ANALYSIS OF REAL TIME STIMULI IN WORLDWIDE TWEETS

R.Lakhmi [1], A.K.Velmurugan[2], sowmya.R [3], Varnikha sree S[4]

Department of Information Technology, MNM Jain Engineering College, Thoraipakkam, Chennai, India

Email:lakshmi.ravi90@yahoo.com

ABSTRACT

In recent years, the increase of interest in using social media as a source has involved in day to day communications and researchers have looked into the development of tools for real-time trend analytics or early detection of newsworthy events as well as the analytical approaches for understanding the sentiment expressed by users towards a public opinion on a specific topic and to identify the disease in specific origin.Conversations on twitter are used to explore as indicators within early warning system to alert the spreading of disease in specific origin. Twitter users have the option to disclose their city-location which would normally be their primary residence. Our objective is to locate the origin of the tweets since the ability to accurately infer the location of affected users and contact with the health care unit to verify whether it is fake or not. This can save lives and helps in crisis management.

KEYWORDS: Twitter, Naive Bayes Classifier, Supervised Classification

I.INTRODUCTION

Social media are increasingly being used in the scientific community as a key source of data to help understand diverse natural and social phenomena, and this has prompted the development of a wide range of computational data mining tools that can extract knowledge from social media for both post-hoc and real time analysis. Thanks to the availability of a public API that enables the cost-free collection of a significant amount of data, Twitter has become a leading data source for such studies. Having Twitter as a new kind of data source, researchers have looked into the development of tools for real-time trend analytics or early detection of newsworthy events, as well as into analytical approaches for understanding the sentiment expressed by users towards a target or public opinion on a specific topic. What motivates the present study is the increasing interest in inferring the geographical location of either tweets or Twitter users in order to identify the disease spreading in their region . The automated inference of tweet location has been studied for different purposes, ranging from data journalism, to public health. As well as numerous different techniques, researchers have relied on different settings and pursued different objectives when conducting experiments

While the approaches summarized in previous work will work well for certain applications, retrieving the tweet history for each user or the profile information of all of a user's followers and followers is not feasible in a real-time scenario. Hence, in this context, a classifier needs to deal with the additional challenge of having to rely only on the information that can be extracted from a tweet of single user. Here, Naive Bayes Classifier and Support Vector Machines are built for each input; hence it comes under the category of supervised classification for the classification of disease spreading in an area.

II.RELATED WORK

O. Ajao et al.[1]proposed that Conversations on Twitter are now being explored as indicators within early warning systems to alert of imminent natural disasters such earthquakes and aid prompt emergency responses to crime. They surveyed a range of techniques applied to infer

the location of Twitter users from inception to state-of-the-art. They found significant improvements over time in the granularity levels and better accuracy with results driven by refinements to algorithms and inclusion of more spatial features.

Hau-Wen Chang et al.[4] studied the problem of predicting home locations of Twitter users using contents of their tweet messages. Using three probability models for locations, they compared both the Gaussian Mixture Model (GMM) and the Maximum Likelihood Estimation (MLE).They proposed two novel unsupervised methods based on the notions of Non-Localness and Geometric-Localness to prune noisy data from tweet messages.

H. Bo, P. Cook et al. [2]proposed an algorithm to predict the location of Twitter users and tweets using a multinomial Naive Bayes classifier trained on Location Indicative Words and various textual features (such as city/country names, #hashtags and @mentions). We compared our approach against various baselines based on Location Indicative Words, city/country names, #hash tags and @mentions as individual feature sets, and experimental results show that our approach outperforms these baselines in terms of classification accuracy, mean and median error distance.

J. Bollen et al.[5] performed a sentiment analysis of all tweets published on the microblogging platform Twitter in the second half of 2008. they used a psychometric instrument to extract six mood states (tension, depression, anger, vigor,fatigue, confusion) from the aggregated Twitter content and compute a six-dimensional mood vector foreach day in the timeline.

F. Atefeh et a[6]l. provides a survey of techniques for event detection from Twitter streams. These techniques aim at finding real-world occurrences that
unfold over space and time. Event detection techniques presented in literature address these issues by adapting techniques from various fields to the uniqueness of Twitter. This article classifies these techniques according to the event type, detection task, and detection method and discusses commonly used features. Finally, it highlights the need for public benchmarks to evaluate the performance of different detection approaches and various features.

III ARCHITECTURE

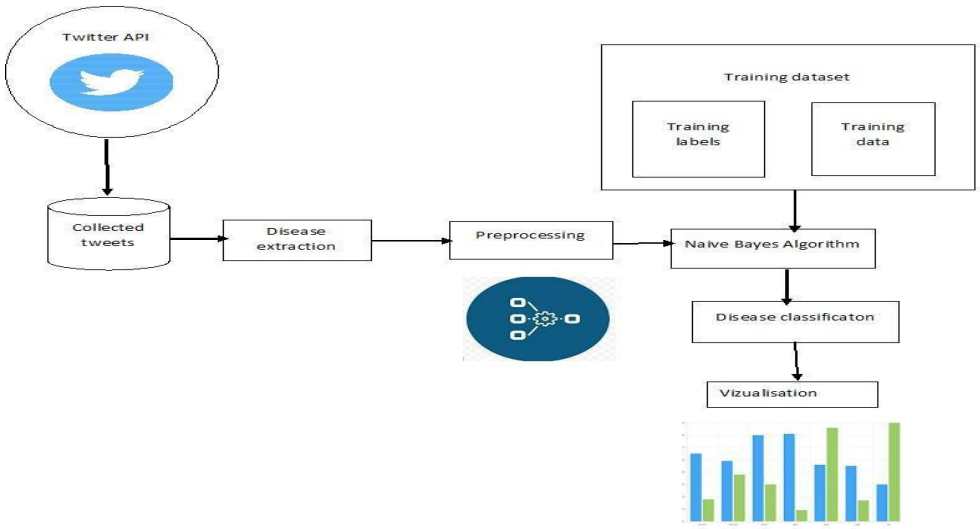

FIGURE 4.2

ISSN NO: 1524-2560
ISSN No: 1945-5534

In fig 4.2, the data's are collected from twitter API, pre-processed in which Naïve Bayes Algorithm is applied to classify the disease type. Here we have trained dataset which include training label and data based on that disease are classified and we can visualize the disease spreading across the country in particular location

IV ALGORITHM

4.2.1.NAIVE BAYES

Naive Bayes is a family of algorithms based on applying Bayes theorem with a strong(naive) assumption, that every feature is independent of the others, in order to predict the category of a given sample. They are probabilistic classifiers, therefore will calculate the probability of each category using Bayes theorem, and the category with the highest probability will be output. we need to convert the probability that we wish to calculate into a form that can be calculated using word frequencies. Here, we adopt the properties of possibilities and Bayes' Theorem to do the conversion. Bayes' Theorem is useful for dealing with conditional probabilities, since it provides a way for us to reverse them.

P(A|B) is the posterior probability of class (A, target) given predictor (B, attributes).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- P(A) is the prior probability of class.
- P(B|A) is the likelihood which is the probability of predictor given class.
- P(B) is the prior probability of predictor

4.2.2 MULTINOMIAL NAIVE BAYES

The multinomial Naive Bayes classifier has been frequently used for various text classification tasks such as sentiment analysis and news categorization.

APPLYMULTINOMINALNB(ℂ,V,prior,condprob,d)
1.W← EXTRACTTOKENSFROMDOC(V, d)
2. for each c ∈ ℂ
3. do score[c] ← log prior[c]
4. for each t ∈ W
5.do score[c] += log condprob[t][c]
return $\text{argmax}_{c \in C}$ score[c]

RULE BASED

We defined a set of rule to classify a tweet based on term frequency. First we extract the features of a tweet and count the term frequency of each feature, the feature having maximum term frequency from all categories. As it cannot be right all time so we maintain the count of categories in which tweet falls, category which is near to tweet will be the next classification.

TRAINMULTINOMINALNB(ℂ, 𝔻)

1. V ← EXTRACTVOCABULARY(𝔻)
2. N ← COUNTDOCS(𝔻)
3. for each c ∈ ℂ
4. do N_c← COUNTDOCSINCLASS(𝔻,c)
5. prior[c] ← N_c /N
6. $text_c$ ←CONCATENATETEXTOFALLDOCSINCLASS(𝔻,c)
7. for each t ∈ V
8. do T_{ct}← COUNTTOKENSOFTERM($text_c$,t)

9. for each t ∈ V
10. do condprob[t][c] ← $\frac{T_{ct}+1}{\sum t^{l}\,(T_{ctl}+1)}$
11. return V,prior,condprob

4.2.2MULTINOMIAL NAIVE BAYES

The multinomial Naive Bayes classifier has been frequently used for various text classification tasks such as sentiment analysis and news categorization.

APPLYMULTINOMINALNB(ℂ,V,prior,condprob,d)
1.W← EXTRACTTOKENSFROMDOC(V, d)
2. for each c ∈ ℂ
3. do score[c] ← log prior[c]
4. for each t ∈ W
1. do score[c] += log condprob[t][c]
return $\text{argmax}_{c} \in_{C}$ score[c]

4.2.3.RULE BASED

We defined a set of rule to classify a tweet based on term frequency. First we extract the

features of a tweet and count the term frequency of each feature, the feature having maximum term frequency from all categories. As it cannot be right all time so we maintain the count of categories in which tweet falls, category which is near to tweet will be the next classification.

TRAINMULTINOMINALNB(ℂ, 𝔻)

12. V ← EXTRACTVOCABULARY(𝔻)
13. N ← COUNTDOCS(𝔻)
14. for each c ∈ ℂ
15. do N_c← COUNTDOCSINCLASS(𝔻,c)
16. prior[c] ← N_c /N
17. $text_c$ ←CONCATENATETEXTOFALLDOCSINCLASS(𝔻,c)
18. for each t ∈ V
19. do T_{ct}← COUNTTOKENSOFTERM($text_c$,t)
20. for each t ∈ V
21. do condprob[t][c] ← $\frac{T_{ct}+1}{\sum t^{l}\,(T_{ctl}+1)}$
22. return V,prior,condprob

4.2.4 MODULES

(a) Login/Registration
(b) Classification
(c) Search Disease
(d) Verification
(e) Notification

4.2.4.1 LOGIN/REGISTRATION

Here, the signup page consists of user name, Password etc. Those details should be stored in database. Login screen contains username and password when the user/admin login it should retrieve the data from database and compare it based on the user input, if it matches

the username and password user/admin will get into his/her homepage and can post tweets otherwise an alert or error message will be shown to the user/admin.

4.2.4.2 CLASSIFICATION

Naive Bayes algorithm which is built on training corpora while containing the correct label for each input, for the classification of disease spreading in an area. The information can be extracted from a tweet of single user and various diseases can be classified and visualized through bar chart then we can identify the location of affected users through search module.

4.2.4.3 SEARCH

After checking the priority of disease, the admin will search the disease having high probability of occurrence. All the tweets related to that disease with tweets origin or location will be retrieved by analyzing the users history without getting any privacy information hence we got the location of affected users

4.2.4.4 VERIFICATION

In this module, the searched result from admin that is the location of affected users will be verified by comparing the results with the actual records maintained by HDO.HDO will give the result back to the admin stating that whether the information is true or not. Hence verification is done by HDO.

4.2.4.5 NOTIFICATION

In this module, after getting the verified information from HDO the admin will send notification to all the twitter users alerting them to be aware of the spreading of disease in that area if that is a correct information else an notification message will be sent to all the users like don't believe them to avoid unnecessary public concerns.

5. CONCLUSION

Twitter is a major social networking service with over 200 million tweets made everyday. Twitter provides a list of trending topics in real time, but it is often hard to understand what trending topics are about. It is important and necessary to classify these topics into general categories with high accuracy for better information retrieval.This Serves as an early warning systems to alert the spreading of disease in specific origin. We can verify with the Health Care Unit to check whether the tweets are fake or not so it will create awareness among the people whether the information about the disease is true or not.Hence the ability to accurately infer the location of affected users can save lives and helps in crisis management.

REFERENCE

1. O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on twitter. Journal of Information Science, 1:1–10,2015.
2. H. Bo, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In Proceedings of COLING, pages 1045–1062, 2012.
3. Xiang.Ji,Soon Ae Chun and James Geller.Knowledge based tweet classification for Disease sentiment monitoring.

4. H.-w. Chang, D. Lee, M. Eltaher, and J. Lee. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In Proceedings of ASONAM, pages 111–118, 2012.
5. J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Proceedings of ICWSM, pages 450–453, 2011.
6. F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. Computational Intelligence, 31(1):132–164, 2015.
7. M. Dredze, M. J. Paul, S. Bergsma, and H. Tran. Carmen: A twitter geolocation system with applications to public health. In HIAI Workshop, pages 20–24, 2013.
8. B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research, pages 451–500, 2014.
9. A. Zubiaga, D. Spina, R. Mart´ınez, and V. Fresno. Real-time classification of twitter trends. JASIST, 66(3):462–473, 2015.
10. A. Zubiaga, I. San Vicente, P. Gamallo, J. R. Pichel, I. Alegria, N. Aranberri, A. Ezeiza, and V. Fresno. Tweetlid: a benchmark for tweet language identification. Language Resources and Evaluation, 50(4):729–766, 2016.

ISSN NO: 1524-2560
ISSN No: 1945-5534