Followers Tell Who an Influencer Is

Dheeman Saha
Department of Computer Science
University of New Mexico
Albuquerque, USA
dsaha@cs.unm.edu

Md. Rashidul Hasan
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, USA
mdhasan@unm.edu

Abdullah Mueen
Department of Computer Science
University of New Mexico
Albuquerque, USA
mueen@cs.unm.edu

ABSTRACT

Influencers are followed by a relatively smaller group of people on social media platforms under a common theme. Unlike the global celebrities, it is challenging to categorize influencers into general categories of fame (e.g., Politics, Religion, Entertainment, etc.) because of their overlapping and narrow reach to people interested in these categories.

In this paper, we focus on categorizing influencers based on their followers. We exploit the top-1K Twitter celebrities to identify the common interest among the followers of an influencer as his/her category. We annotate the top one thousand celebrities in multiple categories of popularity, language, and locations. Such categorization is essential for targeted marketing, recommending experts, etc. We define a novel FollowerSimilarity between the set of followers of an influencer and a celebrity. We propose an inverted index to calculate similarity values efficiently. We exploit the similarity score in a K-Nearest Neighbor classifier and visualize the top celebrities over a neighborhood-embedded space.

KEYWORDS

Influencer, t-SNE, inverted index, Twitter

ACM Reference Format:
Dheeman Saha, Md. Rashidul Hasan, and Abdullah Mueen. 2023. Followers Tell Who an Influencer Is. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543873.3587576

1 INTRODUCTION

Micro-influencers are accounts with anywhere from 1,000 to 100,000 followers, and macro-influencers are accounts with 100,000 to 10 million followers [1]. Micro-influencers are increasingly proven useful in targeted marketing, recommending experts as they can promote items cost-efficiently and easily target the regional audience more effectively than the macro-influencers. As stated in [2], the market share of micro-influencers continues to grow from a 91% share in 2021. On Twitter, there are more than 100,000 users who can be qualified as micro-influences according to the above definition. Consider a micro-influencer `@richardmadden` in Figure-1. The user is popular among his followers because of his sporting reporting.

Figure 1: An Example Profile of a Micro-Influencer who is commonly confused with a celebrity.

In this paper, we focus on how we can find the area of fame (e.g. Politics, Sports, Entertainment, etc.) of an influencer. In general classifying macro-influencers is relatively easier as their identities are already established in the market. But classifying micro-influencers is challenging because of mismatches between their appearances and skills and variation in followers due to multiple skills and identities. Consider the example of `@richardmadden` on Twitter. The user is followed by around two thousand followers, putting him in the micro-influencer class. A text-based classifier would confuse his name with the actor Richard Madden (`@_richardmadden`), who has more than four hundred thousand followers and would categorize the account as an Entertainment influencer. However, the followers of `@richardmadden` follow Jeremy Clarkson (`@JeremyClarkson`, host of Top Gear, the auto show) the most. Because both Jeremy Clarkson and Richard Madden, the micro-influencer, gained their followers through their work as affiliates of BBC News. Thus, the correct categorization of news/media influencer is achieved by tracing the followers back to existing celebrity accounts. Another contrasting example would be of the soccer player `@LuisSuarez9` with 16M followers, and `@luissuarez`, with 2K followers. Although `@luissuarez` is an engineer, his followers are mistakenly following him because of his identical name, and they follow many other soccer players. Thus, the correct categorization of `@luissuarez` is a sports influencer as shown in the Figure- 2.

In order to perform follower-based classification, we face two major challenges. First, we must select a set of celebrity users and manually label them. The users must give us comprehensive coverage of the Twitter users, and the set of celebrities should be small

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587576

Figure 2: (Left) Soccer Player, (Right) Sports Influencer

enough for us to label manually. We have selected the top-1K users with decreasing number of followers. However, this brings us to the next challenge in collecting their followers. The top-1K celebrity users cover more than 99% of the registered users on Twitter; the total number must be more than 1 billion users. In this paper, we have extracted all the followers of the top-1K celebrity users and exploited an inverted index structure to calculate a novel similarity score to categorize influencer accounts. To facilitate the classifier, we have manually labeled the top-1K celebrities across popularity groups.

We consider two separate groups of influencers: macro- and micro-influencers. Although their number of followers can easily identify them, the two groups show very different characteristics in many other ways. Table-1 provides a side-by-side comparison. In this work, we evaluate the classification performance of both groups of influencers.

Description Type	Micro-Influencer	Macro-Influencer
Number of followers	1K to 10K	1M to 5M
Target Audience	Regional	Global
Cost of Marketing	Lower	Higher
Audience Engagement	Higher	Lower
Marketing Type	Sales Driven	Brand Awareness
Promote New Items	Highly Likely	Less Likely
Easier To Collaborate	Yes	No

Table 1: Characteristic Overview of Micro-Infuencer and Macro-Influencer

The rest of the paper describes the data pre-processing, methods, experimental evaluation, and cluster visualization.

2 RELATED WORK

Macro- and micro-influencers are studied in several prior works [3][4]. The authors of this work found that micro-influencers have a higher personal engagement with the products, and consumers are more likely to purchase products due to their close connection with the micro-influencers. Thus, micro-influencers play a significant role in generating revenues and support the idea of sales-driven marketing, as summarized in Table-1.

Several works have looked at the classification of influencers. In [5], the authors proposed a deep learning multi-modal to classify an influencer based on text and image features. However, their method mostly focuses on macro-influencers, who post more frequently and are very specific to target products. Authors in [6] and [7] use image information of the brand to classify influencers. However, such methodology will again encounter challenges in detecting micro-influencers; who simultaneously promote multiple products, often novel ones. Therefore, it is challenging to classify micro-influencers using image and text features.

Another approach to classifying the influencers is using textual information only. The authors in [8] distinguish influencers from the non-influencers using word embedding methods. Authors in [9] used tweets with the Glove [10] word embedding model to identify influencers, arguing that influencers tweet about hot topics more than others. Authors in [11] and [12] classified influencers by exploring the social network about the influencers. Such work scales poorly for the lack of knowledge about the influencers' followers.

Moreover, a few authors worked on *detecting* micro-influencers. In [13], the authors proposed an automated micro-influencers detection algorithm using their tweets related to trending topics and a scoring mechanism to identify them. However, this approach is not relevant to determine the types of micro-influencer. Note that detection is not the same as classification. Furthermore, the authors in [14] specifically classified micro-influencer who are politically affiliated.

Our FollowerSimilarity score enables community detection in a way prior work has not considered. Most prior work [15][16][17][18] considers node density as the characteristics of a community. For example, in [11], the authors use the centrality measure to select relevant customer networks. In another example, in [19], the authors use k-shell decomposition to partition the network and identify the influencers contributing to the information propagation. Unlike these works, our method uses the top celebrities as a projection space to calculate similarity. The FollowerSimilarity is invariant to node density. Authors in [20] create a deep neural network to encode the node similarities, which is less interpretable than our FollowerSimilarity score.

We summarize our uniqueness with respect to the related work below.

- Our work is unique in its approach and robustness. We have collected a sheer number of followers (1.18 billion) of the top celebrities and spent much of the effort organizing the data in a searchable structure. This enables a follower-based classification technique for micro-influencers that no prior work has attempted.
- We manually label the top one thousand celebrities in the three categories with several sub-groups. This enables the nearest neighbor classification of influencers with interpretability.
- Follower-based classification has a unique advantage over text, image, and time-based classification. The technique does not need posts from the influencers; hence, it can classify inactive influencers. Without using any tweet, our approach can identify groups of users with similar language, location, and interest profiles.

3 BACKGROUND

3.1 Twitter API

We use the Python Twitter API library named *tweepy* to collect the description of a given user, including the number of followers. The two main functions we use are *followers_ids* and *lookup_users*. The function *followers_ids* can request a maximum of 5000 follower

IDs, and Twitter API allows 15 such calls in 15 minutes. Thus, it is possible to collect 75K follower IDs within 15 minutes. The other function *lookup_users* can collect profile information for up to 100 users in a single request, and all the users' profile information are collected using this function.

3.2 Data Collection

The list of the top 1000 users (i.e., celebrities) with decreasing number of followers on Twitter is collected from a well-maintained website [21]. We compiled the list on 9th March 2021, after which some celebrity accounts were deactivated/suspended, including Ariana Grande `@arianagrande` and Aamir Khan `@aamir_khan`. The crawling for the user IDs started from the bottom of the list and gradually moved to the top because the listed users at the bottom have much fewer follower numbers than the top ones. Thus, we can easily collect all the follower IDs of a given user. As stated earlier, the Twitter API provides 75K follower IDs every 15 minutes. We used two API keys for the crawling, where the second key is used only when the other is in the 15 minutes waiting period. We collect, on average, 14M follower IDs daily, requiring over 140 days to accumulate about 1.18 billion unique follower IDs. We want to emphasize that a large volume of Twitter user IDs has never been collected by any research project. Note that the number of active users reported by Twitter hovers around 300 million.

3.3 Data Annotation

We have manually annotated each of these top-1K users in three categories among different groups: *Popularity*, *Country* and *Language*. The three authors of this paper extensively performed the annotation task for five days and discussed the conflicting labels to agree on the best labels for each category. Such an extensive categorization of the top celebrities has not been attempted before. The labeled data can be downloaded from the link[1].

In the *Popularity* category, a user is labeled as one of the following six: *Entertainment*, *Sport*, *Politics*, *Corporate/Company*, *News/Media*, and *Religion*. We have chosen these six categories as a first attempt to group the top celebrities in an approximately balanced manner. In the process, we have merged a few smaller groups into one. For example, bloggers (both text and video) are grouped in the *News/Media* group. Although we finalize only six major popularity types, there are instances where we select one among multiple types, e.g., the athlete `@KingJames` (LeBron James) is an actor too. We concluded that his popularity is in the *Sports* category primarily.

The *Country* category is often challenging to determine. Let us consider the user Chicharito Hernandez (`@ch14_`), a Mexican soccer player currently playing for the LA Galaxy, USA club. Therefore, we labeled the *Country* of this user with the USA instead of Mexico.

The *Language* category is simpler to label among the categories. However, some European users have tweeted in multiple languages, e.g., the user Antoine Griezmann (`@AntoGriezmann`) is a native French speaker. In contrast, he has tweeted in Spanish and English in the past. For such users, the recent tweets are collected, and then the frequencies of the languages are calculated. The most frequent language is considered in the *Language* category.

[1]https://sites.google.com/view/twitterinfluence/home

4 METHODOLOGY

4.1 Data Organization

Our dataset has over 1 billion user IDs. The followers of the top-1K users are massively overlapping. The least number of followers among top-1K celebrities is Fifth Harmony (`@fifthharmony`) with 5M followers.

Consider two sports celebrities: Neymar Jr. (57M) and Christiano Ronaldo (101M). They have 13.89M followers in common. On average, users follow ten of the top-1K celebrities. This aggregates to over 1 billion unique user IDs stored in the inverted index. In contrast, if we store the followees of all top-1K users in the forward index, we need to store 110 billion user IDs because of the repetitions of common followees. In terms of memory, the forward index is of size 845.20GB, while the inverted index is 85.48GB. The Figure-3 represents how the information can be organized using those data structures.

The time to create an inverted index is insignificant compared to the gain in search time. The inverted index is created from the raw data collected from Twitter in 17 hours on an off-the-shelf computer.

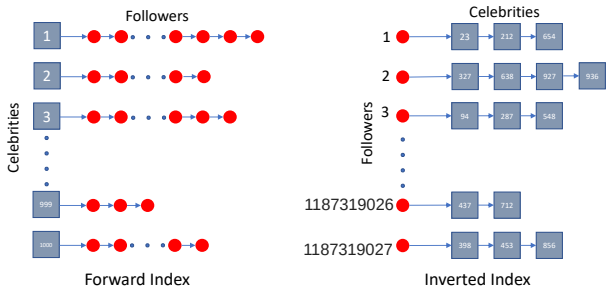

Figure 3: Forward and inverted organization of the followers of top-1K celebrities

4.2 Similarity Measure

In order to classify the category of a user, we use a simple nearest neighbor classifier under a modified Jaccard similarity measure, named *FollowerSimilarity*.

We define the set of followers of a top-1K celebrity as *Top User Followers (TF)*, and the set of followers of an influencer as *Influencer Followers (IF)*.

$$FollowerSimilarity = \frac{|IF \cap TF|}{|IF|} \quad (1)$$

FollowerSimilarity does not use the union of the two sets as the classic Jaccard similarity measure does in the denominator. We have two motivations behind this change. The union between the two sets is always dominated by the celebrity account's large follower set, resulting in very small numbers and reducing the comparability across celebrities. Second, FollowerSimilarity is simply the percentage of shared followers an influencer has with a celebrity. Understandably, higher FollowerSimilarity means the influencer is likely to share common interests with the celebrity.

In order to find the intersecting set of followers between a celebrity and an influencer, we exploit the inverted index. The

inverted index stores a list of celebrity followees for 1.18 billion users. We perform a logarithmic search of the followers in *IF* over the 1.18 billion users to find the intersecting followers. When found, we check if the follower has the celebrity as a followee. This operation costs very little time as the average number of celebrities Twitter users follow is ten. A single similarity computation would need a ten-fold less expensive search if we used the forward index. However, this gain by the forward index diminishes when we consider the classification task.

4.3 Classification

Classifying the category of an influencer collaboratively by the celebrities requires computing similarity scores between an influencer and as many celebrities as possible. Our dataset contains the top-1K celebrities on Twitter. We exploit the inverted index to calculate the similarity score to each of these 1000 celebrities.

We maintain a vector of 1000 counters, one for each celebrity, in the memory to calculate the sizes of the intersecting set of followers. The counters are set to zero and incremented by one whenever a follower of the influencer follows the corresponding celebrity. In contrast, if we used the forward index, we must iterate over 1000 celebrities and need to calculate the similarity scores independently.

Let us consider an example to understand the cost of calculating the sizes of a common set of followers. Consider the user `@kimkardeshian` (note the misspelled name) who has around 2,034 followers. To classify this user, we iterate over these 2,034 followers of `@kimkardeshian` and search for the follower in the index, taking time proportional to $\log(1.18 \times 10^9)$. When found, this user's list of followees (ten on average) is scanned sequentially, and the corresponding counters are updated. The total cost is approximately $2000 \times \log(1.18 \times 10^9) \times 10$. In contrast, if we used a forward index, the cost would be $1000 \times 2000 \times \log(11.68 \times 10^6)$, where 11.68M is the average number of followers for the top-1K users.

The classification of influencer users is produced using a simple K-Nearest Neighbor (KNN) approach. The number of nearest neighbors (K) varied between 1 and 10. In case of any ties, we break them in the order of prior class frequencies.

4.4 Visualization

To visually inspect the similarity space defined by FollowerSimilarity, we classify sets of influencers as described above. If we take a group of influencers, S; the classification process produces a $|S| \times 1000$ matrix of similarity scores to the celebrity accounts. We exploit the t-SNE [22] tool for the neighbor embedding method to visualize the set of influencers with respect to the top-1K users. t-SNE is used as it preserves the local structure of the points which are close to one another in the high-dimensional space. We used cosine similarity as the metric for the embedding space and the FollowerSimiliarity scores to top-1K as features for the embedding.

The outcome of the t-SNE plot represents the different clusters formed, as illustrated in Figure-9. The figure is full of information and even provides detailed explanations within the sub-clusters. In the later section, some of the outcomes of the t-SNE plots will be discussed.

5 EXPERIMENTAL EVALUATION

5.1 Distributions of Celebrities

The Figure-4 shows the overall distribution of the top-1K celebrities in the *Popularity* category. The most frequent category is *Entertainment* containing more than half of the top celebrities. The least frequent category is the *Religion* category. The other categories are very similar in frequencies.

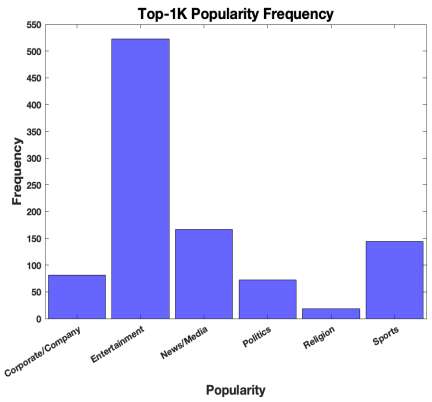

Figure 4: Frequency distribution of popular groups in top-1K

Other than that, the frequency of the labels *Language* and *Country* are calculated. About 20 different types of *Language* are identified, with the English language having the highest frequency of 610 from the top-1K. One interesting find is that several non-English speaking celebrities tweet in English to influence the global audience. The languages that dominate the lists are Spanish (117), Portuguese (65), and Arabic (59).

In the case of the label *Country*, there are 48 different country names, with the USA with the highest frequency of 389. Almost 40% of the celebrities are from the North American region. Other dominating countries on the list are India (116), the United Kingdom (81), and Brazil (61).

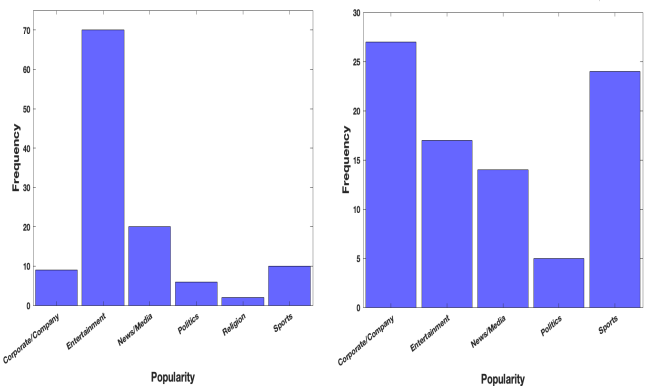

Figure 5: (a) Frequency distribution of popular groups in Macro-influencer (b) Frequency distribution of popular groups in Micro-influencer

We create two different sets of one hundred influencers in each for two follower ranges. We manually label each of the influencers in these sets with an agreement among the three authors of this article. The labeled datasets are available to download from the mentioned Google site.

The first set comprises macro-influencers with followers between 1M and 5M. The second set contains micro-influencers with followers between 1K and 5K. These sets are uniformly sampled from their respective group of users listed in [21]. We label the influencers very similarly to the training data in three categories: *Popularity*, *Country*, and *Language*.

The Figure-5 shows the distribution of *Popularity* category in these two sets. The macro-influencers show a very similar frequency distribution as the celebrities shown in Figure-6. *Entertainment* is the most frequent group and *Religion* is the least frequent group. One of the highest followed users from the *Entertainment* group is `@noah_id`, an Indonesian singer with 4.5M followers. One of the preachers from the *Religious* category is `@omar_almulhem`, having 2.4M followers.

A dramatic shift can be observed in the frequency distribution among micro-influencers. The *Corporate/Company* category has the highest frequency, while other groups have a comparable presence. The *Corporate/Company* type is composed of users from different regions, e.g., `@riscure` is a security system company situated in the Netherlands, having 4.8K followers. This change in frequency distribution matches the characterization in Table-1. A micro-influencer is more self-driven and promotes products and items to the regional or local audience. Hence, we observe a roughly uniform presence in all *Popularity* categories. A macro-influencer is more brand-driven and mostly endorses products by celebrities to the global audience. Therefore, their distribution is more similar to the top-1K users.

5.2 Classification Accuracy

We use the KNN (K-Nearest Neighbor) classifier under the FollowerSimilarity score. We show the results of the KNN classifier on the two influencer sets for various categories while varying the number of nearest neighbors in Figure-6 (left-middle).

To draw a comparison, we show the default accuracy with dashed lines. Default accuracy is defined as the proportion of the most frequent group in the set to mimic a blind classifier that only outputs the most frequent class. We observe a gradual decrease in classification accuracy on both sets as we include more neighbors in decision-making. However, the accuracy numbers are much higher than the default accuracy for each category.

The highest accuracy in the *Language* category suggests that followers are segmented mainly by language. In other words, an influencer's followers also follow celebrities who post in the same language as the influencer post. The lowest accuracy in the *Popularity* category suggests that followers of an entertaining influencer weakly (not as strongly as in *Language*) follow celebrity entertainers.

The classification accuracy drops in all categories for micro-influencers in comparison to that of macro-influencers. This suggests a trend of decreasing capability of FollowerSimilarity for users with low influence.

5.3 Comparison to Text-based Classification

We consider a text-based classifier a baseline to demonstrate the value gained in follower-based classification. We collect the most recent 200 tweets posted by each of the top-1K users. We exclude emojis, stop words, URLs, and mentions from these tweets to produce a bag of words for each celebrity. We convert each word in a bag to a vector using the BERT [23] embedding technique. The vectors in a bag are aggregated to produce one vector for a user. We have experimented with three column-wise aggregation functions: MIN., MAX., and AVG. Next, we performed the KNN classification under Euclidean similarity of the aggregated embedding vectors. The Euclidean similarity is used as it takes into account the magnitude of similarity between the two sentences [24]. The results are reported in Figure-7.

We observe that the text-based classifier is falling short in accuracy for the same classification parameter, K, in classifying *Popularity* and *Country* categories. Since tweets are used to create vectors, we decided not to include the Language category in that figure. In Figure-7a, we see that after the fifth neighboring nodes, the performance of text-based classification is a bit better than the follower-based for the *Popularity* category. This is because more common nodes are identified for that category. However, the classifier does poorly in classifying the Country category. The result is much worst in detecting micro-influencers using text-based information.

5.4 Classification Efficiency

The computational cost is the key challenge of a KNN classifier under the FollowerSimilarity measure. We perform an experiment evaluating the time to classify a macro-influencer using an inverted index by varying the training data size. The results are shown in Figure-6(right). The time to classify one hundred users in the macro-influencer set using the top-10, -50, -100, and -1K celebrities are reported. On average, with top-1K celebrities, the KNN algorithm requires less than half a minute to classify a macro-influencer. We argue that the time to classify a micro-influencer will be much less because they have much fewer followers than the macro-influencers have.

6 VISUALIZATION

We start by introducing our visualization of the top-1K celebrities and the micro-influencers based on FollowerSimilarity scores as features. The visualization is produced using the t-SNE tool that co-locates two users if their similarity scores to top-1K celebrities show similar profiles. It is challenging to depict one thousand labeled dots on a printed page. However, we have included a very high-resolution image in Figure-9 to describe how FollowerSimilarity can produce a never-seen-before visualization of the top-1K celebrities. We recommend zooming in on the figure on the digital screen for better readability. We represent some notable clusters of celebrities formed in the visualization and demonstrate the usefulness of the K-Nearest Neighbor classifier by tracing some of our micro-influencers.

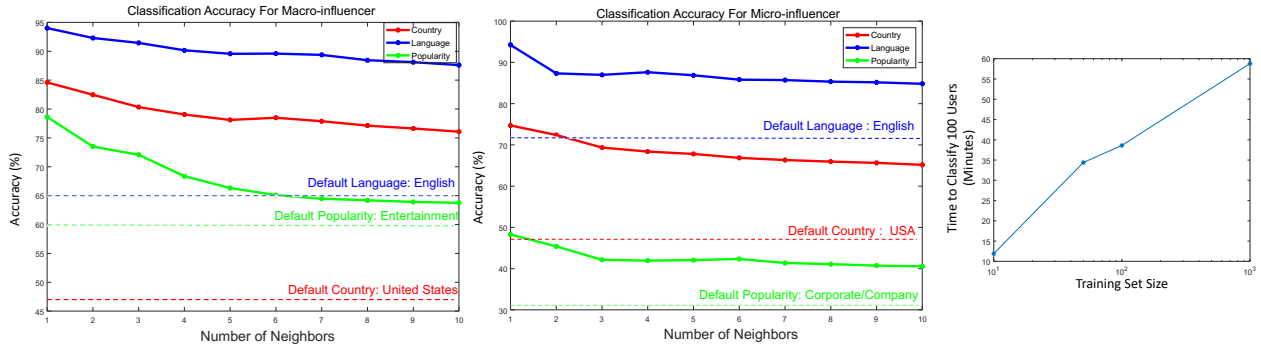

Figure 6: (left) Classification performance on macro-influencers (middle) Classification performance on micro-influencer. (right) Classification time for one hundred macro-influencers with varying training sets.

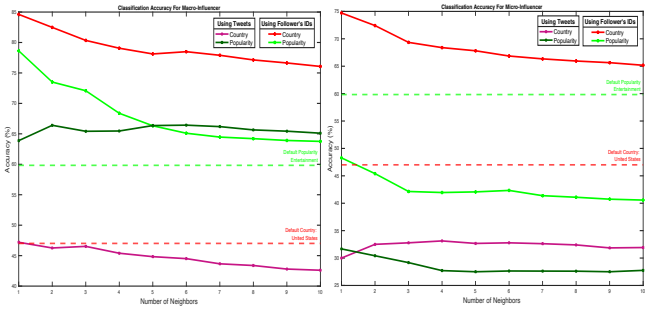

Figure 7: (a) Macro-influencer classification (b) Micro-influencer classification both using the tweets and follower's ID

6.1 Clusters of Celebrities

The Figure-9 consists of two types of dots. The solid dots represent the celebrities in the top-1K set. The unfilled dots represent the users in our micro-influencer set. The color of a dot represents the *Popularity* category as labeled by the authors prior to producing the visualization.

6.1.1 International Soccer Players, Clubs and Leagues. The soccer cluster is formed by the followers who are fans of this sport. The cluster is constituted with soccer players (e.g. `@cristiano`, `@kaka`, etc.), clubs (e.g. `@mancity`, `@fcbarcelona`, etc.), news media (e.g. `@skysports`, `@marca`, etc.) and esports (e.g. `@easportsfifa`). Note that the cluster is truly international with players, clubs, and media from different countries. Small sub-cluster containing players of the Brazil's national soccer team (e.g. `@davidluiz_4`, `@neymarir`, `@10ronaldinho`, etc.) emerge in the larger cluster. Although there are other Brazilian celebrities in the top-1K set, this sub-cluster is formed within the soccer cluster due to the global popularity of the Brazilian soccer team. Soccer leagues form sub-clusters containing players and teams in those leagues. For example, the Spanish soccer club `@fcbarcelona_es` is surrounded by the past and present players such as `@carles5puyol`, `@andresiniesta8`, and `@3gerardpique`. Similarly, `@premierleague` is surrounded by English Premier League clubs `@lfc`, `@mancity`, and `@chelseafc`.

6.1.2 Arabic Language Cluster. The Arabic cluster is composed of celebrities from countries including Lebanon, Saudi Arabia, UAE, and Qatar. There are few sub-clusters within this cluster representing the *Entertainment* (e.g. `@carole_samaha`, `@kadimalsahirorg`), *Religion* (e.g. `@dr_alqarnee`, `@mohamadalarefe`) and *News/Media* (e.g. `alarabiya`, `@france24_ar`) categories.

In addition, we observe Arabic versions of celebrity profiles in this cluster. For example, the Spanish soccer club Real Madrid has an official Twitter page named `@realmadridarab`. Another example is `@france24_ar`, a French news organization providing Arabic news feeds.

6.1.3 Indian Sub-continent. The cluster is mainly composed of Indian celebrities in *Entertainment*, *Sports*, *Politics*, and *News/Media*. Indian sub-continent is diverse in cultures and languages. This is evident in the clusters of celebrities. We observe Hindi language entertainers such as `@aamir_khan`, `@srbachchan`, and `@iamsrk` clustered together. We observe Tamil and Telegu language actors clustering together such as `@MsKajalAggarwal`, `@suriya_offl`, and `@ikamalhaasan`.

Cricket is the most popular sport in India. A cluster of cricket players is formed. Most of these players are past and present members of the Indian national team including `@sachin_rt`, `@msdhoni`, and `@hardikpandya7`. Similar to the soccer cluster, international cricketers playing in the Indian Premier League (IPL), such as `@abdevilliers17` from South Africa, have gained immense popularity among Indians; hence, they are located in this cluster.

Politicians in India are known for their effective use of Twitter. Most political parties have a presence on Twitter, such as `@aamaadmiparty` and `@bjp4india`, and they are clustered together by FollowerSimilarity. A curious exception is `@narendramodi`, the Prime Minister of India, who is situated between three sub-clusters: *Politics*, *Entertainment*, and *Sports*. His popularity in these three different groups has fueled his political success.

6.1.4 English Music Industry. The English music industry is composed of numerous genres and artists from different countries. We observe a sub-cluster for the Hip-hop/Rap genre, including artists `@nickiminaj`, `@eminem`, and `@wizkhalifa`. Rock and Metal music formed a sub-cluster including artists like `@gunsnroses` and `@linkinpark`. We also notice a genre called Techno, composed

of users like @skrillex and @hardwell. This genre is popular in clubs and parties. There are several Pop sub-clusters. One includes the US Pop artists with overwhelming popularity such as @justinbieber, @taylorswift13 and @katyperry. Another one is mostly composed of female artists including @britneyspears, @pink and @kellyclarkson.

Two sub-clusters of celebrities from the United Kingdom are observed. One affiliated with the show @thexfactor, and the other groups composed of music bands and singers including @edsheeran, @onedirection and @littlemix. Because of the large overlapping fan base across the US and the UK, these sub-clusters are close to other US singers.

In addition to the artists, a few related accounts are in this cluster, including @mtv, a television channel for music, and @instagram, a popular social network among these celebrities.

6.2 Clusters based on Tweets

The text-based classification does not achieve comparable accuracy to the follower-based classification. The t-SNE visualization of text-based similarity of the Twitter users in Figure-8 helps our reasoning. High-level clusters of popular groups (e.g., *Entertainment*, *Politics*, etc.) are visible. But they are not representing any regional information and are not very informative. However, the separation among the clusters is not as pronounced as in Figure-9. The micro-influencers (unfilled dots) are frequently around dots of different colors and making it difficult to determine the correct *Category*. An interesting finding is the influencer node, the UK Prime Minister (@10DowningStreet) (marked with an arrow), which is the Political node in the Figure-8 is wrongly classified as News/Media because the neighboring node @AJEnglish were both mourning the death of Queen Elizabeth II. Thus, the tweet-based classification shifts the location of the nodes based on the type of tweets a user posts.

6.3 Classified Micro-influencers

In an ideal world, FollowerSimilarity should place an influencer close to the celebrities in the same category. We observe that most unfilled dots are close to the dots of the same category of *Popularity*, *Country* and/or *Language*. In Figure-9, we mark several micro-influencers with blue underlines in each of the clusters described in this section below.

- Consider the micro-influencers @asroma, a fan page of AS Roma in Italian Serie A. The account has been placed inside the Soccer cluster. Now, consider the micro-influencer @eurolatamsummit), a conference related to sports marketing. Both accounts are correctly placed close to the celebrities in European soccer leagues.
- The micro-influencer @dezsms within the Arabic language cluster. The account represents a calling card company in Kuwait, hence, labeled as a Corporate account. The account is relatively far from the other accounts in that cluster; it can be thought of as a local outlier.
- Consider the micro-influencer @OdishaDHFM in the Indian cluster with 4K followers. The account represents a fan club of the Telegu actor Mahesh Babu. The account is close to other Telegu actors and entertainers in the cluster.
- The account @emilygryson88 with 1K followers is a fan page of the Cuban-American singer @Camila_Cabello having 12.9M followers. The account is almost overlapping with Camila's account within the US Music cluster.

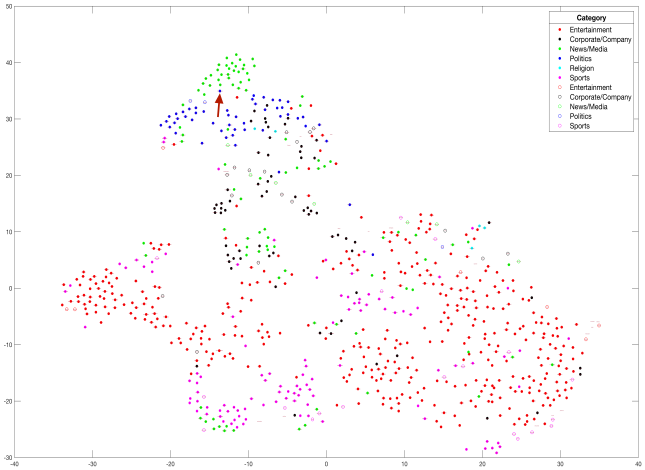

Figure 8: The t-SNE plot for text-based similarity among the top-1K celebrities and the randomly selected micro-influencers. Zoom-in for better viewing.

7 ETHICAL CONSIDERATION

We collect and store user objects of the top-1K celebrities on Twitter. The objects are collected using Twitter API, hence, considered public information. We have collected the followers using two Twitter API keys created by the first and second authors. We have spent a long time (140 days) collecting over one billion user IDs at the rate allowed by Twitter. The user IDs do not associate with any information specific to a profile. The IDs are simply 64-bit numbers organized in lists. In order to classify an arbitrary influencer, one would need a few API calls to collect the IDs of the followers of the influencer and the inverted index described in the paper.

8 CONCLUSION

We present a method to classify an influencer using common followers with celebrity accounts. We devise an inverted index to calculate a novel user similarity score. A simple KNN classifier using our similarity score can achieve a significantly higher classification accuracy than default classifiers. We demonstrate that our similarity scores capture meaningful groups of celebrities of various languages, countries, and cultures. In the future, we would like to extend the classifiers to a more refined set of categories.

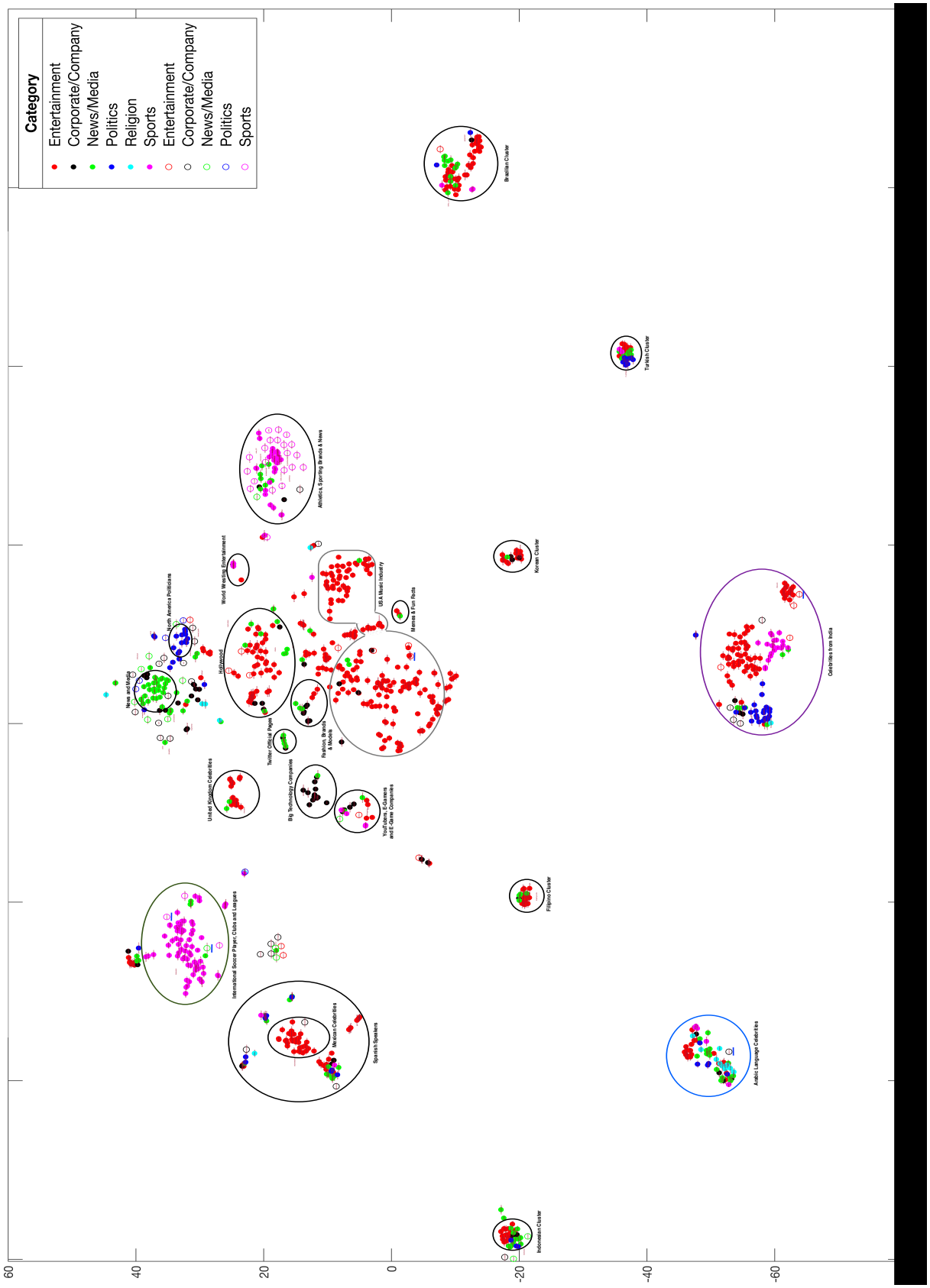

Figure 9: The t-SNE plot is representing the FollowerSimilarity among the top-1K celebrities and the randomly selected micro-influencers. Cluster: International Soccer Players, Clubs & Leagues, Cluster: Arabic Speakers, Cluster: Indian Sub-Continent, and Cluster: United States Music Industry

REFERENCES

[1] Tatum Hunter. What is a micro-influencer, and why do brands use them? https://builtin.com/marketing/micro-influencer, 2022.

[2] Jacinda Santora. Key influencer marketing statistics you need to know for 2022https://influencermarketinghub.com/influencer-marketing-statistics/, 2022.

[3] Samantha Kay, Rory Mulcahy, and Joy Parkinson. When less is more: the impact of macro and micro social media influencers' disclosure. *Journal of Marketing Management*, 36(3-4):248–278, 2020.

[4] Rachidatou Alassani and Julia Göretz. Product placements by micro and macro influencers on instagram. In *International conference on human-computer interaction*, pages 251–267. Springer, 2019.

[5] Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884, 2020.

[6] Young Anna Argyris, Zuhui Wang, Yongsuk Kim, and Zhaozheng Yin. The effects of visual congruence on increasing consumers' brand engagement: An empirical investigation of influencer marketing on instagram using deep-learning algorithms for automatic image classification. *Computers in Human Behavior*, 112:106443, 2020.

[7] Taylor Sweet, Austin Rothwell, and Xuan Luo. Machine learning techniques for brand-influencer matchmaking on the instagram social network. *arXiv preprint arXiv:1901.05949*, 2019.

[8] Benyamin Bashari and Ehsan Fazl-Ersi. Influential post identification on instagram through caption and hashtag analysis. *Measurement and Control*, 53(3-4):409–415, 2020.

[9] Victoria Nebot, Francisco Rangel, Rafael Berlanga, and Paolo Rosso. Identifying and classifying influencers in twitter only with textual information. In *International Conference on Applications of Natural Language to Information Systems*, pages 28–39. Springer, 2018.

[10] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[11] Christine Kiss and Martin Bichler. Identification of influencers—measuring influence in customer networks. *Decision Support Systems*, 46(1):233–253, 2008.

[12] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.

[13] Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. Mining micro-influencers from social media posts. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 867–874, 2020.

[14] Soufia Kausar, Bilal Tahir, and Muhammad Amir Mehmood. Understanding the role of political micro-influencers in pakistan. In *2021 International Conference on Frontiers of Information Technology (FIT)*, pages 31–36. IEEE, 2021.

[15] Punam Bedi and Chhavi Sharma. Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):115–135, 2016.

[16] Kun He, Yingru Li, Sucheta Soundarajan, and John E Hopcroft. Hidden community detection in social networks. *Information Sciences*, 425:92–106, 2018.

[17] Guo-Jun Qi, Charu C Aggarwal, and Thomas Huang. Community detection with edge content in social media networks. In *2012 IEEE 28th International conference on data engineering*, pages 534–545. IEEE, 2012.

[18] Daniel López Sánchez, Jorge Revuelta, Fernando De la Prieta, Ana B Gil-González, and Cach Dang. Twitter user clustering based on their preferences and the louvain algorithm. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 349–356. Springer, 2016.

[19] Chiara Francalanci and Ajaz Hussain. Discovering social influencers with network visualization: evidence from the tourism domain. *Information Technology & Tourism*, 16(1):103–125, 2016.

[20] Yu Xie, Maoguo Gong, Shanfeng Wang, Wenfeng Liu, and Bin Yu. Sim2vec: Node similarity preserving network embedding. *Information Sciences*, 495:37–51, 2019.

[21] Trackalytics. The most followed twitter profiles https://www.trackalytics.com/the-most-followed-twitter-profiles/page/1/, 2021.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[24] Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*, 2014.