

Source credibility analysis on Twitter users

Malith Wijesekara*

Faculty of Information Technology
University of Moratuwa, Sri Lanka
malith.sandagiri@outlook.com

Gamage Upeksha Ganegoda

Department of Interdisciplinary Studies
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

Abstract: Social media has gained impressive popularity all around the world in the last decade. Social networks such as Twitter, Facebook, LinkedIn, and Instagram have acquired their user's attraction by maintaining their identity with very similar features. With the popularity of these platforms, now a day most of the users tend to rely on the information published on social media. Therefore, the credibility of social media information is playing a major role in the present cyberspace. As an example, the Twitter platform is handling 500 million tweets per day. Most of the twitter messages are truthful, but the twitter platform is also used to spread rumors and misinformation. Truthfulness or reliability is depending on the source's credibility. Twitter profiles can be identified as the information source on the twitter platform. In this paper, a user reputation-based prediction method is proposed to analyze the twitter source credibility. The proposed solution is mainly based on the k-means clustering model. Another two models namely, news category analysis and sentiment analysis are deployed to generate novel features for the clustering method. The objective of this paper is to introduce a credibility rating method to visualize the user credibility of twitter user profiles. So that followers can have an understanding about the trustworthiness of the information published on that profile. Producing the agreement score for a specific twitter user is one of a novel experiment in this research. Achieved accuracy by the system is 0.68 according to the evaluations conducted.

Keywords: Credibility analysis, News category analysis, Sentiment analysis, Source credibility, Twitter source

I. INTRODUCTION

Credibility can be defined as "A perceived quality composed of multiple dimensions". In other words, "Offering reasonable grounds for being believed" [1]. On the other hand, source credibility can be defined as a communicator or the news content provider's positive characteristics that affect his follower's acceptance of the provided news content. As a result of the popularity of social media, petabytes of data are produced per day. Most of the information is important for specific communities. But the question is actually above information is produced by credible sources. Most of the time if a social media user needed to identify the credibility of another profile, he/she has to visit the source profiles and use their cognitive skills to guess how to extend that source can be trusted. This system is proposed as a solution to that problem.

Twitter social media platform is a microblogging service that allows users to publish 140 characters long message called "tweet". Users can append images and reference URLs along with the tweets [2]. Tweets can be published through the web portal, through the smartphone app, through an email as well as sending an SMS. Twitter profiles that are currently active are providing news content for its' followers. There

are twitter sources recognized as reputed and credible by the community. On the other hand, there may be accounts that produce non-credible information to the community. Identifying the credibility of a specific account is a complicated and time-consuming task for individuals. It will be a collection of subtasks such as check the information on a specific profile is matched with other sources, the impression of the followers about the source profile, is it a verified account by twitter, identifying the information categories published often and recognize the suspicious activities of that account [3]. It will be ideal if there is an intelligent application to check the credibility of a specific twitter profile and provide a rough idea about how to extend that account can be trusted. In this paper, a method is proposed to evaluate and forecast the credibility level of a specific twitter profile that is currently active in the platform. This system is a subpart of the research project with the title "Fake News and Social Media Platform Manipulation Detection". This system is mainly focused on the sources that are published newsworthy information and that has the strength to influence a large number of people. The system proposed by this paper can work as an independent module and the inputs will be screen name twitter user and outputs are credibility score of relevant twitter profile. In addition to that, the author's contribution is to implement a system that is capable of generating a score representing all the followers, identify the most interesting information category and predict the source credibility level for a specific twitter profile.

Organizing the rest of the paper as follows. The next section is described the Related work of the subsystem proposed by this paper. The methodology of the system is presented in Section – III. Finally, Discussion is presented in Section – IV.

II. RELATED WORK

The system proposed by this paper is responsible for three main sub-tasks. Those are achieved using the components mentioned below.

- A. News category analysis
- B. Sentiment analysis for tweets
- C. Twitter Source credibility analysis

Related work for the above models discussed below.

A. News category analysis

Twitter is a microblogging service. Therefore, identifying and classifying tweets is hindered by traditional documents. The reason for the difference is short text has no sufficient word occurrences for the feature extraction [4].

A paper written by Bharath Sriram et al. proposed a method to classify short text using a set of domain-specific features extracted from the short text and the user profile which published that text. A paper published by Thorsten Joachims was explored the text categorization using Support Vector Machines (SVM) [5]. It was thoroughly discussed on why Support Vector Machines work well for text categorization. Furthermore, the techniques that can be used for Natural Language processing such as Tokenization and Vectorization is discussed in a survey which is published by Mohd Suhairi and Md Suhaimin [6].

B. Sentiment analysis

Sentiment analysis is a popular field in machine learning which is gained tremendous attention from the researchers. Users' opinions on social media posts can be identified as discovery and recognition of the type of expression that is positive or negative on diverse subject matters of interest. The survey conducted by Albert Bifet and Eibe Frank discussed the techniques that can be used for sentiment knowledge mining [8]. The implementation was based on Kappa statistic and the final output for a given text is Positive or Negative sentiment. Similar research conducted by Hao Wang et al. also discussed the real-time analyzing twitter data to extract the opinions on election candidates. Another research by Kelvin Kiema Kiilu et al. discussed applying the Naïve Bayes classifier for sentiment data. The research was achieved 64.47% accuracy for unigram and 70% accuracy for Bigram. In addition to that, the ability to classify opinion into separate categories such as "positive" or "negative" or "neutral" is discussed in a survey conducted by S. M. Vohra et al. [10].

C. Source credibility analysis

The information in the social networks, cannot be trusted as the conventional news media because there may not be an undergone verification process with the news published in social networks. In addition to that, those social media users who report information may have their own biases, agendas, and perceptions. A survey conducted by DWI H. Widyantoro and Yudiwibisono introduced a way to use the sentiment data to determine the credibility of the published tweet [11]. The proposed method is analyzing the replies for a particular tweet and check the support content and oppose content from the followers. Another paper published by Mohammad-Ali Abbasi and Huan Liu state that information credibility is "The quality of being trustworthy" and information credibility has three major parts namely, source credibility, media credibility, and message credibility. Furthermore, a method based on the CredRank algorithm is proposed for evaluating source credibility in the above research [7, 12]. In addition to that, it is a clustering-based method and it is problematic to analyze the source credibility only depending on the followers' reviews regarding the tweet which is published by a particular source. Interesting research was conducted by Gupta et al. that assign a credibility score to each news or event [13]. The more information is credible; the credibility score will be higher.

An approach based on a PageRank-like credibility propagation method is presented by the author. Another research conducted by Krzysztof Lorek et al. proposed a method with creating a data set manually tagged by checking the twitter profiles and their information [14]. They used their own rules to label each account is credible or non-credible. Below mentioned items were taken into consideration for manually tagging the credibility of the profiles.

- Visible characteristics of the profile photo, account name, availability of the verified badge.
- Tweet text syntax and sentence correctness.
- Content of the images appended with tweets.
- Availability of the external links.

The proposed solution in this paper also included the above features and in addition to that added an extremely impotent feature which represents the opinion of the followers about the content of the source.

III. METHODOLOGY

The system proposed by this paper can be separated into four components. Those can be work as separate entities and some of the outputs of the component will be input to the others.

1. News category analysis component
2. Sentiment analysis component
3. Source credibility analysis component
4. Source credibility visualization component

The flow diagram for the system with the above four components is shown in Fig. 1. Input for the system is a username which needs to be analyzed the source credibility. The data required for the analysis should be acquired using the screen-name through the Twitter API [9]. The account timeline of a specific user and the user data has to get separately. The user timeline means the tweets that are published by the user so far. It is playing a major role in analyzing user activities. Only tweet text is required for news category analysis and sentiment analysis components. The output of the above two components will be inputted to the credibility analysis component. Although there are plenty of systems for sentiment analysis and text classification, it is better to implement them both as the accuracy of these components will directly affect to the Credibility analysis component.

The other input for the Credibility analysis component is user data that are requested through the twitter API. This component is responsible for generating the credibility score. It is a probability number that shows how to extend that account can be trusted. In addition to that, recent activities of the relevant account are also analyzed by the Credibility analysis component. Finally, the summarized data and information is visualized by the Source Credibility visualization component. Data visualization techniques are used to interpret information discovered by the system and highlight the follower's reactions for recent activities of the relevant twitter profile.

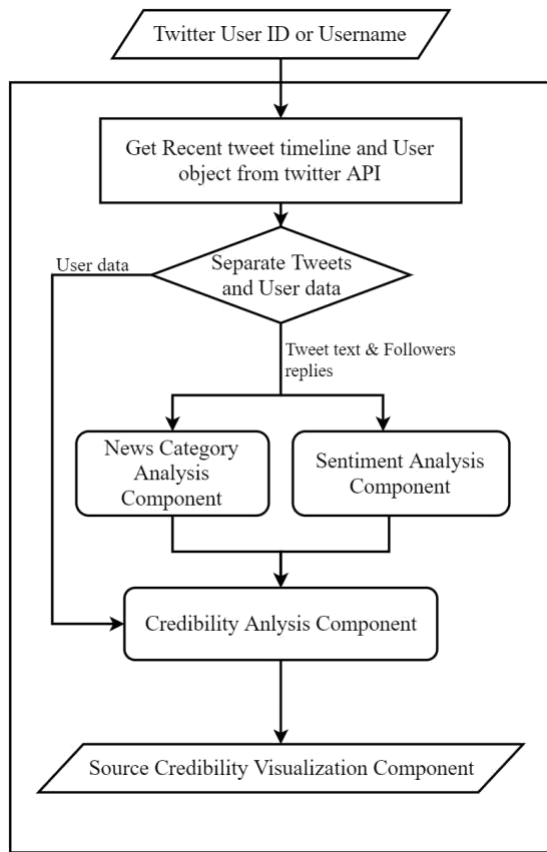


Fig. 1. Flow of the proposed solution

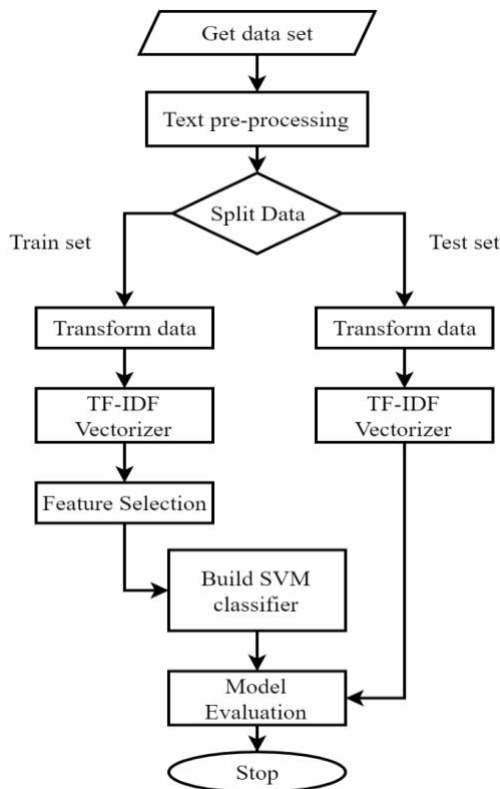


Fig. 2. Flow of the News Category Analysis Component

As mentioned above this system is proposed as a combination of four separate components. The final results are highly dependent on the accuracy of the sentiment analysis and category analysis components. Therefore, the implementation of those components also included in the research. The four subsections are described in detail as follows. The four subsections are described in detail as follows.

A. News category analysis component

News category analysis is responsible for identifying more relevant news category for a given tweet message. this process is based on semantic analysis. Text classification problem has to be solved using a supervised machine learning approach [15]. A comparative study is conducted to identify the best classifier for that data set. Suitable machine learning algorithms are Naïve Bayes classifier, Decision Tree (Hoeffding Tree) and Support Vector Machine. The flow chart of the proposed methodology is shown in Fig.2. According to the Fig.2. Data set consists of tweets and the relevant category is preprocessed to remove blank records, convert text to lowercase. After splitting the dataset into train and test sets, the text is tokenized or separated the sentences into tags. In the next phase, they are vectorized to input into the classifier to create the model. Finally, the best classifier is identified by comparing the results.

B. Sentiment analysis component

Opinion mining or sentiment analysis has to be performed to evaluate the follower’s responses. Opinion mining is the automated extraction of opinions, attitudes, and emotions from the text. Mainly there are two methods for sentiment analysis.

- I) Lexicon-based methods
- II) Machine Learning-based methods.

The lexicon-based method is based on predefined lexicons which are similar to a dictionary. Previously mentioned lexicons are arranged as a list with defining they are positive or negative with a sentiment score. As an example, the word ‘nice’ is defined as a positive word having sentiment scores of +2. The sentiment score can be changed according to the researcher’s point of view who defined the lexicons. If a phrase is an input to the algorithm it will combine individual polarity and calculate the sum of the sentiment scores to output the sentiment of that phrase.

The machine learning method is similar to the automatic mode which is discussed in the previous section: News Category Analysis Component. Naïve Bayes classifier, maximum entropy, and SVM are achieved tremendous success in sentiment analysis. According to the conducted analysis of opinion mining, it is better to choose a machine learning method for sentiment analysis. Because supervised machine learning techniques are relatively performed better than lexicon-based methods. Most of the researchers experienced that SVM has a higher accuracy over the above-mentioned algorithms [16].

But the drawback of the SVM, the smaller the training data set it will reduce the accuracy of the model. So, a large amount of data is required for better results. Sentiment data sets can be found online in large varieties.

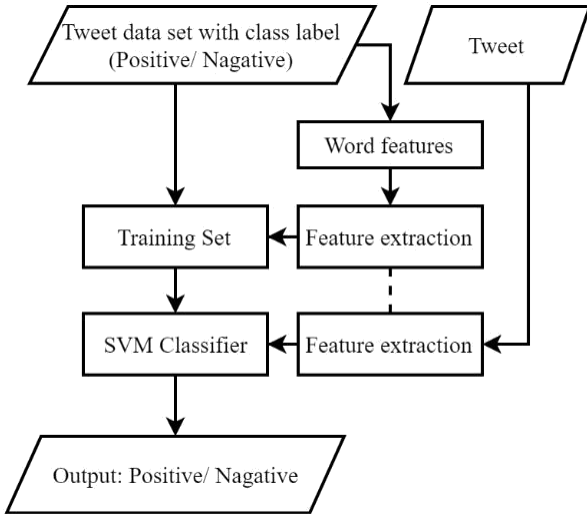


Fig. 3. Flow of the Sentiment Analysis Component

The SVM classifier is selected as the classifier for the model. Fig. 3. is illustrated the flow of the proposed solution for sentiment analysis component. The sentiment data set is needed to be preprocessed and after feature extraction SVM machine learning classifier is applied. In this process, patterns in the text are learned by the classifier concerning class labels. So, the classifier model can classify an opinion or feeling of a given tweet as positive or negative. The sentiment analysis model is used to analyze the follower’s expressions and generate a feature for the credibility analysis model.

C. Credibility analysis component

In any social media, there will be a large number of news and posts which are posted by unknown sources. If the user is not sure about a specific source, there is no way to verify the source without viewing and checking the profile. The credibility analysis component is the solution for that manual process of analyzing the credibility of the sources. The outcome of the credibility analysis is mainly relying on opinion mining. As an example, if someone reads the replies on a particular tweet, he/she can have a rough idea of whether the followers agreed with that tweet or not according to the sentiment of the phrases. First, the tweet should be analyzed and then the individual replies to that comment should be analyzed. This process is called agreement modeling and it is a parameter for the credibility analysis model [17]. The overall sentiment of the followers’ replies to the source profile is calculated in agreement modeling and the output of the process is agreement score. Availability of the verified badge from twitter, responses of the followers, and profile information such as the number of followers and lifetime of the profile are also can be identified as factors to determine source credibility [11, 13]. Already conducted researches identified that Statuses count, followers count, length of the user description, length of username, number of days from account creation date, availability of verified badge, availability of an external URL for the profile, availability of profile location and protected mode are can be used for analyzing source credibility of a twitter user. A novel feature, agreement score and interested news category of the user are added to the data set. All the features and their effect on the credibility of the source are summarized in Table I.

TABLE I. FEATURE SET FOR CREDIBILITY ANALYSIS

	Feature	Description (a tweet is more credible when...)
1	Agreement score	Source has a higher agreement score
2	Interested news category	Will be depending on the other features
3	Statuses count (No. of tweets)	Source has a higher number of previously published tweets
4	Followers count	Source has a higher number of followers
5	Length of the description	The description length is more than '0' characters Length and higher lengths
6	Length of screen name	Source screen name differs from the username
7	The lifetime of the profile	Hight lifetime
8	Is it a verified account?	True
9	Has the "Website" parameter set in the profile?	True
10	Is the user's Twitter profile linked to another social service?	True
11	Is the profile location set?	True
12	Is profile protected?	True

Methodology for Calculating Agreement Score is based on sentiment analysis. A large number of follower's replies for a twitter user are required as the input for this process. Each replay has to be input to the sentiment analysis component and recorded the sentiment. Finally, the below-mentioned rule is applied to get the agreement score.

$$\text{Agreement score} = \frac{\text{Number of positive followers' responses}}{\text{Total number of followers' responses}}$$

The proposed methodology for identifying interested news categories is getting the first 100 tweets published in the twitter user profile and record input them to the news category analysis model discussed previously. According to the output frequently repeated categories will be identified as the interested news category.

However, the collected data set will not be having a ground truth or class label. Therefore, it is obvious that a supervised learning method cannot be applied to the data set. Therefore, the clustering technique is chosen to analyze source credibility [18]. Artificial Neural Network (ANN) also can be used for creating the credibility model, but the higher complexity and requirement of higher processing power is the barrier for working with the ANN method.

The K-means clustering method is a well-known agglomerative partitioning clustering method [19].

K-means clustering technique is not used previously to implement a source credibility related model. Therefore, the

K-mean clustering method is proposed to create a model for the collected dataset. K-means algorithm is an algorithm that identifies the similarities of different data vectors and groups them together into predefined numbers of clusters.

The above data set can be clustered using both k-means clustering or hierarchical clustering. But according to the requirement, five clusters need to be identified as a star rating method is proposed to identify the level of source credibility. Therefore, k-means clustering is the best algorithm because the number of clusters can be predefined. K-means clustering is one of the simplest classifiers and general preprocessing and transformation can be used to build the model. It can be applied to the data set without knowing much about the dataset.

The main challenge of the model is identifying the correct cluster and naming them. K-mean clustering it is common to use a manually annotated small dataset for identifying the properties of the clusters extracted by the algorithm. Therefore, a small set of twitter user profiles was selected and annotated them manually by looking at the feature and rules mentioned in Table 1. Then the manually annotated dataset and trained dataset with predicted values are compared using two bar charts. Hence the proportionally matched rating is considered as correct clusters. Correct clusters for each rating are identified using predicted values and manually annotated values. When it needs to get the credibility level of a new user, a User credibility rating can be acquired by applying the user data into the model and convert the output of the model into the correctly identified rating.

D. Source credibility visualization component

The purpose of this component in the system is visualizing the credibility rating of a twitter user when someone enters a screen name of the profile to the proposed system. In addition to that recent activities, user patterns such as how often tweets are published and highlighted follower’s opinions need to be visualized. The impact that can be happened by a specific account to the community also identified using the category analysis component and sentiment analysis component. The impact that can be happened by an account is determined using the categories of the most recent tweets and followers’ attitudes about the published tweets. The impact that can be happened by the source is identified using predefined rules. Proposed rules are,

- I. If the sentiment of twitter message and the overall sentiment of the replies are matched with each other it is an agreement. (Negative- Negative, Positive - positive).
- II. Otherwise, it is a disagreement. (Negative - Positive, Positive - Negative).

If the followers’ agreement is low, it is identified as a negatively impacted tweet. Otherwise, it will be a positively impacted tweet. In this part, the system should be able to get the recent tweets in the relevant profile timeline a label them as positively impacted or negatively impacted. It helps the system used to have a rough understanding of a specific profile.

IV. EXPERIMENTS

The main consideration for implementing the system is choosing the correct programming language. As most of the things are towards the data science and Python and R languages were the best selection. R language is mainly used for statistical analysis. But the Python is popular as a scalable language and it has several useful highly developed libraries such as Pandas, NLTK and Scikit-learn which are very useful for data analysis. Another plus point of python is it has matured libraries for Twitter APIs than the R language. Therefore, Python is selected for implementing the solution. In addition to that this system will be web-based and has to be handle API endpoints as well. So, the Django REST framework is selected for that purpose.

News category analysis is designed using NLP techniques to identify the relevant category of a twitter message. A data set consist of 124,989 tweets and labeled with the news category is found online for the machine learning process. Experiments were conducted with the above Naïve Bayes classifier, Decision Tree (Hoeffding Tree) and Support Vector Machine classifiers. Support Vector machine was the best classifier with the accuracy of 0.61.

Experiments for the sentiment analysis component also conducted similar to the news category analysis component. A data set with 1.6 Million tweets with the class label is found from a reliable source. The class label has only two distinct values that are positive or negative, which refers to the feeling or sentiment hidden in the text.

The main part of this solution is predicting user credibility.

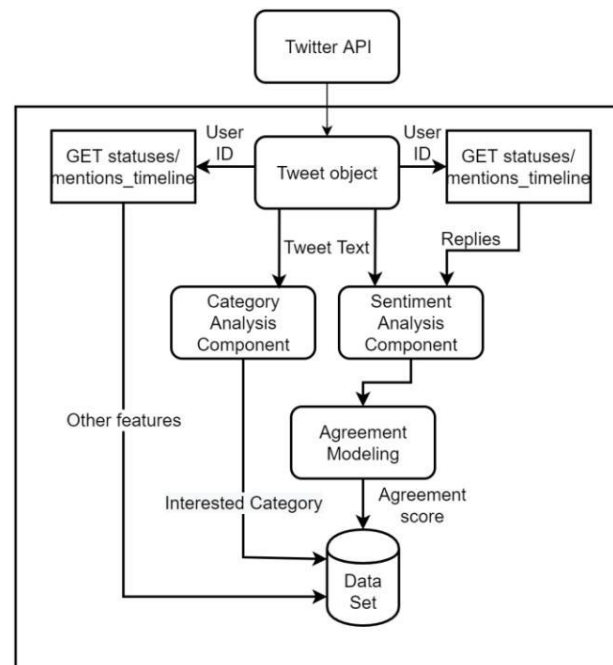


Fig. 4. Flow of the data collection method for credibility analysis component

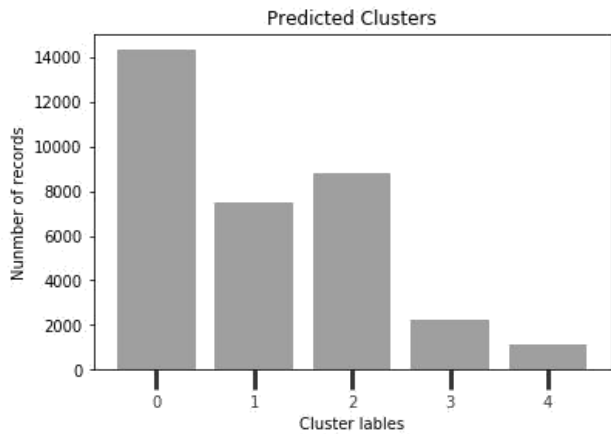


Fig. 5. Predicted labels for the train set

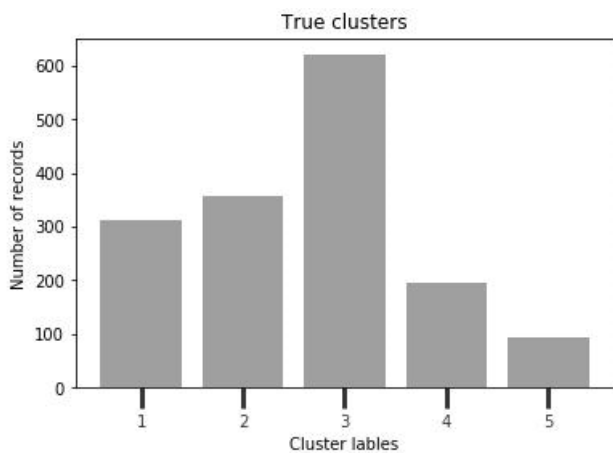


Fig. 6. Manually annotated labels for test set

The data collection process for the credibility model is a tedious task as a new feature has to generate while collecting the data. The flow chart for the data collection process is shown in Fig. 4. Twitter API allows us to get tweets and user profile data up to a certain extent. According to Fig. 4., tweets are collected through the twitter live search API. The tweets object has the details about the source profile. In the meantime, the profile timeline is getting through the API and sent to the category analysis model to identify the most interested category of each user. Meanwhile, Calculating the description length, screen name length and number of days from create date also executed. Profile Data such as status count, a number of followers are separated and saved. At the same time, overall followers' sentiment also calculated in the agreement modelling process.

K-means model is implemented easily by applying the k-means algorithm on the collected dataset. First outliers are removed by analyzing the data set. Encoding, scaling and fitting the dataset to the k-means classifier is needed to repeat several times to match the output of the model with the manually annotated dataset. Predicted labels for the trained dataset and manually annotated labels for the test dataset are illustrated in Fig. 5 and Fig. 6. Results were not satisfactory in the starting experiments. As the optimizations, data scaling and encoding method of the dataset and number of iterates of the K-means algorithm is needed to be changed several times. Minmax scaler improved the accuracy and max iteration value is set to 1000 for better results.

V. EVALUATION

According to the experiments conducted for tweet categorization, the SVM classifier is the best-performed algorithm for the selected data set. Multinomial Naïve Bayes classifier treats the feature as an independent way but SVM considers the interaction between features to some extent. K-fold Cross-validation is applied for each classifier with the K = 10. KNN classifier achieved 0.26, Multinomial naïve Bayes classifier 0.54 and SVM classifier achieved 0.62 accuracy levels. K-fold cross-validation is conducted for the sentiment analysis model as well. The accuracy achiever was 0.77. The general flow for the model evolution for category analysis and sentiment analysis is shown in Fig. 7.

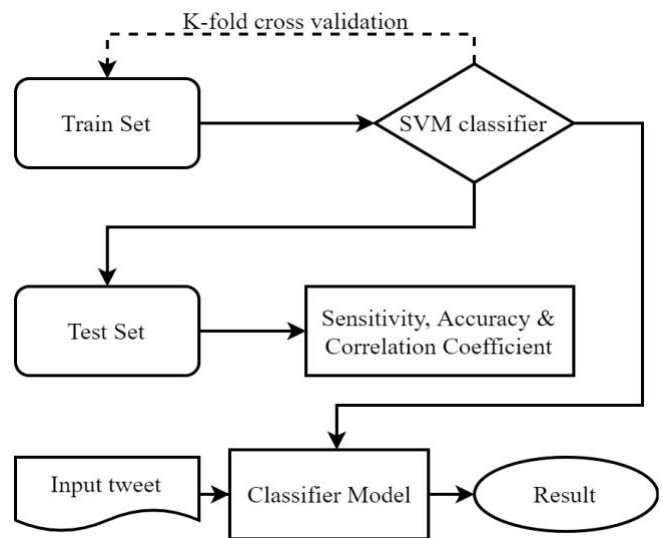


Fig. 7. Evaluation flow for the SVM classifiers

The credibility analysis model is capable of predicting the credibility rating for a user with an accuracy of 0.68. Accuracy was calculated by comparing the manually annotated test set. Precisions for credibility rating for rating 1, 2, 3, 4, 5 are 0.82, 0.61, 0.55, 0.72, 0.70 respectively. The first test set was input to the model and predict the class label. Then the predicted label was compared with the manually annotated class label.

VI. CONCLUSION

As mentioned in the introduction this system is a part of an ongoing research project. The proposed system has to be implemented using mainly Natural Language Processing (NLP) techniques. Research work has to be conducted on text categorization, opinion mining, agreement modeling, and credibility analysis areas. According to the results, support vector machine classifiers have the highest accuracy for both News category analysis model and sentiment analysis. Credibility analysis is a critical task in this system because the data set has to be created using after implementing the above two components. A python program to get and save the live tweet is also created to test the ability to create their own data set.

It was a success but there are some limitations with the twitter API. The main limitation was the number of requests was limited for a specific time period. Sentiment analysis and category analyzing models are evaluated using K-fold cross-validation. For the credibility analysis model credibility

tagging based on a human point of view is required for the evaluation process. It was very rare to find researches for twitter source credibility analysis. This system can be compared with a system proposed by Krzysztof Lorek et al. [14] their solution is a scoring system for each user and mostly based on manually created rules. They predicted the credibility for three categories with a precision of 89%, 87%, and 84% respectively. Our proposed solution It is difficult to compare with the above system as they predicted credibility for three classes. But their accuracy is higher than our proposed system as they predicted for lower number of classes and mostly rely on completely manually annotated data.

REFERENCES

- [1] Castillo, M. Mendoza and B. Poblete, "Information credibility on twitter", 2011.
- [2] Louisa Shu Ying Ha, Mary E. Benedict, Gi Woong Yun and Michael Albert Horning, "Assessing source credibility on social media-An electronic word-of-mouth communication perspective", Etd.ohiolink.edu, 2015.
- [3] R. Guha, R. Kumar, P. Raghavan and A. Tomkins, "Propagation of trust and distrust", 2004. [Proceedings of the 13th international conference on World Wide Web]
- [4] Mengen Chen, Xiaoming Jin and Dou Shen, "Short Text Classification Improved by Learning Multi-Granularity Topics" *IJCAI*, 2011
- [5] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features", *Machine Learning: ECML-98* pp 137-142, 1998.
- [6] Mohd Suhairi, Md Suhaimin, Mohd Hanafi Ahmad, Hijazi Rayner Alfred and Frans Coenen, "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts", *8th International Conference on Information Technology (ICIT)*, 2017
- [7] Miriam J. Andrew J. Flanagan, Keren Eyal, Daisy R. Lemus & Robert M. Mccann, "Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment", *Annals of the International Communication Association*, pp. 293-335, 2016.
- [8] A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", Springer, 2010.
- [9] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, "A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle", *DL.acm.org*, 2012.
- [10] S. M. Vohra and Prof. J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques" Volume – 02, pp. 313-317
- [11] Dwi H. Widiantoro, Yudi Wibisono "Modeling Credibility Assessment and Explanation for Tweets Based On Sentiment Analysis", *Jatit.Org*, 2019.
- [12] M. Abbasi and H. Liu, "Measuring User Credibility in Social Media", Springer, 2013.
- [13] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, Patrick Meier and Carlos Castillo, "TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter", arXiv, 2014.
- [14] L.Krzysztof, J. S Wiciński, M.J Lorek and A. Guptha, "Automated Credibility Assessment On Twitter," *Computer Science* 16 (2) pp. 157–168, 2015.
- [15] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2013, *Aclweb.org*
- [16] L. Rokach and O. Maimon, "Clustering Methods", *Data Mining and Knowledge Discovery Handbook* pp 321-352, 2005
- [17] Matt Thomas, Bo Pang, and Lillian Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts", *f EMNLP 2006*, pp. 327–335.
- [18] Gabriella Schoier, Giuseppe Borruso "A Clustering Method for Large Spatial Databases", Springer, 2004.
- [19] Weizhong Zhao, Huifang Ma, and Qing He, "Parallel K-Means Clustering Based on MapReduce", *CloudCom 2009: Cloud Computing*, pp 674-679