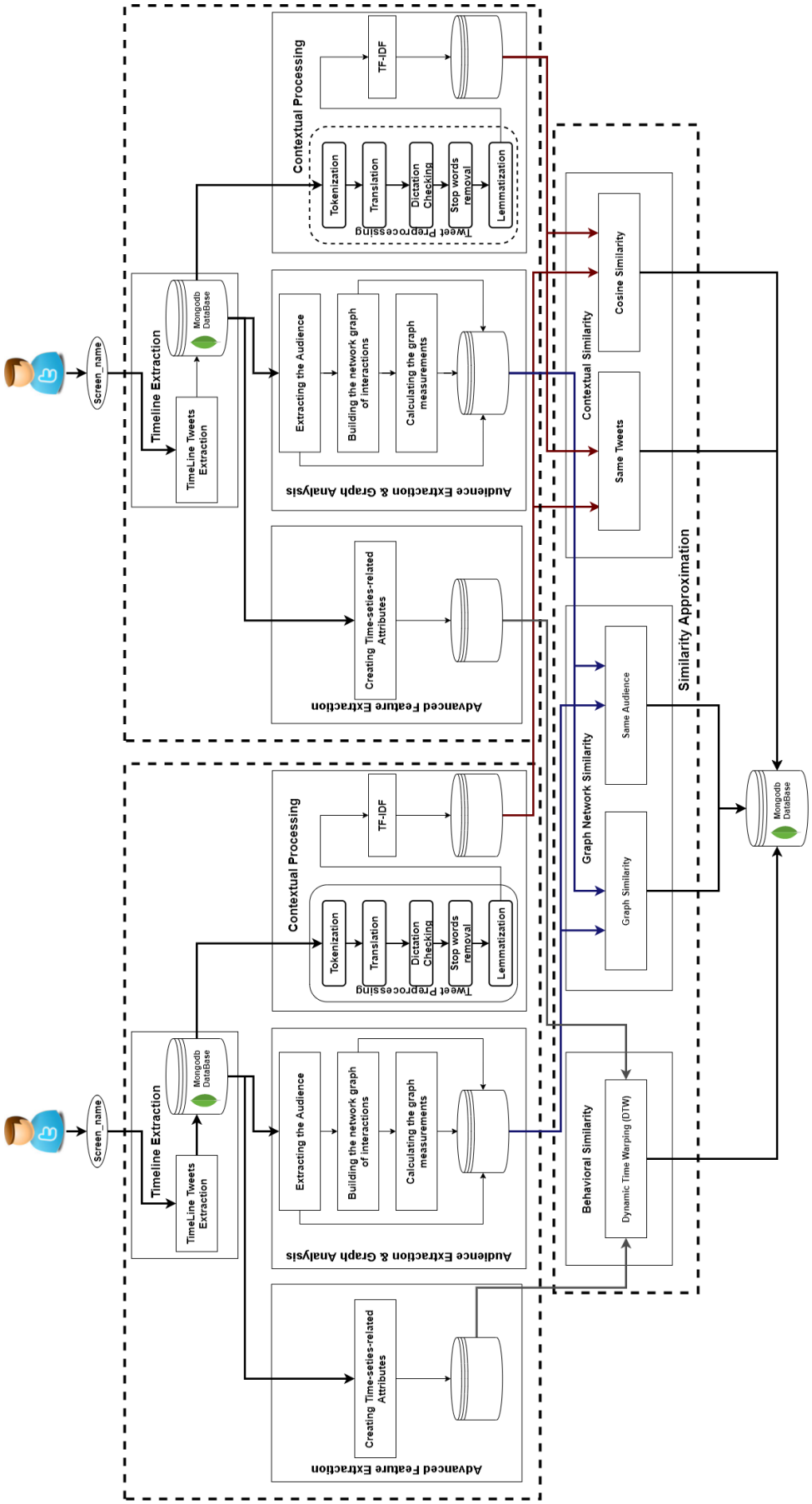

Figure 1. The proposed model for similarity measurement of the profiles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

The first step is extracting the recent tweets of a profile by extracting the user's timeline and storing it as a list of JSON files. Then, in the Advanced Feature Extraction component, the data is restructured. More advanced features are created from the selected primary features, including ratios indicating behavioral features like the number of tweets posted each day. At the same time, from the JSON file extracted from the user's timeline, the audience is the people who are interacting with the chosen profile, defined by the replies and the retweets. In parallel, all tweets of a profile are selected by its language, and not English ones are translated to English; they are pre-processed (first, it is tokenized, then the stop-words are removed, and then each token will be lemmatized and will be transferred to its root). The output of all these components will be handed to the Similarity Measurements component. For each aspect of these new features, a respective similarity measurement has been applied that will be explained in detail each relative sub-section. The outputs of this model give a better understanding of how similar two profiles are by calculating the distance between the users from the mentioned points of view.

This architecture is optimized because it has been designed in the most parallel way possible and consists of four components described in the following subsections.

3.1. Timeline Extraction

Social media platforms like Twitter enable people to distribute and utilize news by interacting with each other and following some policies. The way they share and spread news contains remarkable meaningful hidden information that is interpreted into complex conclusions, from law obligation, [31] to the marketing point of view [32].

In this research, the aim is to calculate the similarity between different Twitter profiles. The first step is to extract the timeline of the user using a component called Timeline Extraction. In this component, the official Twitter API, which is provided by the Twitter development team, has been utilized [33]. It extracts the timeline of the given screen_name of the profile. The 3400 recent tweets on the timeline are extracted using the official Twitter API. The only thing that is mandatory to consider is the API limitations [34]. There are many ways to deal with it. The output of this component is a list of tweet objects containing all the information of the tweets in JSON format.

3.2. Advanced Feature Extraction

In this component, data is restructured, and more complex concepts are derived from the primary features existing in the tweet objects. Especially the behavioral inclination of the user is defined by considering the ratios of activities during the time. Extracting these time-series-related features make us enable to extract behavioral patterns. In other words, calculating these features gives extra information about the user, which is a reliable measure for comparing the similarity of the profiles; for example, *user 1* tends to post on the mornings, however *User 2* has a higher ratio of activities during the night. The output of this component is the set of advanced new features during the user's recent activities. These features have been presented in Table .

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

Table 1: The advanced features extracted from the user's most recent 3240 posts.

Feature	Description
Tweets per day	The number of statuses the user has posted in the recent posts (Tweets + Retweets)
Posted retweets per day	The number of retweets the user has posted in the recent posts
Likes received per day	The number of likes the user has received from the recent posts
Retweets received per day	The number of likes the user has received recently
Replies per day	The number of replies the user has done in the recent posts
Mentions per day	The number of times the user has mentioned others, in the recent posts
Average Tweet Polarities per day	The mean of polarity of the tweets of the user in recent posts , showing how positive or negative the user is posting
Average Tweet Subjectivity per day	The mean of subjectivity ratio of the tweets of the user in recent posts, showing the user is posting the facts or its own ideas

After measuring the time-series-related features, the distance between the same time series features in two profiles will be calculated. Dynamic Time Warping (DTW) [35] is a suitable distance similarity measure that allows the comparison of two time-series sequences with different lengths and speeds. This algorithm is a perfect choice because the time series's length in various features varies from one another and depends on how much the user was active recently. In other words, a user with a high ratio of activities can make the 3240 tweets in one month; however, another user does this amount of posts in three months. Algorithm 2 shows the details of how DTW works. This algorithm computes and returns a dynamic time warping (DTW) similarity measure between (potentially multivariate) time series [36]. In the proposed model, the distance between two time-series features, for example, the distance between the mean number of posted statuses per day, exposes how similar two profiles behave during the extracted timeline.

Algorithm 2 Dynamic Time Warping Algorithm [**?**].

Inputs: TSF1, TSF2: Time series Features.
Output: D: The distance betweenn TSF1, TSF2.
Begin

M(length(TSF1), length(TSF2)) : Cost matrix;
M(1,1) ← (TSF1(1)-TSF2(1))2;
for i ← 2 to length(TSF1)
cost ← (TSF1(i)-TSF2(1))2;
M(i,1) ← cost + D(i-1,1);
end for
for j ← 2 to length(TSF2)
cost ← (TSF1(1)-TSF2(j))2;
M(1,j) ← cost + D(j-1,1);
end for
for i ← 2 to length(TSF1)
for j ← 2 to length(TSF2)
cost ← (TSF1(i)-TSF2(j))2;
M(i,j) ← cost + minD(i-1,j), D(i,j-1), D(i-1,j-1);
end for
end for
D ← sqrt(M(length(TSF1),length(TSF2)));
Return D;

End

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

Table 2 is an example of the similarity of two different features. Understandably, the @User_4 and @User_5 are more similar in the behavioral ratios (likes received, posted statuses, etc.). The results of applying the Dynamic Time Warping algorithm have been presented in Table 2.

Table 2: The advanced features extracted from the user's most recent 3240 posts.

Users	Posted Statuses	Posted Retweets	Posted Tweets	Likes Received	Retweets Received	Polarity	Subjectivity
@User_1 & @User_2	104.50	10.23	10.23	634612.10	73611.20	7.16	7.36
@User_2 & @User_3	114.26	6.78	12.13	48.90	611.73	27.65	16.35
@User_4 & @User_5	102.44	0.40	0.05	0.10	0.00	0.20	0.02

3.3. The Graph of Audience

The attitude of any human being towards their environment and others can tell a lot about them. One of the most significant sources of information is by achieving information about the users who are connected to a specific user. An audience is a group of users interacting with the user via retweeting, quoting, replying, and mentioning. Mapping this information into a directed graph makes analysts derive further information by simply considering the nodes as users and the edges as the connection. That is one of the possible ways of retweeting, quoting, replying, and mentioning. Categorizing the audience based on the frequency of links, which is called weights of digraph, is a way to measure the acquaintanceship of the audience of the profile. A graph sample has been depicted in Fig. 2. In this sample case, the scenario is as below:

- @User_1 has mentioned @User_2, 4 times
- @User_3 has retweeted from @Use_2, twice
- @User_3 has quoted a tweet from @User_1, 2 times
- @User_4 has replied on the statuses of @User_1 and @User_3, once each

The relationship matrix of this scenario is shown in Table 3.

Table 3: The Graph Relationship Sample.

Source	Target	Weight
@User 1	@User 2	4
@User 3	@User 2	2
@User 3	@User 1	2
@User 4	@User 1	1
@User 4	@User 3	1

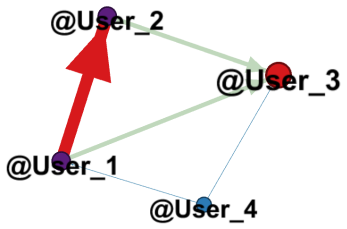

Figure 2. Sample graph of the user relationships.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

In the "Graph of Audience" component, first, relationships between people and the list of the screen names of the audience in contact with the primary user are extracted. After these steps, the directed graph showing the relationships of the user and its audience is built. By utilizing a directed graph, some hidden information about the nodes is derived, like the importance of each node, tightly connected nodes that signify close friendship. These features have been presented in table 4.

Table 4: The advanced features extracted from the user's graph of audience.

Feature	Description
Eccentricity	The maximum shortest distance of one node from another node. To compute this metric, you need a strongly connected graph. The smaller the eccentricity, the greater the force with which a node affects other nodes.
Clustering Coefficient Centrality	Depending on the degree of nodes in the network, nodes that tend to be in the same cluster.
Closeness Centrality	By retrieving the average distance from one vertex to another, it shows how close the node is to other nodes in the network.
Betweenness Centrality	It indicates the degree of influence of a node. The higher the value of link centrality, the more important the node is to the shortest path in the network. Therefore, if this node is deleted, many links will be lost.
In-Degree Centrality	This centrality shows its significance by the number of edges included in the node.
Out-Degree Centrality	This centrality is validated by the number of edges leaving the node.
Degree Centrality	It measures the number of connections a node has. In other words, it is the sum of a node's in-degree and out-degree, and it is a measure of a node's effectiveness in terms of the number of connections it has.

For calculating the similarity between the two graphs, Jaccard's similarity has been applied to the graph of the audience of different profiles. It's a classical measure of the similarity between two sets, introduced by Paul Jaccard in 1901 [37]. Given two graphs of audiences of @User_A and @User_B, Jaccard's similarity is measured by the division of the number of nodes from the audience of @User_A that exists in the audience of @User_B of a total number of nodes minus the same nodes in both graphs.

$$jaccard(Graph\_1, Graph\_2) = \frac{|Graph\_1 \cap Graph\_2|}{|Graph\_1 \cup Graph\_2|} \quad (1)$$

The intersection of two graphs is pointing out the nodes that are common between two sets of graphs, and the union of the graphs means to sum the number of the audience of each profile but remove the ones that are repeated (the intersection).

3.4. Content Processing

In this step, the similarity of tweets is calculated in two aspects; the number of the same tweets and the likeness of the content. The number of the same tweets is easy to calculate by checking the content of two timelines. However, for calculating the similarity in context, the text of all tweets extracted before is preprocessed; by going through the translation, tokenization, checking the dictation of each token, removing stopwords, and lemmatization. The aim is to calculate how similar are the tweets of the two profiles. These tweets are posted in different languages. For unifying these tweets, they need to be turned into the same language, English in this case. Google translate API [38], covers a very vast range of different languages.

After unifying the tweets in different languages, *tokenization* is performed. Tokenization is an essential step of working with text data. It is the act of separating the words in a sentence, i.e., tweet, into smaller units called *token*s. After applying this step,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

the sentence of words is turned into a list of tokens. It is necessary to check the dictation of the tokens due to the limitation of the number of characters that are possible to post as a tweet, which is 280 characters; users usually abbreviate the words to add more information to the tweet. Hens, returning these abbreviations to the original terms is necessary.

Then, the tokens are compared to a set of words called *stop words* [39] in English, which are words filtered out of the text data because they are the most frequently repeated words not carrying so much information. Therefore, in this step, the number of tokens is decreased.

The final set of tokens goes through the *lemmatization* [40] method, which is the process of returning the word to its root. For instance, the root of "plays," "playing," and "played" is played. By employing this method, different words with the same context are considered the root, one unique expression. Next, the tokens of all the tweets on the user's timeline are merged as a big cleaned-up text ready to be compared with the same cleaned-up content of the other profile.

TF-IDF stands for Term Frequency Inverse Document Frequency. It is a very well-liked algorithm for converting text into a meaningful representation of numbers to adapt machine learning algorithms for prediction. The count vectorizer provides the frequency count for the word index, and tf-idf considers the overall word weight document [41]. After applying this vectorization method, the distance between these vectors is calculated using the Cosine similarity algorithm.

Cosine similarity is a measure used to assess the similarity of documents, regardless of their size. Mathematically, it measures the cosine of the angle between two vectors projected in multidimensional space. Cosine similarity is helpful because two similar documents can be separated by a Euclidean distance (due to the size of the document) but oriented closer to each other. The smaller the angle, the greater the cosine similarity [42].

4. Case Study and Results

As a case study, three profiles that Twitter verifies have been selected. Two of them are politicians, and one is a famous singer. Table 5 represents some information about these profiles.

Table 5: The information related to each selected user's profile.

Users	Followers No.	Followings No.	Gender	Sector
@User_1	28.9M	47	Male	Politician
@User_2	83.9M	120.4K	Female	Singer
@User_3	30.9M	914	Female	Politician

Below the performance of the proposed model is presented. It is expected that the two politician profiles have a shorter distance, meaning a higher similarity than the other profile. However, two of these profiles are females, so there might be a hidden pattern correlated with gender.

Fig 3 represents a comparison of the activity level of each profile. As shown, @User_1 tends to post original tweets more than retweeting others; his posts have a higher user engagement by having higher retweet and favorite ratios. And he is more tending to post neutral content. On the other hand, @User_3, the female politician, makes more retweets and has a stronger tendency to post her opinion than the @User_1. On the contrary, @User_3, who is the singer, tends to post positive content, mostly her ideas than facts. Approximately she has a steady behavior in posting tweets and retweets, and her fans also have a constant engagement in comparison to the other two profiles. Also, both politicians have a higher user engagement on the weekends.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

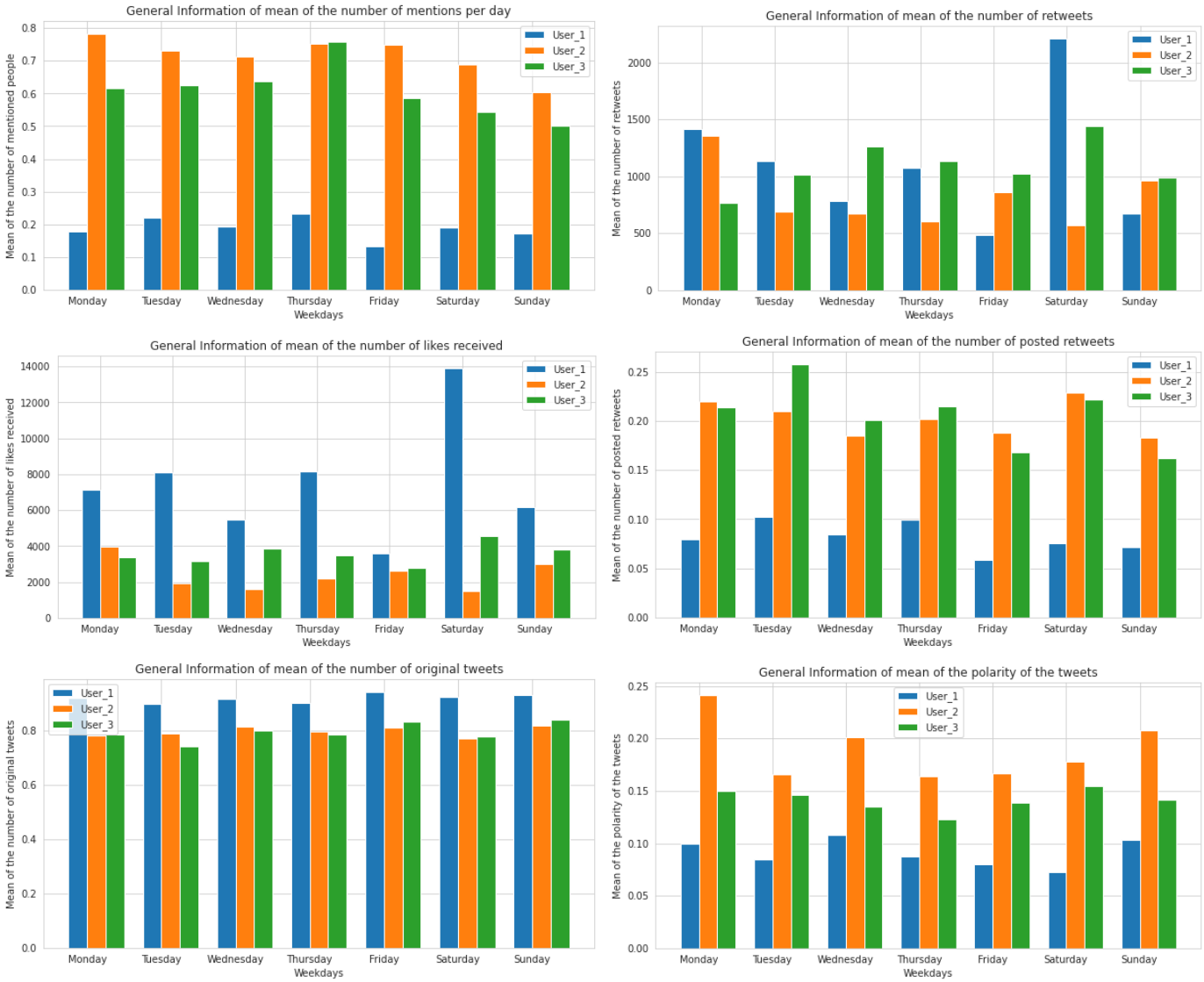

Figure 3. The general information of the time series related features of the three users based on the weekdays.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

The time-series-related features have been depicted in Fig 4 for having a detailed look at the performance of the @User_1.

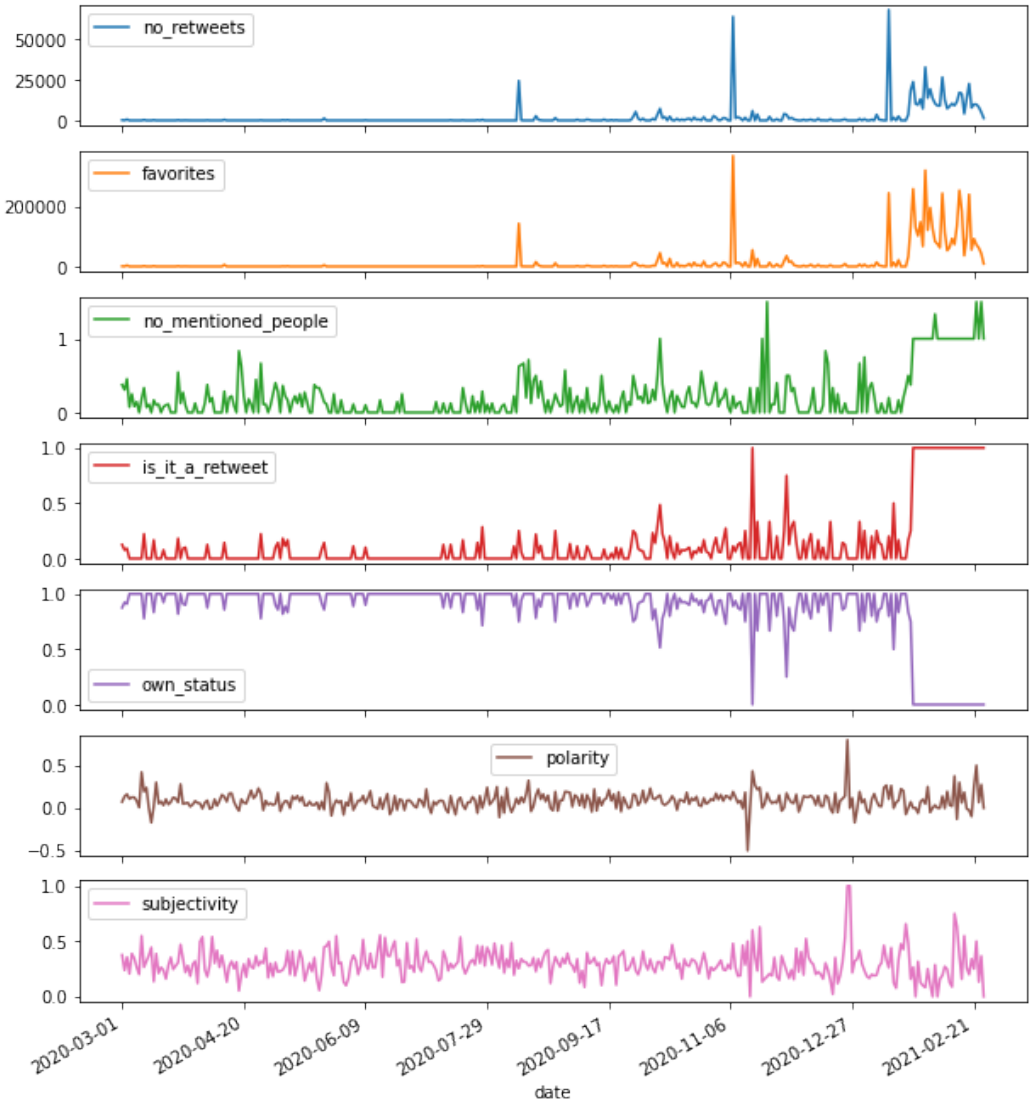

Figure 4. The ratios of time-series features of User 1.

Fig 5 shows the performance of the @User_1 based on daily, weekly, and monthly activity ratios. These plots show the user's activity level and the engagement of the user's posts.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

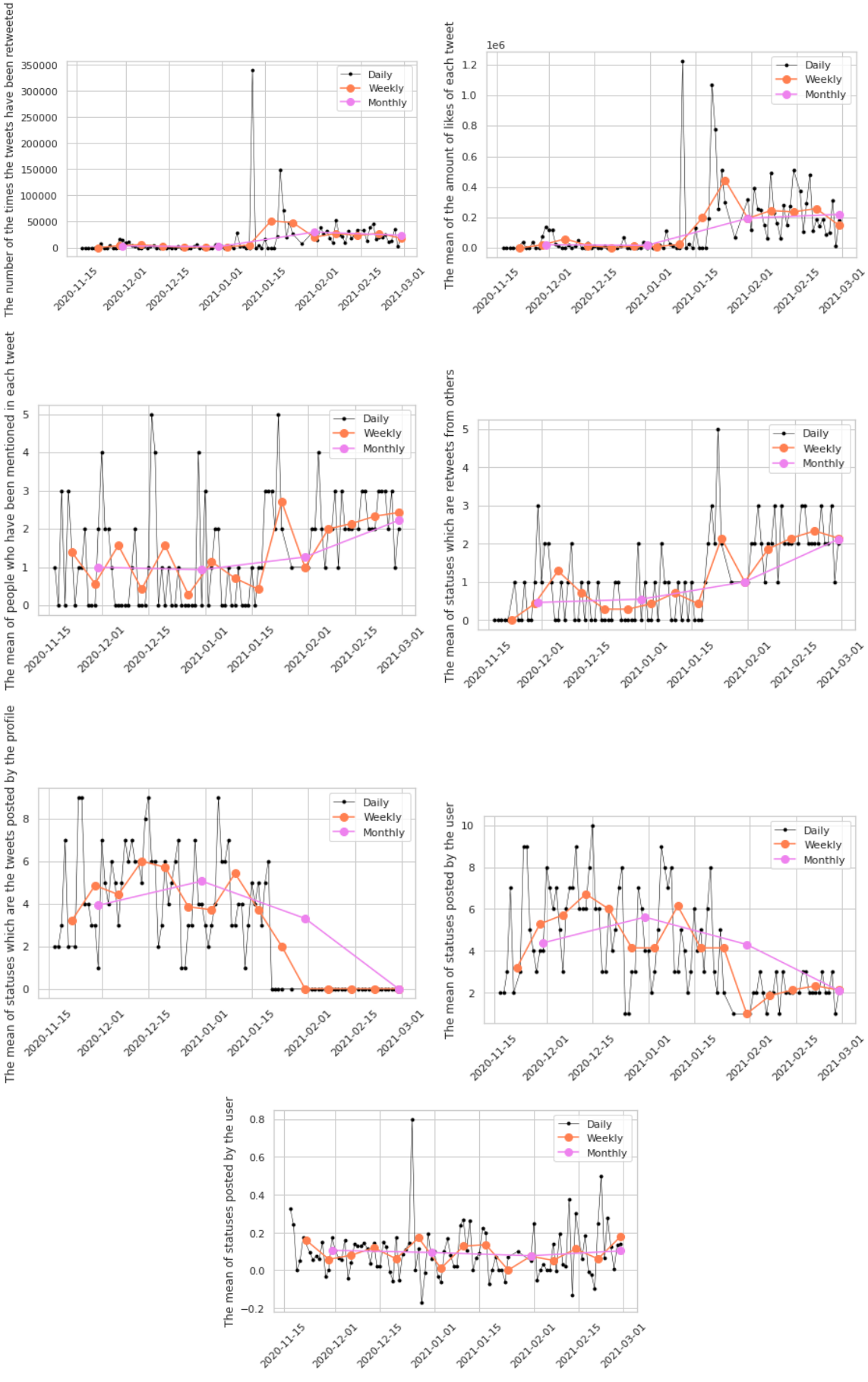

Figure 5. The ratios of time-series features of User 1 in daily, weekly, and monthly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

Fig. 6 is a graph sample of the audience of the @user_1.

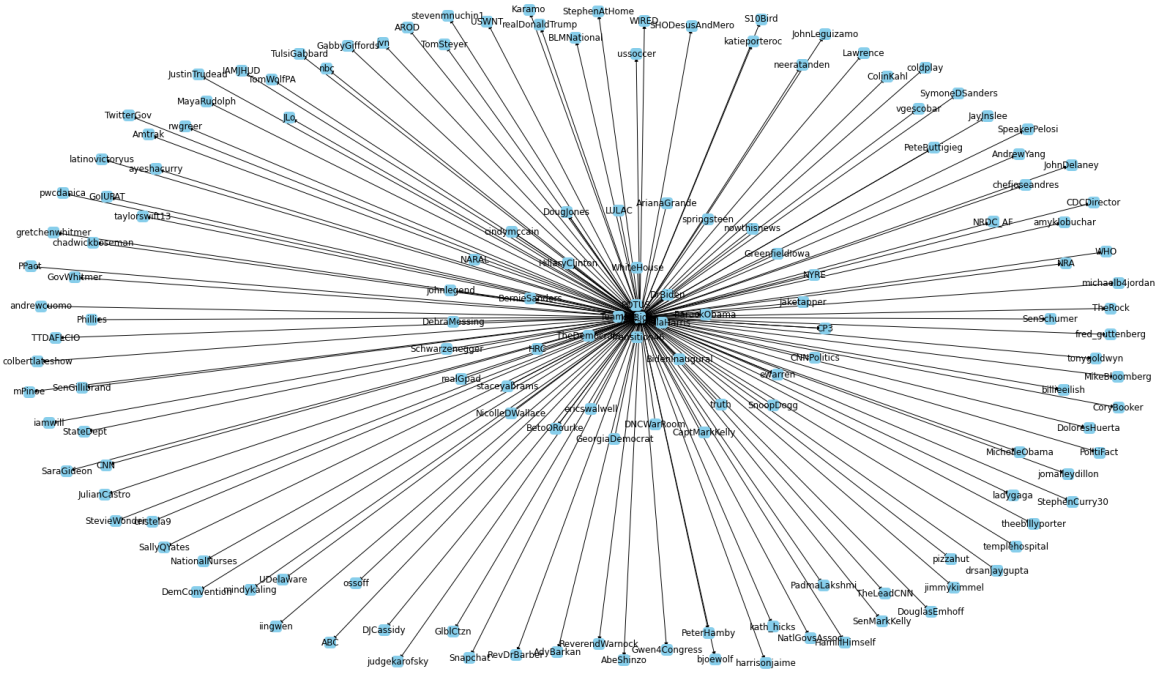

Figure 6. The graph of audience of the @User_1.

There are two ways of comparing two graphs with each other, comparing the graph's structure and reaching the audience of the primary node. Table 6 shows the number of same nodes existing in the graph of audiences of the three selected users.

Table 6: The audience similarity comparison of the three selected users.

	@User_1	**@User_2**	**@User_3**
@User_1	1.00	0.32	0.65
@User_2	0.32	1.00	0.46
@User_3	0.65	0.46	1.00

Results showed that @User_1 and @User_3 have seven same tweets and don't have any tweets the same as @User_2. On the other hand, after content processing, the language is unified to English, and all tweets are preprocessed. Then all tokens are merged to create a large text. Then the Tf-idf vectorization is applied to the text. Finally, the cosine similarity is computed to understand the conceptional similarity of different tweets even in different languages. Table 7 and Fig. 7 show the similarity of these three profile with each other's content.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 June 2021 doi:10.20944/preprints202106.0196.v1

Table 7: The content similarity of the three chosen users calculated by cosine similarity.

Users	Similarity
@User_1, @User_2	38.1%
@User_2, @User_3	54.6%
@User_1, @User_3	74.5%

As expected, users 1 and 3 that both are politicians are more similar in posted content; apart from seven same tweets, they have a 0.745 similarity score. On the one hand, @User_2, a musician, is less similar to the two other users. On the other hand, its content is more similar to @User_3; gender may play a role in the similarity of the users.

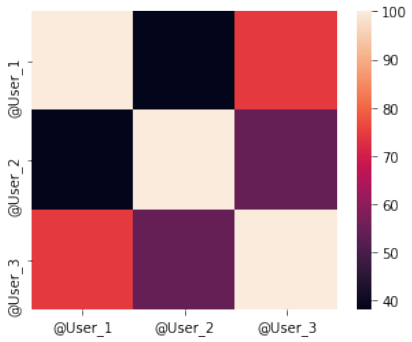

Figure 7. The heatmap plot of the content similarity of the 3 chosen users.

5. Conclusion and Future work

Many studies have been done to quantify the behavior of human beings interacting through social media platforms. Still, the question is that how can we compare these profiles with each other? From which aspects can a profile expresses itself? What kind of analysis can be used for comparing these aspects? In this paper, a method has been proposed to calculate the distance of two profiles. The distance of the profiles can be measured in three different ways; the behavioral ratios, the graph of the audience, and the contents they post. After extracting the recent 3240 tweets on the timeline of each profile using official Twitter APIs, the data is preprocessed in three ways regarding each aspect of similarity measurement.

First, the behavioral ratios are calculated using Dynamic time warping (DTW), an algorithm that calculates the distance between time series features and the distance. This measurement enables us to understand better the difference between the behavioral ratios of activity like the number of posts in the day, retweets posted by the user, etc. Moreover, there is extra information about the ratios of user engagement of the profile because it considers the number of likes and retweets of the profile's audience per day. On the next aspect, the user's audience is extracted by defining the relationship between the profiles, which are replies, retweets, quotes, and mentions. In this step, the directed graph of the user's audience that is interacting with each other is built, and using the Jaccard similarity, the similarity between the two graphs of the two selected users is calculated. The results show that the users in the same sector category have more similarities in the audience graph. Finally, in the content similarity measurement, the number of the same tweets will be calculated. Then, based on the content, all the tweets are unified into the same language, English. The text is preprocessed by employing natural language processing techniques; tokenized, stop words removed, lemmatized. Then by applying TF-IDF, tweets will turn into vectors, and by using cosine similarity, the similarity between two vectors is calculated.

In the future, investigation of the role of gender and similarity of users on Twitter is on the plan, also, combining the information from different social media platforms to create a general purpose social meida analytical platform based on the deepint.net platform, supporting all types of data and contains a full suite of artificial intelligence techniques for data analysis, including data classification, clustering, prediction, optimization, and visualization techniques [43] is on the plan. The abilities provided by deepint make it a perfect choice for implementing the proposed model.

Acknowledgement

This research was partially Supported by the project "Computación cuántica, virtualización de red, edge computing y registro distribuido para la inteligencia artificial del futuro", Reference: CCTT3/20/SA/0001, financed by Institute for Business Competitiveness of Castilla y León, and the European Regional Development Fund (FEDER).

1. Allman-Farinelli, M.; Nour, M. Exploring the role of social support and social media for lifestyle interventions to prevent weight gain with young adults: Focus group findings. *Journal of Human Nutrition and Dietetics* **2021**, *34*, 178–187.
2. Thelwall, M. Word association thematic analysis: A social media text exploration strategy. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **2021**, *13*, i–111.
3. Osorio-Arjona, J.; Horak, J.; Svoboda, R.; García-Ruíz, Y. Social media semantic perceptions on Madrid Metro system: Using Twitter data to link complaints to space. *Sustainable Cities and Society* **2021**, *64*, 102530.
4. Alamsyah, A.; Rahardjo, B.; others. Social network analysis taxonomy based on graph representation. *arXiv preprint arXiv:2102.08888* **2021**.
5. Li, Z.; Zhang, Q.; Du, X.; Ma, Y.; Wang, S. Social media rumor refutation effectiveness: Evaluation, modelling and enhancement. *Information Processing & Management* **2021**, *58*, 102420.
6. Choudhary, A.; Arora, A. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications* **2021**, *169*, 114171.
7. Derhab, A.; Alawwad, R.; Dehwah, K.; Tariq, N.; Khan, F.A.; Al-Muhtadi, J. Tweet-based Bot Detection using Big Data Analytics. *IEEE Access* **2021**.
8. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A.; Abayomi-Alli, A. A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications* **2021**, *173*, 114762.
9. Albalawi, R.; Yeap, T.H.; Benyoucef, M. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence* **2020**, *3*, 42.
10. Dhiman, A.; Toshniwal, D. An Approximate Model for Event Detection From Twitter Data. *IEEE Access* **2020**, *8*, 122168–122184.
11. Wu, W.; Chow, K.P.; Mai, Y.; Zhang, J. Public Opinion Monitoring for Proactive Crime Detection Using Named Entity Recognition. IFIP International Conference on Digital Forensics. Springer, 2020, pp. 203–214.
12. Martyniuk, H.; Kozlovskiy, V.; Lazarenko, S.; Balanyuk, Y. Data Mining Technics and Cyber Hygiene Behaviors in Social Media. *South Florida Journal of Development* **2021**, *2*, 2503–2515.
13. Sushama, C.; Kumar, M.S.; Neelima, P. Privacy and security issues in the future: A social media. *Materials Today: Proceedings* **2021**.
14. Ge, L.; Chen, S. Exact Dynamic Time Warping calculation for weak sparse time series. *Applied Soft Computing* **2020**, *96*, 106631.
15. Hernández-Nieves, E.; Parra-Domínguez, J.; Chamoso, P.; Rodríguez-González, S.; Corchado, J.M. A Data Mining and Analysis Platform for Investment Recommendations. *Electronics* **2021**, *10*, 859.
16. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2020**, *10*, e1355.
17. Di Minin, E.; Fink, C.; Hausmann, A.; Kremer, J.; Kulkarni, R. How to address data privacy concerns when using social media data in conservation science. *Conservation Biology* **2021**, *35*, 437–446.
18. Rivas, A.; Chamoso, P.; González-Briones, A.; Casado-Vara, R.; Corchado, J.M. Hybrid job offer recommender system in a social network. *Expert Systems* **2019**, *36*, e12416.
19. Akhtar, N.; Ahamad, M.V. Graph tools for social network analysis. In *Research Anthology on Digital Transformation, Organizational Change, and the Impact of Remote Work*; IGI Global, 2021; pp. 485–500.
20. Chen, G.; Shi, X.; Chen, M.; Zhou, L. Text similarity semantic calculation based on deep reinforcement learning. *International Journal of Security and Networks* **2020**, *15*, 59–66.
21. Chicco, D. Siamese neural networks: An overview. *Artificial Neural Networks* **2021**, pp. 73–94.
22. Chandrasekaran, D.; Mago, V. Evolution of Semantic Similarity–A Survey. *arXiv preprint arXiv:2004.13820* **2020**.
23. Mueller, J.; Thyagarajan, A. Siamese recurrent architectures for learning sentence similarity. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, Vol. 30.

 doi:10.20944/preprints202106.0196.v1

24. de Souza, J.V.A.; Oliveira, L.E.S.E.; Gumiel, Y.B.; Carvalho, D.R.; Moro, C.M.C. Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages. International Conference on Computational Processing of the Portuguese Language. Springer, 2020, pp. 357–367.
25. Peinelt, N.; Nguyen, D.; Liakata, M. tBERT: Topic models and BERT joining forces for semantic similarity detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7047–7055.
26. Park, K.; Hong, J.S.; Kim, W. A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence* **2020**, *34*, 396–411.
27. Chatterjee, M.; others. Detection of Fake and Cloned Profiles in Online Social Networks **2019**.
28. Sowmya, P.; Chatterjee, M. Detection of Fake and Clone accounts in Twitter using Classification and Distance Measure Algorithms. 2020 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2020, pp. 0067–0070.
29. Punkamol, D.; Marukatat, R. Detection of Account Cloning in Online Social Networks. 2020 8th International Electrical Engineering Congress (iEECON). IEEE, 2020, pp. 1–4.
30. Haustein, S. Scholarly twitter metrics. In *Springer handbook of science and technology indicators*; Springer, 2019; pp. 729–760.
31. Zahra, K.; Imran, M.; Ostermann, F.O. Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management* **2020**, *57*, 102107.
32. Sheth, J.; Kellstadt, C.H. Next frontiers of research in data driven marketing: Will techniques keep up with data tsunami? *Journal of Business Research* **2021**, *125*, 780–784.
33. Twitter API Documentation | Docs | Twitter Developer.
34. rate limits | docs | twitter developer, url=https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits journal=Twitter, publisher=Twitter.
35. Lahreche, A.; Boucheham, B. A fast and accurate similarity measure for long time series classification based on local extrema and dynamic time warping. *Expert Systems with Applications* **2021**, *168*, 114374.
36. Berndt, D.J.; Clifford, J. Using dynamic time warping to find patterns in time series. KDD workshop. Seattle, WA, USA:, 1994, Vol. 10, pp. 359–370.
37. Gosliga, J.; Gardner, P.; Bull, L.; Dervilis, N.; Worden, K. Foundations of Population-based SHM, Part II: Heterogeneous populations–Graphs, networks, and communities. *Mechanical Systems and Signal Processing* **2021**, *148*, 107144.
38. Vollmer, S. Google Translate. In *Figures of Interpretation*; Multilingual Matters, 2021; pp. 72–75.
39. Kumar, G.K.; Rani, D.M. Paragraph summarization based on word frequency using NLP techniques. AIP Conference Proceedings. AIP Publishing LLC, 2021, Vol. 2317, p. 060001.
40. Hilario, M.; Esenarro, D.; Petrlik, I.; Rodriguez, C. Systematic Literature Review of Sentiment Analysis Techniques. *Journal of Contemporary Issues in Business and Government* **2021**, *27*, 506–517.
41. Subba, B.; Gupta, P. A tfidfvectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes. *Computers & Security* **2021**, *100*, 102084.
42. Aljuaid, H.; Iftikhar, R.; Ahmad, S.; Asif, M.; Afzal, M.T. Important citation identification using sentiment analysis of In-text citations. *Telematics and Informatics* **2021**, *56*, 101492.
43. Corchado, J.M.; Chamoso, P.; Hernández, G.; Gutierrez, A.S.R.; Camacho, A.R.; González-Briones, A.; Pinto-Santos, F.; Goyenechea, E.; Garcia-Retuerta, D.; Alonso-Miguel, M.; others. Deepint. net: A Rapid Deployment Platform for Smart Territories. *Sensors* **2021**, *21*, 236.