DATA MINING APPROACH FOR CLASSIFYING TWITTER’S USERS

Mashael Saeed Alqhtani
Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia

M. Rizwan Jameel Qureshi
Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia

ABSTRACT

Social networks are the most important communication channels in recent years, which popular among the different social groups. These networks affected the ideas and policies of individuals, groups and communities. Every day, millions of tweets on Twitter are being published. These tweets reflect opinions and beliefs of their publishers and affect others as well. Therefore, it is important to analyze these tweets and identify and classify trends of different users.

This research aims to classify social network to anomaly groups such as: Terrorist and dissident; by analyzing tweets data on the Twitter; then identify an anonymous user’s affiliation to these groups. To address this problem, we first extract a set of features to characterize each group using different data mining techniques and store these features in the database. Text mining, sentiment analysis, and opinion mining techniques will be used to accomplish this extraction. The objective of data extraction is to measure the similarity of selected user tweets with respect to extracted features. It will enable to determine high percentage of similarity between the user tweets and group characteristics to expose his/her affiliation to this group.

Key word: Data mining, Social network, Twitter, Analysis, Classification.


1. INTRODUCTION

Social networks such as LinkedIn, Twitter, Google+ and Facebook is a web-based services that present a social structure of individual related to each other directly or indirectly through a joint relation or common interest e.g. trust and friendship, and improved by the technology and the concept of Web 2.0. It allows individuals to communicate and connect with other users on the network (Srivastava, 2008)(Adedoyin-Olowe, Gaber, & Stahl, 2013). Social
Data Mining Approach for Classifying Twitter’s Users

networks can be simply illustrated graphically as nodes and links that used to clarify social relations on social networks. The nodes show entities and the relations between them form the links as presented in Fig. 1 (Borgatti, 2009).

![Figure 1 A Graphical representation of a social network structure](image)

These social networks are useful resources of online contents sharing and interactions, approaches, opinions, feelings, affiliation, and sentiments expressions; taken out from posts, tweets, reactions, blogs, news, discussions, or other documents (Adedoyin-Olowe et al., 2013).

In last recent years, social network services have accelerated information interchange between many users around the world. So, many government departments, organization, companies and individual follow users’ activities on the social network to detect and obtain useful and valuable knowledge from data generated on these social networks; by using effective data mining tools.

Data mining is a technique that used to extract useful, valuable and accurate information from massive data by analyzing and processing the data. This technique used for statistical modelling, information retrieval, and machine learning. And requires huge datasets to mine marked patterns and discover useful information from the data; so, social network sites are powerful resources for applying data mining techniques (Kacprzyk, 2010)(Gundecha & Liu, 2012). Due to the rapid growth of social network sites and communications data mining techniques have become a very powerful and important technique to obtain valuable information from such huge datasets such as rules, trends and patterns. Also, help in clear understanding of social data for organizational and research functions (Aggarwal, 2011). Major data mining techniques used to obtain and extract the information and knowledge are: classification, generalization, association rule mining, clustering, data visualization, and fuzzy logic. Classification involves finding rules that partition the data into disjoint groups. There are several classification discovery models, they are: the decision tree, neural networks, genetic algorithms and some statistical models. (Angulakshmi & ManickaChezian, 2014)(Tewary, 2015).

Social networks have spread rapidly in the Arab region since 2011 and have continued to do so on the past year. Facebook boasted 1.28 billion monthly active users globally, by the end of 1st quarter of 2014. Twitter also, has strong and active increase around the world with 255 million monthly active users as of end of first quarter of 2014. LinkedIn has too also (Mourtada & Salem, 2014).

Saudi Arabia is the country with the highest number of the Twitter active users in the Arab World with 2.4 million users from 5,797,500 active users in all 22 Arab countries. So the twitter becomes the major communication channels for individual, groups and
organizations. More 40% of all active users in the Arab World are from Saudi Arabia. Saudi Arabia also produced 40% of all tweets in the Arab region, alone, while Egypt produced 17% (Mourtada & Salem, 2014). These enormous tweets that have been produced daily show and reflect the beliefs, opinions, behavior, responses and trends of their publishers and affect other users as well. Twitter also has created many opportunities for people to express their opinions to the public, sometimes thus making serious problem. Indeed, some persons and social groups with different interests and beliefs used them tweets to attract followers and then influence them. So, Twitter played a key role in organizing groups as well as reporting about political events. For all these issues, it is important to analyze, study, and examine this data.

This research aims to classify Saudi Twitter users according to anomaly groups; by analyzing them tweets’ in Twitter. This will be achieved by using different data mining techniques and tools for social networks.

The rest of this paper is structured as follows. Section 2 discusses the related researches. Section 2, explains the problem statement. Section 3 explains the proposed system. Section 4 describes the experimental design and the expected results. Section 5 provides concluding remarks.

2. LITERATURE REVIEW

Recently, data mining techniques are widely used in the field of social network; so, there was a rapid increase in the interest with respect to this field. Several studies have been published that aim to extract knowledge from massive data on the online environment. Kathy Lee and others colleagues on (Lee et al., 2011), have used two different data classification models Network-based and Text-based to classify Twitter trending topics into 18 general classes. Aim to help searching on Twitter by aid users to look at smaller subgroups of trending topic which will improve information retrieval. The key contribution of their study is the use of the structure of the social network, rather than using social information exclusively. Considering tweets are not grammatically structured as regular document, and the limitation on the number of characters, no more than 140 characters in generated tweet; text-based classification will provide fair results. But, it is still useful in cases where we cannot be able to apply network-based analysis.

A hierarchical classification approach to recognize emotions in Twitter regarding Brazilian Soccer League 2011 is proposed in (Esmin, Jr., & Matwin, 2012). Six emotions are chosen: happiness, anger, sadness, surprise, disgust, and fear. The authors have used a three-level hierarchal classification. The first step is to determine if the tweets are neutral or emotional. If the tweets identified as emotional, then they further classified depending on their polarity as either positive or negative. Among the six chosen emotions, happiness has the positive polarity, while the other five emotions belong to the negative polarity class. Finally, take the negative tweets from the previous step and classify them in another five negative emotions classes. The main advantage of this study is the consideration of hierarchical information while applying classification. Also, it can be improved by using the features of the emotional lexicon.

Anna Jurek and others in (Jurek, Bi, & Mulvenna, 2014), analyzing the sentiment Twitter contents by developing a lexicon-based sentiment analysis algorithm to estimate the level of disorder and disruption during public events. This algorithm differs from other existing model in the way in which it aggregates the values of positive and negative sentiment words within a Twitter message. They also increase the accuracy of the algorithm by proposing evidence-
based combination function to be applied in cases when mixed of positive and negative words appear in a message. The best aspect of this study is multi-dimensional sentiment analysis rather than just label positive and negative only, but further evaluation is still needed.

In (Akaichi, 2013) the author concentrates on the utilization of text mining techniques for sentiment classification of Tunisian users’ statuses on Facebook during the “Arabic Spring”. He aims to identify users’ behaviors and sentiments during this significant and critical period. For this, he creates his own dataset and then applied two machine algorithms on it: Naive Bayes, because of its performance and simplicity, and Support Vector Machine (SVM), because of its performance and adoption by many previous studies. Those algorithms show high accuracy for classifying sentiment, so it can be applied on different datasets for many purposes.

A clustering is a data mining technique that improves and assists the matching process in social networks. In (Alsleh, Nayak, & Xu, 2011), the authors have used this technique for grouping users with same characteristics together into communities, then matching different users to these communities. This paper addresses the two major problems of matching users to users on social networks, namely, matching accuracy and computational complexity. The clustering process of users in social networks is divided into two phases. The first is the data pre-processing phase and the second is the data mining process. The proposed system reduces the computational complexity by limiting measuring the similarity to assigned communities instead of all users. But additional data techniques should be explored such as association rules that could be developed and assessed to improve the matching of users in social networks.

A technique that analyzes Twitter short message to recognize the user identity in twitter is proposed in (Keretna, Hossny, & Creighton, 2013). The authors aim to authenticate real accounts versus forged accounts using text mining technique. The technique at first analyses short messages from the original accounts, then extracts a set of linguistic features that characterize the writing style of the account owner. These extracted features will be used as input to the learning model to generate a classification model. Finally, this classification process can predict the author of the Twitter message taken as input. This study helps overcoming harmful activities against social network users and introduces a new method to recognize a user’s writing style that can be applied in different domains. Now, the Twitter users can authenticate them accounts through the Twitter Company; so this study lost his important regarding identifying user identity.

A novel class detection framework was described in (Abrol, Khan, Khadilkar, Thuraisingham, & Cadenhead, 2012) that detects trends and evolving patterns of social networks. This proposed approach differs from traditional techniques in the ability to detect the presence of the new class. The key concept of the class detection technique is that each class should have an important characteristic: the data points belonging to the same class must be close to each other this known as cohesion, and must be too apart the data points that belonging to other classes this known as separation. This study produces important contributions. First, it provides a detailed and a deep understanding of the novel class characteristics, and then proposes a new technique to detect new classes. Second, develops a framework that uses this novel class detection technique for identifying Twitter new trends. On the other side, this technique has some limitations, it cannot address the emerging multiple novel classes simultaneously. Also, it does not handle the high-dimensional feature spaces problem. This technique needs refinement to achieve more powerful detection technique.
Our study is different from previous studies in integrating different data mining techniques: text mining, opinion mining, sentiment analysis and clustering to propose a new classification technique that can identify the user’s affiliation to some social network groups. Also, the originality of this research will present on the collected dataset that leads not only to the users’ affiliation, but also to them trends regarding public events. This research will provide a gateway to a new range of applications on social network to improve safety and security of citizens before an accident is occurred.

3. PROBLEM STATEMENT
The modern world is characterized by singularity in the tremendous progress in communication technology, and the pages of social communication of various types and forms. These tools emerged as a cultural achievement created by the giant and creative minds, through the advanced technology; these types of communication have great positive and negative effects.

Some benefits of social media are:
- Used for share and exchange experiences in multiple life areas.
- Gain important information, multiple skills in life and in various fields, and creative experiences in diverse cultural fields.
- Acquire appropriate awareness of the events and news.

Despite the many benefits we derive from social networking sites, it has many disadvantages such as:
- Social networks are an open society, in front of all cultures, including the promotion of the values of corruption and decay.
- It is a suitable place to plan for the spread of crime and extremism sometimes, where they represent a fertile opportunity to meet the extremists and strengthen their and criminal experiences.
- A suitable entrance for enemies to follow the youth and follow their various activities.
- They have a strong opportunity to spread and promote rumors, and in various aspects of life.

To overcome the negative effects of these networks many studies have been provided. The main aims of this research are to classify and identify Twitter users according to different anomaly groups by analyzing users’ tweets on Twitter. This can be achieved as follows:
- Study, analyze and compare the current data mining techniques and machine learning algorithm on social networks to perform a comparative study.
- Extract a set of features that distinguish each group using text mining techniques.
- Propose a new machine learning algorithm that measures the similarity of the user’s tweets with respect to all features in the database or use an existing one according to the comparative study results.
- Evaluate the gained results using qualitative and quantitative techniques.

4. PROPOSED METHODOLOGY
In this paper, we propose a technique to detect user affiliation to some anomaly groups by analyzing short text extracted from different selected Twitter accounts. We use a supervised machine learning technique to learn the features from every user, which have been extracted
from a training data set collected from different Twitter accounts. In order to achieve this, the work is divided into six phases as presented in Figure 2.

**Figure 2** The Overall System Phases

### 4.1. Data Collection

To collect information from social networks, we need to use a data crawler. The crawler that used in our experiment is called Python Twitter Tools (PTT) called Tweepy, which is according to the tool’s website; is “an easy-to-use Python library for accessing the Twitter API,” (“Tweepy,” 2016).

For our experiments’ we use a crawler to collect Twitter massages from 45 different accounts, which were classified into three different groups according to the nature of their tweets:

- Daesh or ISIL (ISIS) members, whom tweets are very aggressive and containing graphic visuals of ISIS vicious actions.
- Opponents member whom tweets are politically opposing the government.
- The Supporters, who react to the incidents with positive comment and no such interest in opposing the government nor supporting ISIS actions.

<table>
<thead>
<tr>
<th>Group name</th>
<th>Accounts number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daesh</td>
<td>15</td>
</tr>
<tr>
<td>Opponents</td>
<td>15</td>
</tr>
<tr>
<td>Supporters</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
</tr>
</tbody>
</table>

### 4.2. Preprocessing

Due to the special nature of the Twitter tweets, some preprocessing steps are needed. As an example, Twitter does not supports embedding audio, video or pictures in its tweets. It just hosting them externally and attaches the link of such media to the text message. So, this requires excluding the URL attachments. Also, the limited size of the Twitter tweet length, that is 140 characters only, has issues some concerns like:

- there is no enough data to be extracted comparing to a long article;
- the short messages are always in slang and there is a high probability of spelling mistakes;
- there are some users who deal with size limitation by using different abbreviations or short words.

In addition to these challenges associated with Twitter, The Arabic language has many different dialects, sometimes in the same country. These dialects will lead to difficulties in the analysis of the message contents, and may need a glossary to translate them into the general
meaning. So, all these issues must be addressed before the feature extraction and data analysis.

4.3. Data Analysis and Feature Extraction

Feature extraction plays a key and critical role in the classification phase of this research. Before feature extraction, there is a need for analyze each group data tweets carefully to select the appropriate features that are closed to each other and different from other group features. To carry out this phase the author used the Weka filter; StringToWordVector, by importing its class from the Weka libraries. This filter, is an important feature in order to categorize the text, that is Weka software tool supports tokenizing, i.e. breaking text utterances into indexing terms called Word Stems, and assigning to them weights in a step of creating a feature vector, a required step before Classification. This tokenizing and indexing process is achieved by using this filter, either in the Java program or in the Weka software tool itself. The feature vector is sent to the learning module to generate a classification model.

4.4. Learning

The learning algorithm will take the feature vector as input to produce a set of trained classification models to predict the user affiliation (group) as shown in fig. 3 graphically. After applying the filter StringToWordVector, the feature vector was written by the constructor of the class TextCategorization into an ARFF file named. The feature vector is sent to the learning module to generate a classification model.

4.5. Classification

In this phase, the following Classification Models were applied to the Training Data stored in the ARFF file, which was obtained in Phase Four: 1) Naïve Bayes, 2) Decision Trees using C4.8, and 3) Random Trees classifiers.

4.5.1. Naïve Bayes

Is a set of algorithms for training classifiers, it assumes a common principle; that is all naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable, regardless of any possible correlations between features.
An advantage of naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. In this project the text of the all extracted tweets is considered to be small in terms of Data Mining, thus naïve Bayes is suitable for the data this project worked on.

The authors applied the naïve Bayes classifier in both paths, Figure 4 shows the results obtained from running this classifier on the training data in Weka Software tool.

The algorithm was run with ten-folds Cross Validation, i.e. it was given an opportunity to make a prediction for each instance of the dataset with different training folds, and the presented result is a summary of those predictions.

The Classification Accuracy shows that the model achieved a result of 61/70 correct or 87.14%. The Confusion Matrix shows a table of actual classes compared with predicted classes, it shows that only three Daesh class members have been classified as being other classes members, only two Opponents class members have been classified as Daesh class members, and four Supporters class members have been classified as other classes members, total of nine errors.

**4.5.2. Decision Trees**

J48 class, imported from Weka libraries, is an implementation of the C4.8 algorithm in Java (“J” for Java, 48 for C4.8). This algorithm is used to generate decision trees. Figure 5 shows the results of running C4.8 classification algorithm on the filtered data.

Similar to what was produced by naïve Bayes classifier, the C4.8 algorithm was run with ten-folds Cross Validation, the following is a summary of those predictions.

The classification accuracy shows that the model achieved a result of 61/70 correct or 87.14%. The Confusion Matrix shows that only two Daesh class members were classified as
Supporters class members, four Opponents class members were classified as members of other classes, and three Supporters class members were classified as other classes’ members, total of nine errors.

Figure 5 Results obtained in Weka Software tool after applying C4.8 classifier

4.5.3. Random Trees
Weka uses this term to refer to a decision tree built on a random subset of columns. With k random features at each node, this classifier could be referred to as a tree drawn at random from a set of possible trees, that is each tree in the set of trees has an equal chance of being sampled. In this project, using another decision tree classifier is just to double assure the accuracy of the results that were produced by the previous classifier.

Figure 6 shows the results obtained from Weka Software tool when the Random Trees classifier was applied to the filtered data, with less accuracy than the C4.8 classifier, the Classification Accuracy of Random Trees classifier shows that the model achieved a result of 59/70 correct or 84.29%, while, as shown before, it was 61/70 correct or 87.14% for the C4.8 classifier. The Confusion Matrix shows that three Daesh class members were classified as members of other classes, seven Opponents class members were classified as members of other classes, and only one Supporters class member was classified as a Daesh class member, producing the total of eleven errors, thus reducing the accuracy of the results.

The previous, shows that the consistency of the results produced by naïve Bayes and C4.8 classifiers, unlike the Random Trees classifier.
4.6. Testing and Evaluation

The F-score (or F-Measure) is calculated based on the Precision and Recall. The calculation is as follows:

\[
\text{Precision} = \frac{t_p}{(t_p + f_p)}
\]

\[
\text{Recall} = \frac{t_p}{(t_p + f_n)}
\]

\[
F - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Where \( t_p \) is the number of true positives, \( f_p \) is the number of false positives, and \( f_n \) is the number of false negatives. Precision is defined as the fraction of elements the algorithm correctly classified as positive out of all the elements the algorithm classified as positive, whereas Recall is the fraction of elements correctly classified as positive out of all the positive elements.

In the multiclass case, which is the case of this system, each class \( A \) have a respective Precision and Recall, in which a "true positive" is an element predicted to be in \( A \) is really in it, and a "true negative" is an element predicted to not be in \( A \) that is not in it. Thus, each class can have its own F-score by doing the same calculation as in the binary case. This is what Weka produced in for the classes in this project.

The model created in the previous phases was tested in Weka Software tool to make sure that it classifies Twitter accounts to their correct classification or not. The Evaluation Measures displayed in the results of each applied classifier showed the accuracy of the created model.
Precision, Recall, and F-Measure for Daesh class members in naïve Bayes was 0.87, Precision for Opponents class members in naïve Bayes was 0.783, while Recall for the same class was 0.9, and F-Measure was 0.837. Finally, Precision for Supporters class members in naïve Bayes was 0.958, while Recall was 0.852, and F-Measure was 0.902.

Precision, Recall, and F-Measure for Daesh class members in C4.8 was higher than it was in naïve Bayes, they were equal to 0.913, the three measures for the Opponents class members in C4.8 classifier were also higher than they were in the naïve Bayes for two out of the three classes, that is Precision was equal to 0.889, Recall was less than it was in naïve Bayes, it was equal to 0.8, while F-Measure was higher, it was equal to 0.842. Finally, two of the measures in C4.8 were less than they were in naïve Bayes for the Supporters class members, that is Precision became less, it was equal to 0.828, Recall became higher, it was equal to 0.889, and F-Measure became less, it was equal to 0.857.

5. CONCLUSION AND FUTURE WORK

Every day, millions of tweets on Twitter are being published. These tweets reflect opinions and beliefs of their publishers and affect others as well. Therefore, it is important to analyze these tweets and identify and classify trends of different users. This may provide valuable information and help to observe what is displayed on social sites to overcome any unusual events.

In this paper, we have introduce a new idea to detect user affiliation to some anomaly groups by analyzing short text extracted from different selected Twitter accounts. The primitive results are very promising. The next step is to apply this technique to a larger number of accounts and investigate the effectiveness and robustness of the proposed approach.

REFERENCES


Malpani Radhika S and Dr. Sulochana Sonkamble, A Data Mining Approach To Avoid Potential Biases. International Journal of Computer Engineering and Technology, 6 (7), 2015, pp. 27-34.