Semantic-based Follower Recommendations on Twitter Network

Niranjan G[1], Praveen Kumar T[2], Suresh Kumar T[3], Dr. N. Saravanan [4]

[1,2,3]Student, Department of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchengode-637215, Tamilnadu, India, Email: niranjangopalan948@gmailcom[1], praveentp8807@gmail.com[1], sureshv0071@gmail.com[3]

[4]Associate Professor, Department of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchengode-637215, Tamilnadu, India, Email: n.saran@gmail.com

ABSTRACT-Analysis and size in large networks is very challenging assignment for community managers. Cheating Bandwidth performs a Twitter is an interesting platform for the dissemination of news. The real-time nature and reference of the tweets are conducive to sharing of data associated with important events as they unfold. One of the greatest challenges is to find the tweets that we can characterize as news in the block of tweets. In this paper, we proposed a method for detecting and tracking breaking news from Twitter in real-time. We filter the stream of incoming tweets to get rid of junk tweets employing a text classification algorithm. Then, we rank the news using a dynamic scoring system which also allows us to track the news over a period.

Keywords-IBK, SVM, tweets classification, secure and insecure data, data visualization, encrypted data.

1. INTRODUCTION

The real-time and shortness of the tweets encourages the user to communicate real-time events using the least amount of text used Twitter for the early detection of earthquakes in the path of sending words about them before they hit. In fact, thanks to this real-time nature, Twitter are often used as a sensor to collect up-to-date information about the state of the planet. The goal of this paper is to style a system to be used for detecting and tracking breaking news in real-time on Twitter.

The paper proposes an approach to detect and track the breaking news in the presence ofsome data streams without relying on traditional style news publishers. We evaluate different algorithms that classify tweets as either news or junk. We also show how a traditional density-based clustering algorithm can be used for detecting clusters in a streamof streaming data. We also propose a singular technique to parallelize the classification of tweets using RabbitMQ. Finally, the paper also proposes a novel dynamic scoring system for ranking and tracking news.

CLASSIFICATION OF TWEETS

Most of users share opinions on various topics using micro-blogging every day. For such reason, it becomes a rich source for sentiment analysis and data mining. The aim of this paper is to identify and create such a functional classifier which can correctly and automatically classify the sentiment of an unknown tweet. This project introduce two methods as sentiment classification algorithm (SCA) based on k-nearest neighbor (KNN) and the other one is based on support vector machine (SVM). This project also evaluate their performance based on real tweets. These days, social networks, blogs, and other media produce a huge amount of data on the World Wide Web.

For both SCA and SVM this project calculate weights based on different features. Then in SCA, this project build a pair of tweets by using different features. From that pair, this project measure the Euclidian distance for every tweet with its counterpart. From those distance this project only consider nearest eight tweets label to classify that tweet. On the other hand in SVM, build a matrix from the calculated weights based on different features and by applying PCA (principal component analysis), this project try to find k eigenvector with the largest Eigen values. From this transformed sample dataset this project try to find the best c and best gamma by using grid search technique to use in SVM. Finally, this project apply SVM to assign the sentiment label of each tweet in the test dataset. In both algorithms, this project use confusion matrix to calculate the accuracy. Later, this project compare our two techniques in respect to an accuracy level of detecting the sentiment accurately. This project found that Sentiment Classifier Algorithm (SCA) performs better than SVM.

2. METHODOLOGY

2.1 REAL-WORLD EVENT IDENTIFICATION IN TWITTER

The work proposes User-contributed messages on social media sites such as Twitter have emerged as powerful, real-time means of information sharing on the Web. They tend to reflect a varieties of events in the real time and makes twitter particularly well suited as a source of real-time event content. Our approach relies on an upscale family of aggregate statistics of topically similar message clusters. Large-scale experiments over many Twitter messages show the effectiveness of our approach and for surfacing real-world event contents on Twitter Social media sites have emerged as powerful means of communication for people looking to share and exchange information on a good sort of real-world events. Twitter messages reflect useful event information for a spread of events of various types and scale. These event messages can provide a set of unique perspectives, regardless of the event, reflecting the points of view of users who are interested or participate in an event.

The users often posts varieties of messages in anticipation of the event. Identifying events in real time on Twitter may be a challenging problem, thanks to the heterogeneity and immense scale of the info. The Twitter users post messages with a variety of content types, including personal updates and various bits of information .While most of the content on Twitter is not related to any real-world event, messages, by design, contain little textual information, and sometimes exhibit inferiority.

2.2 SUPPORT VECTOR MACHINE

2.2.1TEXT CATEGORIZATION WITH SUPPORT VECTOR MACHINES

The utilization of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyses and checks the particular properties of learning with text data and identifies why SVMs are appropriate for this task and performance. Empirical results support the theoretical findings. SVM technology achieve substantial improvements over the currently best performing methods and behave robustly over a spread of various learning tasks. Further, they are fully automatic, eliminating the need for manual parameter tuning. With the rapid climb of online information, text categorization has become one among the key techniques for handling and organizing text data. Text categorization techniques are used to classify news stories, to find interesting information on the web, and to guide a user's search through HTTP.

The building text classifiers by hand is hard and time-consuming. It is advantageous to learn classifiers from examples. They are often founded in terms of computational learning theory and very open to theoretical understanding and analysis. After reviewing the standard feature vector representation of text, we will identify the particular properties of text in this representation. We will argue that SVMs are very well suited for learning in this setting. The empirical results in will support this claim. Compared to others like state method, SVM show substantial performance gains. Moreover, in contrast to conventional text classification methods SVM will prove to be highly robust, eliminating the need for expensive parameter tuning.

2.3 A COMPARISON OF EVENT MODELS FOR NAIVE BAYES TEXTCLASSIFICATION

In this paper, Andrew McCallum [2012] theory proposed recent approaches to text classification have used two different first-order probabilistic models for classification, both of which make the naive Bayes assumption. Most use a multi-variate Bernoulli model, that is, a Bayesian Network with no dependencies between words and binary word features (Sahami 1997). Other devices uses a multi model, that is, a unique language model with integer word counts (e.g. Lewis and Gale 1994; Mitchell 1997). This topic clarifies the confusion by showing the differences and details of these two models, and by comparing their classification performance on five text corpora. This paper finds that the Bernoulli performs with less vocabulary sizes, but the multi-nominal performs usually performs even better at larger vocabulary sizes, providing on average of 27% reduction in mistakes over the multivariate Bernoulli model at any vocabulary size.

Bayesian classifiers are emerging popularity and are found to perform surprisingly good. These probabilistic approaches make strong assumptions about how the data is generated and posit a probabilistic model that embodies these assumptions; then they use a collection of labelled training the parameters of the generative model. The naive Bayes classifier is that the simplest of those models, therein it assumes that each one attributes of the examples are independent of every other given the context of the category. This is the called naive Bayes assumption. This assumption is clearly states that, in most real-world tasks, naive Bayes often performs classification. It is due to the independence assumption, the parameters for every attribute are often learned separately, and this greatly simplifies learning, especially when the amount of attributes is large.

2.4 TOPICAL CLUSTERING OF TWEETS

In this paper, the emerging field of micro-blogging and social communication services, users post millions of short messages every day. Keeping track of all the messages posted by your friends and therefore the conversation as an entire can become tedious or maybe impossible. This paper presented a study on clustering and classifying Twitter messages, also referred as tweets, into different categories, inspired by the approaches taken by news aggregating services like Google News. The following results says that the clusters produced by simple unsupervised methods can often be unsophisticated from a topical perspective, but utilizing a supervised methodology that utilize the hash-tags as indicators of topics produce surprisingly good results. This paper also offers a discussion on temporal effects of our methodology and training set size considerations. Finally, this paper describe a simple method of finding the most representative tweet in a cluster, and provides an analysis of the results.

Some recent researches in social media and natural language processing have focused on interesting uses of Twitter messages, or tweets as they are more colloquially known, and other communicated messages. An interesting problem in tweet analysis is the automatic detection of topics being discussed in tweets. This paper propose that the hash-tags that appear in tweets can be viewed as approximate indicators of a tweets topic. The first discuss past work on tweet and microblogging message analysis. Next, this paper formulate our approach to Twitter message topic detection, target topics and describe our data set. Then this paper describe a set of experiments and results. Finally, this paper offer a discussion of our results and suggest research future directions.

3.MODULE DESCRIPTION

3.1 PREPROCESSING

The division of sentiments as positive and negative is inappropriate, because some diseases are generally classified as negative. Positive emotions could arise as a result of relief about an epidemic subsiding, but this project ignore this possibility. Use negative as the name of the first category and non-negative for the second one. The problem reduces to a two-class classification problem, and a Trends tweet can either be a Negative tweet or a Non-Negative tweet. The Twitter messages were formed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity. This project use negative as the name of the first category and non-negative for the second one. Thus, the problem reduces to a two-class word alignment problem, and a Trends review can either be a Negative review or a Non-Negative review.

3.2 CLUE-BASED REVIEW LABELI1NG

The clue-based classifier parses each review into a set of tokens and matches them with a corpus of Trend clues. There is no available corpus of clues for Trends versus News classification. The MPQA corpus contains a total of 7229 words, including 3340 adjectives, 469 adverbs, 1346 any-position words, 2287 nouns, and 1322 verbs. As for the sentiment polarity, among all 8221 words, 4912 are negatives, 570 are neutrals, 2744 are positives, and 21 can be both negative and positive. In terms of strength of subjectivity, among all words, 5089 are strongly subjective words, and the other 1602 are weakly subjective words. Social media users tends to express their trends and opinions in a more casual way compared with other documents, such as News, online reviews, and article comments. It is expected that the existence of any profanity might lead to the conclusion that the review is a Trend review.

3.3 MACHINE LEARNING CLASSIFIERS FOR TRENDS REVIEWCLASSIFICATION

This combined the high configuration of clue-based classification with Machine Learning-based classification in the Trends vs. News classification. The two classes of data [p and n] from the clue-based label model is used as the training datasets to train the Machine Learning models. The three popular models are Tri Model, polynomial-kernel Support Vector Machine. After the Trends vs. News classifier is trained the classifier is used to make predictions which is the pre-processed tweets.

The dataset of low recall in the clue-based approach, this project combined the high precision of clue-based classification with Machine Learning-based classification in the Trends versus News classification. After the Trends versus News classifier is trained, the classifier is used to make predictions on each twitter in p, which is the pre-processed reviews dataset. The goal of Trends versus News classification is obtain the separate Labels.

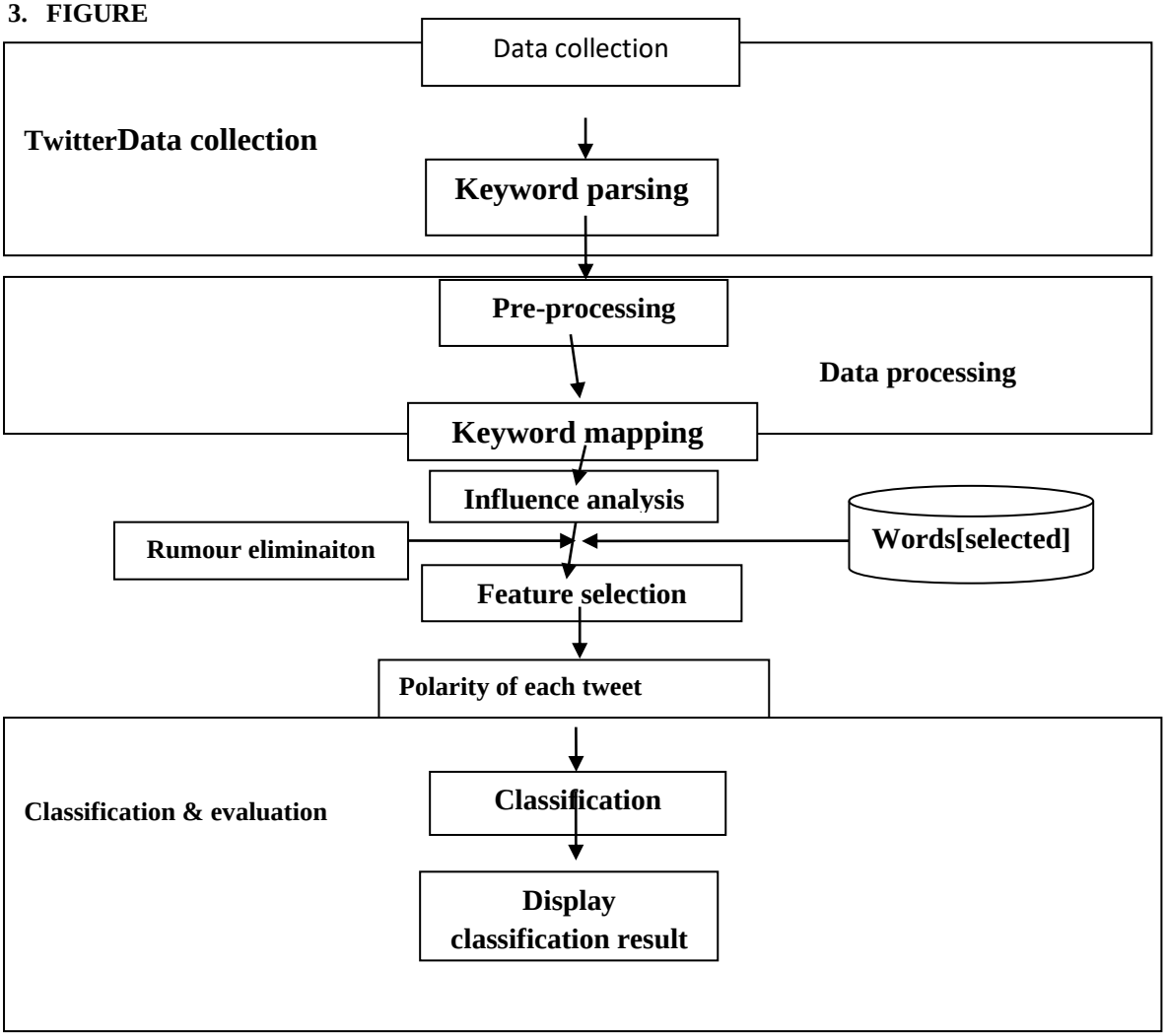
4.1.4TOPIC CLASSIFIACTION AND IDENTITY TWEETS:

The topic classification of the well-known ‘Bag-of-Words approach’ for text classification and network-based classification.. This paper propose that the hash-tags that appear in tweets can be viewed as approximate indicators of a tweets topic. In network-based classification method, this project identify top 5 similar topics for a given topic based on the number of common influential users. The categories of the topics and therefore the number of common influential users between the given topics and their similar topics are wont to classify the given topic employing a C 5.0 decision tree learner. Outputs on a database of selectively selected 564 trending topics [over 18 classes] shows that classification and accuracy up to 68% and 70% are often achieved.

This proposed project is to monitor the public health concern from the reviews and them as positive and negative opinion. To find accuracy a two-step sentiment classification approach is implemented: In the first step, classify health reviews into Personal disease inference reviews versus News reviews. It uses a subjective clue-based lexicon and News stop words to automatically extract training datasets labelling Personal disease inference disease inference reviews and News reviews. The self-generated training datasets are then used to train Machine Learning models to classify then, a review is Personal disease inference disease inference or News.

In the second step, we utilized an emotion-oriented clue-based method to automatically extract training datasets and generate another classifier to predict whether a Personal disease inference review is negative or non-negative. In sentiment classification, by combining a clue-based method with a machine learning method, good accuracy can be achieved. This overcomes the drawbacks of the clue-based method and the Machine Learning methods when used separately.

3. FIGURE



REFERENCES

[1] Ajmal Shahzad and Alex X. Liu, ―Accurate and Efficient Per-Flow Latency Measurement Without Probing and Time Stamping‖, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 24, NO. 6, DECEMBER 2016, P3477-3492

[2] IEEE, Yongzheng Zhang, Member, IEEE, and Yu Zhou, Member, IEEE, "A Semantics-Aware Approach to the Automated Network Protocol Identification", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 24, NO. 1, FEBRUARY 2016, P583-595.

[3] UPnP and Secure Group Communication Technique for Zero–configuration Environment construction using IncrementalClustering‖, International Journal of Engineering Research &Technology (IJERT), Vol. 2 Issue 12, December – 2013, ISSN: 2278– 0181, pp. 2095–2101.

[4] P. C. Sethi, C. Dash: ―High Impact Event Processing using Incremental Clustering in Unsupervised Feature Space through Genetic algorithm bySelective Repeat ARQ protocol‖, ICCCT– 2nd IEEE Conference – 2011, pp. 310–315.

[5] P.K. Behera, ―Secure Packet Inspection matching implemented Using Incremental Clustering Algorithm‖, December–22–24, ICHPCA–2014 (IEEE International Conference)

[6] P. C. Sethi, P. K. Behera, ―Internet Traffic Classification for Faster and Secured Network Service‖, International Journal of ComputerApplications (IJCA), Volume 131 – No.4, December2015, pp. 15–20.

[7] P. C. Sethi, P. K. Behera, ―Methods of Network Security and Improving the Quality of Service – A Survey‖ International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) Volume 5, Issue 7, July 2015, pp. 1098–1106.

[8] P. C. Sethi, P. K. Behera, ―RSA Cryptography Algorithm Using linear Congruence Class‖, International Journal of Advanced Research (2016), Volume 4, Issue 5, 1335-1347.

[9] Luigi Grimaudo, Marco Mellia, Elena Baralis and Ram Keralapura,SeLeCT: Self-Learning Classifier for Internet Traffic‖, IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENOL. 11, NO. 2, JUNE 2014.