Multi-Class Tweet Categorization Using Map Reduce Paradigm

Mohit Tare[1], Indrajit Gohokar[2], Jayant Sable[3], Devendra Paratwar[4], Rakhi Wajgi[5]

[1] *(Department of Computer Technology, Nagpur University, YCCE Nagpur, India)*

[2] *(Department of Computer Technology, Nagpur University, YCCE Nagpur, India)*

[3] *(Department of Computer Technology, Nagpur University, YCCE Nagpur, India)*

[4] *(Department of Computer Technology, Nagpur University, YCCE Nagpur, India)*

[5] *(Department of Computer Technology, Nagpur University, YCCE Nagpur, India)*

ABSTRACT: *Twitter is one of the most popular micro-blogging website in today's globalized world. Twitter messages can be mined to gain valuable information. Although Twitter provides a list of most popular topics people tweet about known as Trending Topics in real time, it is often hard to understand what these trending topics are about. Therefore, various efforts are being made to classify these topics into general categories with high accuracy for better information retrieval. We propose the use of one of the classification algorithm called Naïve Bayes for the categorization of tweets which has been discussed in this paper. It then proposes how the Map – Reduce paradigm can be applied to existing Naïve Bayes algorithm to handle large number of tweets.*

Keywords *- Categorization, Map-Reduce, Trending Topics, Tweet, Twitter*

1. INTRODUCTION

Twitter is a social networking site launched in July 2006. There are mostly social or otherwise informational relationships among the users of Twitter. The users follow other users to maintain social links and for gaining access to interesting information generated by others as well [1][2]. Many people make use of social networking sites like Twitter to share their emotions, sentiments as well as providing latest information which is evident from the reactions of people on the events encompassing the Egyptian revolution of 2011 [3].

As Twitter is very much popular site in all around the globe, huge amount of population in the world uses Twitter and generate millions of tweets each day. These users are also overwhelmed by the massive amount of information available and the huge number of people they can interact with. If a user wants to search tweets about a particular topic then the categorization of tweets must be done first to make any sense out of the vast amounts of tweets in Twitter. For this purpose we propose the use of Naïve Bayes supervised learning classifier and the Hadoop Map Reduce framework for the categorization of tweets. Before going into the details, the necessary information related to twitter in general are reviewed as follows.

TWEET - This means posting a message of up to 140 characters, known as tweets. People post messages about their various daily activities via these tweets. News are also been posted by the news channels, tweeting them via Twitter to alert the users. These messages also include URLs to web pages or hash tags to relate tweets of similar topics together. Each hash tag is a keyword prefixed by a # symbol. #Obama & #Romney for example have been used extensively during the recent elections in the US.

@USERNAME - Using the '@' symbol people can address their tweets to some person who they want to communicate with. A person can also address to multiple users for example, @aaronpaul @bryancranston

#HASHTAG – The popular or the trending topics on Twitter are highlighted using the hashtags. People can also contribute towards the trending topics by attaching their personal message to them for example, #rooney 'How much is he really worth after signing the new deal?'

RT - A retweet is a popular feature of Twitter using which a user can forward a tweet of another person like a celebrity to his own followers.

Before diving deep into all the research work currently being done related to tweet text classification or categorization it is imperative to have some knowledge of classification methods or algorithms being used for the particular research work.

The remainder of the paper is structured as follows. Section 2 deals with all the different research work done related to classification or categorization based on twitter content. Section 3 deals with the collection followed by the preprocessing of tweets. Section 4 deals with the use of the Naïve Bayes algorithm along with the Map Reduce paradigm for categorization of tweets and the paper is finally concluded in the section 5.

2. RELATED WORK

Currently, there is a lot of research going on in the area of user classification and sentiment detection. Many papers have been proposed in the area of text classification.

Kathy Lee et al. [4] proposes the classification of Twitter trending topics into 18 general categories by using Bag-of-Words approach and network based classification.

Zahan Malkani and Evelyn Gillie [5] discuss a study of various supervised learning methods and evaluate them on their performance. The evaluation is based on two sources of data namely an attitudes dataset that classifies tweets of the users according to their attitude and a topics dataset that classifies tweets into limited domain topic set.

Naaman et al. [6] categorizes Twitter messages based on their content. Their results show some fascinating observation such as, 80% of the tweets belong to user to user communication whereas the rest 20% has news characters.

Sankaranarayanan et al. [7] proposes a system called 'Twitter Stand' for capturing Tweets regarding worldwide breaking news. For this the categorization is done into two classes namely 'news' and 'junk'.

All of these approaches provide a valuable insight into Tweet classification but none proposes an approach to handle large number of tweets.

3. DATA

3.1 Collection of Tweets

The preliminary step deals with collection of tweets for categories like sports, politics and technology. Twitter provides two API's for gathering tweets namely the Twitter Streaming API and the Twitter REST API. We make use of the Twitter REST API to gather our tweets.

We have used the Twitter4j library to gather tweets which internally uses twitter REST API. The Twitter4j library requires OAuth support to access the API. Twitter uses OAuth to provide authorized access to its API. We have used the Application Only Authentication where the application makes API requests on its own behalf, without a user context. API calls are still rate limited per API method, but the pool each method draws from belongs to your entire application at large, rather than from a per-user limit. We have generated OAuth settings by using a twitter account, as we cannot use the settings unless you do not have a registered account.

3.2 Preprocessing

Before using the tweets collected from Twitter as training data, preprocessing of the tweets is done to remove redundant and unnecessary information. The preprocessing steps include:

1. Removing links (URLS): This step involves removing of URLS from the gathered tweets as they do not give us any significant information.

2. Removing usernames: Twitter lets the users send private messages to other users by using the '@' character followed by a username at beginning of the tweet. The usernames are not of importance to us in the classification process. So we remove the '@Username'.

3. Removing special symbols: This step constitutes the removal of special characters (like #, . , ^ , $) which are unnecessary.

4. Removing emoticons: This step removes the various emoticons in the tweets.

5. Removing Stop words: This step removes the various stop words that constitutes the grammar of the sentence as they do not give us significant information (like the, for, who etc).For this we gathered a corpus of stopwords which is used for removing the stopwords from the tweets.

6. Removing of retweets: Retweets look like normal Tweets with the author's name and username next to it, but are distinguished by the Retweet icon and the name of the user who retweeted the Tweet. Retweets contain the copy of the original tweet, so it adds to redundancy of information. So in this step we remove the retweets so that training data contains the unique tweets only.

3.3 Labeling

The final step after preprocessing of tweets is the labeling of tweets based on categories namely politics, sports and technology.

Table 1: Distribution of dataset as per the category

Category	Percent of total tweets
Politics	0.34
Sports	0.29
Technology	0.37

4. ALGORITHM

4.1 Naïve Bayes Classifier

It is a probabilistic classifier that is based on applying Bayes' theorem with strong (naive) independence assumptions.

In simple terms, this classifier is based on the assumptions that the presence or absence of a particular feature, given the class variable, is unrelated to the presence or absence of any other feature. We have used the Naïve Bayes classifier for the categorization of tweets due to its simplicity and also because it can be easily implemented by using the Map-Reduce metaphor. The probability model for a Naïve Bayes classifier is a conditional model.

Now *p (C|F1, ……, Fn)* over a dependent class variable **C** with a small number of outcomes or classes, is conditional on several feature variables **F1** through **Fn.**

Using Bayes' theorem, this can be written as:

$$p(C|F1,\dots,Fn) = \frac{p(C) * p(F1,\dots,Fn|C)}{p\ (F1,\dots,Fn)} \qquad (1)$$

In general language, we can represent it as:

$$\text{Posterior} = \frac{\text{Prior} * \text{Likelihood}}{\text{Evidence}} \qquad (2)$$

In simple terms, Bayes's rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then [8]:

$$\Pr[H|E] = \frac{\Pr[E|H] * \Pr[H]}{\Pr[E]} \qquad (3)$$

In case of Text classification Naïve Bayes classifier would model a document on basis of presence or absence of words on that document. We have employed a Multinomial Naïve Bayes classifier which considers frequency of words. In our case it can be denoted as:

$$\mathrm{P}\ (\mathrm{c}\,|\,\mathrm{t}) \propto \mathrm{P(c)} * \prod_{1<k<nd} \mathrm{P(Wk}\ |\mathrm{c}) \qquad (4)$$

Where, P (c | t) is the probability of tweet t being in category c, P(c) is the prior probability of category c (obtained from the training data) P (Wk | c) is the posterior probability of word belonging to category c.

Despite of the assumption that features in the text being classified are independent of each other, we observe that this classifier performs reasonably well on our data set and also is significantly faster. We observed that the accuracy of the classifier is about 75% for the test dataset, consisting of 100 sample test tweets, which we collected form Twitter.

4.2 Strategy for Naïve Bayes using Map – Reduce paradigm

As millions of tweets are being generated traditional classification approaches have fallen short of efficiency and speed. To address this problem, the power of the multicore technology and machine learning can take help of parallel programming method for large data sets to potentially speed up the operations. Cheng – Tao Chu, Andrew Ng et al. [9] have adapted the Map - Reduce approach and demonstrated the speed up of many machine learning algorithms.

Map-Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The data are fed into the map function as key value pairs to produce intermediate key/value pairs. The client does not have to deal with the splitting of main job into the Mappers. The partitioning of jobs into correct number of maps is handled by the framework. For example, in case of Apache Hadoop framework, which we have used in our proposed strategy the correct number of maps, is driven by the number of input blocks in the DFS (Distributed File System) to be used by the job. All nodes will do same computation in the Mapper phase. It uses Data Locality to increase performance. Once the mapping is done, all the intermediate results from various nodes are reduced to create the final output.

The proposed strategy uses Apache Hadoop framework, an open source java framework, which relies on Map – Reduce paradigm and a Hadoop Distributed File System (HDFS) to process data. Our proposed Map – Reduce strategy for classification of tweets using Naïve Bayes classifier relies on two Map-Reduce passes.

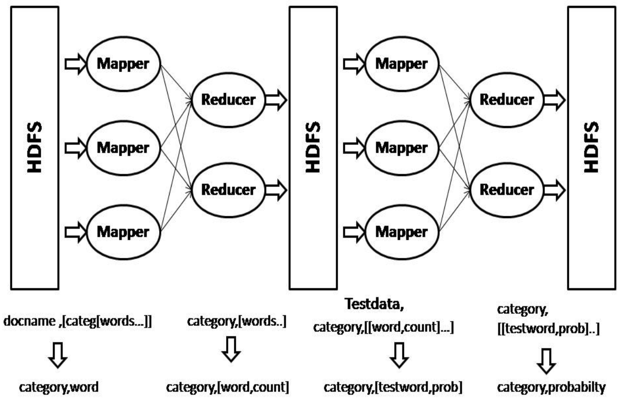

Fig. 1: The Naïve Bayes classifier using map reduce

In the first Map-Reduce pass, the mapper takes the labeled tweets from the training data and outputs category and word as key value pair. The Reducer then sums up all instances of the words for each category and outputs category and word-count pair as key-value. The Map-Reduce thus deals with formation of model for the classifier. The next Map-Reduce pass does the classification by calculating conditional probability of each word (i.e. feature) and outputs category and conditional probability of each word as key-value pair. Then final reducer calculates the final probability of each category to which the tweet may belong to and outputs the predicted category and its probability value as key-value pair.

5. CONCLUSION

Twitter research so far has become an exciting area for research projects related to machine learning and data mining. The huge amounts of data being generated on social networking sites such as Twitter can be very helpful to big multinational companies or politicians in making important strategic decisions based on user classification, sentiment or geographical distribution. This paper thus helps in this topic by proposing how Map-Reduce paradigm can be applied to the existing Naïve Bayes classifier for tweet classification. This will help to analyze large dataset much easily using multiple node clusters.

REFERENCES

[1] Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. "Why we twitter: understanding microblogging usage and communities." In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56-65. ACM, 2007.

[2] Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?." In Proceedings of the 19th international conference on World wide web, pp. 591-600. ACM, 2010.

[3] Choudhary, Alok, William Hendrix, Kathy Lee, Diana Palsetia, and Wei-Keng Liao. "Social media evolution of the Egyptian revolution." Communications of the ACM 55, no. 5 (2012): 74-80.

[4] Lee, Kathy, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. "Twitter trending topic classification." InData Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 251-258. IEEE, 2011.

[5] Malkani, Zahan, and Evelyn Gillie. "Supervised Multi-Class Classification of Tweets." (2012).

[6] Naaman, Mor, Jeffrey Boase, and Chih-Hui Lai. "Is it really about me?: message content in social awareness streams." In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pp. 189-192. ACM, 2010.

[7] Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. "Twitterstand: news in tweets." In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42-51. ACM, 2009.

[8] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[9] Chu, Cheng-Tao, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun. "Map-reduce for machine learning on multicore." In NIPS, vol. 6, pp. 281-288. 2006.