MINING POSTS AND COMMENTS FROM ONLINE SOCIAL NETWORKS

JongoniSrikant, Dr. Prasadu Peddi, Dr. M. Laxmaiah
Research Scholar, ShriJagdishprasadJhabarmalTibrewala University, Jhunjhunu, Rajasthan.
Assistant Professor, Dept of CSE,ShriJagdishprasadJhabarmalTibrewala University, Jhunjhunu, Rajasthan.
Professor, Dept of CSE and III Cell Head, CMR Engineering College, Kandlakoya, Hyderabad,Telangana.

ABSTRACT: The comment thread as well as article extraction from newspaper's web pages are performed by Python scripts. One script will extract the newspaper article as well as another script that extracts comments related with the story. A comprehensive outline of the steps to be followed during the extraction process is provided. Newspaper websites are dynamic websites in that they collect content from multiple sources. They employ scripts that carry out functions like database information articles, comments on articles hyperlinks to other news sites, links to different news categories or topics as well as. They also populate the page with information from various sources now of visualization. The extraction of posts and comment threads of Web social media (YouTube as well as Facebook) is accomplished with two free online tools. An in-depth description of the steps to follow during the extraction process is presented in this paper.

KEYWORDS: Social Media, Facebook, Youtube, Extractor, Tokenization and stemming.

I. INTRODUCTION

Facebook is a surprisingly simple social network and communication tool that is accessible to all. Studies show that Facebook is used by over 1.7 billion people who use it to connect, connect and communicate with others and to learn about important things to people. Facebook is an amazing instrument for creating business. Communicating through Facebook and other similar social media platforms focus on two key aspects: first, reaching out to as many people as is possible and the second is allowing them to interact with users on a regular basis. Accepting that other channels for advertising at the very least, where users is able to communicate are able to compete with these numbers.

So, many sellers and sellers decide to use social media websites to market their products to remain competitive. Marketing through Facebook assists businesses as well as sellers find new customers and establishing a long-lasting relationship with them. This is reason why every company, regardless of size, must advertise on Facebook. Advertisement on Facebook is a simple method to find and choose those who are important to you, draw their attention, and see results. The people you target can be identified according to demographics, behaviors or contact details. Facebook pages can be a huge value for both Business-to-Customer (B2C) and Business-to-Business (B2B) marketing to display and promote their products or company. It is essential for companies to have a presence on social media specifically on Facebook. Facebook is a fantastic platform to businesses in reaching out to clients from all around the globe. Marketers quickly recognize the importance of promoting their brand making use of Facebook. They implement marketing campaigns using Facebook's

capabilities. Effective marketing strategies aid them to establish a relationship with current customers and draw in new customers.

Facebook as a tool for business is a success because it includes groups and a follower page that can be explored. It is affordable, accessible, and efficient. It's free to set up an account on Facebook. Numerous renowned businesses A significant percentage of their potential customers are on Facebook. So, regardless how big your business, all businesses need to have a Facebook profile to allow them to connect with their customers and receive feedback on their operations and products. It allows business owners and sellers to invite potential customers follow their Facebook pages and let them interact with them, hold contests, organize polls, or any other activity that people want to participate in. The primary goal is to create a community of fans who share positive feedback with their friends. This is ultimately profitable for any company. For many small-sized businesses, a Facebook is an essential part of their business. Businesses are keen on exchanging and disseminating information, transferring products and services, while remaining in contact with current and prospective customers, and gaining a better understanding of their clients through Facebook. Facebook provides a medium that allows marketers to be a part of conversations with their customers. It is a transparent, open and a sociable platform that allows information to be exchanged between consumers and marketers.

Many sellers have realized that Facebook can cut down on the expense of advertising and allows in promoting their businesses at a lower cost. Communication costs have dropped dramatically through Facebook which has opened up opportunities for businesses to speak directly, rapidly and regularly with thousands of potential clients. Facebook is a powerful direct, two-way feedback marketing platforms that can offer quick and precise responses to marketing queries and issues within the shortest period. A significant portion of users have a positive and rational approach to Facebook products that are marketed on Facebook. While Facebook serves as a platform for requirement for information exchange and social media marketing was found as more efficient when it is also enjoyable for the user. It was discovered that Facebook was able to create an awareness of the need and served as an information source, while it influenced consumption among consumers.

Facebook is a central platform for producers and buyers that allows interaction and sharing of content mode. Facebook allows consumers to analyze the products and assists marketers in creating their online image. Brand impressions are created by customers through Facebook. Customers can reach a large number of sellers within a single region and even across the globe. It's helpful for buyers to leave feedback that can assist sellers in improving their the quality of service. Customers will be able to immediately access the details of the Facebook page that has been subscribed online. For instance, what kind of product is being launched by the company, the promotional activity is being in place, where you can purchase the product, etc. can be found easily. People are encouraged to test different products and services based on recommendations from friends on Facebook. This is a clear indication that more internet users spend more time on Facebook as opposed to any other platform or channel. This is a sign of the need for encouraging users to be brand ambassadors.

II. FEATUREEXTRACTION

The initial step is feature extraction is the first step. The process begins with documents which were originally used as inputs. It then it processes these

unstructured documents in order to split them into smaller pieces and identify the most crucial elements that can be found in the document. The result of feature extraction is typically the form of lines in which every single segment of a document is left, and each line has its own distinct characteristics.

Features Extraction is a part of the sub-steps:

- Stop-words removed
- Stemming
- Tokenization

STOP WORDSREMOVING

Stop words that do not transmit information. Stop words are eliminated to remove these non-information-carrying words from archives, and to reduce the background noise.

One of the best characteristics of stop words is the fact that they're extremely popular words. Meanings of phrases remains even after stop words are removed. Most search engines don't record stop words in order to reduce the space required and to improve the results. To reduce massive amounts of data, stopping words may offer similar benefits. First it can reduce the space. Additionally, it assists identify the sounds and retain those words essential to them. This makes the handling of details more productive and efficient.

Stop words are a part of the natural language, and they have distinct stop word list. Examples:

- English A, in it is that me you and he and, in addition, almost before, and following
- Danish: en, et, det, jeg, du, Hun, han, igen, senere

In the end there are three types of stop-words that are generic such as misspelling stop-words, and domain stop-words. Stop-words with a generic meaning could be found in archives, and the other types must be stored in the same location until all archives in the corpus are scanned and the estimates have been done.

GenericStop-Words

Stop words that are generic general non-information-bearing words in a specific language, and can be eliminated without having to think about any specific domain. When it comes to English the majority of the phrases include "a", "an", "the" and so on.

A method to find and eliminate stop-words that are not specifically defined is to create an alphabetized listing of words which could be used to describe the culture of that particular. When you read an article, if the word that is mentioned in the report and is included on the list of stop-words the word can be removed instead of being included in later processing.

Misspelling Stop Words

Stop-words which have been mispelled are not necessarily words but could be spelling mistakes for words. There are some people who use words that aren't included in the lexicons such as, "world," for example "world" is spelled "world". In this instance one could recognize that the spelling error yet be able

to discern the proper meaning of the word. It can take some time for computers to verify the correct spelling, even though search engines can spot incorrect spellings and provide suggestions for correcting it. One method to correct spelling errors is to view them as a last word and eliminate the word for further investigation.

One way to determine the spelling mistake is to analyze the frequency of the word throughout the entire report. In the vast collection of reports that comprises greater than 10,000 pages words that appear in fewer than one case might be spelling mistakes. Certain words that appear several times throughout the entire report aren't actually spelling stop words.

Domain Stop Words

Stop words that refer to domains, typically speaking, don't seem to be popular, but they can become stop words in specific fields of study or in content. For instance, when you examine an archive of documents that comprise documents from different categories like animals, cars and geography along with computers and computers in general "computer" isn't one that could be considered a final word because it's not a common term in any other category and distinguishes computer-related documents from other types of documents like animal or geographical documents. When you take a look at the collection, in which all archives talk about various aspects related to computers such as hardware, software and computer-related software, the words "computer" are not sufficient to be considered part of the final analysis.

STOP WORD ELIMINATION

In the search for facts, the reflections of words are usually of some importance because of their usage in texts. Stop words can affect the results of these checks due to fact that they are usually used whenever possible within the texts. Elimination or removal of any words could be applied to documents being processed before being processed to get rid of the issue. Because these words aren't able to differentiate sentences from each other, the removal of them is crucial for improving the efficiency of the system.

STOP WORDS FILTERING

They are able to gather information and stop word filter is the most common separating method which is a rapid advancement in the method of processing. Stop Word Filters are used to eliminate words that aren't organized or don't contain pertinent information. The first time it was used was in 1958, which was it was the very first instance. Hans Peter Lunn used the term "stop words" to describe words that were not related to keywords in his concept of Keyword-in-Context (KWIC) and also an order method.

Stop words are terms used to create phrases, but they are of no significance. They can create a substantial content division in reports, which causes the report to divide into several parts of the report's contents. The stop-word filter is used to remove the terms from reports. Therefore when an expression is listed into the list, it's created, and can be entered by the person using it or a framework is able

to automatically combine it. Certain methods are suggested to calculate the time for stop-word records.

STEMMING

Stemming is a technique to alter a word the original. In different dialects the syntactic patterns of the words are a representation of the same concept. When we speak of English terminology, words are in plural and singular forms. Verbs employ exhibit or past or plural forms. Words with different spellings could cause problems in the examination of content because they may have different spellings yet they have the same meaning. In English the numerous spellings for"learn. "learn" could mean learned (current condition) studying to become master (showing that the verb is used) as well as learned (past the moment of learning).

TOKENIZATION

The first step in breaking down the basic elements that compose the text is break it down into elements of the document. This can be accomplished using Alexey. Alexey can sort characters into various categories of information. It can then break them down into tokens that generally include words, a method called tokenization. One way to achieve this is to eliminate characters that aren't alphanumeric.

Tokenization is the first step in processing text. The first written text is the outcome of an alphabet of characters. Each process of mining content is connected to the dialogue. Words are the arrangement of significant characters. In this way, it is important to transform these images in words that are suitable to further manipulation. Tokenization is a method to distinguish all words in text.

III. ERT MODEL

The most well-known processing steps that aid in mining text can be combined to create the framework. There are three major steps in the ERT Framework i.e., Expand removal, Tokenization, and Expand.

Figure 3.1 ERT Framework

ExpansionPhase

Corpus includes acronyms, abbreviations Short forms, Polysemy icons, and misspellings. Short text does not provide enough details. When expanding it's mostly focused on locating the script that is smaller, and then expanding the text using the font that is small.

Acronym

An acronym can be defined as a pronouncing word that originates from first letters of the title or name but the term is not used in all cases. It's also known as an acronym that is used in the Language of Internet the following is a list of most frequently used abbreviations (Table 3.1) which are commonly utilized on the internet, as well as in emails. The list isn't exhaustive, but it does contain abbreviations. They can help increase the effectiveness of messages that are text-based. Abbreviations and acronyms are generally written with capital letters.

Table 3.1 Acronym

Acronym	**Equivalent Meaning**
KIT	Keep In Touch
OMG	Oh My God
PM	Private Message
IDK	I Don’t Know
AWOL	Absent Without Leave
MIA	Missing In Action
APB	All-Points Bulletin
TY	Thank You
NP	No Problem
WTG	Way To Go

Shortforms

The short form is used to describe the scenario where the word which is long is substituted with an equivalent shorter word (Table 3.2). There are numerous acronyms that have been discovered and only a few short forms can be utilized in chat conversations among participants. Some short forms are very sensitive to the specifics of chat and discussion.

Table 3.2 Short Form

Short Form	**Equivalent Meaning**
Gr8	Great
Sry	Sorry

Str8	Straight
Btw	Between
U	You
Sth	Something
W8	Wait
B4	Before
K	Okay
L8r	Later
Contrasts	Congratulations
Pls	Please

Polysemy

Polysemy refers to a word that can have multiple meanings (Table 3.3). In contrast the polysemy must be composed of either a word or an extremely small structure.

Table 3.3 Polysemy

Polysemy	**Equivalent Meaning**
Tech	Technology, Technical
Sec	Second, Secondary

Mis–spelling (ChatterSlang)

Misspelling is more prevalent and occurs at a higher speed in chat documents than in standard document with text in discrete format (Table 3.4). This mispelled word causes a many confusion.

Table 3.4 Misspelling

Misspelling	**Equivalent Meaning**
Sssssss	Yes
Hiiiiiii	Hai
Ccccuuuuuu	See You
Okie	Okay

Icons

In dialogues, icons are utilized as images mixed with texts (Table 3.5). On the basis of the graphics pleasing icons, they are able to be separated into two distinct categories, that are text-based and non-text-based. Text icons are

portrayed through similes (graphical texts). Non-text icons include a bit of textual information.

Table 3.5 Icon

Icon	**EquivalentMeaning**
:), :)))))), :)),	Laughing
b-)	Cool
:D, d	Laughing loudly
:|	Straight Face
:(, :((, :(((	Unhappiness
:-?	Thinking

Abbreviation

Abbreviations are the condensed version of an expression or word which is typically used in scripts to indicate the whole structure. The most common way to complete it is by a full-stop. The truncations usually are printed in uppercase. The structure is not large enough to allow for a precise declaration. (Table 3.6).

Table 3.6 Abbreviation

Abbreviation	**EquivalentMeaning**
approx.	Approximate

RemovalPhase

The words in the corpus do not carry a cent percent importance. Certain words have more meaning and some words are useless and meaningless. This method was employed to remove meaningless words from the vast input corpus.

Stemming

A stemmer can examine an image that has been affected by the search table. The benefits of this method is that it's fast and simple, it is easy to regulate exclusion. However, one drawback is that the words that are inflected have to be clearly defined in tables. A language that isn't used but if they're accepted.

Stop word removal

The stop-word catalog is loaded into and studied to answer full-text queries using the character set of the server. False hits or misspellings can be detected for stop word lookups by using Stop word file files, or columns..

Tokenization Phase

Tokenizing can be accomplished by breaking your text in white spaces and punctuation marks, which don't have abbreviations which could have been located in the earlier step.

ERT PROCESS

Documents (or) corpus is used as an input to an ERT framework. The initial step in expanding is to go through the document. It could include acronyms, short form Polysemy Misspelling, polysemy Icons and abbreviations. The expansion alters the text more readable and moves into an elimination stage.

Next, you must remove the suffixes and prefixes of words, in addition to words that aren't keywords. The result of this process includes an index of keywords and the word's roots.

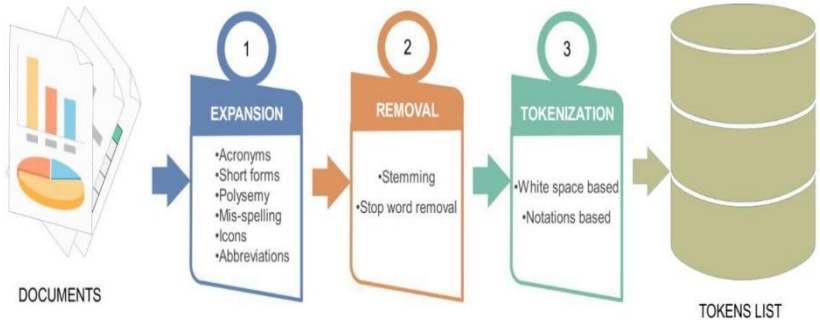

Figure 3.2 ERT Process

The final phase is constantly changing word collections based to the list of words known as marks, and is kept in databases.

The rise of Social Media as News Platforms Merriam-Webster Dictionary defines the term social media to mean: "Forms of electronic communication (such as web sites for social networks and microblogging) that allow people to build online communities for sharing information such as ideas, personal notes, as well as other media (such in the form of video clips)". According to the Oxford Dictionary (2011) defines social media as "websites and applications that are used to facilitate the social network'. In other words, social media refers to Interactive Web based technologies and applications that allow people to form online communities through the sharing ideas and information. Social media help to create relationships that connect individuals of interest across physical borders. As the technology of social media continues to develop, they serve diverse functions and uses. Social Networking Sites (Facebook, Twitter), websites for sharing content (YouTube, Flickr, Instagram), Wikipedia, blogging websites, social bookmarking sites (red it) and more are all part of the social media options of currently. Social media, in actual has brought news-related publishing to the streets and gives the average person the ability to exchange ideas and thoughts with others. Thanks to the generous contributions of users to the internet's resources today, User Generated Content (UGC) represent a highly important resource, despite its limitations. Jacka as well as Scott (2011) describes social media in terms of "a collection of web-based broadcasting technologies that allow the decentralization of content and allows people to move from being consumers of content into publishers'. The increasing impact of social media can be seen from the information provided in this table.

Table 3.7 Frequency paramters

Name of Site	Users in Millions	Frequency of Use			Use as a news platform
Facebook	1871	76	15	7	66
Twitter	317	42	24	33	59
Instagram	600	51	26	22	23
LinkedIn		18	31	51	19

According to the IAMAI-IMRB information from June, Facebook (96%), Google Plus (61%), Twitter (43%) and LinkedIn (24 percent) are the most popular social media platforms in India. The data from IAMAI-IMRB also show that in India the most popular actions on social media websites include: Maintain your profile- 59% Update Status-58% Comment on blogs of others-55 Update self-made videos and music 53 Read blogs, tweets and customer reviews, or watch video. 49 percent of people publish articles, blogs, personal website-42 % Contribute to online forums, modify articles on Wikipedia 40 % Gradually but surely, social media is now a news-related platform that is useful for both the public and journalists. It allows users to share their stories and opinions; while user-generated content (UGCs) provides journalists with an essential source of news to share news stories. Social media platforms offer an essential feature - #hashtag (#), that aids journalists discover what's being discussed by those who are interested in a specific subject. It also provides the latest trends in the news that is being discussed around the world, which assists journalists in making decisions regarding stories that should be covered. Social media platforms, specifically Twitter and Facebook in fact, have transformed journalism by altering the method by which news is gathered and stories are discovered. As the complexity of the social media grow new tools for social media are created to find news leads, verifying facts, sharing stories and generating curiosity. These helpful tools for journalism facilitate the integration of social media into newsrooms, and can ultimately influence the structure as well as the functional and ethical aspects of professional journalism.

Research Design and Methods

The basis of the research is a survey of selected Indian journalists regarding using social networks in their newsmaking process. The study's sample was chosen randomly from journalists who have had prior experience working on the traditional as well as digital news platforms. The research assumed that someone who was familiar with both platforms would better understand the shift. A list of reporters in Delhi as well as NCR was first compiled by the researcher. Thirty of them were chosen for the study. A well-structured questionnaire that contained 10 questions related to the use of social media by journalists was employed as the main instrument for data collection. The questionnaire was distributed to participants who were asked to provide their responses. The responses that were collected are codified and arranged in charts and tables. The results were analyzed using simple statistical tools. The results have been evaluated in relation to the research environment and general conclusions were drawn to allow for more general application of the findings from the research in the sections to are listed below.

IV. Data Analysis and Interpretation

The first question designed to determine the goal of using social media by journalists. Respondents had three options to select from: using social media to fulfill personal goals as well as its use in professional lives, using social media both for professional and personal use. (All tables are provided in appendix, and the analysis is provided below). Table 1 shows that all Journalists who were surveyed use social media for professional and personal purposes. It suggests that Indian media is embracing social media in a massive manner. Table 2 shows that Twitter is used primarily to fulfill professional purposes, which is then followed by Facebook. Twitter is utilized by 87.10 percent Journalists . It is followed by Facebook which is used by 67.74 Percent of Journalists. YouTube is utilized by 41.94 percent of Journalists. The most notable result of research is YouTube is primarily utilized for TV Journalists compared to Print and Web Journalists. Additionally, Print Journalists are less dependent on social media for professional reasons. Web Journalists are heavily using professional social media and they use many social media platforms as well. Table 3 indicates that Facebook is the most popular choice by Journalists who are personally followed by WhatsApp. Facebook is extensively used for 83.87 per cent of journalist, and WhatsApp is employed in 77.42 percent. It's interesting to see it's true that YouTube along with Instagram are more commonly used to do personal work than Twitter. Table 4 indicates that the main purpose of using social media for journalists in their personal lives is to gather information and to establish relationships. Journalists face such a high level of pressure from their jobs that they use social media personal to search for facts.

Very few journalists utilize social media to entertain themselves or to increase their passion. The table 5 shows that journalists utilize the social networks as source of news to locate leads to stories. 96.77 percent of journalists agree that the use of social media in their job to search for information as well as leads to stories. Following this journalists have also endorsed on the importance of sharing their work and sharing links from other sources on social media. The table 6 shows how frequently social media platforms are utilized by journalists, either professionally or personally. They both Facebook and WhatsApp rank high in the list of journalists' tools of choice. 96.77 percentage of reporters are on Facebook. Twitter is not as popular on this list , with a percentage of 87.10 however LinkedIn along with Flickr are the least used social media platforms. The table-7.1 indicates that when it comes to finding opportunities for reporting, Twitter and Facebook are the most popular among Journalists. Additionally, social bookmarking websites are beneficial for journalists. Journalists have accepted they YouTube as well as Instagram are better than WhatsApp to locate leads for stories. Flickr and LinkedIn don't stand a chance. The information in table-3.2 illustrates which social media are most effective in disseminating work by journalists. Facebook is at the top of the social media platforms as does WhatsApp to share the journalistic work.

The table-3.7 illustrates that Twitter is the most used social tool for journalists to keep track of newsmakers. Instagram is second in this regard, closely followed by Facebook. According to table-7.4 journalists mainly use Facebook to share hyperlinks and is followed by WhatsApp followed by YouTube. YouTube Facebook and Twitter are the most popular social media platforms used by Journalists. They are used primarily to find leads for stories. The second use of Facebook is to disseminate the journalism. This is due to fact that Facebook is generally used to fulfill personal needs (Analysis from Table 3). WhatsApp is

also utilized for personal use It's also after Facebook to distribute the journalism. Another use for Twitter is to track newsmakers. Table 8 shows how many news agencies have their own guidelines for social media that are adhered to by journalists. 54.84 per cent of reporters have agreed that their companies have their specific guidelines for social media, and 41.94 percent have denied. It's apparent that more than half of news organizations do not have their own guidelines for social media. The table-9 reveals which journalistic job has been made simpler by making use of social media. It is clear it is 64.52 percentage of reporters are aware that the process of finding information has been made simpler thanks to social media. Distributing News and Networking with people/audiences are the second and third most popular choices.

V. CONCLUSION

The principal goal of this research is to provide an understanding of how much social media's content is embedded in news. The study sought to determine how much news outlets on internet use Facebook and Twitter to provide information. It also examined the features among mainstream and less reliable media that use social media for news gathering. But the study isn't without limitations. For example, we focused on Facebook as well as Twitter in the context of social media platforms. There are numerous other popular websites including Instagram, YouTube, Tumblr etc. We do believe Facebook as well as Twitter is the two most popular social networks that are used today. Our method of identifying quotations and sources is easy, yet precise according to the section on methods. These methods could be improved using machine-learning and natural methods of processing language. We will continue to overcome these limitations, and while doing so, we will also investigate other avenues of research.

REFERENCES

1. *Bettman, JR , Luce, MF, John W &Payne, JW 1998, "Constructive Consumer Choice", Journal of Consumer Research, vol 25, issue. 3, pp. 187-217.*
2. *Besley 2008, "Cyberbullying: An Emerging Threat to the always on", Generation in Canadian Teacher Magazine, 18-20.Gross, vol. 25, no. 6, pp. 633-649.*
3. *Budhiano, Pawar. S., Hole, Y., Bhaskar, M. (2020). Service quality, its importance in marketing and competitive strategies. International Journal of Control and Automation, 13(2s), 27-35.*
4. *Agarwal, S, Sureka, A 2014, "A focused crawler for mining hate and extremism promoting videos on youtube". In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 294-296.*
5. *Akrimi, Y &Khemakhem, R 2012, "What Drive Consumers to Spread the Word in Social Media? Journal of Marketing Research & Case Studies", pp. 1-14.*
6. *Ba, S &Pavlou, PA 2002, "Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior", MIS Quarterly, vol. 26, no. 3, pp. 243-268.*
7. *Scott Jensena, XiaozhongLiub, YingyingYuc& StasaMilojevicd 2016, "Generation of topic evolution trees from heterogeneous bibliographic networks", Journal of Informetrics vol. 10, pp. 606–621.*
8. *Prasadu Peddi (2018), Data sharing Privacy in Mobile cloud using AES, ISSN 2319-1953, volume 7, issue 4.*
9. *Prasadu Peddi (2019), "AN EFFICIENT ANALYSIS OF STOCKS DATA USING MapReduce", ISSN: 1320-0682, Vol 6, issue 1, pp:22-34.*

10. *Kaleel, S. B., &Abhari, A. (2015). Cluster-discovery of twitter messages for eventdetection and trending. Journal of Computational Science, 6, 47-57.*
11. *Liang, W., Shi, Y., Chi, K. T., Liu, J., Wang, Y., & Cui, X. (2009). Comparison of co-occurrence networks of the Chinese and English languages. Physica A: Statistical Mechanics and its Applications, 388(23), 4901-4909.*
12. *Sayyadi, H., Hurst, M., & Maykov, A. (2009, May). Event Detection and Tracking in Social Streams. In ICWSM.*
13. *Ling Hu;Qiang Ni ; Feng Yuan, 2018, "Big data oriented novel background subtraction algorithm for urban surveillance systems", ISSN: 2096-0654, Volume: 1 , Issue: 2 , PP: 137- 145.*
14. *Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121*