Classification of Streaming Big Data

Prof. Jagannath Nalavade[1] Aditya Kshirsagar[2] Chaitrali Kardile[3] Durgarani Vyavhare[4] Tushar Kharmate[5]

[1]Professor [2,3,4,5]Student

[1,2,3,4,5]Department of Computer Engineering

[1,2,3,4,5]S.I.T. Lonavala, India

Abstract— There is huge communication taking place in social networking site due to that many users came across various issues like cyber bullying and grooming [2]. There are some political issues which tend to riots. Usually a social networking site wall is flooded with lots of posts; it's difficult for the user to view important post. That's why our aim is to focus on classifying those posts into various categories. Post can be classified into various categories like: Entertainment post, Political post, Social post, Historical post, Educational post, Sports post etc.

Key words: K-Means Clustering, Post Classification, Streaming Big Data, Social Networking Site, Supervised Learning, Machine Learning

I. Introduction

On social networking site there are huge amount of posts are generated and it's become very complicated for the user to view those post [1]. For that purpose it is important to give some structure to these posts for user to have user-friendly environment.

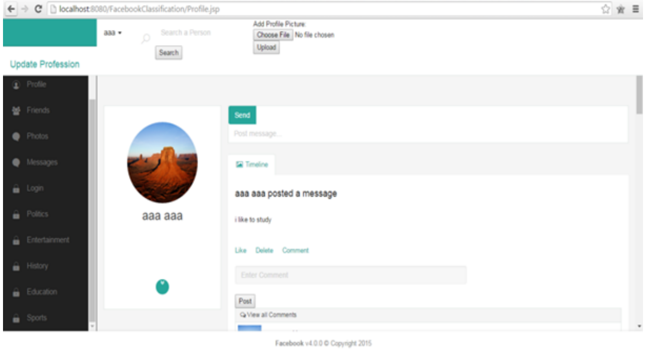
Post may include political issues, entertainment issues, social issues and much more daily issues flooded in their surroundings.

Classified social networking site post column1 Educational post, column2 Social posts, column3 Entertainment posts, column4 Political posts, column5 Historical posts. Sentiments such as sad, happy, angry, neutral are displayed for all incoming posts on users wall.

In present world users are more interested in social networking site status, comments and posts which are related to personal life.

Social networking site does not belong to any particular type of user, category.

It's open for all, so that everyone can use it and its not related to any particular country.-news feed on social networking user post about entertainment, jokes, photos as well as their views opinions on different topics.



aaa aaa posted a message

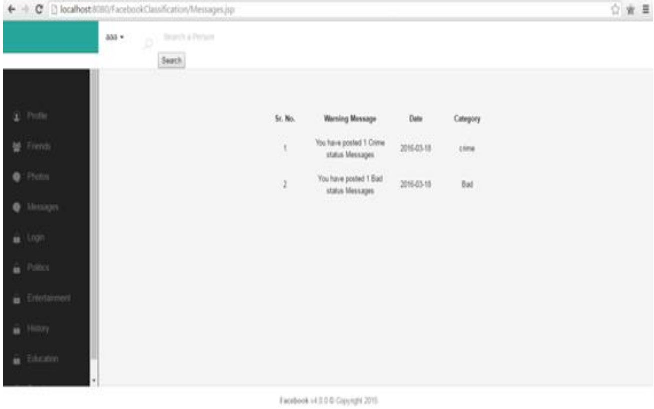
i like to study

Fig. 1: User profile and User Post

Sentiment analysis [3] is use for block post, comments, and opinions. What people actually think, in what sense its positive way or negative way it can be classified.

By using sentiment analysis [3] we can find out what user trying to say.

Fig.1 Depicts the user profile updates and different categories of post. Left side of Fig1. Represents different categories of post like historical, political, sport, educational, Entertainment.



Sr. No.	Warning Message	Date	Category
1	You have posted 1 Crime status Messages	2016-03-18	crime
2	You have posted 1 Bad status Messages	2016-03-18	Bad

Fig. 2: Warning message

Fig2. Represent result of our posts. If any offensive post or any threat which are posted by user then it will give warning message to that user. It gives remark about post like bad, very bad, good, and neutral. And also show date of that warning message.

In this paper, we are proposing a system for classification of social networking site posts which will automatically classify the posts into various categories. In Yahoo, news feeds are automatically classified into different streams like Health, Entertainment, and Relationship. Statistics may provide rough idea about daily social networking site usage. In this paper we are trying to show graph relation based on the reports.

To generate approximate result or approximate graph on users behaviour using sentiment analysis.

II. Proposed System

The objectives used to design a new system.

Give user details to database server to fetch data to particular user. Fetch data from social networking site database, using graphical API's.

For this system we use database server to store user information and data which belongs to categories

This user information will be used for authentication of post which is posted by user. It will store in database. Then this data will be fetched by the classifier which classifies the post into five categories. Using k-means algorithm classified post goes into different five clusters which are Historical, Political, Educational, Entertainment, Sports. Using NLP API's we perform sentimental analysis on different categorical post which gives sentiment of that user post. Result of all these processes will represent in graphical form using R programing language Due to graphical representation it will be easier for user to understand which category have more number of posts in a day.

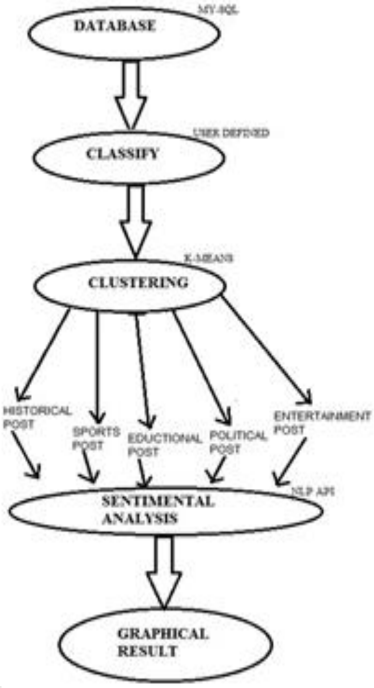


Fig. 3: Proposed System

Classify data into different categories using user defined classifier. Graphical representation of sentiment analysis using R programming. In data mining it fetched data of posts from social networking site.

III. Related Work

In this section we give brief survey of proposed system for text classification and sentiment analysis. In paper [1] author seeks about the Facebook post classification like, like page post, friends post, entertainment post, life event post.

In paper [2] author try to say about the cyber bulling, cyber stalking, online grooming.in this paper they mainly focus on this threats and it can be determine by image analysis, data mining techniques and social media analytics to avoid attacks.

In paper [3] is done at three levels i.e. document level, sentence level and attributes level. It uses two types of techniques machine learning and semantic orientation.

In document sentiment detection NLP is used.

A. Data Fetch

In this we fetched the user’s information when the user authenticate. Whenever user login that time system creates a call to the graph API. For instance when the user wants a information about name, birth date, language, collage that time data can be fetched from the social networking site database.

B. Sentiment Analysis

Opinion mining is also known as sentiment analysis. Sentiment analysis uses the natural language processing, text analysis to identify and extract information. A basic task of sentiment analysis is to classify the post negative, positive or neutral and emotional status angry, sad and happy.

1) NLP API

NLP API is natural language processing is easy to use mechanism to recognize positive and negative statements from any post. This API is useful for to extract positive and negative posts from any webpage, image, text.

C. Graphical Representation

Graphical representation is used to visualize the data using bar graph and charts .Graphical representation tries to show sort and clear data view it is used when large amount of posts are generated.

1) R Programming

The R programming language is interpreted language. It is used to convert data into graph. It uses the command line interface

D. Supervise Learning

Supervised learning it comes under the machine leaning task which belongs to supervised classification. This supports to the text and image classification techniques. Supervised leaning classification is based on two sides training set and test set. Training set automatically classifies data into text and images and in test set we perform the operations on input data.

E. Classification

The various classes considered are as follows:

1) Historical: Most of the users share some thoughts on historical events and post their photos, some of posts might hurt someone’s feelings and may create riot.
2) Entertainment posts: Posts related to entertainment industry such as related to any film actors, actresses or any TV personality. And also updates like movies, technology, poems and new released books.
3) Sport posts: These posts are related to sports person or events, like sportsman’s personal and professional life or information related to sport events.
4) Educational posts: information related to education like technologies, digital education or post related to colleges events and post related to job information.
5) Political posts: User can share post information related to political person or their photos or any political events like elections.

F. Image Processing

1) Adult Image Detection:

In image processing, we mainly focus on adult image detection. If the Image is adult content or nudity detection it can be classify further. Image can classify into different classes like close-up face, skin color.

IV. Algorithm

A. Clustering

Clustering is the process to collect the data into similar and dissimilar groups. Document should be similar to same cluster and dissimilar to another cluster.

B. K-Means Algorithm

This is the one of the simplest algorithm to determine number of cluster k. It is NP hard problem.

$$\boxed{E_i = \int_{x \in R_i} \rho(x) \| x - z_i \|^2 dx \,,}$$

Steps for finding k number of cluster is as given below

1) Find the centroid co-ordinate.
2) Estimate distance of each object to the centroid.
3) Objects which have minimum distance from the centroid should group together.
4) Repeat 1,2,3 until centroid of cluster remains same.

C. Classification

Classification is the task to identify set of categories. Classification is considered as a supervised learning. It classifies data based on training set and set values. It predicts class labels.
Two steps process:

– Model construction: describing a set of predetermined classes
– Model usage: for classifying future or unknown objects

1) Naïve bays

This classification supports supervised learning method as well as statistical methods for classification. It requires small amount of training data set.

V. Future Scope

– Victim can immediately file case against the offender.
– Find location and trace the offender.
– Social networking site authorities.
– There must be compulsion for user to enter mobile number in social networking site user details.

VI. Conclusion

– Research in data streams is still in its early stage. If the problems are addressed or solved and if more efficient and user-friendly mining techniques are developed for the end users.
– Most algorithms handle data with only numerical or ordinal attributes
– Present algorithms not dealing with missing data.
– It is likely that in the near future data stream mining will play an important role in the business world as the data flows continuously.
– To create graphical analysis for user.

References

[1] Shankar Setty, Rajendra Jadit, Sabya Shaikh, Chandan Mattikalli, Vma Mudenagudi,” Classification of Facebook News Feeds and Sentiment Analysis”, B. V. Bhoomaraddi College of Engineering and Technology, Hubli-lndia,2014.

[2] Marlies Rybnicek, Rainer Poisel and Simon Tjoa,” Facebook Watchdog: A Research Agenda For Detecting Online Grooming and Bullying Activities”, Institute for IT Security Research St. P¨olten University of Applied Sciences 3100 St. P¨olten, Austria, 2014.

[3] G.Vinodhini and RM.Chandrasekaran,” Sentiment Analysis and Opinion Mining: A Survey”, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002 India.

[4] Jiaqi Ge ,Yuni Xia, Jian Wang, “A Naïve Bayesian Classifier in Categorical Uncertain Data Streams”, IEEE 2014.

[5] Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, and Li Guo,” E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams”, IEEE Feb-2015..

[6] Fatemeh Sheikholesalmi, Morteza Mardani ,Georgios B. Giannakis, “Classification of Streaming Big Data with Misses”, IEEE 2014.

[7] Wikipedia “adult image detection using Filtering Objectionable Image Based On Image Content ”

[8] Yin, X.Xu,L.Ye “Big Skin Regions Detection for Adult Image Identification”,China.

[9] Wei Zeng,Wei-Qiang Wang,Wen Gao,”Shape-based Adult Image Detection”,2005

[10] Facebook Inc., “Facebook’s latest news, announcements and media esources,” https://newsroom.fb.com/, 2013, Accessed April 17th, 2013.

[11] J. K. Ahkter and S. Soria, "Sentiment analysis: Facebook status messages," The Stanford NLP Group, Stanford University, Natural Language Processing, Final Project Report, 2010.

[12] Gang Li, Fei Liu, “A Clustering-based Approach on Sentiment Analysis”, 2010, 978-1-4244-6793-8/10 ©2010 IEEE.