Social Media Post Analyzer

[1]Neha Bhondwe, [2]Nupur Chillal, [3]Priyanka Sutar, [4]Snehal Wakade, [5]Shital Jadhav

[1,2,3,4,] UG Students, [5]Professor

[1, 2,3,4,5] Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, India

Abstract: **Detection of emerging topics is interest motivated by the rapid growth of social networks. Posts include not only text but also images, URLs, and videos. We focus on emergence of topics signaled by social aspects of these networks. In this system we are taking a Facebook as Social media to perform the operation, specifically. We proposing the NLP algorithm of the mentioning behavior of a social network user, and propose to detect the emergence of a new topic from the Aggregating facebook data from hundreds of users by using the Hadoop Framework. We demonstrate our technique in several real data sets we gathered from Twitter. The experiments show that the proposed KNN approaches can detect new topics at least as early as text-anomaly-based approaches, and in some cases much earlier when the topic is poorly identified by the textual contents in posts.**

Keywords: **Topic detection, social network such as facebook, Hadoop framework.**

I. INTRODUCTION

The growing popularity and spread of social media has changed the design of Lifestyle of Human being. Nowadays, there are over millions of users of social media. Daily there is arrival of new data or posts in the form of different media or document files. Therefore, there is creation of large amount of data on social media. As these data sets are coming from different place it's not an easy task to deal with it. Also there is presence of large amount of data on social media account on any individual. This data contains all type of data i.e. it may contain document files or media files such as jpg or mp3 or mkv.

There is no limit on where this data is coming from. There are many users on social media therefore we get different type of data from different users. But this data is not all what we want, there is lots of data in which we are not interested but as the posts or data on social media is random or we can say it is not as per the user interest. The data coming on social media sites are as per the modifications when some new posts come it modify it again. Therefore, there is lot of need to visualize these posts as per user interests which will use in achieving efficiency for social media.

II. SCOPE

In earlier days on social media the number of posts is stored on home page and we need to refer all those irrespective of our interest. Nowadays, People don't get enough time to refer all these post to refer and therefore as per some people consideration social media had became wastage of time for them. Therefore to make efficient usage of Social media we need to visualize posts as per user requirements and interests.

The main objective of this project is to help us to filter out posts appeared on social Media such as Facebook ,twitter, LinkedIn - the posts are related to the sports or religion or general knowledge or entertainment or technology must be filtered out i.e. we are sorting out the posts of similar category.

III. PROPOSED METHOD

The overall flow of the propose system is shown in fig.1. Each step in the flow is described in the corresponding sub section. We assume that the

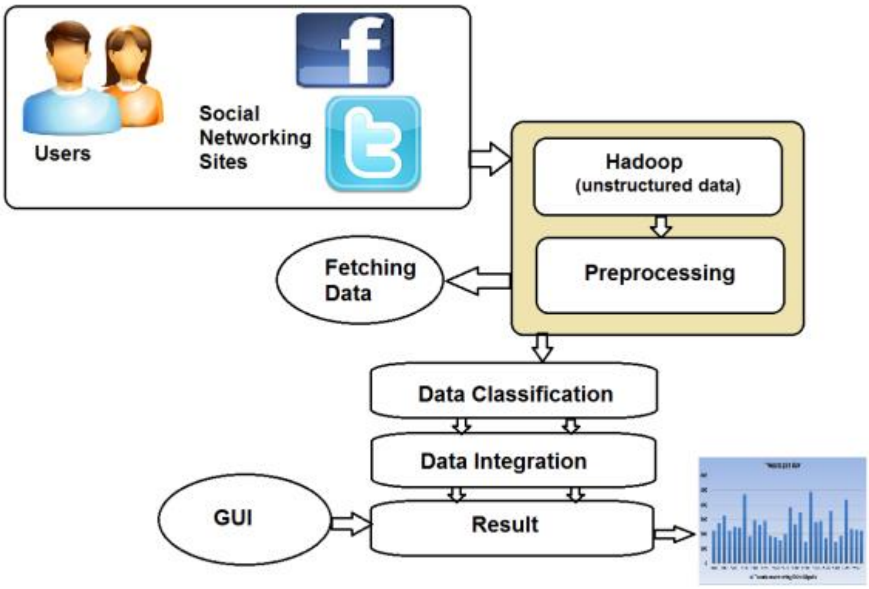

Fig.1: Proposed system

Data arrive from a social network service in a sequential manner through some API. We are storing data in Hadoop system (dynamically) (Step1). Preprocessing on fetch data from Hadoop And for storing fetch data using MYSQL server (step 2).After this we are classifying the data as per our requirement by KNN algorithm and integrating data for specific field by using Lucene method (step 3).Generating graph on the integrated data with help of BI tool.

IV. METHOD

i. KNN Algorithm:

An object is classified by majority votes of its neighbor. It is simplest algorithm for classification.

K-nearest neighbor algorithm is a non-parametric method.

1. Determine parameter K=Number of nearest

Neighbour.2.Calculate the distance between the query instance & all the training samples.3.Sort the distance & determine nearest neighbour based on the K'th minimum distance. 4. Gather the category Y of the nearest neighbour. 5. Use simple majority of nearest neighbour as the prediction value of the query instance.

ii. NLP:

Learning algorithms — often, although not Modern NLP algorithms are based on machine learning especia lly statistical machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general always, grounded in statistical interface— to automatically learn such rules through the analysis of large *corpora* of typical real-world examples. A *corpus* (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned.

iii. Lucene Method:

Lucene is an open-source Java full-text search library which makes it easy to add search functionality to an application or website. Lucene is an open source java based search library. Lucene is very popular and fast search library used in java based application to add document search capability to any kind of application in a very simple and efficient way. Indexing process is one of the core functionality provided by Lucene. Following diagram illustrates the indexing process and use of classes. Index writer is the most important and core component of the indexing process.

Search System:

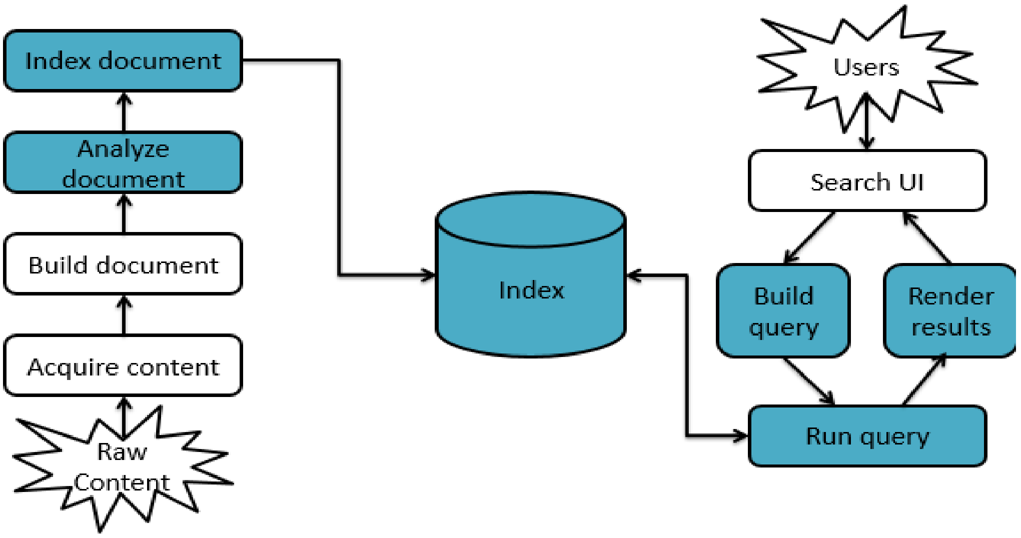

Fig.2: Lucene system

Indexing process is one of the core functionality provided by Lucene. Following diagram illustrates the indexing process and use of classes. Index Writer is the most important and core component of the indexing process. We add Document(s) containing Field(s) to Index Writer which analyzes the Document(s) using the Analyzer and then creates/open/edit indexes as required and store/update them in a Directory. Index Writer is used to update or create indexes. It is not used to read indexes.

V. CONCLUSION

In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. We have proposed NLP algorithm captures both the number of mentions per post and the frequency of mentioned. We used clustering KNN approach for clustering the post, hence the system will generate the approximate graph or the result.

REFERENCES

[1] *J. Kleinberg, "Bursty and Hierarchical Structure in Streams," Data Mining Knowledge Discovery, vol. 7, no. 4, pp. 373-397, 2003.*

[2] *Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-Time Change-Point Detection Using Sequentially Discounting Normal- ized Maximum Likelihood Coding," Proc. 15th Pacific-Asia Conf. Advances in Knowledge Discovery & Data Mining (PAKDD' 11), 2011.*

[3] *S. Morinaga and K. Yamanishi, "Tracking Dynamics of Topic Trends Using Finite Mixture Model," Proc. 10th ACM SIGKD Int'l Conf. Knowledge Discovery And DATA Mining pp 811-816 2004*

[4] *Q. Mei and C. Zhai, "Discovering Evolutionary Theme Pattern from An Exploration of Temporal Text Mining," Proc. 11Th ACM SIGKDD Int'l Oonf. Knowledge Discovery in Data Mining, pp. 198-207, 2005.*