iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** V **Month of publication:** May 2018

DOI: http://doi.org/10.22214/ijraset.2018.5470

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

Stock Market Credible Twitter Users Discovery

Bhavya Babuta[1], K. Abhijit[2], Kusuma.N[3], Lovey[4] Mrs. Vani KA[5]
[1,2,3,4] Undergraduate Students
[5] Assistant Professor Dayananda Sagar Collge Of Engineering Department Of Information Science And Engineering

Abstract: The face of traditional information flow is now changed with social networks. We talk less and tweet, text, blog more. This provides us with an opportunity to discover people with specialties and use their content full data on social network to extract relevant information. This paper presents an approach to discover the credible twitter users in the stock market domain by mining the recent tweets of their followers by applying $cash-tag based filtering and naïve baye's classification methods.

I. INTRODUCTION

Social networks have changed the face of traditional information flow on the global level. People today share their own opinion about everything on the Web. Twitter, is one of the most popular microblogs in the world covering a wide range of active and passive users from ordinary people to experts. Many professionals actively use Twitter such that their profile and timeline are visible to the public. This provides a valuable source of information to learn implement a prediction model for stock market . But before we go ahead one question needs to be answered yet i.e. : how can we find these credible users to follow them in a prediction task? The study presented here is conducted with the main aim of investigating a technique to identify credible Twitter users, who might also be experts in the stock market, according to their social network, in particular, their follower connections. It is hypothesized that expert users follow and might be potentially pursued by relevant and other domain experts and in this case, other stock market experts. The user connections (number of followers) might also be a measure of user expertise in a particular stock market, or in another word, a signal to measure credibility. The merit of the proposed method is that it avoids analyzing each user's time-line. This is a challenging issue since accessing to historical Twitter data is not always an easy task.

II. OVERVIEW

Stock market plays a vital role in the country's economy and growth. There have been a number of efficient ways in predicting stocks. The versatility of this paper is we use twitter to find the users who are stock market experts and such expertise would be willing to share their knowledge on social media such as twitter. Hence, following these credible stock market related users is an efficient way in taking wise decision on stocks.

A. Proposed Approach

A new approach is presented to find these experts based on their follower networks. The approach is as follows:

1) *Step 1:* Extraction of stock market related twitter users information.
2) *Step 2:* From the information gathered, filtering is done on the following three criteria.
   a) Users with large number of followers (more than 1million who belong to news agencies or celebrities) are not considered.
   b) Users who are handled by organizations or companies are ignored.
   c) Users whose last tweets were created before 2015 meaning they are no longer active.
3) *Step 3:* If the above mentioned conditions are met, then the followers information of the selected seed users is extracted(ID, screen name and their last 200 tweets) .
4) *Step 4:* The tweets of these followers were then analyzed by a filter which identified whether these users are related to stock market.

In the final step, the following features are used for classification: number of followers. number of stock market-related followers.
ratio of stock market-related follower to the total number of followers.
the number of seed users tweet.
A Naive-Bayes approach, which is a non-domain-specific standard model, would be implemented to classify users based on their social network and the results will be compared with the score of each user prediction of stock market changes.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

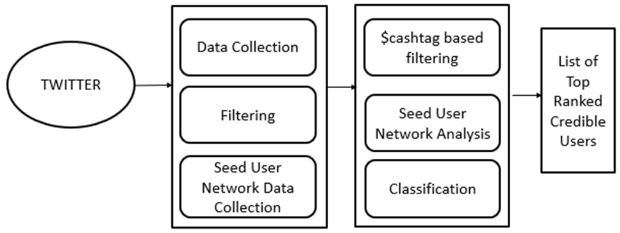

Fig 1:Architecture of Discovering credible users.

B. System Design

Key components of this paper are:

1) Data Collection
2) Filtering and Seed User Network Data Collection
3) Seed User Network Analysis
4) Naïve Bayes Classifier

C. Data Collection

Data collection is the process of gathering information about targeted variables in a systematic fashion. The data thus collected is then used answer relevant questions and evaluate outcomes. Here, we use Tweepy which is open sourced, and hosted on GitHub and enables Python to communicate with Twitter platform and use its API.

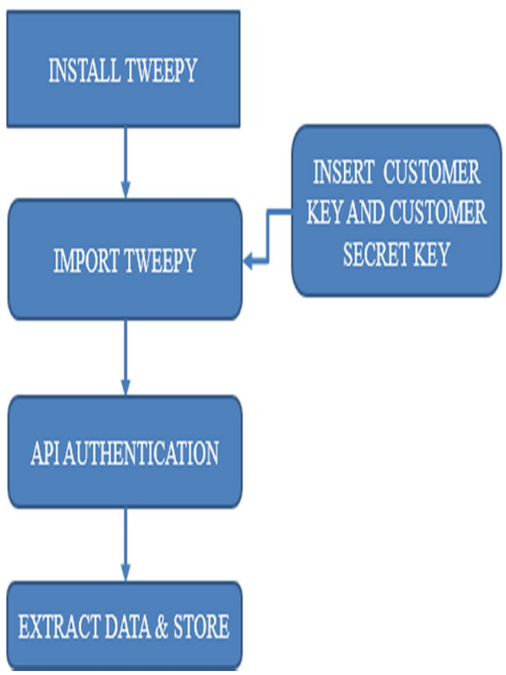

Fig 2: Data collection process.

D. Filtering and Seed User Network Data Collection

From the information gathered, filtering is done only if the above mentioned conditions are met, the followers information of the selected seed users is extracted(ID, screen name and their last 200 tweets) .

E. $Cashtag-Based Filtering

1) Cashtags are stock ticker symbols that are prefixed with a dollar sign.
2) The use of cashtags is a mechanism to denote a financial theme in a tweet.
However, the presence of $Cashtag is not required for a tweet to be considered valid from an investment standpoint.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

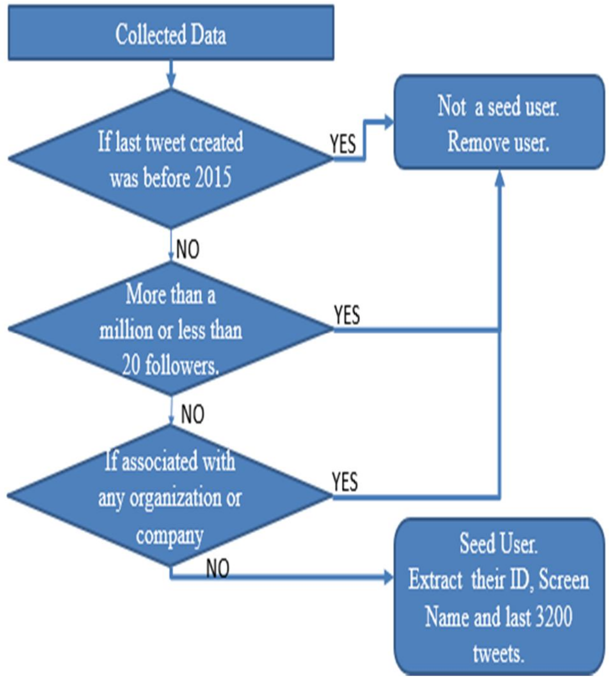

Fig 3: Filtering and Seed User Network Data Collection

F. Seed User Network Analysis

In this step, the following features are used for classification:

1) number of followers.
2) number of stock market-related followers.
3) ratio of stock market-related follower to the total number of followers.
4) the number of seed users tweet.

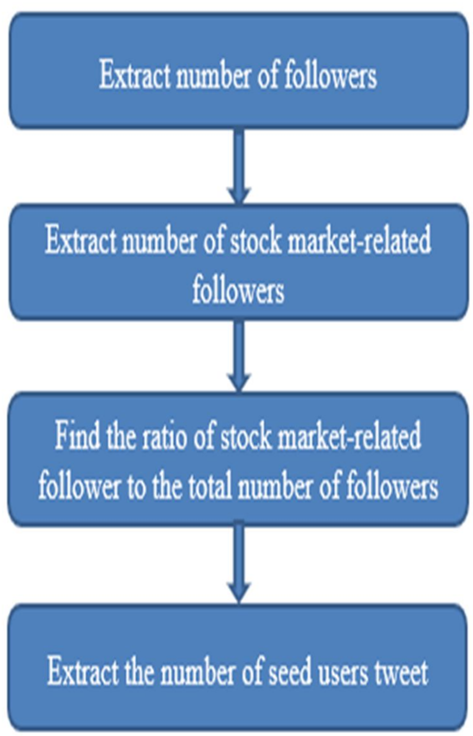

Fig 4: Seed User Network Analysis

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

G. *Naïve Bayes Classifier*

A Naive-Bayes approach, which is a non-domain-specific standard model, would be implemented to classify users based on their social network and the results will be compared with the score of each user prediction of stock market changes. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even in highly sophisticated classification methods. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature i.e. the two predictors are independent.

H. *Pros*

1) It is easy and fast to predict class of test data set also it performs well in multi class prediction.
2) Assuming the predictor independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression.
3) It performs well in case of categorical input variables compared to numerical variable(s).

I. *Cons*

1) If categorical variable has a category (in test data set), that was not observed in the training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction.
2) Since in real life, it is almost impossible that we get a set of predictors which are completely independent the assumption of independent predictors is another limitation of Naive Bayes.

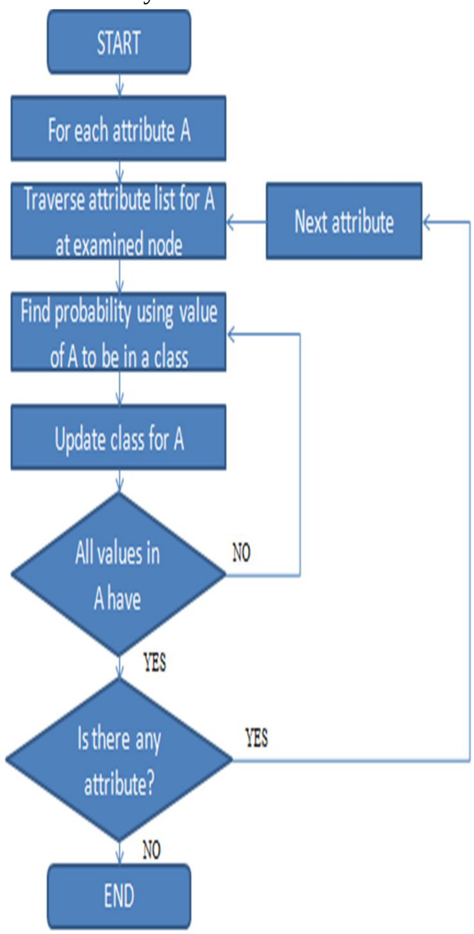

Fig 5: Naïve Bayes Classifier

J. *Data Flow Diagram*

A DFD shows the flow of data through a system. It views a system as a function that transforms the inputs into desired outputs. Any complex systems will not perform this transformation in a single step and a data will typically undergo a series of transformations before it becomes the output.

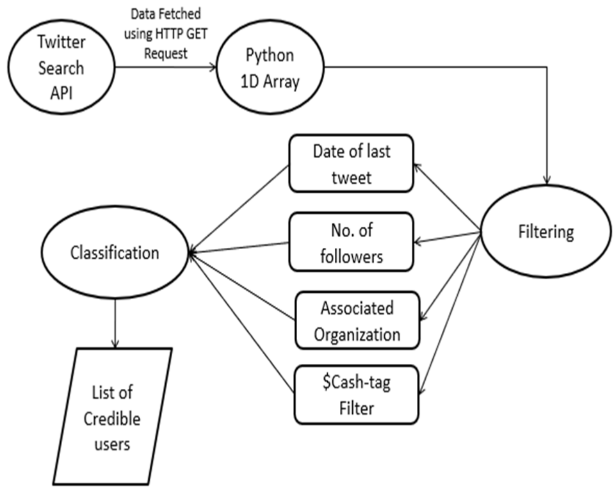

Fig 6: Data flow diagram

K. Use-Case DIAGRAM

The external objects that interact directly with the system are called ***actors.*** The important thing about actors is that they are not under control of the application. In this project, user of the system is the actor.

L. Sequence Diagram

A sequence diagram in a Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It shows the participants in an interaction and the sequence of messages among them; each participant is assigned a column in a table.

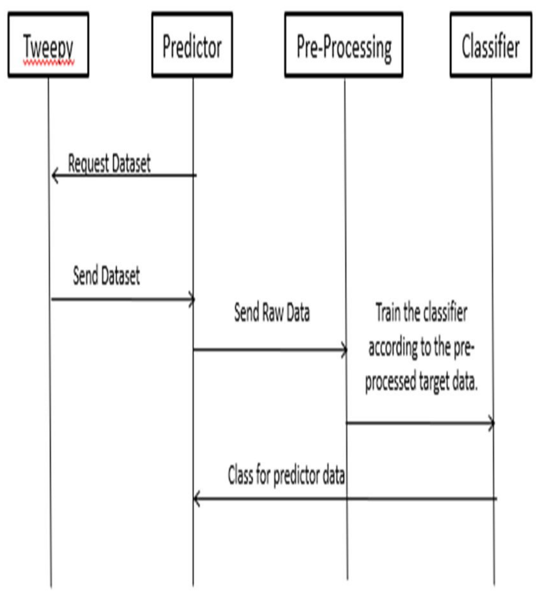

Fig 7: Sequence diagram.

III. RESULTS

We train the dataset containing the twitter stock market related users for credibility analysis. Then, enter the user on whom the credibility is to be found, the number of user's followers is displayed. Also, the ratio of stock market-related followers to the total number of followers is displayed. If the user entered is a credible stock market related user, the result shows 1.0 else it shows 0.0

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

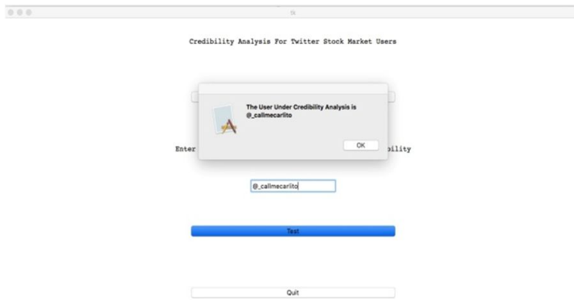

Fig 8: testing for twitter user

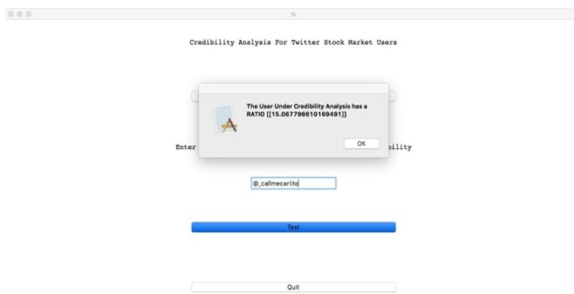

Fig 9: Ratio of stock market-related followers to the total number of followers.

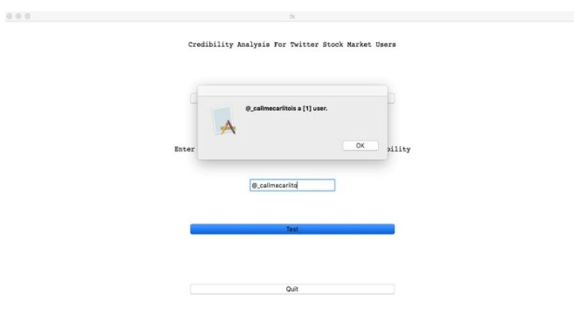

Fig 10: Result for credible user

IV. CONCLUSION AND FUTURE WORK

The proposed approach identifies that, in order to find the most credible expert users in the stock market field as well as other domains of interest, analyzing profiles and social networks, does not need to address natural language processing challenges. This method is more reliable and less complex than analyzing users' posts and comparing them with stock market changes. From the presented results, a correlation between user credibility and follower networks was observed, confirming the proposed hypothesis of a direct relationship between user social networks and their credibility. Explore more data sources like spinn3r, twitter fire hose etc. to gather more data than what is available with the twitter API. Implement a multi-threaded GUI to enable parallel testing for each trained data in the dataset. Implement a section to find the individuals who are the highest influencers of your related stock. Gather data from more sources like facebook, google +,linkedin, etc.

REFERENCES

[1] Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users." ICWSM, vol. 11, p. 5th, 2011.
[2] E. D. Brown, "Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market," Proc. of SAIS, 2012.
[3] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen, "Mining expertise and interests from social media," in Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013, pp.
[4] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012
[5] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, "Finding news curators in twitter," in Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.
[6] P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly, and K. P. Gummadi, "Deep twitter diving: Exploring topical groups in microblogs at scale," in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014

[7] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Twitter improves seasonal influenza prediction." in HEALTHINF, 2012
[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1,2011.
[9] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using twitter content to predict psychopathy," in Machine Learning and Applications (ICMLA) 2012.
[10] Tweepy Documentation [Online]. Available: [https://www. http://docs.tweepy.org/]