Efficient Anomaly Monitoring System In Micro blog Platform

Kanchana R, Dr. Shashikumar D R

M. Tech Student, Dept. of CSE , Professor Dept. of CSE
Computer Science Engineering

Cambridge Institute of Technology , Bangalore, Karnataka

***Abstract*—** The paper describes Anomaly detection on Microblog platform using the big data area. We cover the primary elements behind detecting Anomaly,Correlated and Evolution by processing structured Framework. In this paper, the Big Data Processing is referenced for analysing the huge volume of data and some social media data in Real time. Our aim is to provide a efficient way of detecting the anomaly and concentrate on non functional feature such as performance and reliability.

IndexTerms **-** Big data,Anomaly Detection,Microblog,Analysis,Real-time.

I. INTRODUCTION

Microblog platform has become an important platform for exchanging message in social media. This platform convey information in viral form due to its popularity. A wide rage of Anomaly event will come across this platform will range from from low information to a political information. These platform report an important event such as earthquakes and accident .

As an example it will discover and reply to emergency event in timely manner[1] and summarize popular trending events [4]While the net event observation state of affairs demonstrate quick dynamic topics and need high out corne of knowledge process. Nor are they ready to perform early detection of rising abnormal events in time period. A subject model for brief texts[5] has been developed, which has good performance on small texts.

Neither of them will establish efficient suspicious events before they are in a large scale. The paper[3] could detect the potential abnormal events before they're in style and spreading at scale.

The Big Data is an important role in galactic organization. Data is exploding rapidly in different areas of of population growth and technology developments. Cost-effective methods are required to manage the Big Data Analysis. In order to procedure Spark helps to simplify the challenging and compute-intensive task of processing flooding volumes of real-time both structured and unstructured seamlessly integrating relevant complex capabilities such as machine learning and graph algorithms[2]. Spark bring forward Big Data processing to the masses.

Different data streams could have own features. Processing frameworks compute over the data in the system by reading from non-volatile keeping Computing over data is the process of extracting information and insight from large quantities of individual data points. In the same time, the data stream for sensors depends on sampling[6] and so, existing a sample of the entire population. Sometimes, data streams could be buzzing. Spatial and temporal attributes could dramatic work important role in data streams processing. In some cases we have to pay attention the limited resources for data streams processing. Hadoop is an open source softer framework for storing data and runing applications on cluster of commodity hardware. It support distributed computing environment for processing large data set. The Hadoop technologies is used for real time analysis for huge volume of data from web,social-media,audio and machine generated data.

They wanted to return web search results faster by distributing data and calculations across different computers so multiple tasks could be accomplished simultaneously. It was based on the same concept storing and processing data in a distributed, automated way so that relevant web search results could be returned faster. Hadoop's distributed computing model processes big data fast[7]. The more computing nodes you use, the more processing power you have. Unlike traditional relational databases, you don’t have to preprocesses data before storing it. You can store as much data as you want and decide how to use it later.

II. PROPOSED SYSTEM

In this paper, we present over microblog text streams. Emerging anomaly monitoring has attracted much attention from the research domain. Here we aim to monitor emerging anomalous events on microblog platforms. Our emerging anomaly monitoring methods are based on graph mining techniques, which provides unique opportunities to detect anomaly. In the system, emerging anomaly monitoring includes early detection, correlation analysis and temporal evolution tracking of anomalous events. Early detection would capture emerging events before they go viral. Correlation analysis would automatically reveal multiple aspects of the anomalous event, or the causality of anomalous events, or categorical structure of related anomalies. A user friendly interface is also provided to facilitate the analysis of emerging events with visualization. We provide a scalable anomaly monitoring approach meeting all the listed requirements. Especially, we are among the first to provide detailed correlation analysis of anomalies under the real-time emerging anomaly monitoring scenario.

III. IMPLEMENTATION

A. Data Collection module

A distributed crawler is constructed to gather information from Weibo, the biggest microblog platform in China. The crawler unendingly collects the most recent microblogs printed by users ideally with an outsized range of followers, i.e., opinion leaders. The crawler encompasses a master/slave design. The master node utilizes key/value store to perform task programming. Slave nodes get the assignments from the master and crawldata. A task would monitor the reports and comments of an artless tweet and retrieve the report and comment list, from that we will construct the forward graph of every tweet. The tasks square measure regular in keeping with posts' priority, that is weighed by the quantity of reports and comments .

B. Indexing Module

A distributed on-line index system for temporal microblog knowledge. The complete index is split into fine-grained time vary partitions to supply neighborhood for knowledge access consistent with temporal approximate. For higher synchronous access, whenever vary partition is split into sub-partitions by hash functions. every term's inverted tweet list is simply mapped to corresponding sub-partitions. With these structures, given question|a question a question} with a selected time vary the time vary info will be accustomed navigate to corresponding time vary partitions and so utilize query to quickly navigate to corresponding sub-partitions.

C. Processing Data Module

To with efficiency update graph structure, we tend to introduced a hash based graph partitioning technique to support find-grained and fast update. To support progressive computation on evolving graphs, we tend to designed associate application hardware to coordinate applications with totally different computing request frequency and management the employment of calculate resources.

D. Early prediction module

The anomaly detection method consists of two steps: trending keyword detection and community detection over keywords. Our intuition is that the essential keywords about an event would have similar trends and show a burst in usage compared to their own history. Trending keywords are detected as an anomaly with abrupt usage increase. An event is represented as a "bag" of keywords that co-occur and correlate with each other, with its detection time and representative tweet extracted for better comprehension. Then the trending keywords and their co-occurrence relationship is a strong combination to distill emerging anomalous events.

IV. EMERGING ANOMALOUS EVENT MONITORING

The idea that you can build applications to draw real-time insights from data before it is persisted is in itself a big change from traditional ways of handling data. Even machine learning models are being developed with streaming algorithms that can make decisions about data in real time and learn at the same time. Fast performance is important in these systems, so in-memory processing methods and technologies are attracting a lot of attention.

These architectures as shown in Fig:1 we need to handle very large volumes of data, so the tools used to implement them need to be highly scalable throughout the system. It's also important to design systems that can handle data from multiple data sources, making it available to a variety of data consumers.

First of all, notice that the results output from the real-time application now goes to a message stream that is consumed by the dashboard rather than reaching the dashboard directly. In this way, the results can easily be used by an additional component, such as the anomaly detector shown in this hypothetical example. One nice feature of this style of design is that the anomaly detector can be added as an afterthought. The flexible system design lends itself to modifications without a great deal of administrative hassles or downtime.

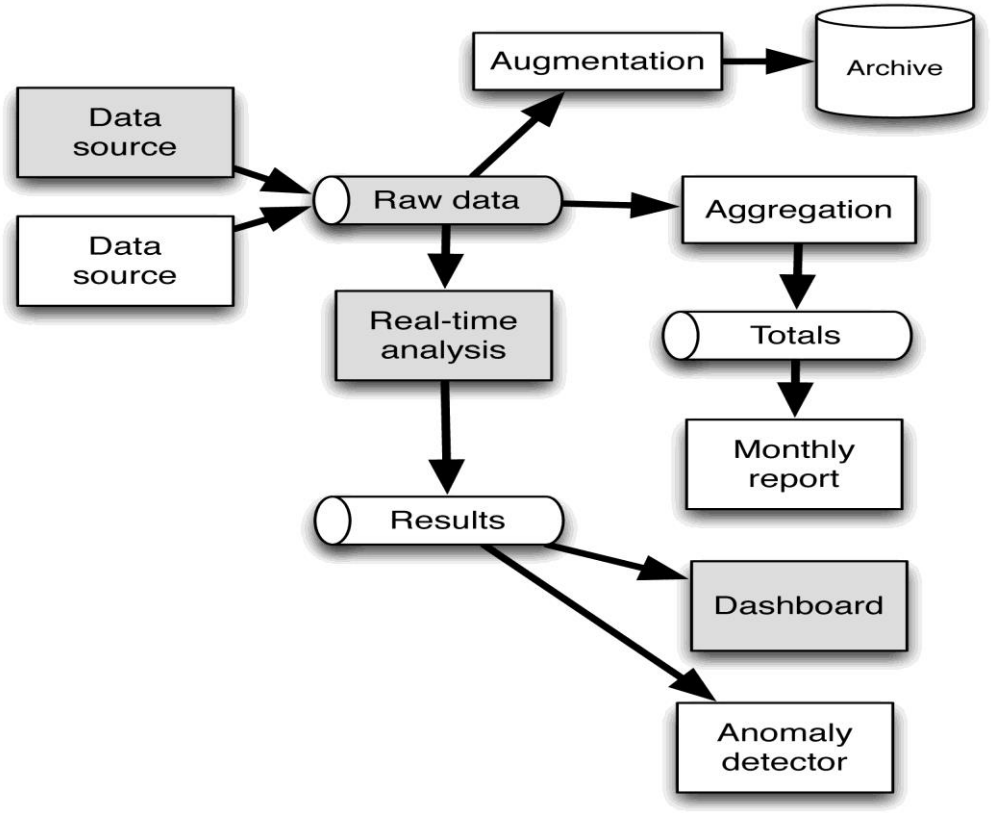

Fig 1:System Architecture

Our overall design also takes into account the desire to use multiple data sources. Since the consumers of messages don't depend on the producers, they also don't depend on the number of producers. The messaging system also makes the raw data available to non–real time processes, such as those needed to produce a monthly report or to augment data prior to archiving in a database or search document. This happens because we assume the messaging system is durable.

1. Term frequency inverse document frequency

Term frequency inverse document frequency appearance at however often a word seems in an exceedingly document and its importance relative to the full set of documents. "Words that seem often terribly ton of documents might not be very helpful like 'the', 'a'. however if there area unit words that show up often in stories regarding the Greek debt crisis however not regarding one thing else just like the elections, for instance, then those area unit helpful words to stay track of. And that's what Term frequency inverse document frequency captures," Sullivan explained.

This can be wont to build classifiers or prognostication models, he said. For instance, a corporation that has regarding ten years' price of client center dialogue that has been transcribed into text might faucet into this knowledge and work out what it all says. To do this, Sullivan aforesaid the calls may well be classified into 'conversations with customers on the brink of leave', 'conversations with customers downgrading their service', 'conversations with customers upgrading their service'.

This algorithmic rule is helpful once you have a document set, particularly an outsized one, that must be classified. It is price noting the variations between Term frequency inverse document frequency and sentiment analysis. though each may well be thought-about classification techniques for text, their goals area unit distinct. On the opposite hand, Term frequency inverse document frequency classifies documents into classes within the documents themselves. This could offer insight regarding what the reviews area unit regarding, instead of if the author was happy or sad. If we tend to analyzed product review knowledge from associate degree e-commerce web site mercantilism laptop elements, we'd find yourself with teams of documents regarding 'laptop', 'mouse', 'keyboard', etc. we'd gain an outsized quantity of information regarding the categories of reviews that had been written, however wouldn't learn something regarding what the users thought of these merchandise. Though the algorithms area unit similar in this they classify text.

```
[1] "The sky is blue."

[2] "The sun is bright today."
```

[3] "The sun in the sky is bright."

[4] "We can see the shining sun, the bright sun."

One thing you can see is that the word “bright”, which appeared only in 3 out of the 4 documents is a given really low score across all the documents. A word should be representative of a document if it shows up a lot, but if that word occurs too often across all the documents, then it is most likely a meaningless indicator.

2. stopwords

In each language, some words are not capably common. whereas their use within the language is crucial, they don’t typically convey a selected that means, particularly if taken out of context. This is often the case of articles, conjunctions, some adverbs, etc. That are normally called stop-words. within the example on top of, we will see 3 common stop-words – to, and on. Stop-word removal is one vital step that ought to be thought-about throughout the pre-processing stages. One will build a custom list of stop-words, or use on the market lists.

The algorithm is implemented as follows The target document text is tokenized and individual words area unit hold on in array. Next a single stop word is scan from stopword list. Then the stop word is compared to focus on text in style of array victimization subsequent search technique. Then if it matches , the word in array is removed , and also the comparison is sustained until length of array. After removal of stopword utterly, another stopword is scan from stopword list and once more formula follows step a pair of. The formula runs unceasingly till all the stopwords area unit compared. Finally resultant text destitute of stopwords is displayed, also required statistics like stopword removed, number of stopwords from target text, total count of words in target text, count of words in resultant text, individual stop word count found in target text displayed.

V. RESULTS AND DISCUSSION

In this section the Efficient Anomaly Monitoring System In Micro blog Platform system is determined in term of Anomaly event detection.

Dataset used here is the product review. Dataset is taken as an input for our system which consist of product review. We have to consider only the negatives so the positive review should not be included in output file and all the bad words are removed by stop-word concept.

Dataset file is taken as input which contain dataset has all combination of positive ,negative and bad comments as shown in Fig:2. We have to identified the negative anomaly leaving the positive review and remove all the bad words.

```
Sony Xperia Z2  Rs. 25,000      Askmebazaar     simply awesome.
Sony Xperia E4 Dual     Rs. 7,699       smartprice      Good phone.
Sony Xperia Z5 Premium Dual     Rs. 53,729      Smartprice      fuck off.
Sony Xperia T3  Rs. 10,499      Smartprice      Satisfied with  phone.
Sony Xperia C5 Ultra Dual       Rs. 23,669      Paytm   Good phone.
Sony Xperia E1 Dual     Rs. 4,790       Paytm   not-satisfied with phone.
Sony Xperia Tipo        Rs. 5,026       Paytm   Not good. Battery poor.
Sony Xperia Tipo Dual   Rs. 5,000       Paytm   Battery and llok and feel is good.
Lenovo K3 Note (Black, 16 GB)   Rs. 9,999       Paytm   Good phone.
Lenovo K3 Note (White, 16 GB)   Rs. 9,999       Paytm   fuck off.
Lenovo VIBE P1m (Black, 16 GB)  Rs. 7,999       Paytm   Not good. Battery poor.
Lenovo A2010 (Black, 8 GB)      Rs. 4,990       Paytm   Good phone.
Lenovo A2010 (White, 8 GB)      Rs. 4,990       Paytm   Good looks & design and its defficiencies...
Lenovo A6000 Plus       Rs. 7,499.00    Paytm   Better battery life.
Lenovo X2 45 AP Gold    Rs. 16,299.00   Paytm   Good looks & design and its defficiencies...
Lenovo S60(Graphite Grey)       Rs. 10,000.00   Paytm   Arrived on time and good packing.
Lenovo A6000 Plus (Black, 16 GB)        Rs. 7,499.00    Paytm   simply awesome.
Sony Xperia Z3 Plus +-(Xperia Z4)       Rs. 36,075      Smartprice      Good phone.
Sony Xperia L   Rs. 6,790       Smartprice      Performaing and good to have.
Sony Xperia Z2  Rs. 25,000      Smartprice      Satisfied with  phone.
Sony Xperia E4 Dual     Rs. 7,699       Smartprice      Good phone.
Sony Xperia Z5 Premium Dual     Rs. 53,729      Smartprice      not-satisfied with phone.
Micromax Canvas Xpress 2 E313   Rs. 6,199       Paytm   Not good. Battery poor.
Micromax Canvas Juice 2 Rs. 6,199       Smartprice      Battery and look and feel is good.
Micromax Canvas Pulse 4G E451   Rs. 9,999       Smartprice      Good phone.
Micromax Canvas Fire 4 A107     Rs. 5,069       Amazon India    fuck off.
Micromax Canvas Mega E353       Rs. 7,224       Amazon India    Not good. Battery poor.
Motorola Moto G (3rd Gen)       Rs. 12,999      smartprix       Good phone.
Motorola Moto E (2nd Gen)       Rs. 5,299       Snapdeal        Good looks & design in this range.
Motorola Moto G Turbo Edition   Rs. 14,499      Snapdeal        Arrived on time and good packing.
Motorola Moto X Play    Rs. 19,999      Snapdeal        Best deal ever in mid range!!
Motorola Moto X Style   Rs. 29,999      Snapdeal        Good but poor battery life.
Motorola Moto X (2nd Gen)       Rs. 14,999      Snapdeal        Good looks & design and its
defficiencies...
Motorola Moto X Force   Rs. 49,999      Infibeam.com    Amazingly smooth and has a much better battery
life.
```

Fig 2:Sample Input of anomaly event detection.

The result file contain only the negative anomaly from the product review. We have only the records in the output file as shown in Fig:3. The output file indicates the absence of bad word and positive review.

```
Brand New Sony Xperia E4:Rs. 8,619.00 Comments(1)       not bad and battery is good. ||
Lenovo A2010 (White, 8 GB):Rs. 4,990 Comments(1)        Good but poor battery life.||
Lenovo K3 Note (Black, 16 GB):Rs. 9,999 Comments(1)     Poor backbone side. ||
Lenovo K3 Note (White, 16 GB):Rs. 9,999 Comments(1)     poor performance||
Lenovo S60(Graphite Grey):Rs. 10,000.00 Comments(1)     not satisfied with phone. ||
Motorola Moto X Style:Rs. 29,999 Comments(1)    Good but poor battery life.||
Samsung Galaxy J2:Rs. 8,149 Comments(1) Good look and design. But poor service. ||
Samsung Galaxy S Duos 2 S7582:Rs. 8,990 Comments(1)     not-satisfied with phone. ||
Sony Xperia E1 Dual:Rs. 4,790 Comments(2)       Service is not so great and poor performance||not-
satisfied with phone. ||
Sony Xperia Z3 Plus (Xperia Z4):Rs. 36,075 Comments(1)  not-satisfied with phone. ||
Sony Xperia Z5 Premium Dual:Rs. 53,729 Comments(1)      not-satisfied with phone. ||
Xiaomi Mi 4 (White, 16GB):Rs. 14,999.00 Comments(1)     Service is poor and heating effect ||
```

Fig 3:Sample Output of anomaly event detection.

VI. CONCLUSIONS

We have demonstrated Efficient Anomaly Monitoring System monitor the event and relevant event in more efficient way. The system present the advantage compare with the existing system. In the future ,we have to do research on text streaming and automatic event detection by using the evolving knowledge for more better performance.

REFERENCES

[1] T. sakaki, m. okazaki, and y. matsuo, “earthquake shakes twitter users: real-time event detection by social sensors,” in www, 2010.

[2] C. li, a. sun, and a. datta, “twevent: segment-based event detection from tweets,” in cikm, 2012.

[3] X. yan, j. guo, y. lan, and x. cheng, “a biterm topic model for short texts,” in www, 2013.

[4] E. schubert, m. weiler, and h.-p. kriegel, “signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds,” in kdd, 2014.

[5] P. lee, l. v. lakshmanan, and e. milios, “cast: a context-aware story-teller for streaming social content,” in cikm, 2014.

[6] A. saha and v. sindhwani, “learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization,” in wsdm, 2012.

[7] Y. chen, h. amiri, z. li, and t.-s. chua, “emerging topic detection for organizations from microblogs,” in sigir, 2013.