International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

TRENDS OF DIFFERRNT CLUSTERS ARE EFFECTIVELY MINED USING MACHINE LEARNING APPROACHES

AyubBaig[1],Raya Swathi[2],T.Shivani[3],ValasaniGirija[4]

[1]Assistant professor, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad, Telangana, India

[2, 3, 4]UG Scholar, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad, Telangana, India

ABSTRACT:

One of the major sources of trending news, events and opinion in the current age is micro blogging. Twitter, being one of them, is extensively used to mine data about public responses and event updates. This paper intends to propose methods to filter tweets to obtain the most accurately descriptive tweets, which communicates the content of the trend. It also potentially ranks the tweets according to relevance. The principle behind the ranking mechanism would be the assumed tendencies in the natural language used by the users. The mapping frequencies of occurrence of words and related hash tags is used to create a weighted score for each tweet in the sample space obtained from twitter on a particular trend.

Keywords:Twitter, filter, trend, mining.

1. INTRODUCTION:

The exponential development of computer science and technology provides us with one of the greatest innovations of the "Internet" of the 21st century, where one person can communicate to another worldwide with the help of a mere smartphone and internet.In the initial days of the internet, people used to communicate with each other through Email only and it was filled with spam emails. In those days, it was a big task to classify the emails as positive or negative i.e. spam or not - spam. As time flows, communication, and flow of data over the internet got changed drastically, especially after the appearance of social media sites. With the advancement of social media, it becomes highly important to classify the content into positive and negative terms, to prevent any form of harm to society and to control antisocial behavior of people. In recent times there have many instances

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

where authorities arrest people due to their harmful and toxic social media contents[1]. For example, one 28-year-old man was arrested in Bengal for posting an abusive comment against Mamata Banerjee on Facebook and one man from Indonesiawas arrested for insulting the police of Indonesia on Facebook. Thus, there is an alarming situation and it is the need of the hour to detect such content before they got published because these negative contents are creating the internet an unsafe place and affecting people adversely. Suppose there is a comment on social media “Nonsense? Kiss off, geek. What I said is true”, it can be easily identified that the words like Nonsense and Kiss off are negative and thus this comment is toxic. But to mine the toxicity technically this comment needs to go through a particular procedure and then classification technique will be applied on it to verify the precision of the obtained result.Different machine learning algorithms will be used in the classification of toxic comments on the Data set of Kaggle.com. This paper includes six machine learning techniques i.e. logistic regression, random forest, SVM classifier, naive bayes, decision tree, and KNN classification to solve the problem of text classification. So, we will apply all the six machine learning algorithms on the given data set and calculate and compare their accuracy, log loss, and hamming loss.

2. LITERATURE SURVEY:

1. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.

AUTHORS: boyd, danah, Scott Golder, and GiladLotan.

Twitter - a microblogging service that enables users to post messages ("tweets") of up to 140 characters - supports a variety of communicative practices; participants use Twitter to converse with individuals, groups, and the public at large, so when conversations emerge, they are often experienced by broader audiences than just the interlocutors. This paper examines the practice of retweeting as a way by which participants can be "in a conversation." While retweeting has become a convention inside Twitter, participants retweet using different styles and for diverse reasons. We highlight how authorship, attribution, and communicative fidelity are negotiated in diverse ways. Using a series of case studies and empirical data, this paper maps out retweeting as a conversational practice.

2 What people study when they study Twitter: Classifying Twitter related academic papers

AUTHORS:irley Ann Williams, Melissa Terras, Claire Warwick

Purpose - Since its introduction in 2006, messages posted to the microblogging system

Twitter have provided a rich dataset for researchers, leading to the publication of over a thousand academic papers. This paper aims to identify this published work and to classify it in order to understand Twitter based research. Design/methodology/approach - Firstly the papers on Twitter were identified. Secondly, following a review of the literature, a classification of the dimensions of microblogging research was established. Thirdly, papers were qualitatively classified using open coded content analysis, based on the paper's title and abstract, in order to analyze method, subject, and approach. Findings - The majority of published work relating to Twitter concentrates on aspects of the messages sent and details of the users. A variety of methodological approaches is used across a range of identified domains. Research limitations/implications - This work reviewed the abstracts of all papers available via database search on the term "Twitter" and this has two major implications: the full papers are not considered and so works may be misclassified if their abstract is not clear; publications not indexed by the databases, such as book chapters, are not included. The study is focussed on microblogging, the applicability of the approach to other media is not considered. Originality/value - To date there has not been an overarching study to look at the methods and purpose of those using Twitter as a research subject. The paper's major contribution is to scope out papers published on Twitter until the close of 2011. The classification derived here will provide a framework within which researchers studying Twitter related topics will be able to position and ground their work.

3 User interactions in social networks and their implications.

AUTHORS: C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao

Social networks are popular platforms for interaction, communication and collaboration between friends. Researchers have recently proposed an emerging class of applications that leverage relationships from social networks to improve security and performance in applications such as email, web browsing and overlay routing. While these applications often cite social network connectivity statistics to support their designs, researchers in psychology and sociology have repeatedly cast doubt on the practice of inferring meaningful relationships from social network connections alone. This leads to the question: Are social links valid indicators of real user interaction? If not, then how can we quantify these factors to form a more accurate model for evaluating socially-enhanced applications? In this

paper, we address this question through a detailed study of user interactions in the Facebook social network. We propose the use of interaction graphs to impart meaning to online social links by quantifying user interactions. We analyze interaction graphs derived from Facebook user traces and show that they exhibit significantly lower levels of the "small-world" properties shown in their social graph counterparts. This means that these graphs have fewer "supernodes" with extremely high degree, and overall network diameter increases significantly as a result. To quantify the impact of our observations, we use both types of graphs to validate two well-known social-based applications (RE and SybilGuard). The results reveal new insights into both systems, and confirm our hypothesis that studies of social applications should use real indicators of user interactions in lieu of social graphs.

2.4 RT^2M: Real-Time Twitter Trend Mining System

AUTHORS: Min Song,MeenChul Kim

The advent of social media is changing the existing information behavior by letting users access to real-time online information channels without the constraints of time and space. It also generates a huge amount of data worth discovering novel knowledge. Social media, therefore, has created an enormous challenge for scientists trying to keep pace with developments in their field. Most of the previous studies have adopted broad-brush approaches which tend to result in providing limited analysis. To handle these problems properly, we introduce our real-time Twitter trend mining system, RT ^{2} M, which operates in real-time to process big stream datasets available on Twitter. The system offers the functions of term co-occurrence retrieval, visualization of Twitter users by query, similarity calculation between two users, Topic Modeling to keep track of changes of topical trend, and analysis on mention-based user networks. We also demonstrate an empirical study on 2012 Korean presidential election. The case study reveals Twitter could be a useful source to detect and predict the advent and changes of social issues, and analysis of mention-based user networks could show different aspects of user behaviors.

5 Opinion mining and trend analysis on twitter data

AUTHORS:AvneeshJha, Ajay Singh Chahar, Abhishek Singh Chauhan

ABSTRACT: With the rise in internet users across the globe, there has been tremendous increase in the data available online. People use

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

various social media apps and web platform to express their views and opinion regarding every possible aspects of their lives from politics to entertainment, Sports to economics etc. The project "Trend Analysis and Opinion Minion" is developed with the motive to gather all public opinions from twitter and analyze the current trends which may be helpful in determining marketing strategy, certain campaigning and spreading awareness. This can also be helpful in sensing cyber-bullying activities online.

3. METHODOLOGY

It is assumed that the tweets in themselves are not highly descriptive. One of the constraints that Twitter allows a maximum of 140 characters, alphanumeric and special characters that can be used in a tweet. This constraint makes tweets be in a telegram format where key words are prioritized over grammatical correctness. The scope of the methods proposed in this paper extends to assign scores to the tweets taken as sample space so as to be able to rank them as highest descriptive with the highest score. This can be used to accustom a new user to get acquainted with the trend content.

PROBLEM DEFINITION

Some topic will trend at some point in the future and others will not. We wish to predict which topics will trend. And apply algorithm to find out what public opinion about that topic which use to predict mood. Trend analysis and based on that predicting public opinions. It plays important role, many researcher working on automatic technique of extraction and analysis of huge amount of twitter data.

OBJECTIVE OF PROJECT

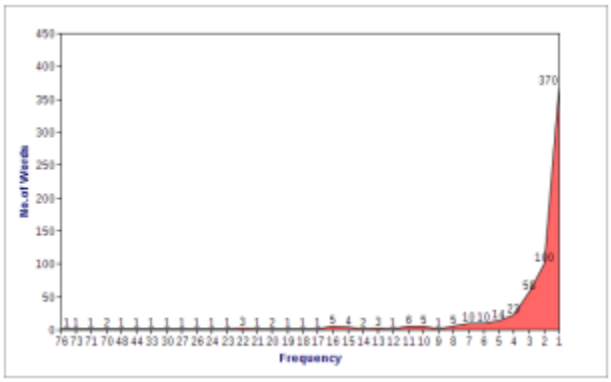
Twitter is most popular social media that allows its user to spread and share information. It Monitors their user postings and detect most discussed topic of the movement. They publish these topics on the list called "Trending Topics". It show what is happening in the world and what people's opinions are about it. For that it uses top 10 trending topic list. Some topic will trend at some point in the future and others will not. We wish to predict which topics will trend.

Working:

The algorithm takes as input a sample space of 'n' predefined number of tweets. It also takes the highest trending 'x' number of trends. The output of running the algorithm is the tweets of the sample space ranked with a deceasing description index. The algorithm also uses two dictionaries. The first dictionary contains the list of the words, which have less significance to the content description and are more grammatical

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

tools, namely articles, prepositions and conjunctions. The second dictionary consists of all common nouns, adjectives, adverbs, verbs and their derivatives. The former will be called 'filter' and the latter 'cnfilter' hereon. The used sample space is placed in a file, separated by an end of tweet character, like '%%'. Once the tweets are acquired, the frequency of every word that is used in the file containing the tweet sample space is found. This would exclude the '#' tags and the '@' tags. The URLs in the tweets are also ignored while finding the frequencies. Hence the list of words and their corresponding frequencies is prepared and stored. It is now to check for association of the highest trending tweet with the other high trending tweets. Tweets about the same event, or person, hold useful content and can be assumed to contain more relevant data. The tweets with a high trending hashtag along with the highest trending hashtag for the second time are used to collect the frequency so as to update the previously generated frequency table. Once the frequency list is obtained we perform a rating on the words to find its weighted score. This weighted score is used to get the cumulative score of each tweet, which can be used to rank the tweets according to its content relevance.



CONCLUSION

Considering the results generated by the algorithms produced by this study, the Tweets can successfully be used to describe the content of the trends. Since the ranking algorithms logically hold true for most of the tweet types, it is still not a good way to detect spam or RT, ReTweets. For implementing those features Image and URL checking algorithms needs to be in place. Also these algorithms do not take into considerations the 'favorite' attribute that is associated with each tweet. A 'favorite' attribute is a counter which is incremented whenever a user up votes a particular tweet.

Future Enhancement

In the current implementation we do not consider the length of the tweet. As per restrictions, a tweet may not exceed 140 characters, but when we run the proposed algorithms we do not consider whether the tweet is using the maximum available character

space.For the target high frequency words to contribute more towards the score of a Tweet its frequency should ideally be substantially more than the more common words, whose contribution should be insignificant.

REFERANCES

1.D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In Proceedings of the ACM 2009 international conference on Supporting group work. ACM, 2009.

2. boyd, danah, Scott Golder, and GiladLotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." HICSS-43. IEEE: Kauai, HI, January 6

3. Shirley Ann Williams, Melissa Terras, Claire Warwick (2013). "What people study when they study Twitter: Classifying Twitter related academic papers". Journal of Documentation, 69 (3).

4. Twitter Search API. http://apiwiki.twitter.com/TwitterAPI-Documentation.

5. R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. Proc. of the National Academy of Sciences, 105(41):15649–15653, 2008.

6. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proc. of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

7. C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In Proc. of the 4th ACM European conference on Computer systems. ACM, 2009.

8. M. E. J. Newman and J. Park. Why social networks are different from other types of networks. Phys. Rev. E, 68(3):036122, Sep 2003.

9. Jin O., Liu N.N., Zhao K., Yu Y., Yang Q. Transferring topical knowledge from auxiliary long texts for short text clustering 2011 International Conference on Information and Knowledge Management, Proceedings