CATEGORIZATION OF GEO-LOCATED TWEETS FOR DATA ANALYSIS

Harshita Kumar[1], Deepanshu Chandel[2], Dr. M.L Sharma[3]

[1,2] Students, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India
[3]Head of department, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

---***---

Abstract - *Microbloging websites are rich sources of data for data analysis and extracting trending topic. Twitter is one of the best microbloging website which provide authorization to excess its tweet data to it's users. In this research, we focus on using Twitter for extracting trending tweets about events, products, people and use it for classification of topics. We will classify trending topics into 3 categories I.e sports, politics and technology. This includes a new approach for classifying Twitter trends by adding a layer of feature ranking. A variety of feature ranking algorithms, such as bag-of-words, TF-IDF are used to facilitate the feature selection process. Naive Bayes text classifiers, backed by these sophisticated feature ranking techniques, is used to successfully categorize Twitter trends.*

***Key Words*: Naive Bayes classifier, Tweets classification, Feature extraction, Unigram,Term Frequency – Inverse Document Frequency (TF-IDF)**

1.INTRODUCTION

In recent years, with the sudden increase in popularity of various social networks, the way we generate and extract information has changed dramatically. Microbloging has become a very popular communication tool among Internet users.

Millions of messages are generated daily in popular web-sites that provide services for microbloging such as Twitter, Facebook. Users of these services write about their life, share opinions on variety of topics and discuss current issues. Twitter is one of the most popular of these social networks and from past few years it has been at the hub of most of the discussion going around the world. People frequently tweet on recent issues or events, topics which become popular on twitter are known as trending topics. Twitter provide a list of all the trending topics in the world. These trends can be related to music, sports, technology etc. It is very interesting to know trends in world and know people opinion about it. Now days trends are identified with hashtags ,@,a special character or a word. People tend to use them a lot to make something popular such as #Metoo,@nariendermodi etc.

Twitter only provide information about trending topic but not about its domain. The trending topic names may or may not be expressive enough to tell messages or information about it to unfamiliar user's, until and unless they explicitly read about it. It is a very important aspect about any topic as it defines its relives with the tweets. For example #Me Too where people tweeted about the sexual and physical abuse they have been through.We find that trend names are not indicative of the information being transmitted or discussed either due to obfuscated names or due to regional or domain contexts.

The first step to organize this information is to categorize them. We classified trending topic into 3 categories such as sports,politics and technology. Our goal is to provide accurate information to user regarding any trending topic in specified domains. To classify we have taken navie bayes classifier as an approach. In order to achieve this aim, we used supervised machine learning to train a Naïve Bayes Classifier to classify twitter's trending topics

We mined the textual data in the tweets (associated with each trend) to train our classifier. The experimental setup involves three major steps: 1.) Cleaning and preparing the data in the right format, 2.) Feature ranking i.) Bag-of-words and TF-IDF , 3.) Training and testing the classifier to successfully classify the trends into the three categories.

The rest of this paper is organized as follows. Section 2 describes some of the related works. Section 3 provides an overview of proposed system. Section 4 presents details of the methodology of twitter trending topic classification system. Section 5 describes experimental results. Finally, the conclusion and some future directions are presented in Section 6 and 7 respectively.

2. RELATED WORK

Data analysis has always been of interest to researchers. It got some more attention with the introduction of social networks. However, the textual data on these social networks is in a form of natural language, which includes a lot of slangs and abbreviations. Therefore, understanding them and taking out relevant information from it is a bit of challenging task .

We chose to work with twitter firstly because is it used worldwide, people are exceptionally active on it and tweet's frequently. Moreover the response on twitter is more prompt and also more general. Secondly, it provides authorization to its users to extract data for personal usage without any charges.

Twitter is also useful to get acquainted with trending topics of a country or in a world with respective to their countries. Twitter itself provide service to easily extract trending topic but those results are not up to required accuracy. So in this project we tried to get output with better accuracy which specifically defined the domain of the trending tweet and also the focused event or person they are talking about through tweets.

This research draws its inspiration from the various papers. Paragraph from *twitter trending topic classification* by (Lee et.al,2011) classify the twitter trends in real time. They used 2 approaches for classification of tweets into 18 categorizes which are text-based classification and network based classification. Through labeling they divided tweets into different categorizes manually and then performed data modeling on the organized data.This paper helped us to understand text classification and manual labeling of tweets where as in our research we have automated the labeling process.

Paragraph from *real time classification of twitter trends* by (Zubiaga et.al,2013) they have used support vector machine (SVM) for classification. They used 2 different representation of classification of process i.e twitter feature analyzed and bag of words. 15 features were considered for analyzing such as retweets, hashtags, length, exclamation etc. For textual content they relied on bag of words. This paper helped us in understanding working and importance of bag-of words which plays important role in feature ranking.

Paragraph from *Classification of Twitter Trends using Feature ranking and Feature Selection* By (Shah,2015) explained the importance of approaches of feature ranking such as bag-of words and TF-IDF. He categorized data with the usage of training and testing data. These two types of data were different in nature training data contains predefined tweets in perspective of their category where as testing data was random in nature without categorization. This paper built the knowledge about feature ranking in TF-IDF and how can we categorize data using training and testing data.

Paragraph from *Is Naïve Bayes a Good Classifier for Document Classification?* By (Ting et.al,2011) discussed about naive bayes in detail. Though naive bayes is less accurate than SVM but still because of its effectiveness people prefer to use it. Through this paper we gained motivation to use naive bayes as our primary approach for classification of tweets.

3. RESEARCH METHODOLOGY

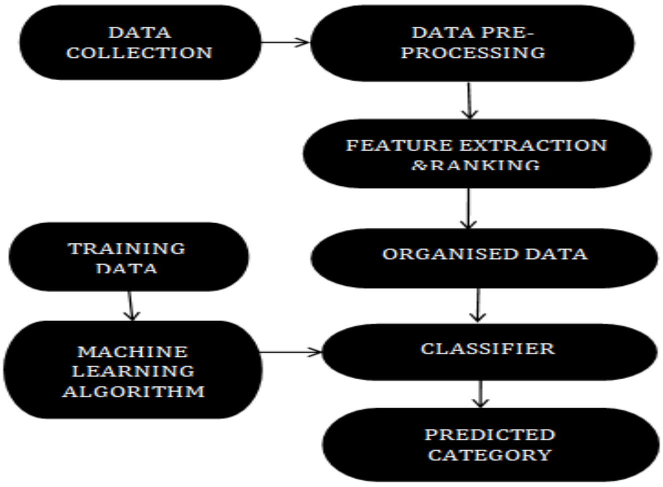


Fig1- Classification of tweets

In order to generate tweet classifier figure 1 depicts methodology of it. We begin our process by collecting two types of data i.e training and testing. As data is unstructured and contains useless content it is important to remove it, so in pre-processing we removed noncontributing content from tweets. After cleaning and organizing the data we extracted its feature using bag-of-words or TID-IF. Simultaneously we trained the classifier using training data. Once classifier is trained we inserted input i.e organized data which provided us with predicted categorizes.

3.1 DATA COLLECTION

The data was gathered using twitter's publicly available API. Twitter contentiously updates its top trending topic list. As such there is no information how these trends are identified and make up to the list. However, one can request up to 1500 tweets for a given trending topic. We extracted 2 types of list of tweets i) tweets with defined domain ii) random tweets. Tweets with defined domain were used for training the classifier where as random tweets was the input for the system.

All the tweets containing a trending topic constitutes a document. For example, while the topic "T20" is trending, we keep downloading all tweets that contain the word "T20" from Twitter, and save the tweets in a document called "T20". In case a tweet contains more than two trending topics, the tweet is saved in all relevant documents. For example, if a tweet contains two trending topics "T20" and "hawkings", the same tweet is saved in two different documents. Likewise we downloaded tweets of 3 different categorizes for training the classifier and save in respective files.

3.2 DATA PRE-PROCESSING

The main approach involved in this project were the various data pre-processing steps, the machine learning classifiers and feature extraction. The main machine learning algorithm used were Naive Bayes. The main data pre-processing steps include filtering, twitter slang removal, stopwords removal and stemming.

Filtering :-

In this we removed URL,user-names, duplicate or repeated characters.

Twitter slang removal :-

As mentioned people use casual language in tweets which can include abbreviation short forms. Here we removed slang like these.

Stopword removal:-

In information retrieval,there were many words added as conjunctions such as and,before,that which didn't help in classifying the tweet in category as they are present in every type of tweet. So we remove such word from data.

Stemming:-

It is process of retrieving the first form of verb of word. Example walked, walking all are derived from word walk.

3.3 FEATURE EXTRACTION & RANKING

A feature is a piece of information that can be used as a characteristic which can assist in solving a problem . The quality and quantity of features is very important as they are important for the results generated by the selected model. Selection of useful words from tweets is feature extraction.

- Unigram features – one word is considered at a time and decided whether it is capable of being a feature.
- N-gram features – more than one word is considered at a time.
- External lexicon – use of list of words with predefined positive or negative sentiment.

After feature extraction we used these unigrams and bigrams and apply TF-IDF or bag-of-words on it to find the weight of particular feature in a text. Here we have used TF-IDF.

3.4 EXPERIMENTATION

After the pre-processing and feature extraction steps are performed, we worked towards training and validating the model's performance. The collected dataset was divided in two– training set and testing set. The training set was used to train the classifier (machine learned model) while the testing set was the one on which the experimentation is performed.

Divided the set as training set containing 21000 tweets while testing set 3900 tweets (approx. 93% and 7%) while used 75% data for training set and used approx. 83% for training. Manual labeling was avoided as classification work is topic based and adaptive in nature.

3.5 NAIVE BAYES ALGORITHEM

Here we will applied naive algorithm which will actually identify the category of the tweets.

P(c | t)=P(t| c)P(c)/P(t)

Where, P (c | t) = Probability of trend t belonging to class c (Posterior)

P (t | c) = Likelihood of generating trend t given class c

P (c) = Probability of occurrence of class c

As one can see that while each word is related to the trending topic category, they are independent of each other. In other words, the words (features) are not related to each other. This feature independence is at the core of every Bayesian Network.

After this we will get predicted categories.

4. EXPERIMENTAL SETUP

4.1 ALGORITHM

Naive Bayes classification algorithm of Machine Learning is a very interesting algorithm. It was used as a probabilistic method and was a very successful algorithm for learning/implementation to classify text documents. Any kind of objects can be classified based on a probabilistic model specification. This algorithm is based on Bayes' theorem.

In bayes's theorem we find probability of label with some of given features.We can compute directly by applying this formula

P(L | features)=P(features | L)P(L)P(features)

All we need now is some model by which we could compute P(features | Li)P(features | Li) for each label.

This was where the "naive" in "naive Bayes" comes in: if we made very naive assumptions about the generative model for each label, we can find a rough approximation of the generative model for each class, and then proceed with the Bayesian classification. Here we were considering Gaussian naive Bayes.

P(c | t)=P(t| c)P(c)/P(t)

Where, P (c | t) = Probability of trend t belonging to class c (Posterior)

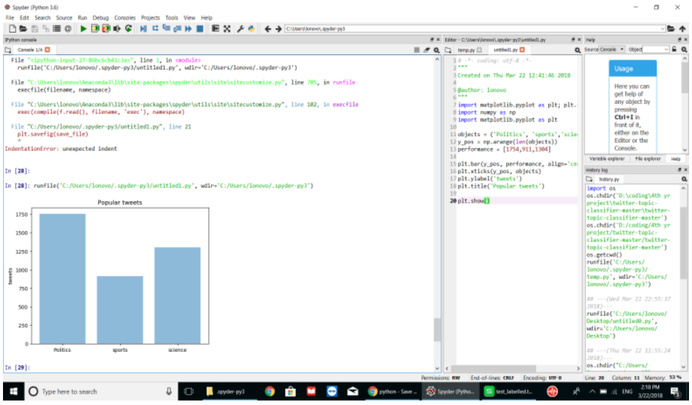
P (t | c) = Likelihood of generating trend t given class c

P (c) = Probability of occurrence of class c

4.2 IMPLEMENTATION

For developing a trending topic classifier we decided to use supervised machine learning algorithms. For implementing those algorithms in form of a code we chose python as our core language. As python has evolved in past few years by introducing new packages which provided us needed environment and features for this project. Here we have used python 3.0 and sypder IDE for developing the classifier.

5. RESULT



```
  File "<ipython-input-27-86bc6cb41c3a>", line 1, in <module>
    runfile('C:/Users/lonovo/.spyder-py3/untitled1.py', wdir='C:/Users/lonovo/.spyder-py3')

  File "C:\Users\lonovo\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 705, in runfile
    execfile(filename, namespace)

  File "C:\Users\lonovo\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 102, in execfile
    exec(compile(f.read(), filename, 'exec'), namespace)

  File "C:/Users/lonovo/.spyder-py3/untitled1.py", line 21
    plt.savefig(save_file)
    ^
IndentationError: unexpected indent


In [28]:

In [28]: runfile('C:/Users/lonovo/.spyder-py3/untitled1.py', wdir='C:/Users/lonovo/.spyder-py3')

In [29]:
```

```
# -*- coding: utf-8 -*-
"""
Created on Thu Mar 22 12:41:46 2018

@author: lonovo
"""
import matplotlib.pyplot as plt; plt.
import numpy as np
import matplotlib.pyplot as plt

objects = ('Politics', 'sports','scie
y_pos = np.arange(len(objects))
performance = [1754,911,1304]

plt.bar(y_pos, performance, align='ce
plt.xticks(y_pos, objects)
plt.ylabel('tweets')
plt.title('Popular tweets')

plt.show()
```

Usage

Here you can get help of any object by pressing **Ctrl+I** in front of it, either on the Editor or the Console.

```
import os
os.chdir('D:\coding\4th yr project\twitter-topic-classifier-master\twitter-topic-classifier-master')
os.chdir('D:/coding/4th yr project/twitter-topic-classifier-master/twitter-topic-classifier-master')
os.getcwd()
runfile('C:/Users/lonovo/.spyder-py3/temp.py', wdir='C:/Users/lonovo/.spyder-py3')

## ---(Wed Mar 21 22:55:37 2018)---
runfile('C:/Users/lonovo/Desktop/untitled0.py', wdir='C:/Users/lonovo/Desktop')

## ---(Thu Mar 22 11:55:24 2018)---
os.chdir("C:/Users/
```

Fig 2- BAR GRAPH

Here graph shows the no. Of tweets belong to different categorizes which we have predicted using naive bayes classifier.

From this it is clearly visible that now days trending topics are more about politics than any other category.

6. CONCLUSION

In this research, we have explored the top conversations shown as trending topics on the site. We have introduced a system to organize Twitter's trending topics according to their categories. This system includes the following 3 types of trending topics: sports, politics and technology.

We have performed classification experiments using NAVIE BAYES classifiers to study the usefulness of these features to discriminate types of trending topics. The proposed method provides an immediate way to accurately organize trending topics using a small amount of features.

7. FUTURE SCOPE

Based on the performance of the proposed system, some changes and extensions can be made. Future work would include adding more categories, which would help some classifiers further distinguish features. Examples of good categories to pursue next include real-time concerts and posts about music in general. Another extension could be finding the exact trending topic of most popular domain. Exploring new machine learning algorithm to get more accurate results.

ACKNOWLEDGEMENT

Every successful venture has the blessings and the support of many behind it. It would be unjust if every such support is not revealed and thanked in person as well as in the published report.

First and foremost, We would like to thank Dr. M.L Sharma for giving me an opportunity to work under his constant guidance throughout the project.

Finally, an honorable mention goes to my family for their support to help me do this project. Without the help of the particular's mentioned above, we would have faced many difficulties while pursuing this project.

REFERENCES

[1] A. Shah," Classification of Twitter Trends using Feature ranking and Feature Selection",2015.

[2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," 2009.

[3] A. McCallum, and K. Nigam, "A comparison of event models for naïve Bayes text classification", in Journal of Machine Learning Research, Vol. 3,2003, pp. 1265–1287.

[4] A. Zubiaga , D. Spina , R. Mart´ınez , V. Fresno ,"Real-Time Classification of Twitter Trends "in Journal of the American Society for Information Science and Technology,2013.

[5] B. David," A lot of randomness is hiding in accuracy", in Engineering Applications of Artificial Intelligence,2007 , pp. 875 – 885.

[6] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," Machine learning, vol. 6, no. 1, 1991,pp. 37–66.

[7] IBM SPSS Modeler, http://www-01.ibm.com/software/analytics/spss/products/modeler/.

[8] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,2009, pp. 42–51.

[9] K. Lee, D. Palsetia, R. Narayanan, Md. M.A. Patwary, A Agrawal, and A. Choudhary,"Twitter Trending Topic Classification",in 11th IEEE International Conference on Data Mining Workshops,2011,pp. 251-258.

[10] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification", in Proceedings: IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, pp. 61–67.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update", in SIGKDD Explorations, Vol. 11, No. 1,2009, pp. 10-18.

[12] M. N. Hila Becker and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in Proceedings of AAAI, 2011.

[13] M. Naaman ; H. Berker , and L. Gravano . Hip and trendy, ”Characterizing emerging trends on twitter”, in Journal of the American Society for Information Science and Technology, may 2011,pp.902–918.

[14] R. Rifkin and A. Klautau,”In defense of one-vs-all classification”, in The Journal of Machine Learning Research, vol .5 , December 2004,pp.101-104.

[15] S.J. Delany, P. Cunningham, and L. Coyle, “An assessment of case-based reasoning for spam filtering”, in Artificial Intelligence Review Journal, Vol. 24, No. 3-4,2005, pp. 359-378.

[16] S. Kinsella, A. Passant, and J. G. Breslin, “Topic classification in social media using metadata from hyperlinked objects,” in Proceedings of the 33rd European conference on Advances in information retrieval, 2011,pp. 201–206.

[17] S.L. Ting, W.H. Ip, Albert H.C. Tsang,”Is Naïve Bayes a Good Classifier for Document Classification?”, in International Journal of Software Engineering and Its Applications, Vol. 5, No. 3,July 2011,pp.37-46.

[18] Y. S. Yegin Genc and J. V. Nickerson, “Discovering context: Classifying tweets through a semantic transform based on wikipedia,” in Proceedings of HCI International, 2011.

[19] Zubiaga, A. ; Mart´ınez, R. , and Fresno, V.,” Getting the most out of social annotations for web page classification” ,in Proceedings of the 9th ACM symposium on Document engineering, vol. 09,2004, pp. 74–83.

[20] Zubiaga, A. ,Spina, D. , Amig ´o, E. , and Gonzalo, J.,” Towards real-time summarization of scheduled events from twitter streams”, in Proceedings of the 23rd ACM conference on Hypertext and social media, vol. 12, pp. 319–320,2012.