Twitter Trend Analysis by using Latent Dirichlet Allocation (LDA)

Prof. Mrs. Laxmi R. Sisode[1] Shubham Deshpande[2] Shubham kamble[3] Vikram Damase[4] Akash Kamerkar[5]

[1,2,3,4,5]P.E.S Modern College of Engineering, India

Abstract— The community of users participating in social media tends to share about common interests at the same time, giving rise to what are known as social trends. A social trend reflects the voice of a large number of users which, for some reason, becomes popular in a specific moment. Through social trends, users therefore suggest that some occurrence of wide interest is taking place and subsequently triggering the trend. In this work, we explore the types of triggers that spark trends on the microblogging site Twitter. The user will be allowed to search for the latest trends by inputting a keyword into search field. Based on user provided keyword, the system will search for similar keywords in database and summarize the total count to provide the trending tweets on twitter. The trending tweets with hashtag () will be displayed first and then the rest words will be displayed. By clicking on every trending tweet, the user commented tweets will be displayed. User can view all the tweets from the searched keyword.

Key words: Pre-Processing, LDA, Stemming, Tokenization, Stop-Word Removal, Twitter

I. Introduction

Social platforms like Twitter, Facebook, Instagram has been established as an information source for different trend identification.

Twitter is one of the most famous social media networks; it allows users to post tweets, messages etc on its social network. Twitter as one of the most visited and used social network, it is a very important resource for data about people interest. It also gives information about different trends. Also important are the determination of the topics behind such trends. As a result, a number of areas have grown in interest of summarizing the based information including topic generation and keyword based information extraction. Traditional techniques have not performed well in supporting topic generation and classification, due to non-standard words as well as the overall high amount of noise that is present in such communications. The different issues are extensive use of symbols, abbreviations, emogis etc. Latent Dirichlet Allocation (LDA) is a fully generative model for describing the latent or specific topics of documents. LDA models every topic as a distribution over the words of a vocabulary and every document over the sampled topics from a Dirichlet distribution. We consider the application of LDA for the purpose of topic generation by evaluating the different topic categories of interest of the user. Preprocessing is using techniques for preparing data for the analysis process.It includes several steps, each step produces data ready for the next step until transforming process done. It mainly includes tokenization, stop word removal and stemming.

II. Literature survey

LDA (Latent Dirichlet allocation) is a very widely used model that can represent the similarity of the data. For example, in the case that an observation is a collection of words in a document, LDA retrieves a set of keywords that are likely to be able to describe the document and uses this set of keywords to represent the document for the purpose of analysis.

LDA is a type of topic modellig whichis a mixture idea from computer science, mathematics and other fields, which uses Bayesian statistics and machine learning and various other concepts to discover the latent or specific topics in a document. Asked on this knowledge, topic models can also be very useful to give prediction about the future related documents. Because of all these features of LDA , topic models are powerful tools, which helps to understand the information among topics which might not be related with each other.

Generally the LDA gives two types of probability distributions:

Probability distribution over words in the same topic. Probability distribution over topics in the document. For example, in the topic of “department”, there will be some words, such as “computer”, “electrical”, “mechanical”, occur together very often, then LDA gives the probability distribution over these words.

III. System model

In this paper, we discussed the different modules of the proposed system which are as follows:

A. Data Collection

Twitter is third party applications and developers use it to get access to the enormous amount of data generated by users .The Twitter Streaming API works by making a request for a specific type of data filtered by keyword, user, geographic area, or a random sample, and then keeping the connection open as long as there are no errors in the connection. The API that we have used is the Twitter4j API for data collection.

B. Data Preprocessing

1) Tokenization

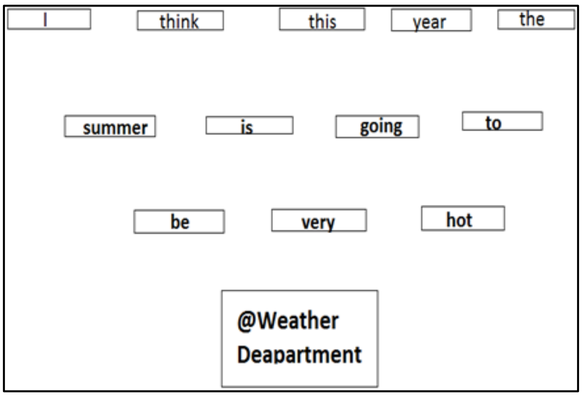
In order to transform a tweet into tokens the tweet passes through two phases: Word segmentation and Cleaning.

Word segmentation is the process of separating the statements of written language to its words, which compose the sentences structure. The proposed system, analyzes tweets written in English. Tweets are short sentences, so the proposed algorithm divides the sentence into words and symbols (which are separated by spaces), and stores each word or symbol in a separate row in a table. The main purpose of tokenization is to break the tweet into different parts which are known as tokens. It helps to understand the tweet in an easy way without any complications.

An example of tokenization is shown below:

For example if the tweet is; “I think this year the summer is going to be very hot, @WeatherDepartment”.

The tokens are:

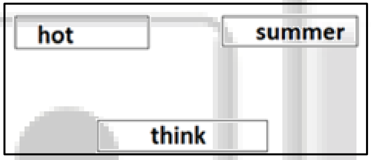


Here the word "hot" and the comma "," are considered together as a single token, because they are not separated by a space.

2) Stop-word removal

Till now we have seen that the data stored in the table is large data set, and it contains many words that is not useful for the analyzing system which known as stop words.

The stop words of English language are stored in a table in this step, items in the tokens table compared with each word in the stop word table in order to delete the stop words from tokens table for each tweet. There are different algorithms that are used for stop-word removal. If we consider the above example then after stopword removal it will give the following result: the remaining tokens are:

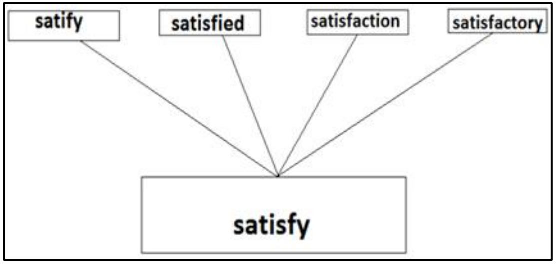


The words "is", "will", "be", "very", are removed as these words are stop words.

3) Stemming

After the above steps we have data that has meaning and values, but the amount of data still large Also we have many words which are having the same meaning. Stemming is nothing but the process which is used to find out the root/stem of a word.

For example, the words satisfy, satisfactory, satisfied, and satisfaction all could be stemmed to the word "satisfy".



The purpose of this step is to remove various suffixes, to have exactly matching stems, to save memory space and time. The system which is going to be used uses the potter stemming algorithm for performing the stemming process.

For example:

Consider the following sentences:

It was a lovely movie.

I loved the movie.

In the above sentences the words lovely and loved can be stemmed to its root i.e. "love".

C. LDA

LDA is an unsupervised Machine Learning algorithm which identifies latent or specific topic information among large document collections.

The steps applied for document collection are as follows:

1) For each document, select a topic from its distribution over topics.
2) Sample a word from the distribution over the words associated with the chosen topic.
3) The process is repeated for all the words in the document.

The important equations used are as follows:

P(ti|d)=P(ti | zi=j)P(zi=j|d)

In the above equation, P (ti|d) is the probability of the ith term for a given document d and zi is the latent topic. Also P(ti | zi=j) is the probability of Ii within topic j and P(zi = j | d) is the probability of picking a term from topic j in the document.

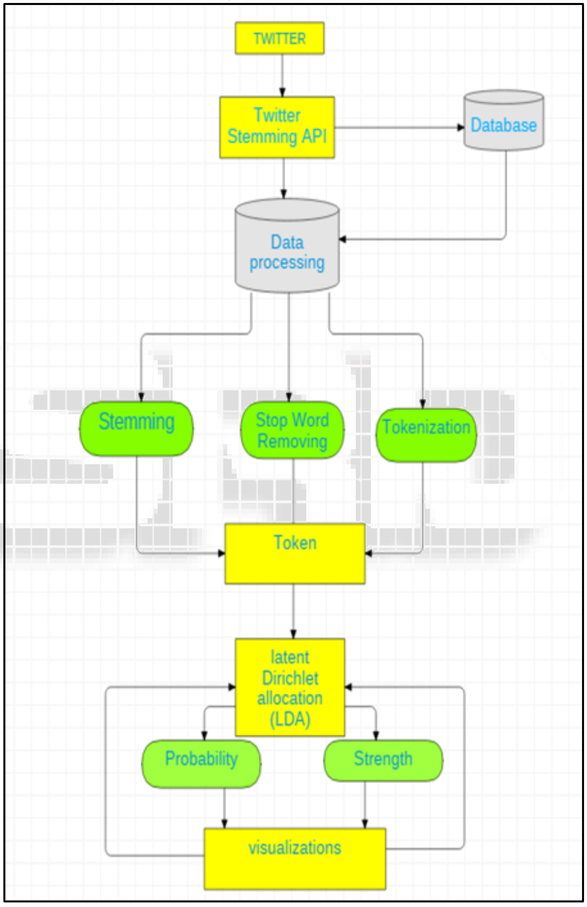


Fig. 1: System Architecture

IV. CONCLUSION

In this paper, we have given the information of a Twitter Trend Mining system that is designed for real-time twitter data to:

1) Store every tweet produced in Twitter.
2) Keep track of trending tweets.
3) Visualize the trending topics.

The major contribution of the study is making it possible to mine social trends and content that is generated in Twitter through adequate integration of state-of-the-art techniques. We have evaluated LDA for the application of topic modeling for the purpose of supporting a greater understanding of trends in Social Media. In our project we were able support an unsupervised classification on a set of Twitter information. From our results, it has demonstrated potential as a complementary technique that could serve as a means to support the derivation of information about trends as well as assist in identifying new trends. Future work includes the integration of LDA in conjunction with our supervised techniques.

References

[1] David Alfred Ostrowski,"Using Latent Dirichlet Allocation for TopicModelling in Twitter"in IEEE 9th International Conference on Semantic Computing, 2015, pp. 493-497.

[2] Blei, David M., Andrew Y. Ng, and Michael L Jordan,"Latent dirichlet allocation" the Journal of machine Learning research 3 (2003).

[3] Min Song,Meen Chul Kim,"RT2M : Real-time Twitter Trend Mining System" in International Conference on Social Intelligence and Technology,2013,pp.64-71.

[4] Chaney, A., Blei, D,"Visualizing topic models". Proceedings of the 6th International AAAI Conference on Weblogs and Social Media 2012.

[5] Dr. Hussein K. Al-Khafaji, Areej Tarief Habeeb, "Efficient Algorithms for Preprocessing and Stemming of Tweets in a Sentiment Analysis System" IOSR Journal of Computer Engineering (IOSR-JCE), 2017.