This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Seeing and Believing: Evaluating the Trustworthiness of Twitter Users

TANVEER KHAN, ANTONIS MICHALAS
Tampere University, Tampere, Finland
(e-mail: {tanveer.khan, antonios.michalas}@tuni.fi)

ABSTRACT Social networking and micro-blogging services, such as Twitter, play an important role in sharing digital information. Despite the popularity and usefulness of social media, there have been many instances where corrupted users found ways to abuse it, as for instance, through raising or lowering user's credibility. As a result, while social media facilitates an unprecedented ease of access to information, it also introduces a new challenge - that of ascertaining the credibility of shared information. Currently, there is no automated way of determining which news or users are credible and which are not. Hence, establishing a system that can measure the social media user's credibility has become an issue of great importance. Assigning a credibility score to a user has piqued the interest of not only the research community but also most of the big players on both sides - such as Facebook, on the side of industry, and political parties on the societal one. In this work, we created a model which, we hope, will ultimately facilitate and support the increase of trust in the social network communities. Our model collected data and analysed the behaviour of 50,000 politicians on Twitter. Influence score, based on several chosen features, was assigned to each evaluated user. Further, we classified the political Twitter users as either trusted or untrusted using random forest, multilayer perceptron, and support vector machine. An active learning model was used to classify any unlabelled ambiguous records from our dataset. Finally, to measure the performance of the proposed model, we used precision, recall, F1 score, and accuracy as the main evaluation metrics.

INDEX TERMS Active Learning, Influence Score, Credibility, Trust, Sentiment Analysis, Fake News, Twitter, Machine Learning

I. INTRODUCTION

An ever increasing usage and popularity of social media platforms has become the sign of our times – close to a half of the world's population is connected through social media platforms. The dynamics of communication in all spheres of life has changed. Social media provide a platform through which users can freely share information simultaneously with a significantly larger audience than traditional media.

As social media became ubiquitous in our daily lives, both its positive and negative impacts have become more pronounced. Successive studies have shown that extensive distribution of misinformation can play a significant role in the success or failure of an important event or a cause [1], [2]. Barring the dissemination and circulation of misleading information, social networks also provide the mechanisms for corrupted users to perform an extensive range of illegitimate actions such as spam and political astroturfing [3], [4]. As a result, measuring the credibility of both the user and the text itself has become a major issue. In this work, we assign a credibility score to each Twitter user based on certain extracted features.

Twitter is currently one of the most popular social media platforms with an average of 10,000 tweets per second [5]. Twitter-enabled analytics do not only constitute a valuable source of information but provide an uncomplicated extraction and dissemination of subject specific information for government agencies, businesses, political parties, financial institutions, fundraisers and many others.

In a recent study [6], 10 million tweets from 700,000 Twitters accounts were examined. The collected accounts were linked to 600 fakes news and conspiracy sites. Surprisingly, authors found that clusters of Twitter accounts are repeatedly linked back to these sites in a coordinated and automated manner. A similar study [7] showed that 6.6 million fake news tweets were distributed prior to the 2016 US elections.

Globally, a number of social and political events in the last three years have been marred by an ever-growing presence of misleading information provoking an increasing concern about their impact on society. This concern translated into an immediate need for the design, implementation, and adoption

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

of new systems and algorithms that will have the ability to *measure* the credibility of a source or a piece of news. Notwithstanding, the seemingly unencumbered growth of social media users is continuing[1]. Coupled with the growth in user numbers, the generated content is growing exponentially thus producing a body of information where it is becoming increasingly difficult to identify fabricated stories [9]. Thereupon, we are facing a situation where a compelling number of unverified pieces of information could be misconstrued and ultimately misused. The research in the field is therefore currently focusing on defining the credibility of the tweets and/or assigning scores to users based on the information they have been sharing [10]–[17].

A. OUR CONTRIBUTION AND DIFFERENCES WITH PREVIOUS WORKS

We would like to draw your attention to the areas in which this work builds on our previous one [18] and where, we believe, it expounds it and offers new insights. In this work we used *additional ML models*, such as Multi-Layer Perceptron (MLP) and Logistic Regression (LR). Since the MLP model outperformed the LR, we only present the findings for the MLP model. For MLP, we performed the experiments for Tanh, ReLU and Logistics. Moreover, unlike [18], where just one evaluation metric, "Accuracy", was used to evaluate the model's performance, in this work, here, we measure the model's performance by using four evaluation metrics – "Precision", "Recall", "F1" score, and "Accuracy" (see table 5). Furthermore, we provide the descriptive statistics of the features (see table 4) as well as their correlation with the target (see figure 3) and compare our work with other similar works as SybilTrap [19] (see table 2). Finally, we conduct a comparative review of the user characteristics primarily used in the literature so far, and the ones used in our model and provide supplementary information to help with stratifying trusted and untrusted users (see table 3).

Our main contribution can be summarized as follows:

- First, we gathered a 50,000 Twitter users dataset where for each user, we built a unique profile with 19 features (discussed in Section III). Our dataset included only users whose tweets are public and have non-zero friends and followers. Furthermore, each Twitter user account was classified as either trusted or untrusted by attaching the trusted and untrusted flag based on different features. These features are discussed in detail in Section IV.
- We measured the social reputation score (Section III-C), a sentiment score (Section III-C), an h-index score (Section III-C), tweets credibility (Section III-C) and the influence score (Section III-D) for each of the analyzed Twitter users.
- To classify a large pool of unlabelled data, we used an active learning model – technique best suited to the situation where the unlabelled data is abundant but manual labelling is expensive [20], [21]. In addition, we evaluated the performance of various ML classifiers.

We hope that this work will inspire others to further research this problem and simultaneously kick-start a period of greater trust in social media.

B. ORGANIZATION

The rest of paper is organized as follows: Related work is discussed in Section II, accompanied by a detailed discussion of our proposed approach in Section III. In Section IV, the active learning method and the type of classifier used are discussed. The data collection and experimental results are presented in Section V. Finally, in Section VI, we conclude the paper.

II. RELATED WORK

Twitter is one of the most popular Online-Social-Networks (OSNs). As data aggregator, it provides data that can be used in research of both historical and current events. Twitter, in relation to other popular OSNs, attracts significant attention in the research community due to its open policy on data sharing and distinctive features [22]. Although openness and vulnerability don't necessarily go hand in hand, on a multiple occasions malicious users misused Twitter's openness and exploited the service (e.g. political astroturfing, spammers sending unsolicited messages, post malicious links, etc.).

In contrast to mounting evidence towards the negative impact of fake news dissemination, so far, only a few techniques for identifying them in social media have been proposed [3], [4], [22]–[24].

Among the most popular and promising ones is evaluating Twitter users and assigning them a reputation score. Authors in [3] explored the posting of duplicate tweets and pointed that this behaviour, usually not followed by a legitimate user, affects the reputation score. Posting the same tweet several times has a negative effect on the user's overall reputation score. The authors presented research that supports the above by calculating the edit distance to detect duplications between two tweets posted from the same account.

Furthermore, users have used an immense amount of exchanged messages and information on Twitter to hijack trending topics [25] and send unsolicited messages to legitimate users. Additionally, there are Twitter accounts whose only purpose is to artificially boost the popularity of a specific hashtag thus increasing its popularity and eventually making the underlying topic a trend. The BBC investigated an instance where £150 was paid to Twitter users to increase the popularity of a hashtag and promote it into a trend[2].

In an attempt to address these problems, researchers have used several ways to detect the trustworthiness of tweets and assign an overall rank to users [24]. Castillo *et al.,* [26] measured the credibility of tweets based on Twitter features by using an automated classification technique. Alex Hai

[1]In 2020, an estimated 3.23 billion people were using social media worldwide, a number projected to increase to almost 3.64 billion in 2024 [8].

[2]https://www.bbc.com/news/blogs-trending-43218939

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

Wang [3] used the followers and friends features to calculate the reputation score. Additionally, Saito and Masuda [27] considered the same metrics while assigning a rank to Twitter users. In [28], authors analysed the tweets relevant to Mumbai attacks[3]. Their analysis showed most of the information providers were unknown while the reputation of the remaining ones was very low. In another study [29] that examined the same event, the information retrieval technique and ML algorithm used found that mere 17% of the tweets were credibly related to the underlying attacks.

According to Gilani *et al.,* [30], when compared to normal users, bots and fake accounts use a large number of external links in their tweets. Hence, analysing other Twitter features such as URL is crucial for correctly evaluating the overall credibility of a user. Although, Twitter has included tools to filter out such URLs, several masking techniques can effectively bypass Twitter's existing safeguards.

In this work, we evaluate the users' trustworthiness and credibility [31], [32] by analysing a wide range of features (see Table 1). In comparison to similar works in the field, our model explores a number of factors that could be signs of possible malicious behaviours and makes honest, fair, and precise judgements about the users' credibility.

III. METHODOLOGY

In this section, we discuss the model and main algorithms we used to calculate the user's influence score. Our first goal is to enable the users to identify certain attributes and assess a political Twitter user by considering the influence score that is the outcome of a proper run of our algorithms. Figure 1 illustrates the main features we used to calculate users' influence score. We also compare our work with state-of-the-art work in this domain (see Table 2). Secondly, the political Twitter users are classified into either trusted or untrusted based on features as social reputation, the credibility of tweets, sentiment score, the h-index score, influential score etc. Accounts containing abusive and/or harassment tweets, low social reputation and h-index score, and low influential score are grouped into untrusted users. The trusted users category envelops more reputable among the users with high h-index score, more credible tweets as well as those having high influential score. We will discuss this in more detail in Section IV.

In addition, we also present the approach used to calculate the Twitter users' influence score based on both their context and content features. For the user evaluation we took into consideration only the Twitter features that can be extracted through Twitter API. We used the outcome of that evaluation and derived more features to help us provide a better rounded and fair evaluation (Section III-C). The features, as well as the relevant notation used throughout the paper, are given in Table 1.

[3]https://www.theguardian.com/world/blog/2011/jul/13/mumbai-blasts

Table 1: Features Considered to Calculate the Influence Score

Notation	Description
$N_{fri}(u_i)$:	Number of friends of the user
$N_{fol}(u_i)$:	Number of followers of the user
N_{ret}:	Number of retweets for a tweet
$R_{ret}(u_i)$:	Retweet ratio of the user
N_{lik}:	Number of likes for a tweet
$R_{lik}(u_i)$:	Liked ratio of the user
$U_R(u_i)$:	Tweet of the user containing URLs
$R_{url}(u_i)$:	URLs ratio of the user
$L(u_i)$:	List count of the user
$N_T(u_i)$:	Total number of tweets or Status of the user
$R_{ori}(u_i)$:	Original content ratio of the user
$R_s(u_i)$:	Social reputation score of the user
$h_{ind}(u_i)$:	h-index of the user
$R_{hind}(u_i)$:	Retweet h-index of the user
$L_{hind}(u_i)$:	Liked h-index of the user
$Twt_{cr}(u_i)$:	Tweets credibility of the user
$Sen_s(u_i)$:	Sentiment score of the user
$N_{neu}(u_i)$:	Neutral tweets
$N_{pos}(u_i)$:	Positive tweets
$N_{neg}(u_i)$:	Negative tweets
$R_{has}(u_i)$:	Hashtag ratio of the user
$Inf(u_i)$:	Influence score
I_t:	Tweet Index

A. FEATURES SELECTION AND COMPARISON WITH PREVIOUS MODELS

The features used for calculating the influence score were based on extensive study of the existing literature. The selected features were used for detection purposes [33]–[35], assigning a score [24] or classification purposes [36]. We used the features given in Table 1 to assign an influence score to a u_i. Table 2 provides a comparative overview of existing models based on feature selection.

B. TWITTER FEATURES EXTRACTION

The pivotal step in the process of assigning a score to a Twitter user is to extract the features linked to their accounts. The features can be either user account specific, such as the number of followers, friends, etc., or user tweet specific, such as the number of likes, retweets, URLs, etc. In our model, we considered both and used them to calculate some additional features. We then combined them all to assign an influence score to a Twitter user. Below we provide more detailed information on features used in our model.

Number of Friends

Friend is a user account feature indicating that a Twitter user (u_i) has subscribed to the updates of another u_i [37].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

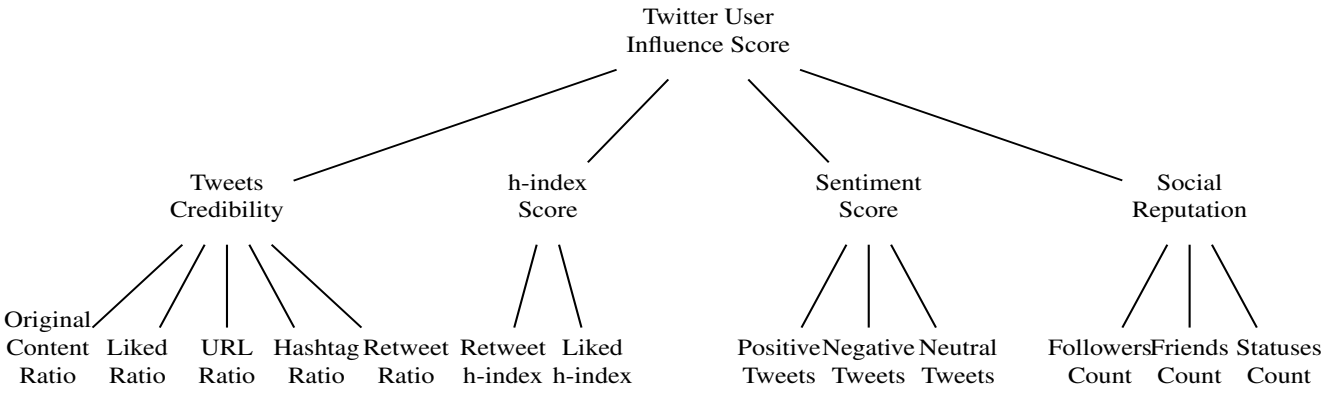


Figure 1: Twitter Users Influence Score Calculation

Table 2: Models Comparison using Features

Papers	$R_s(u_i)$			h_index		$Sens_s(u_i)$	$Twt_{cr}(u_i)$						URLs, List and Mentions			
	$N_{fol}(u_i)$	$N_{fri}(u_i)$	$N_T(u$	$R_{hind}(u_i)$	$L_{hind}(u_i)$		$R_{ret}(u_i)$	$R_{lik}(u_i)$	$R_{has}(u_i)$	$R_{url}(u_i)$	$R_{ori}(u_i)$	$N_T(u_i)$	$R_{men}(u_i)$	$N_M(u_i)$	$U_R(u_i)$	$L(u_i)$
[5]	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	✓	✗
[19]	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗
[24]	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
[33]	✓	✓	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓	✗
[34]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓	✗
[35]	✓	✓	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓	✗
[36]	✓	✓	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗
Proposed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Following users who are not part of interpersonal ties yields a lot of novel information. One of the important indicators for calculating the $Inf(u_i)$ is the $follower/following$ ratio. The $follower/following$ ratio compares the number of u_i's subscribers to the number of the users, u_i is following. Users are more interested in updates if the $follower/following$ ratio is high [38]. The ideal $follower/following$ ratio is 1 or close to 1. In our model, we use the Number of Friends $N_{fri}(u_i)$ as one of the indicators for assigning User's Social Reputation $R_s(u_i)$.

Number of Followers

$N_{fol}(u_i)$ is another user account feature showing the number of people interested in the specific u_i's tweets. As discussed in [39], $N_{fol}(u_i)$ is one of the most important parameters for measuring u_i's influence. The more followers a u_i has the more influence he exerts [40]. Preussler *et al.,* [41] correlates the $N_{fol}(u_i)$ with the reputation of a u_i. According to their study, the credibility of a u_i increases as the $N_{fol}(u_i)$ increases. Based on the above we consider the $N_{fol}(u_i)$ an important parameter and use it as input to calculate the $R_s(u_i)$.

Number of Retweets

A tweet is considered important when it receives many positive reactions from other accounts. The reactions may take the form of likes or retweets. Retweets act as a form of endorsement, allowing u_i to forward the content generated by other users, thus raising the content's visibility. It is a way of promoting a topic and is associated with the reputation of the u_i [42]. Since retweeting is linked to popular topics and directly affects the u_i's reputation, it is a key parameter for identifying possible fake account holders. As described in [30], bots or fake accounts depend more on retweets of existing content than posting new ones. In our model, we consider the N_{ret} as one of the main parameters for assigning the $Inf(u_i)$. We calculate the $R_{ret}(u_i)$ (used by Twitter grader) for each tweet by considering N_{ret} divided by $N_T(u_i)$, as given in equation 1.

$$R_{ret}(u_i) = \frac{N_{ret}}{N_T(u_i)} \tag{1}$$

Number of Likes

The N_{lik} is considered a reasonable proxy for evaluating the quality of a tweet. Authors in [36] showed that humans receive more likes per tweet when compared to bots. In [43], the authors used likes as one of the metrics to classify Twitter accounts as a human user or automated agent. As mentioned in [5], if a specific tweet receives a large N_{lik}, it can be safely concluded that other u_i's are interested in the tweets of the underlying u_i. Based on this observation, we calculate the $R_{lik}(u_i)$ by using the N_{lik} for each tweet and dividing it with $N_T(u_i)$ as shown in equation 2.

$$R_{lik}(u_i) = \frac{N_{lik}}{N_T(u_i)} \tag{2}$$

URLs

URL is a content level feature some u_i's include in their tweets [44]. As tweets are limited to a maximum of 280 characters, it is common that u_i's cannot include all relevant information in their tweets. To overcome this issue, u_i's often populate tweets with URLs pointing to a source where more information can be found. In our model, we consider the URL as an independent variable for the engagement measurements [45]. We count the tweets that include a URL

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

and calculate the $R_{url}(u_i)$ by considering the $U_R(u_i)$ over the $N_T(u_i)$ as given in equation 3.

$$R_{url}(u_i) = \frac{U_R(u_i)}{N_T(u_i)} \tag{3}$$

Listed Count

In Twitter, a u_i has the option to form several groups by creating lists of different u_i's (e.g. competitors, followers etc.). Twitter lists are mostly used to keep track of the most influential people[4]. The simplest way to measure the u_i's influence is by checking the $L(u_i)$ that the u_i is placed on. Being present in a large number of lists is an indicator that the u_i is considered as important by others. Based on this assumption, we also considered the number of lists that each u_i belongs to.

Statuses Count

Compared to the other popular OSNs, Twitter is considered as a service that is *less* social[5]. This is mainly due to the large number of inactive u_i's or users who show low motivation in participating in an online discussion. Twitter announced a new feature "Status availability", that checks the $N_T(u_i)$[6]. The status count is an important feature closely related to reporting credibility. If a user is active on Twitter for a longer period, the likelihood of producing more tweets increases, which in turn may affect the author's credibility [46], [47]. To this end, for the calculation of the $Inf(u_i)$, we also took into account how active users are by measuring how often a u_i performs a new activity[7].

Original Content Ratio

It has been observed that instead of posting original content, most u_i retweet posts by others [38]. As a result, Twitter is changing into a pool of constantly updating information streams. For u_i's with high influence in the network, the best strategy is to use the $30/30/30$ rule: 30% retweets, 30% original content, and 30% engagement [48]. Having this in mind, in our model, we look for u_i's original tweets and add them to their corresponding influence score. We calculate the $R_{ori}(u_i)$ by extracting the retweeted posts by others from the total tweets of u_i as given in equation 4.

$$R_{ori}(u_i) = \frac{N_T(u_i) - Retweeted\ other\ tweets}{N_T(u_i)} \tag{4}$$

C. DERIVED FEATURES FOR TWITTER USERS

Following the considerations for the selection of the basic features for calculating the $Inf(u_i)$, in this section we elaborate on the extraction of the extra ones. Additionally, we discuss the sentiment analysis technique used to analyse u_i's tweets.

[4]https://www.postplanner.com/how-to-use-twitter-lists-to-always-be-engaging/

[5]https://econsultancy.com/twitter-isn-t-very-social-study/

[6]https://www.pocket-lint.com/apps/news/twitter/146714-this-is-what-twitter-s-new-online-indicators-and-status-updates-look-like

[7]https://sysomos.com/inside-twitter/most-active-twitter-user-data/

By using the basic features described earlier, we calculated the following features for each u_i:

- Social reputation of a user;
- Retweet h-index score and liked h-index score;
- Sentiment score of a user;
- Credibility of Tweets;
- Influence score of a user.

User's Social Reputation

The main factor for calculating the $R_s(u_i)$ is the number of users interested in u_i's updates. Hence, $R_s(u_i)$ is based on the $N_{fol}(u_i)$, $N_{fri}(u_i)$ and $N_T(u_i)$ [3], [38].

$$R_s(u_i) = 2\log(1 + N_{fol}(u_i)) + \log(1 + N_T(u_i)) - \log(1 + N_{fri}(u_i)) \tag{5}$$

In equation 5 we utilized the log property to make the distribution smoother and minimize the impact of outliers. In addition to that, since $log0$ is undefined, we added 1 wherever log appears in equation 5. In equation 5, $R_s(u_i)$ is directly proportional to $N_{fol}(u_i)$ and $N_T(u_i)$. Based on several studies [3], [5], [38], $R_s(u_i)$ is more dependent on $N_{fol}(u_i)$ hence we give more importance to $N_{fol}(u_i)$ in comparison to $N_T(u_i)$ and $N_{fri}(u_i)$. If a u_i has a large $N_{fol}(u_i)$ then the u_i is more reputable. In addition, if a u_i is more active in updating his/her $N_T(u_i)$ there are more chances that u_i's tweets receive more likes and get retweeted. While $N_{fol}(u_i)$ and $N_T(u_i)$ increase, $R_s(u_i)$ also increases and vice versa. Alternatively, if a u_i has less $N_{fol}(u_i)$ in comparison to the $N_{fri}(u_i)$ then, the $R_s(u_i)$ is smaller. As can be seen from equation 5, there is an inverse relation between $R_s(u_i)$ and $N_{fri}(u_i)$.

h-Index Score

The h_{ind} score is most commonly used to measure the productivity and impact of a scholar or scientist in the research community. It is based on the number of publications as well as the number of citations for each publication [49]. In our work, we use the h_{ind} score for a more accurate calculation of $Inf(u_i)$. The h_{ind} of a u_i is calculated considering N_{lik} and N_{ret} for each tweet. To find the h_{ind}[8], we sort the tweets based on the N_{lik} and N_{ret} (in decreasing order).

Algorithm 1 describes the main steps for calculating the h_{ind} of a u_i based on the N_{ret}. The same algorithm is used for calculating the h_{ind} of a u_i based on N_{lik} by replacing N_{ret} with N_{lik}. $R_{hind}(u_i)$ and $L_{hind}(u_i)$ are novel features used for measuring the relative importance of a u_i. A tweet that has been retweeted many times and liked by many users is considered as attractive for the readers [5], [50]. For this reason, we use $R_{hind}(u_i)$ and $L_{hind}(u_i)$ for measuring the $Inf(u_i)$. The higher the $R_{hind}(u_i)$ and $L_{hind}(u_i)$ score of a u_i, the higher will be the $Inf(u_i)$.

[8]https://gallery.azure.ai/Notebook/Computing-Influence-Score-for-Twitter-Users-1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

Algorithm 1 Calculating h-index score based on retweets

1: **procedure** H-INDEX SCORE(h_{ind})
2: Arrange N_{ret} for each tweet of a u_i in decreasing order
3: **for** I_t `in list:` **do**
4: **if** N_{ret} of a tweet $< I_t$ **then**
5: return I_t
6: **end if**
7: **end for**
8: return N_{ret}
9: **end procedure**

Twitter User Credibility

The credibility is actually the believability [26] – that is, providing reasonable grounds for being believed. The credibility of a u_i can be assessed by using the information available on the Twitter platform. In our approach, we use both the $Sen_s(u_i)$ and $Twt_{cr}(u_i)$ to find a credible u_i.

Sentiment Score: It has been observed that OSNs are a breeding ground for the distribution of fake news. In many cases even a single Twitter post significantly impacted [51] and affected the outcome of an event.

Having this in mind, we used sentiment analysis and the TextBlob [52] library, to analyze tweets with the main aim to identify certain patterns that could facilitate identification of credible news. The sentiment analysis returns a score using polarity values ranging from 1 to -1 and helps in tweet classification. We classified the collected tweets as *(1)* Positive *(2)* Neutral, and *(3)* Negative based on the number of positive, neutral and negative words in a tweet. According to Morozov *et al.,* [53], the least credible tweets have more negative sentiment words and opinions and are associated with negative social events, while credible tweets, have more positive ones. Hence we classified positive tweets as being the most credible followed by the neutral, and finally the least credible negative tweets.

Following the tweets classification we assign a $Sen_s(u_i)$ to each u_i [5] using the following equation:

$$Sen_s(u_i) = \frac{\sum N_{neu}(u_i) + \sum N_{pos}(u_i)}{\sum N_{neu}(u_i) + \sum N_{pos}(u_i) + \sum N_{neg}(u_i)} \tag{6}$$

Tweets Credibility: Donovan [54] focused on finding the most suitable indicators for credibility. According to their findings, prime indicators for a tweet's credibility are mentions, URLs, tweet length and retweets. Gupta *et al.,* [29] ranked tweets based on tweets credibility. The parameters used as an input for the ranking algorithm were: tweets, retweets, total unique users, trending topics, tweets with URLs, start and end date. Based on the existing literature, we compute the $Twt_{cr}(u_i)$ by considering $R_{ret}(u_i)$, $R_{lik}(u_i)$, $R_{has}(u_i)$, $R_{url}(u_i)$ and $R_{ori}(u_i)$ (see equation 7):

$$Twt_{cr}(u_i) = \left(\frac{R_{ret}(u_i)+R_{lik}(u_i)+R_{has}(u_i)+R_{url}(u_i)}{4}\right) \cdot R_{ori}(u_i) \tag{7}$$

To begin, we consider the $R_{ori}(u_i)$ (tweet) by a u_i and for each $R_{ori}(u_i)$ we collect $R_{ret}(u_i)$, $R_{lik}(u_i)$, $R_{has}(u_i)$ and $R_{url}(u_i)$. These four features are linked with the $R_{ori}(u_i)$ such as $R_{ret}(u_i)$ and $R_{lik}(u_i)$ specify the number of times the $R_{ori}(u_i)$ has been retweeted and liked while $R_{has}(u_i)$ and $R_{url}(u_i)$ return only $R_{ori}(u_i)$ having URLs and hashtags. Hence, to calculate the credibility of tweets, we first calculate the average of these four parameters and then multiply it with $R_{ori}(u_i)$.

D. INFLUENCE SCORE

The $Inf(u_i)$ is calculated based on the evaluation of *both* content and context features. More precisely, we consider the following features described earlier: $R_s(u_i)$, $Sen_s(u_i)$, $Twt_{cr}(u_i)$ and $h_{ind}(u_i)$. After calculating the values of all of these features we use them as input to Algorithm 2 line 7 which calculates the $Inf(u_i)$.

Equation Formulation: In order to ascertain how influential a u_i is, researchers have taken into consideration one, two or more of the following characteristics:

- Social reputation [55] and weight-age of his tweets [5];;
- Tweets credibility [5], [54];
- His ability to formulate new ideas, as well as his active participation in follow-up events and discussions [56].

An influential u_i must be highly active (have ideas that impact others' behaviours, able to start new discussions etc.,). Additionally, the tweets must be relevant, credible and highly influential (retweeted and liked by a large number of other u_i's). If the tweets of highly influential u_i's are credible and the polarity of their tweets' content is positive, they are considered as highly acknowledged and recognized by the community. In short, for a u_i to be considered influential, we combine the efforts of [5], [54]–[56] and calculate the $Inf(u_i)$ using equation 8.

$$Inf(u_i) = \frac{Sen_s(u_i)+Twt_{cr}(u_i)+R_s(u_i)+R_{hind}(u_i)+L_{hind}(u_i)}{5} \tag{8}$$

Algorithm 2 Influence score Calculation

1: **procedure** INFLUENCE SCORE($Inf(u_i)$)
2: For i^{th} User
3: Calculate $R_{hind}(u_i)$ and $L_{hind}(u_i)$, using Algorithm 1
4: Calculate $R_s(u_i)$ using equation 5
5: Calculate $Sen_s(u_i)$ using equation 6
6: Calculate $Twt_{cr}(u_i)$ using equation 7
7: Compute $Inf(u_i)$ using equation 8
8: **end procedure**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

IV. ACTIVE LEARNING AND ML MODELS

In line with the existing literature, the classification of a u_i is performed on a manually annotated dataset. The manually annotated dataset gives a ground truth, however, manual labelling is an expensive and time-consuming task. In our approach, we used active learning, a semi-supervised ML model that helps in classification when the amount of available labelled data is small. In this model, the classifier is trained using a small amount of training data (labelled instances). Next, the points ambiguous to the classifier in the large pool of unlabelled instances are labelled, and added to the training set [21]. This process is repeated until all the ambiguous instances are queried or the model performance does not improve above a certain threshold. The basic flow of active learning approach[9] is shown in Figure 2. Based on the proposed model, we first trained our classifier on a small dataset of human-annotated data. Following this step, it then further classified a large pool of unlabelled instances efficiently and accurately.

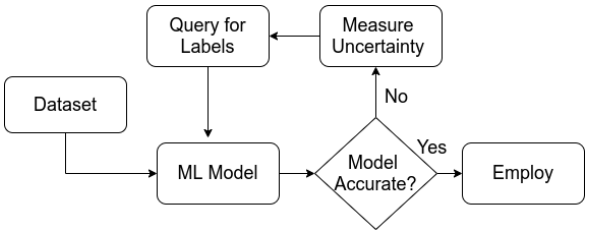


Figure 2: Active Learning Flow

The steps in our active learning process were as follows:

- **Data Gathering:** We gathered unlabelled data for 50,000 u_i's. The unlabelled data was then split into a seed – a small manually labelled dataset consisting of 1000 manually annotated data – and a large pool of unlabelled data. The seed was then used to train the classifier just like a normal ML model. Using the seed dataset we classified each political u_i as either trusted or untrusted.
- **Classification of Twitter Users:** Two manual annotators in the field classified 1000 u_i's as trusted or untrusted based on certain features. Out of 1000 u_i's, 582 were classified as trusted and the rest 418 as untrusted. For feature selection, we employed the feature engineering technique, and selected the most important features among those presented in Table 1. Based on the existing literature [57]–[60] and correlation among features, certain features were considered the most discriminatory for u_i's classification. We did not include the discriminatory features because they serve as an outlier and are biased. In addition, certain features were distributed almost equally between the trusted and untrusted users, as shown in Table 3. We discarded both as they do not add any value to classification. However, certain features were good candidates for differentiating trusted and untrusted users such as high $R_{hind}(u_i)$, $L_{hind}(u_i)$, $Inf(u_i)$, $Sen_s(u_i)$, $Twt_{cr}(u_i)$, $R_s(u_i)$. In Table 3, the features marked with $*$ were used for classification in the existing literature [3], [58], [61] while the features marked with $\cap$ were based on the correlation among the features. The impact of the individual feature is shown in Figure 3. The figure indicates that among the features, the $L_{hind}(u_i)$ and $N_{fol}(u_i)$ are very relevant for assessing $Inf(u_i)$. In addition, all the features except $R_{ret}(u_i)$ and $R_{has}(u_i)$ have a positive impact on the user's $Inf(u_i)$(see Figure 3).
- **Choosing Unlabelled Instances:** A pool based sampling with a batch size of 100 was used in which 100 ambiguous instances from the unlabelled dataset were labelled and added to a labelled dataset. Different sampling techniques were employed to select the instances from the unlabelled dataset. For the new labelled dataset, the classifier was re-trained and then the next batch of ambiguous unlabelled instances to be labelled was selected. The process was repeated until the model performance did not improve above a certain threshold.

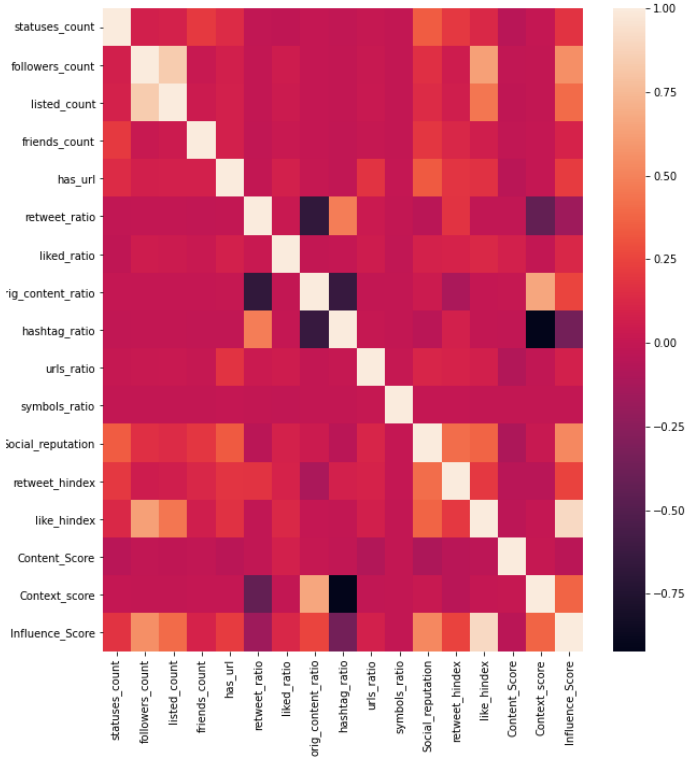


Figure 3: Features Correlation

Among unlabelled instances, active learning finds the most useful ones to be labelled by human annotators. In general, the unlabelled instance which confuses the ML model the most will be the most valuable instance. The following sam-

[9]https://github.com/modAL-python/modAL

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

Table 3: Feature Engineering: All values greater than or equal to 0.5 are considered high, whereas those below 0.5 are considered low.

Discriminatory Features	Equally Distributed	Potential Features	Trusted	Untrusted
$\bullet\, ^{\cap}N_T(u_i)$	$\bullet\, ^{*}R_{url}(u_i)$ $\bullet\, ^{*}U_R(u_i)$	$\bullet\, ^{*}N_{fol}(u_i)$	High	Low
		$\bullet\, ^{\cap}L(u_i)$	High	Low
		$\bullet\, ^{*}N_{fri}(u_i)$	Low	High
		$\bullet\, ^{\cap}R_s(u_i)$	High	Low
		$\bullet\, ^{\cap}Twt_{cr}(u_i)$	High	Low
		$\bullet\, ^{\cap}Inf(u_i)$	High	Low
		$\bullet\, ^{\cap}R_{hind}(u_i)$	High	Low
		$\bullet\, ^{\cap}R_{ori}(u_i)$	High	Low
		$\bullet\, ^{*}R_{has}(u_i)$	Low	High
		$\bullet\, ^{\cap}L_{hind}(u_i)$	High	Low
		$\bullet\, ^{*}R_{lik}(u_i)$	High	Low
		$\bullet\, ^{\cap}Sen_s(u_i)$	High	Low
		$\bullet\, ^{*}N_{lik}$	High	Low
		$\bullet\, ^{*}N_{ret}$	High	Low
		$\bullet\, ^{*}R_{ret}(u_i)$	Low	High

pling techniques were employed to select instances from the unlabelled dataset[10]:

- **Uncertainty Sampling:** It is the most common method used to calculate the difference between the most confident prediction and 100% confidence.
  $U(x) = 1 - P(\hat{x}|x)$
  where $\hat{x}$ is the most likely prediction and x is the instance to be predicted. This sampling technique selects the sample with greatest uncertainty.
- **Margin Sampling:** In margin sampling, the probability difference between the first and second most likely prediction is calculated. Margin sampling is calculated using equation:
  $M(x) = P(\hat{x_1}|x) - P(\hat{x_2}|x)$,
  where $\hat{x_1}$ and $\hat{x_2}$ are the most likely instances. As the decision is unsure for smaller margins, in this sampling technique, the instance with the smallest margin is selected.
- **Entropy Sampling:** It is the measure of entropy and is defined by the equation:
  $H(x) = -\sum_k p_k \log(p_k)$
  where p_k is the probability of a sample belonging to class k. Entropy sampling measures the difference between all the predictions.

Details of the three classifiers we used and their performance characteristics are given below:

- **Random Forest Classifier (RFC):** An ensemble tree-based learning algorithm [62] that aggregates the votes from various decision trees to determine the output class of the instance. RFC runs efficiently on large dataset and is capable of handling thousands of input variables. In addition, RFC measures the relative importance of each feature, and produces a highly accurate classifier.
- **Support Vector Machine (SVM):** SVM models are commonly used in classification tasks as it achieves high accuracy with less computation power. The SVM finds a hyperplane in N-dimensional space (N represents the number of features) to classify an instance [63]. The goal of SVM is to improve classification accuracy by locating the hyperplane that separates the two classes.
- **Multilayer Perceptron (MLP):** A supervised ML algorithm that learns a nonlinear function by training on a dataset. The MLP network is divided into an input layer, hidden layer(s), and output layer [64]. Each layer consist of interconnected neurons transferring information to each other. In our proposed model the MLP consisted of one input and output layer and 50 hidden layers. In addition, the activation functions used in MLP are *Tanh*, *ReLU* and *Logistics*. We do not provide the plots for ReLU activation function as its performance is not as good as Tanh and Logistics (see Table 5).

[10]https://modal-python.readthedocs.io/en/latest/content/query$_s$trategies/uncertainty$_s$ampling.html

V. EXPERIMENTAL RESULTS AND MODEL EVALUATION

Experimental Setup: We used Python 3.5 for features extraction and dataset generation. The python script was executed locally on a machine having configuration: Intel Core i7, 2.80 GHZ, 32GB, Ubuntu 16.04 LTS 64 bit. For training and evaluating the ML models, Google Colab is used. In addition, the modAL framework [65], an active learning framework for python is used for manually labeling the Twitter users. It is a scikit-learn based platform that is modular, flexible and extensible. We used the pool-based sampling technique for the learner to query the labels of instances, and different sampling techniques for the query strategy. For classification purposes, we used different classifiers, implemented using the scikit-learn library.

A. DATASET AND DATA COLLECTION

We used tweepy – the Twitter's search API for collecting u_i's tweets and features. Tweepy has certain limitations, as it only allows the collection of a certain number of features. Additionally, a data rate cap is in place, which prevents the information collection above a certain threshold. Our main concern was to select a sufficient number of users for our dataset. In our dataset, we analysed the Twitter accounts belonging to 50,000 politicians. This dataset was generated in 2020.

The main reason for choosing to evaluate politicians' profiles is their intrinsic potential to influence the public opinion. The content of such tweets originates and exists in the sphere of political life which is, unfortunately, often surrounded by controversial events and outcomes. During the selection, we only considered politicians with a *public profile*. Users that seemed to be *inactive* (e.g. limited number of followers and activities) were omitted. In addition, because duplicate data might influence model accuracy, we used the "max ID" parameter to exclude them from the data set. Firstly, we requested the most recent tweets from each user (200 tweets at a time) and kept the smallest ID (i.e. the ID of the oldest tweet). Next, we iterate through the tweets and the value of the max ID now will equal the ID of the oldest tweet minus one. This means in the next requests (for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access

Table 4: Dataset Descriptive Statistics of only Four Features

	Status Count	Follower Count	Listed Count	Friends Count
Total	473152	28347960	39977	451852
Mean	1112.02	5964.66	7.14	465.04
Standard Deviation	8174.28	199066.10	228.97	2586.83

tweets collection), we got all the tweets having an ID less than or equal to a specific ID (max ID parameter). For all the subsequent requests, we used the max ID parameter to avoid tweet duplication.

For each u_i, we extracted all the features required by our model. Using the extracted features and tweets we calculated $Inf(u_i)$. Furthermore, we collected data that included 19 features including the influence score for 50,000 u_i's. Table 4 summarizes the statistics of some of the features examined in the dataset. For features which have no upper bound defined and may have outliers values, such as the number of followers, likes, etc., we used a percentile clip. We then normalized our features using min-max normalization, with 0 being the smallest and 1 being the largest value.

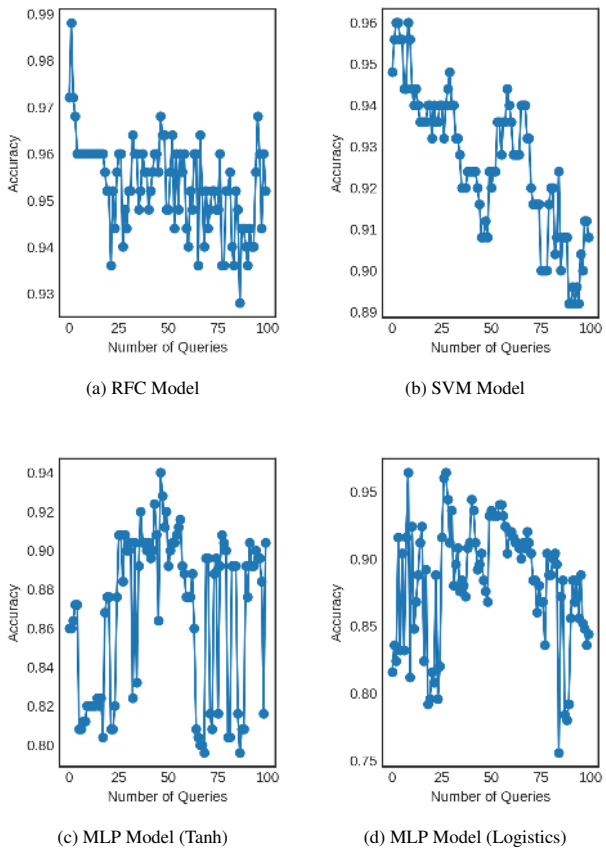
B. PERFORMANCE MEASUREMENTS OF MACHINE LEARNING AND NEURAL NETWORK MODELS

We gathered 50,000 unlabelled instances of u_i's and divided our dataset into three subsets: training, testing, and unlabelled data pools. For the training and testing cohorts, we had 1000 manually annotated data instances. The rest of the data was unlabelled (49,000 instances). The model was trained on the labelled training dataset while the performance of the model was measured on the testing dataset.

For the classification, we used different classifiers (all classifiers were trained on the labelled dataset and predictions are reported using 10 fold cross-validation). The *precision*, *recall*, *F1 score* and *accuracy*, were used as the main evaluation metric for the model performance. Precision is the ratio between true positive and all the positives while recall is the ratio of true positive predictions to the total positives examples. F1 score is the weighted average of precision and recall while accuracy measures the percentage of the correctly classified instances. The precision, recall and F1 score are based on true positive, true negative, false positive and false negative. To define these terms, first we considered that the trusted users are positive (labelled as 1), while the untrusted users are negative (labelled as 0). When the model predicts the actual labels, we categorize them as a true positive and true negative, otherwise false positive and false negative. If the model predicts that the user is trusted but the user is not it is false positive, and if the model predicts that the user is untrusted but the user is not then it is a false negative. The performance of the model (precision, recall, and F1 score) was calculated on the testing dataset. To improve the model accuracy, the active learner randomly selected ambiguous data instances from the unlabelled data pool using three different sampling techniques. These ambiguous data instances were then manually labelled by human annotators. The annotated data was added to the labelled dataset. In our model, the human annotators labelled the 100 most ambiguous instances from the unlabelled dataset returned by the active learner. The respective sampling techniques and the accuracy obtained for the top three classifiers (RFC, SVM and MLP) are discussed below.

Uncertainty Sampling

In uncertainty sampling, the least confidence instance is most likely to be considered. In this type of sampling method, the most probable labels are considered and the rest are discarded. The RFC obtained accuracy of 96% (Figure 4a), the SVM obtained an accuracy of 90.8% (Figure 4b), while the MLP obtained an accuracy of 90% (Figure 4c) for Tanh and 84% for Logistic as given in Figure 4d.



(a) RFC Model (b) SVM Model

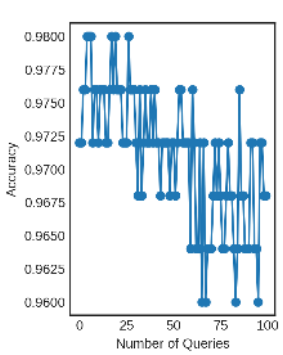
(c) MLP Model (Tanh) (d) MLP Model (Logistics)

Figure 4: Accuracy using Uncertainty Sampling

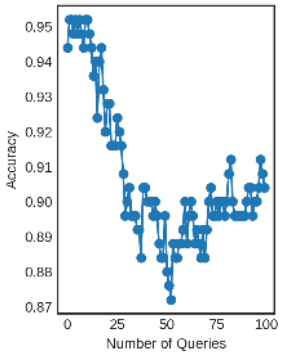
Margin Sampling

In margin sampling, instances with the smallest difference between the first and second most probable labels were considered. The accuracy for RFC, SVM and MLP using margin sampling was 96%, 91.2%, 87% and 88.4% as shown in Figure 5a, Figure 5b, Figure 5c and Figure 5d respectively.

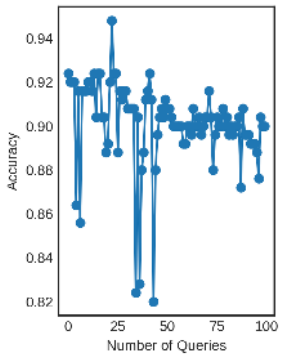
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3098470, IEEE Access



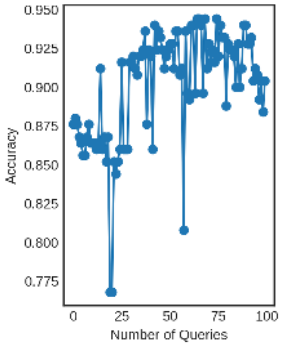
(a) RFC Model



(b) SVM Model

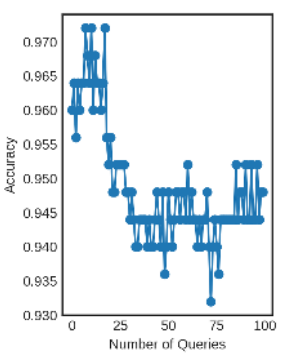


(c) MLP Model (Tanh)

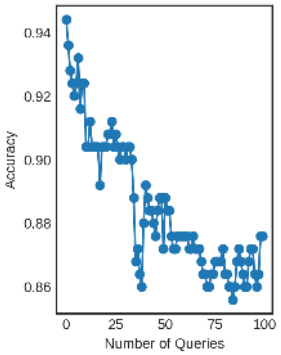


(d) MLP Model (Logistics)

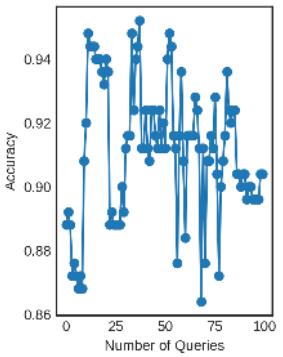
Figure 5: Accuracy using Margin Sampling



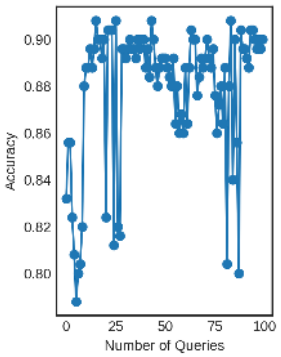
(a) RFC Model



(b) SVM Model



(c) MLP Model (Tanh)



(d) MLP Model (Logistics)

Figure 6: Accuracy using Entropy Sampling

Entropy Sampling

Lastly, the entropy sampling method obtained an accuracy of 95% for RFC, 88% for SVM, almost 90% for MLP (Tanh) and 90% for MLP (Logistic). Obtained results for the RFC, SVM and MLP, are shown in Figure 6a, 6b, 6c and 6d.

Comparison on the performance of our models and different sampling techniques used can be found in Table 5. Precision, recall, F1 score, and accuracy evaluation metrics were used to evaluate the results. Trusted users are represented by 1 while untrusted users are represented by 0 (see Table 5). RFC outperforms the other models in uncertainty sampling, with an F1 score of 96% for both trusted and untrusted users. Similarly, for margin sampling, RFC received an F1 score of 95% for untrustworthy users and 97% for trustworthy users and again outperformed other models. Finally, RFC outperforms in entropy sampling as well, obtaining an F1 score of 95% for both trusted and untrusted users. Overall, RFC was the best performing algorithm, while MLP (ReLU) had the worst performance. The results obtained by RFC were the best due to its superior accuracy and better record when it comes to low-dimensional datasets. Similarly, the improved performance, in the case of margin sampling, can be attributed to the fact that it considers the most probable labels probabilities, unlike the other sampling methods.

Open Science & Reproducible Research

As a way to support open science and reproducible research and give the opportunity to other researchers to use, test and hopefully extend/enhance our models we plan to make both our datasets as well as the code for our models available through the Zenodo research artifacts portal. This does not violate Twitter's developer terms. However, in case the paper gets accepted and in order to keep our anonymity, we will make this available in the camera-ready version .

VI. CONCLUSION

Contemplating the momentous impact unreliable information has on our lives and the intrinsic issue of trust in OSNs, our work focused on finding ways to identify this kind of information and notifying users of the possibility that a specific Twitter user is not credible.

To do so, we designed a model that analyses Twitter users and assigns each a calculated score based on their social profiles, tweets credibility, sentiment score, and h-indexing score. Users with a higher score are not only considered as more influential but also, as having a greater credibility. To test our approach, we first generated a dataset of 50,000 Twitter users along with a set of 19 features for each user. Then, we classified the Twitter users into trusted or untrusted

Table 5: Comparison of various Models using Different Sampling Techniques

Models			Sampling Techniques											
			Uncertainty Sampling				Margin Sampling				Entropy Sampling			
			Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy
Random Forest		0	0.92	0.98	0.96	0.96	0.94	0.97	0.95	0.96	0.93	0.96	0.95	0.95
		1	0.98	0.93	0.96		0.98	0.96	0.97		0.96	0.94	0.95	
Support Vector Machine		0	0.84	0.97	0.90	0.908	0.88	0.94	0.91	0.912	0.83	0.91	0.86	0.88
		1	0.98	0.86	0.92		0.94	0.88	0.91		0.93	0.86	0.89	
Multilayer Perceptron	Logistics	0	0.76	0.98	0.86	0.84	0.83	0.89	0.86	0.884	0.95	0.83	0.89	0.90
		1	0.97	0.71	0.82		0.92	0.88	0.90		0.86	0.96	0.91	
	ReLU	0	0.81	0.87	0.84	0.864	0.81	0.85	0.83	0.84	0.74	0.81	0.77	0.80
		1	0.91	0.86	0.88		0.87	0.83	0.85		0.85	0.80	0.83	
	Tanh	0	0.85	0.93	0.89	0.90	0.89	0.81	0.84	0.87	0.87	0.88	0.88	0.90
		1	0.95	0.88	0.91		0.86	0.92	0.89		0.92	0.91	0.92	

using three different classifiers. Further, we employed the active learner approach to label the ambiguous unlabelled instances. During the evaluation of our model, we conducted extensive experiments using three sampling methods. The best results were achieved by using RFC with the margin sampling. We believe this work is an important step towards automating the users' credibility assessment, re-establishing their trust in social networks, and building new bonds of trust between them.

References

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2):211–36, 2017.

[2] Emily Metzgar and Albert Maruggi. Social media and the 2008 us presidential election. Journal of New Communications Research, 4(1), 2009.

[3] Alex Hai Wang. Don't follow me: Spam detection in twitter. In 2010 international conference on security and cryptography (SECRYPT), pages 1–10. IEEE, 2010.

[4] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security, pages 27–37. ACM, 2010.

[5] Majed Alrubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami, Mohammad Mehedi Hassan, and Atif Alamri. Reputation-based credibility analysis of twitter social network users. Concurrency and Computation: Practice and Experience, 29(7):e3873, 2017.

[6] Matthew Hindman and Vlad Barash. Disinformation, and influence campaigns on twitter. 2018.

[7] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. Nature communications, 10(1):1–14, 2019.

[8] J. Clement. Number of global social network users 2017-2025, July 2020. Surfline.com [Online; posted 27-August-2012].

[9] Muhammad Al-Qurishi, Ryan Aldrees, Majed AlRubaian, Mabrook Al-Rakhami, Sk Md Mizanur Rahman, and Atif Alamri. A new model for classifying social media users according to their behaviors. In 2015 2nd World Symposium on Web Applications and Networking (WSWAN), pages 1–5. IEEE, 2015.

[10] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The tweets they are a-changin: Evolution of twitter users and behavior. Icwsm, 30:5–314, 2014.

[11] Kevin R Canini, Bongwon Suh, and Peter L Pirolli. Finding credible information sources in social networks based on content and social structure. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pages 1–8. IEEE, 2011.

[12] Ramine Tinati, Leslie Carr, Wendy Hall, and Jonny Bentwood. Identifying communicator roles in twitter. In Proceedings of the 21st International Conference on World Wide Web, pages 1161–1168, 2012.

[13] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In Proceedings of the 2012 SIAM International Conference on Data Mining, pages 153–164. SIAM, 2012.

[14] Marie-Francine Moens, Juanzi Li, and Tat-Seng Chua. Mining user generated content. CRC press, 2014.

[15] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents, pages 37–44, 2010.

[16] Hend S Al-Khalifa and Rasha M Al-Eidan. An experimental system for measuring the credibility of news content in twitter. International Journal of Web Information Systems, 2011.

[17] Muhammad Moeen Uddin, Muhammad Imran, and Hassan Sajjad. Understanding types of users on twitter. arXiv preprint arXiv:1406.1335, 2014.

[18] Tanveer Khan and Antonis Michalas. Trust and believe-should we? evaluating the trustworthiness of twitter users. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pages 1791–1800. IEEE, 2020.

[19] Muhammad Al-Qurishi, Sk Md Mizanur Rahman, Atif Alamri, Mohamed A Mostafa, Majed Al-Rubaian, M Shamim Hossain, and Brij B Gupta. Sybiltrap: A graph-based semi-supervised sybil defense scheme for online social networks. Concurrency and Computation: Practice and Experience, 30(5):e4276, 2018.

[20] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. Journal of machine learning research, 2(Nov):45–66, 2001.

[21] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[22] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. In Fifth international AAAI conference on weblogs and social media. Citeseer, 2011.

[23] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1):22–36, 2017.

[24] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In International Conference on Social Informatics, pages 228–243. Springer, 2014.

[25] Nikita Jain, Pooja Agarwal, and Juhi Pruthi. Hashjacker-detection and analysis of hashtag hijacking on twitter. International journal of computer applications, 114(19), 2015.

[26] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web, pages 675–684, 2011.

[27] Annabell Preussler and Michael Kerres. Managing reputation by generating followers on twitter. Medien–Wissen–Bildung Explorationen visualisierter und kollaborativer Wissensräume, pages 129–143, 2010.

[28] Aditi Gupta and Ponnurangam Kumaraguru. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? Technical report, 2012.

[29] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In Proceedings of the 1st workshop on privacy and security in online social media, pages 2–8, 2012.

[30] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. Of bots and humans (on twitter). In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pages 349–354, 2017.

[31] T. Dimitriou and A. Michalas. Multi-party trust computation in decentralized environments. In 2012 5th International Conference on New Technologies, Mobility and Security (NTMS), pages 1–5, May 2012.
[32] Tassos Dimitriou and Antonis Michalas. Multi-party trust computation in decentralized environments in the presence of malicious adversaries. Ad Hoc Networks, 15:53–66, April 2014.
[33] Mohd Fazil and Muhammad Abulaish. A hybrid approach for detecting automated spammers in twitter. IEEE Transactions on Information Forensics and Security, 13(11):2707–2719, 2018.
[34] Amit A Amleshwaram, AL Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang. Cats: Characterizing automation of twitter spammers. In COMSNETS, pages 1–10, 2013.
[35] Chao Yang, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security, 8(8):1280–1293, 2013.
[36] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. An in-depth characterisation of bots and humans on twitter. arXiv preprint arXiv:1704.01508, 2017.
[37] Mark S Granovetter. The strength of weak ties. In Social networks, pages 347–367. Elsevier, 1977.
[38] Isabel Anger and Christian Kittl. Measuring influence on twitter. In Proceedings of the 11th international conference on knowledge management and knowledge technologies, pages 1–4, 2011.
[39] Corren G McCoy, Michael L Nelson, and Michele C Weigle. University twitter engagement: using twitter followers to rank universities. arXiv preprint arXiv:1708.05790, 2017.
[40] Alex Leavitt, Evan Burchard, David Fisher, and Sam Gilbert. The influentials: New approaches for analyzing influence on twitter. Web Ecology Project, 4(2):1–18, 2009.
[41] Annabell Preussler and Michael Kerres. Managing reputation by generating followers on twitter. Medien–Wissen–Bildung Explorationen visualisierter und kollaborativer Wissensräume, pages 129–143, 2010.
[42] Hridoy Sankar Dutta, Aditya Chetan, Brihi Joshi, and Tanmoy Chakraborty. Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 242–249. IEEE, 2018.
[43] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. Classification of twitter accounts into automated agents and human users. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pages 489–496. ACM, 2017.
[44] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. International journal of emergency management, 6(3-4):248–260, 2009.
[45] Xu Han, Xingyu Gu, and Shuai Peng. Analysis of tweet form's effect on users' engagement on twitter. Cogent Business & Management, 6(1):1564168, 2019.
[46] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. Modeling topic specific credibility on twitter. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, pages 179–188, 2012.
[47] Jacob Ross and Krishnaprasad Thirunarayan. Features for ranking tweets based on credibility and newsworthiness. In 2016 International Conference on Collaboration Technologies and Systems (CTS), pages 18–25. IEEE, 2016.
[48] Megan Soltau. Twitter news is a popularity contest., Jan 2020. .
[49] Jorge E Hirsch. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences, 102(46):16569–16572, 2005.
[50] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on twitter: A survey. Information processing & management, 52(5):949–975, 2016.
[51] Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer. Social media and the arab spring: Politics comes first. The International Journal of Press/Politics, 18(2):115–137, 2013.
[52] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing. Secondary TextBlob: simplified text processing, 3, 2014.
[53] Evgeny Morozov and Mourjo Sen. Analysing the Twitter social graph: Whom can we trust? PhD thesis, MS thesis, Dept. Comput. Sci., Univ. Nice Sophia Antipolis, Nice, France, 2014.
[54] John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer, and Sibel Adalii. Credibility in context: An analysis of feature distributions in twitter. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pages 293–301. IEEE, 2012.
[55] David Garcia, Pavlin Mavrodiev, Daniele Casati, and Frank Schweitzer. Understanding popularity, reputation, and social influence in the twitter society. Policy & Internet, 9(3):343–364, 2017.
[56] Gina Masullo Chen. Tweet this: A uses and gratifications perspective on how active twitter use gratifies a need to connect with others. Computers in human behavior, 27(2):755–762, 2011.
[57] Kheir Eddine Daouadi, Rim Zghal Rebaï, and Ikram Amous. Organization vs. individual: Twitter user classification. In LPKM, 2018.
[58] Albert Pritzkau, Steffen Winandy, and Theresa Krumbiegel. Finding a line between trusted and untrusted information on tweets through sequence classification.
[59] Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. arXiv preprint arXiv:1707.00086, 2017.
[60] Supraja Gurajala, Joshua S White, Brian Hudson, Brian R Voter, and Jeanna N Matthews. Profile characteristics of fake twitter accounts. Big Data & Society, 3(2):2053951716674236, 2016.
[61] P. Nyein M. Myo, M. Swe and N. Myo. Fake accounts classification on twitter. International Journal of Latest Engineering and Management Researc, 3(6):141–146, 2018.
[62] Gérard Biau. Analysis of a random forests model. The Journal of Machine Learning Research, 13(1):1063–1095, 2012.
[63] William S Noble. What is a support vector machine? Nature biotechnology, 24(12):1565–1567, 2006.
[64] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7):579–588, 2009.
[65] Tivadar Danka and Peter Horvath. modal: A modular active learning framework for python. arXiv preprint arXiv:1805.00979, 2018.

Tanveer Khan received the Master Degree in Information Security from COMSATS University Islamabad Pakistan. After his Master's, he worked as a Data Analyst on the project CYBER Threat Intelligence Platform at COMSATS University, Islamabad, Pakistan. He also worked as a Junior analyst at Trillium Infosec, Pakistan. Currently, he is working as a Ph.D., Researcher at the Department Computing Sciences, at Tampere University, Finland. He is also a member of Network and Information Security Group (NISEC) at Tampere University, Finland. His interest is in privacy-preserving machine learning, fake news detection in social networks, cyber security, digital forensics and malware analysis.

Prof. Antonis Michalas received his PhD in Network Security from Aalborg University, Denmark and he is currently working as an Assistant Professor at the Department Computing Sciences, at Tampere University, Finland where he also coleads the Network and Information Security Group (NISEC). The group comprises Ph.D., students, professors and researchers. Group members conduct research in areas spanning from the theoretical foun-

dations of cryptography to the design and implementation of leading edge efficient and secure communication protocols. Apart from his research work at NISEC, as an assistant professor he is actively involved in the teaching activities of the University. Finally, his role expands to student supervision and research projects coordination. Furthermore, Antonis has published a significant number of papers in field related journals and conferences and has participated as a speaker in various conferences and workshops. His research interests include private and secure e-voting systems, reputation systems, privacy in decentralized environments, cloud computing, trusted computing and privacy preserving protocols in eHealth and participatory sensing applications.