2019 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 27-28, 2019

A Framework for Detecting and Tracking Religious Abuse in Social Media

Tanvir Ahammad
Dept. of Computer Science & Engineering
Jagannath University
Dhaka, Bangladesh
E-mail: tanvir@cse.jnu.ac.bd

Md. Khabir Uddin
Dept. of Computer Science & Engineering
Jagannath University
Dhaka, Bangladesh
E-mail: monzilahamed321@gmail.com

Abdul Karim
Dept. of Computer Science & Engineering
Jagannath University
Dhaka, Bangladesh
E-mail: karim0725@gmail.com

Sajal Halder
Dept. of Computer Science
RMIT University
Melbourne, Australia
E-mail: sajal.halder@rmit.edu.au

***Abstract*—Religious abuse in social media is a common phenomenon in recent years. This abuse affects our religious morality which generates wrong ideas to the people connected in social media about different religious communities. There are many approaches in the state-of the-art have been used to filter out the desired sentiments from the social media. This paper presents such a framework for identifying and tracking suspicious users who mislead people by spreading conflicting religious information. The framework is designed and trained with diverse religious keywords so as to process real time data stream. We used Twitter social data stream for experiment purposes. Initially, the interested tweets (i.e., Islamic) have been extracted; then the preliminary stages of misinterpreted religious data is detected on the basis of Twitters spam policy, user-based and content-based features. Moreover, a web crawler is constructed with API method provided by the Twitter. Afterward, the processed data is classified to distinguish the mistrustful behaviors from regular events (i.e., the trained classifier is applied to the entire data set). Finally, the performance of the framework is evaluated with different classifiers i.e., Support Vector Machine (SVM), Random Forest and Decision Tree classification algorithms. The result shows much closer accuracy of each in most of the cases. However, we can say that the proposed approach can detect and trace any religious fraud acts in social network.**

***Index Terms*—Social Media Security; Fraud detection; Twitter API; Classifier.**

I. Introduction

Nowadays, social networking and microblogging services are becoming popular day by day. These online social networking medium connect millions of users all over the world, specially with friend circles, meet with new people, make work-oriented connections, share or post their own thoughts and more. Instead of the merits of social networking services, some suspicious users and malicious community use these services for unwanted purposes (e.g., spreading misinformation, spamming news, terrorist activity). Comparing to other social networking services, Twitter is the fastest growing and recently generated much noticing site in research field because of its peculiar characteristics, open policy for data sharing, and tremendous popularity, surging more than 2,800% in 2009 according to the report [1]. The main goal of it is to permit well-wishers to communicate and retain connection by exchanging short messages. Unfortunately, spammers, suspicious users or fraud users use Twitter as a tool to post malicious links, send unsought messages to legal users, and also spreading misinformation about various topics. Their detection and tracking policy was discussed in [2] [3]. However, spreading of misinformation about religious community becomes a popular topic in recent times, as social media grows rapidly. It is a natural trend that people find different ways to abuse it. In this paper, we proposed a framework to identify and track the suspicious users who misguide people by spreading misinformation about various religious community. The major contributions of this paper are as follows:

- The framework is organized and trained with various religious keywords of different community so that we can process it with real time social media data stream.
- A Web crawler (i.e., based on Twitter API method) is used to extract publicly available data stream.
- Novel user based and content-based features are proposed to facilitate the abuse detection based on the spam policy.
- Different classification methods are used to evaluate the performance the model in order to distinguish suspicious behaviors from regular events.

The rest of the paper is organized as follows. In section II, we present related research background regarding fraud detection in social media. Section III, we clearly specify the problem that is trying to solve in this paper. Section IV, represents features extraction from social media data sources. Then, the proposed system is presented in section V. In section VI, we present how we experimented and show results. In section VII, we show performance evaluation of different classifiers. Finally, we conclude by suggesting future directions in section VIII.

II. Related Background

Different data mining techniques have been applied to detect, track, mitigate, resolve many different problems in the social networking and micro-blogging field. A deep study including analysis of Twitter and rules, regulations on Twitter was proposed by Pear Analytics in [1]. It included that what actually Twitter does, what services it provides to the social medias people, what rules and regulations should maintain the user using this platform. A deep study whether tweets are from human, bot, or cyborg [4] was proposed by Chu, Z. et al. Honeycutt [5]. The general concept of social graph model was proposed by Wang A. H., to know about the details definition about friend, follower, mutual friend, stranger and their relationships among them and also know about spam detection analyzing Twitter data [2]. A system was designed by Ratkiewicz et. al. in [3] explained the analysis of data obtained using a mechanism that describes the detection of astroturf campaigns on Twitter. They demonstrated a web service that points political sentiments in Twitter and detects Astroturf, smear campaigns, and other misinformation. As we work with real time data stream from Twitter, we put deep look on [6] by Sakaki et al. Furthermore, a system for detecting suspicious behavior tendency and its subsequent directions was proposed by Meng Jiang in [7] that provides us to know about different detection scenarios, traditional spam and different detection methods etc. We have found a detailed learning about fake news detection on social media using data mining approach that was proposed [8] by Kai Shu in 2017.

III. Problem Statement

People are now consented to share their daily activities, emotions, thoughts or opinions to diverse communities in social media platforms.These make a good social consciousness among the people, but some misleading users usually spread false idea or create false impression on communities. Few of users have become more interested in spreading religious iniquities. As a result, spreading one person to another the falsified information creates religious clashes among the different communities. So it is difficult to extract negative religious sentiments because of big data stream in social medias. As Twitter has open data sharing policy and simple message of less than 140 characters, hence it is a good research platform to analyze this type religious abuse. In this section, we formulate the problem with respect to Twitter data source.

The general concept of social graph model proposed by Wang A. H. [2] represents the concept of friend, follower, mutual friend, stranger and their relationships among them. "Following" is one of the unique feature (i.e., Twitter) which is not a mutual relationship unlike other online social networks. However, the features are formally stated as-

Definition 1. ***(Follower).*** *"Node v_j is a follower of node v_i if the arc a = (j, i) is contained in the set of arcs, A; i.e., followers are the incoming links of a node. Let the set A_i^I denote the inlinks of node v_i, or the followers of user i."*

Definition 2. ***(Friend).*** *"Node v_j is a friend of node v_i if the arc a = (i, j) is contained in the set of arcs, A. Friends are the outgoing links of a node. Let the set A_I^O denote the outlinks of node v_i , or the friends of user i."*

Definition 3. ***(Mutual Friend).*** *"Node v_i and node v_j are mutual friends if both arcs a A_i^M= (i, j) and a = (j, i) are contained in the set of arcs, A. Since a mutual friend is a follower and friend at the same time, the set of mutual friends is the of the set of friends and the set of followers. If A_i^M denote the set of mutual friends of node v_i, then the following holds: $A_i^M = A_i^I \cap A_i^O$."*

Definition 4. ***(Stranger).*** *"Node v_i and node v_j are strangers if neither arcs a=(i,j), nor a=(j, i) is contained in the set of arcs, A."*

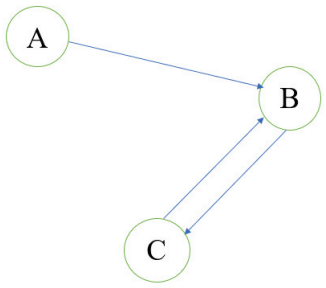

Fig. 1. User interaction graph in Twitter.

Twitter is such a social media platform that exhibits a complex graph model. The Fig. 1 represents a one of its simple model where A is following B, and B and C are following each other. Hence, A is a follower of B; B is a friend of A; B and C are mutual friends; whereas A and C are strangers. Since the proposed framework is about tracking and detecting religious abuse, we mention herein some scenarios of fraudulent incidents.

Case-1: Suppose a user creates nearly same tweet multiple times using third party tweet generating tools in order to get more different audiences in each tweet (e.g., shown in Fig. 2). We may refer it to fallacious act because a legitimate user usually never try it unnecessarily.

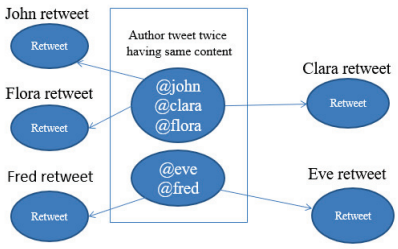

Fig. 2. Tweeting in multiple times for the purpose of suspicious act.

Case-2: If a large percentage of retweeters of a tweet is found fraud, then this post may also be fraud with higher confidence as shown in Fig. 3.

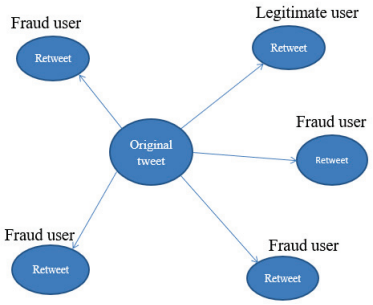

Fig. 3. Status of retweeters involve in Tweeting.

Case-3: If a fraud post containing hash-tag(s) is found on many other posts recently, then the hash-tag seems to be viral among religious community. The Fig. 4 represents this scenario.

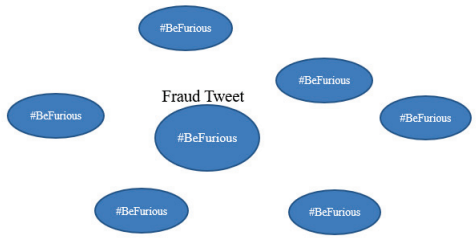

Fig. 4. Specific hashtags found on many user's Tweet.

IV. Features

In this section, the extracted features from user account in social media required to track fraud activity are introduced. The features (e.g., in Twitter) are extracted from different aspects including graph-based and content-based features.

A. Graph-based Features

Following friends and allowing followers to follow of user account is a usual phenomenon in social platform; but in some cases false accounts are being attempted to follow large number of users. From the twitters spam and abuse policy, if a user has a number of followers compared to the numbers of people are following, then it may be considered as a spam account.So we need to formulate a criteria for finding spam user. Reputation is such a term defined as the ratio using the number of friends and the number of followers. If the number of followers is comparatively less than the number of people following, then the reputation is low (approx. to zero). This increases the probability of associated account being spam.

B. Content-based Features

There are also content-based features e.g., Duplicate Tweets and Replies & Mentions described in this subsection.

1) Duplicate Tweets: Legitimate users generally will not post same contents or updates multiple times. If duplicate tweet posting occurred on an author's timeline or a retweeters timeline, the user should be predicted as fraud user. The duplicity is detected by Levenshtein distance (also known as edit distance) measurement between two dissimilar tweets. The distance is zero if and only if the two tweets are homogeneous. However, the tweets are counted as duplicate when the distance exceeds a certain threshold.

2) Replies and Mentions: A user is identified by a unique username and can be mentioned in the '@username' format on Twitter. The @username creates a link to the user's profile automatically. One can send a reply message to another user in "@username+message" format where "@username" is the message receiver. Moreover, the user can mention another "@username" anywhere in the tweet rather than just the beginning. This service is abused by the spammers to gain attention of various users in sending unsolicited replies and mentions. The number of replies and mentions in one user identity is measured with the number of '@' symbol contained in user's some recent tweets.

V. Proposed Religious abuse Detection Framework

Before proceeding to our detection system, we firstly specify what is religious post, keywords, and fraud religious twitters that are incorporated into the proposed framework.

Definition 5. ***(Religious Post).*** *If a person gives opinion in social media based on authentic or divergent sources among the religious communities, then it can make either positive or negative religious impact on the people is considered religious posts.*

Definition 6. ***(Religious Keywords).*** *Any person of any religious community identifies a post whether it is related to Islam, Christianity, Hinduism, or Buddhism is considered as religious keywords e.g., Islam like Allah, Ramadan, Prophet Mohammad(sm.), Namaj; for Christianity like Bible , Church, Crucifix; for Hinduism like Puja , Veda , Ramayan; for Buddhism like Dhama ,Gompa etc.*

Definition 7. ***(Fraud Tweeter).*** *Fraud Tweeters are those people responsible for spreading viral or bad morality on any topics on social media. They usually spread misinformation about the religions.*

A. Architecture

The Fig. 5 shows the overall architecture of our proposed methodology.

1) Information Extraction: The users connected in social networks generate many more data time to time. So extracting the real-time data streaming from user activity is the first step of the proposed methodology. The activity is determined with

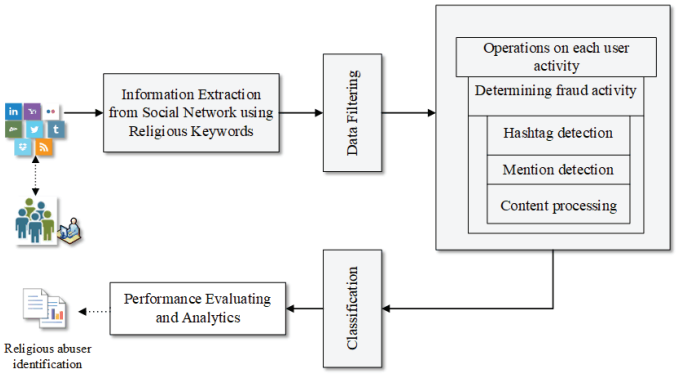

Fig. 5. Architectural diagram of the proposed methodology.

predefined set of different religious keywords. The user post or status that matches the keywords within the time range (e.g., 30 minutes) is considered the targeted dataset.

2) Filtering: As not all activities are involved in religious abuse, hence only interested posts are considered for further processing. So the extracted data is filtered out by calculating interest score based on number of sharing or retweets and likes of a status or tweet. If the post gets more interest score than a specified threshold value, then it is marked as an interesting activity.

3) Operations on each activity: This is the core module of our proposed methodology. The filtered user activity datasets are then processed with various operations to identify behaviors, sources and how the activity actually spreads; to determine whether suspicious acts spreading in a good manner or if it spreading misinformation about various religious communities. Furthermore, additional operations on illegal act including viral #hashtags detection is determined.

- **Fraud activity detection:** The detection is accomplished with reputation score and is defined mathematically

$$R_j = \frac{N_{ij}}{N_{ij} + N_{oj}} \tag{1}$$

where, R_j is the reputation of user j; N_{ij} is the number of followers of user j; and N_{oj} is the number of friends of user j.

We set 0.3333 as threshold value for the reputation score. So if a user involved in Duplicate activity or post same content multiple times, then this can be done using Levenshtein distance as mentioned in section IV. After that, we count the Replies and Mentions appeared in users recent tweets.

- **Operation on Fraud activity:** We now separate the hashtags and mentions included in a activity. A hashtag is detected as viral with a defined third party search operation. We then get the user status (hashtag as keyword in this case) having this hashtag within pre-specified time periods. If the number of interesting user posts or tweets are more than a threshold value (20 or more), the these hashtags are marked to become viral among the communities.

B. Algorithms

This subsection demonstrates algorithms with respect to our proposed system to detect and track abuse. Since we targeted to experiment with Twitter as social media data stream, we designed the algorithms in accordance with Twitter data sources. However, the Algorithm 1 shows how to extract tweet using keywords with specified time range and process them; then interest of each tweet is checked. The tweets having interest more than σ will be processed for fraud detection. Then for each fraud tweet we perform additional operation such as detecting hashtags, processing texts.

Algorithm 1: ReligiousAbuseDetect($Tweets_{keywords}, time$)

Data: $Tweets_{keywords}$: All tweets returned from search API using keywords; $Time$: Time range for tweet creation time;
Result: Information about fraud Tweet involving spreading religious abuse

```
/* Tweet table is initially empty      */
Religious_tweets ← φ
for ∀ tweet ∈ Tweets_keywords do
    if tweet_time > time then
        Religious_tweets ← NoT; /* No Tweets */
    else
        interest=Check_interest(tweet, int_score, σ);
        if interest==1 then
            Fraud =
             Detect_fraud(tweet_id, author, text);
            if Fraud == 1 then
                Processing_hashtags;
            end
        end
        Religious_tweets ← tweet;
    end
end
```

In algorithm 2, we checked if the concerned tweet has sufficient interest.

Algorithm 2: Check_interest($tweet, int\_score, \sigma$)

Data: $tweet$: A single tweet; int_score: Interest score; σ: Threshold for interest;
Result: Determine whether tweet should be processed

```
int_score = tweet_likes + tweet_retweets
if int_score > σ then
    return 1  /* Tweet will be processed  */;
else
    Continue
end
```

In Algorithm 3, we detect a single Tweet is fraud or not. We check the property of the tweet both in retweeters level

Algorithm 3: Detect_fraud($tweet_{id}$, author, text)

Data: $tweet_{id}$: ID of the tweet; $author$: Original creator of tweet; $text$: tweeted text;
Result: Detect whether tweet is fraud or legitimate

$retweeters \leftarrow \phi$ /* Empty retweeters list */;
for $\forall$ $retweeter \in Retweeters_{tweet_{id}}$ **do**
  retweeters $\leftarrow$ retweeter
end
$spammer = Calculate\_spammer(retweeters, \delta, \gamma)$;
$dup_{sign} = Check\_duplicate(tweet, author\_timeline)$;
$rep\_mention_{sign} = Check\_rep\_men(author\_timeline)$
if $spammer > len(retweeters)/2$ ***or*** $dup\_sign == True$ ***or*** $rep\_mention\_sign == True$ **then**
  return 1 /* Fraud detected */;
else
  return 0
end

Algorithm 4: Calculate_spammer(retweeters, δ, γ)

Data: $retweeters$: Retweeters involve in this tweet; δ: Threshold value of reputation; γ: Replies/mentions threshold;
Result: Number of spammer amongst the Retweeters

$spammer \leftarrow 0$
for $\forall$ *user* $\in$ *retweeters* **do**
  Calculate reputation;
  Calculate rep_ment_score;
  dup_tweet : Check using lev. distance;
  if $reputation < \delta$ ***or*** $rep\_ment\_score > \gamma$ ***or*** $dup_{tweet} == True$ **then**
    $spammer = spammer + 1$
  end
end
return *spammer*

Algorithm 5: $Check\_duplicate(tweet, author\_timeline)$

Data: $tweet$: Original tweeted text; $author\_timeline$: set of some other tweet posted by author;
Result: Determines whether author posted the tweet (nearly same) several times

for $\forall$ $post \in author\_timeline$ **do**
  $leven\_dis$: Levensthein distance between post and tweet;
  if $leven\_dis \rightarrow 0$ **then**
    return True /* Nearly same tweets */;
  else
    return False
  end
end

and author level.In Algorithm 4, we calculate the percentage of spammers among the total retweeters of a tweet. Finally, Algorithm 5 determines the duplicate tweets in multiple times.

VI. Experiments and Result Analysis

A. Data-set collection

We used Twitters search API is used to collect users activity, then perform search on twitter against predefined list of 50 keywords from [9]. The search API returns Tweet object with many attributes. We stored only several of them in a table. The features are summarized as

- **Tweet ID:** It is the unique ID of a Tweet.
- **Tweeted Text:** The text contained in a tweeter post.
- **Tweeted ID:** Profile ID of the original creator(Author) from which it is re-tweeted.
- **Spammer percentage:** Percentage of spam/fraud retweeters involve in the tweet. It is calculated by the method describe in section IV.
- **Spammer sign:** It indicates, if the spammer percentage exceeds the threshold or not.
- **Duplicate Sign, ReplyMention Sign:** Calculate used by method described in section IV
- **Fraud:** Identified by a group of expertise whether it is fraud or not. Having two value(TRUE/FALSE).

Moreover, we used 70% data among 1000 data entries for training purpose and rest of 30% data for testing purpose. This data is unlabeled initially. The class label is selected as Fraud.

B. Result Analysis and Discussion

We perform operation on 30% unlabeled data and detect the Tweet; determine whether it is Fraud or none and then label the data. The Fig. 6 shows the percentage of retweeters of each Tweet. When the re-tweets exceed 60% then the tweets are considered abused and rest of them are not religious suspicious tweets. Moreover, in Fig. 7 we showed how religious fraud activities are involved in different tweets of 500 Tweeters. Among these tweets 84 are spamming, 36 duplicate tweets, 4 reply mentions and 110 (approx.) frauds.

VII. Performance Evaluation

To evaluate the performance of our framework various evaluation metrics are used. In this subsection, we reviewed the most widely used metrics for fraud detection.

- **True Positive (TP):** when predicted fraud news pieces are actually annotated as fraud news;
- **True Negative (TN):** when predicted true news pieces are actually annotated as true news;
- **False Negative (FN):** when predicted true news pieces are actually annotated as fraud news;
- **False Positive (FP):** when predicted fraud news pieces are actually annotated as true news.

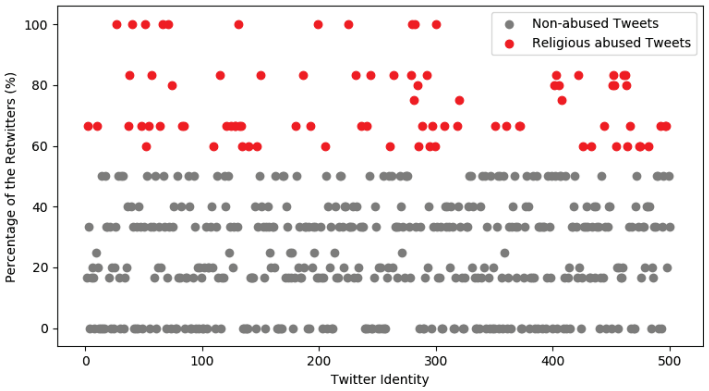

Fig. 6. Percentage of fraudulent re-tweets of each Tweeter.

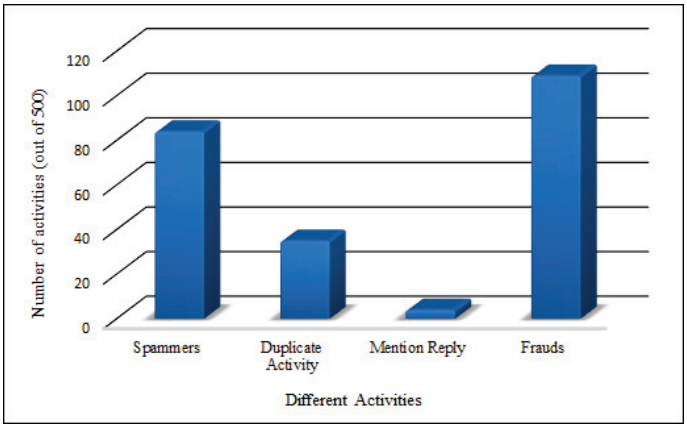

Fig. 7. Number of fraud activities invloved in different tweets.

Since we used different approaches to classify abused activities, we formulate few other metrics to show the performance:

$$Precision = \frac{|\,TP\,|}{|\,TP\,| + |\,FP\,|} \tag{2}$$

$$Recall = \frac{|\,TP\,|}{|\,TP\,| + |\,FN\,|} \tag{3}$$

$$F1 = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{|\,TP\,| + |\,TN\,|}{|\,TP\,| + |\,FP\,| + |\,FN\,|} \tag{5}$$

We evaluate the proposed framework on the basis of three classifiers namely SVM, Decision Tree and Random Forest. Additionally, we label the status of each Tweet by means of three classifier. This is stated as TP, TN, FP or FN. Then we calculate Precision, Recall, F1 and accuracy values as shown in TABLE I.

It is seen form the TABLE I that the Random Forest classifier produces the best results, followed by the Decision Tree and SVM classifier.

TABLE I
EVALUATION OF DIFFERENT CLASSIFIERS

Classifier	Precision	Recall	F1	Accuracy
Decision Tree	0.953	1.00	0.976	0.986
SVM	0.628	1.00	0.772	0.893
Random Forest	0.954	1.00	0.976	0.987

VIII. CONCLUSION

In this paper, we have shown the suspicious behaviors in social networking platforms. As our target is to make an effective technique that works with the real-time data stream from social media (e.g., Twitter) for detecting and tracking religious abuse. To formalize the problem, a social relationship model is proposed to demonstrate the follower and friend relationships. Based on the spam policy, novel user-based and content-based are stated to facilitate the religious abuse detection. A Web crawler using Twitter API methods is also mentioned to extracted real data set from publicly available information based on religious keywords. The extracted different activities are then flirted out and detected various frauds acts. Finally, the performance of the framework is measured with three classifiers. The results showed much closer accuracy of each classifier. So, we claim that the proposed system shows a feasible detecting and tracking mechanism of any religious abuses in social media. However, in future, we will make this work applicable for multilingual fraud detection and tracking purposes.

ACKNOWLEDGMENT

This work was supported by Research Fund of Dept. of CSE, Jagannath University, Dhaka, Bangladesh.

REFERENCES

[1] P. Analytics, Twitter study, 2009. [Online]. Available: http://www.pearanalytics.com/wp-content/uploads/2009/08/.[Accessed: 27/08/2018].
[2] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Security and cryptography (SECRYPT), proceedings of the 2010 international conference on*. IEEE, 2010, pp. 1–10.
[3] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media." *ICWSM*, vol. 11, pp. 297–304, 2011.
[4] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?" in *Proceedings of the 26th annual computer security applications conference*. ACM, 2010, pp. 21–30.
[5] C. Honey and S. C. Herring, "Beyond microblogging: Conversation and collaboration via twitter," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. Ieee, 2009, pp. 1–10.
[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.
[7] M. Jiang, P. Cui, and C. Faloutsos, "Suspicious behavior detection: Current trends and future directions," *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 31–39, 2016.
[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
[9] Religion Spirituality Keywords. [Online]. Available: https://kdp.amazon.com/.[Accessed: 27.08.2018].