ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Word Matching and Tweet Recognition for User Tracking

Mrunali Omprakash Thakare, Prof. Mrudula Nimbarte

M.Tech Student, Department of CSE, B.D.C.E. Sevagram, Wardha, India

Assistant Professor, Department of Computer Engineering, B.D.C.E. Sevagram, Wardha, India

ABSTRACT: Twitter has become one of the most important communication channels with its ability providing the most up-to-date and newsworthy information. Targeted Twitter stream is main usually constructed by filtering tweets and that abused words with predefined selection criteria. Due to its invaluable business value of timely information from these tweets, it's a necessary to understand that abused word's language for a large body of downstream applications, such as named entity recognition (NER), event detecting and summarizing that particular word, opinion mining, sentiment analysis, and etc. In these proposed system, here using two algorithm, String searching algorithm and bubble sort algorithm. Here we used Bubble sort algorithm for fast search. In this application is developed which take tweet is a input and search semantic negative or illegal words from database. Here we using lots of databases as like a negative word database, user's database, post database, comment database. Generate report of those abusing words and send to cyber crime's site. Then depending on the tweet, track the related information of the person. Track all the tweets of that particular person and track that person through identification databases (IP addresses). Then take a action by cyber crime. And after that prevent the tweet we show that exactly achieved in named entity recognition by applying some algorithm.

KEYWORDS: Twitter stream, tweet segmentation, named entity recognition, Database of abused words, Wikipedia.

I. INTRODUCTION

Status messages posted on social media websites such as facebook and twitter present a new and challenging style of text for language technology due to their noisy and informal nature. Social media is an internet-based form of communication. Social media platforms allow user's conversations, share their information and create web contents. There are many types of social media, including blogs, micro-blogs, wikis, social networking sites, photo-sharing sites, instant messaging, video-sharing sites, podcasts, widgets, virtual worlds etc. All people a rounding the world use social media to exchange their information or views and make connections. On a personal level, social media allows you to communication with friends and their family, learn a lot of things, develop your interests, and be entertained. On a professional level, they can use social media's to broaden their knowledge in a particular field and build their professional network by connecting with other professionals in their industry. At the company level, social media allows to have a conversation with their audience, customer's feedback and raise your brand. Here we are using two algorithms. First is string searching algorithm and second is bubble sort algorithm. First, in string search algorithm, string searching is an important component of many problems, including text editing, data retrieval, and symbol manipulation. Despite the use of indices for searching large amounts of text, string searching may help in an information retrieval system. For example, it may be used for filtering of potential matches or for searching retrieval terms that will be highlighted in the output. The string searching or string matching problem consists of finding all occurrences (or the first occurrence) of a pattern in a text, where the pattern and the text are strings over some alphabet. We are interested in reporting all the occurrences. It is well known that to search for a pattern of length m in a text of length n (where $n > m$) the search time is 0(n) in the worst case (for fixed m). Moreover, in the worst case, at least $n - m + 1$ characters must be inspected. This result is due to rivest (1977). However, for different algorithms the constant in the linear term can be very different. For example, in the worst case, the constant multiple in the naive algorithm is m, whereas for the knuth-morris-pratt (1977) algorithm it is two. We present the most important algorithms for string

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

matching: the naive or brute force algorithm, the knuth-morris-pratt (1977) algorithm, different variants of the boyer-moore (1977) algorithm, the shift-or algorithm from baeza-yates and gonnet (1989), and the karp-rabin (1987) algorithm, which is probabilistic. Experimental results for random text and one sample of english text are included. We also survey the main theoretical results for each algorithm. Although we only cover string searching, references for related problems are given. We use the c programming language described by kernighan and ritchie (1978) to present our algorithms. Second is bubble sorting which is used for fastest searching. Bubble sort is a sorting algorithm (duh!), which essentially means that it is an algorithm used to take an unordered list and to put them into a certain order. For learning purposes, this typically means taking a list of numbers and sorting them in non-decreasing order. For example, you could take the list 5, 4, 2, 3, 1, 0 and after sorting it you would get 0, 1, 2, 3, 4, 5. We say "non-decreasing" order instead of "increasing" order because the former is strictly true, while the latter is not. For example, the list 1, 1, 2, 2, 3 is non-decreasing because no number is followed by a number less than it, but not every number is followed by a number that is greater than it. This might feel like nit-picking, but you are likely going to run across both terms and i wanted to explain the difference ahead of time. That said, not all sorting deals with numbers, and it most definitely doesn't always consist of numbers in increasing order. These are mostly just used for teaching, and then in the practice problems i will include a few problems that do not involve sorting numbers in increasing order. Bubble sort isn't the most efficient sorting algorithm, but it does have two primary benefits over many other sorting algorithms. First, it is one of the simplest sorting algorithms to implement, especially for a beginner, so it serves as a very handy learning tool and as a simple way to sort relatively small lists[1]. Second, we can easily detect when a list is already sorted with bubble sort and terminate our code early, making it useful when we are provided with a sorted list but do not know ahead of time that the list is sorted.

How does it work?

Bubble sort works by walking through a list of numbers and comparing each consecutive pair of numbers. When we compare the numbers we will ask "is the first number greater than the second number?" and if the answer is yes, we swap the position of each number in the pair. Regardless of whether or not we swap the numbers, we then move on to the next consecutive pair in the list and repeat the comparison and potential swap steps [2], [5], [7].

Database :

Databases rank among the most significant structural elements of the World Wide Web today. Lying in the basis of the majority of the content-driven websites and applications, databases serve a special mission - to provide a well-organized mechanism for data manipulation. The database approach in website/application development now rules the web by offering a quick and automated way for the information to be stored, managed, deleted or retrieved. Databases' powerful set of capabilities has determined the introduction of dynamic websites, which has opened a new page in the evolution history of the web. The communication between databases and computer programs working with them is executed through a database management system.. All these actions are executed through specific sql commands. Users can also easily add new data categories or attributes to the database without causing any system interruptions. Database management systems work with all basic database models available such as the network model and the relational model. Due to the databases fundamental role in running dynamic websites the database approach is used on practically every new website appearing on the world wide web today. On commercial websites, for example, database are used to stored and manage various data such as visitors log-in information, purchase details, order log, company reports, pricing schemes, etc. In this paper we are using a lots of databases as like a abuse word database, user database, comment database, post database for automatically stored the whole information about the particular person and that person's activity [1] [3].

II. RELATED WORK

1. PROPOSED WORK:

In these proposed system application is develop which take tweet is an input and search semantic negative words from database. Depending on the result track the related information of the person. Track all the tweets of that particular person and track that person through identification databases. Tweet Search Application is the software or may be web application which contain text analyser module where the application fetch the tweets and gives to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

analyser module. Firstly, user has to register on a website and after the registration log in the account and that user's information stored in a database. Fetcher fetch the user's tweet from their chatting then segment the data with the help of n-gram algorithm then processing on a data and search that abusing, terror or negative words from a database which is already saved in a database. Then that word gives to the text analyser then Text analyzer module highlight that word from tweets and track the person's information that is responsible for that tweet from database through the IP addresses and send the warning message. As shown in below figure.

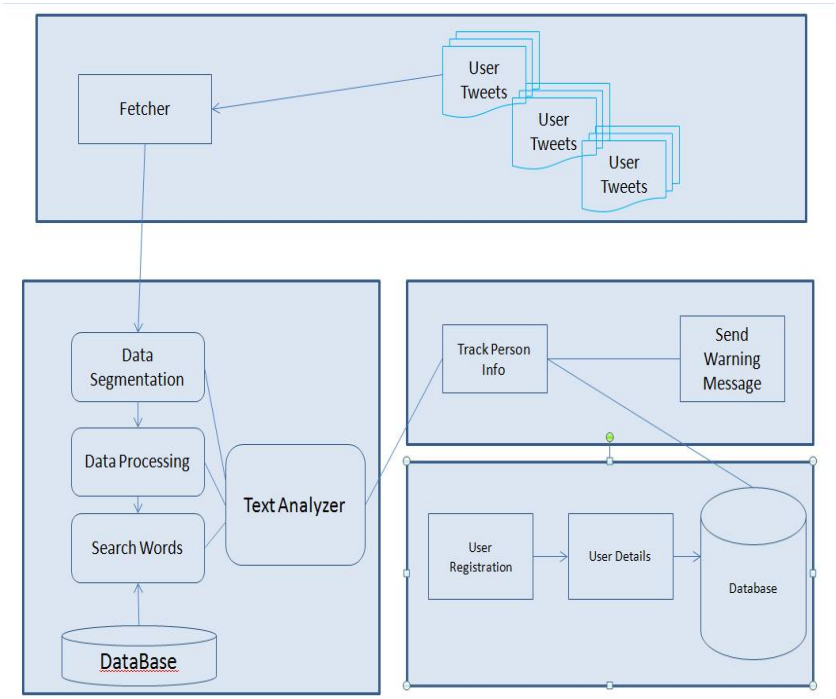

Fig 1: Proposed architecture of user tracking using tweet segmentation and word recognition

N-grams algorithm:

An n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes as like bad, pad, etc, part of a word, letters, words or base pairs according to the application. The n-grams typically are sorted a words from a text or speech. Two benefits of n-gram algorithm are simplicity and scalability with larger n, and model can store more contexts with a well understood space time trade off.

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$$

$$P(w_8|w_1^{8-1}) \qquad P(w_8|w_{8-2+1}^{8-1})$$

E.g. for bigrams,

$P(w_{n-1},w_n) = P(w_{in} \mid w_{n-1})\, P(w_{n-1})$

$P(w_{8-1},w_8) = P(w_8 \mid w_7)\, P(w_7)$

By the Chain rule we can decay a joint this probability, e.g. $P(w_1,w_2,w_3)$ as follows

$P(w_1,w_2,\ldots,w_n) = P(w_1|w_2,w_3,\ldots,w_n)\, P(w_2|w_3,\ldots,w_n) \ldots P(w_{n-1}|w_n)\, P(w_n)$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

To Start the process, firstly user registered on a website and that user's data saved in a database, then here fetcher fetch the users tweet, segment the particular that tweet with the help of n-gram algorithm then processing on a tweet, if found the negative or abusing word from the database then save the tweet in a database then after that track the person who is the responsible for that tweet from the database and if not found then fetch the another tweet. And after that information found in a database then send warning message to that person but if not found information then track information using IP address

III. PROPOSED ALGORITHM

1. **String searching algorithm:**

In computer science, string searching algorithms, sometimes called string matching algorithms, are an important class of string algorithms that try to find a place where one or several strings (also called patterns) are found within a larger string or text. Let Σ be an alphabet (finite set). Formally, both the pattern and searched text are vectors of elements of Σ. The Σ may be a usual human alphabet.

```
Function brute_force(text[], pattern[])

    {
    // let n be the size of the text and m the size of the
    // pattern

    for(i = 0; i < n; i++) {
    for(j = 0; j < m && i + j < n; j++)
    if(text[i + j] != pattern[j]) break;
    // mismatch found, break the inner loop
    if(j == m) // match found
    }
    }
```

The "naive" approach is easy to understand and implement but it can be too slow in some cases. If the length of the text is n and the length of the pattern m, in the worst case it may take as much as (n * m) iterations to complete the task

IV. EXPERIMENTAL ANALYSIS AND RESULT

Following graph shows the performance measurement of Existing System and Proposed System. In Existing System only the tweet recognition and user prevention can be 75% done whereas the proposed work 95% work on tweet recognition, user identification and tweet prevention. In proposed work the identification of abuse word and the abuse word user must be detected string searing and N-gram algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

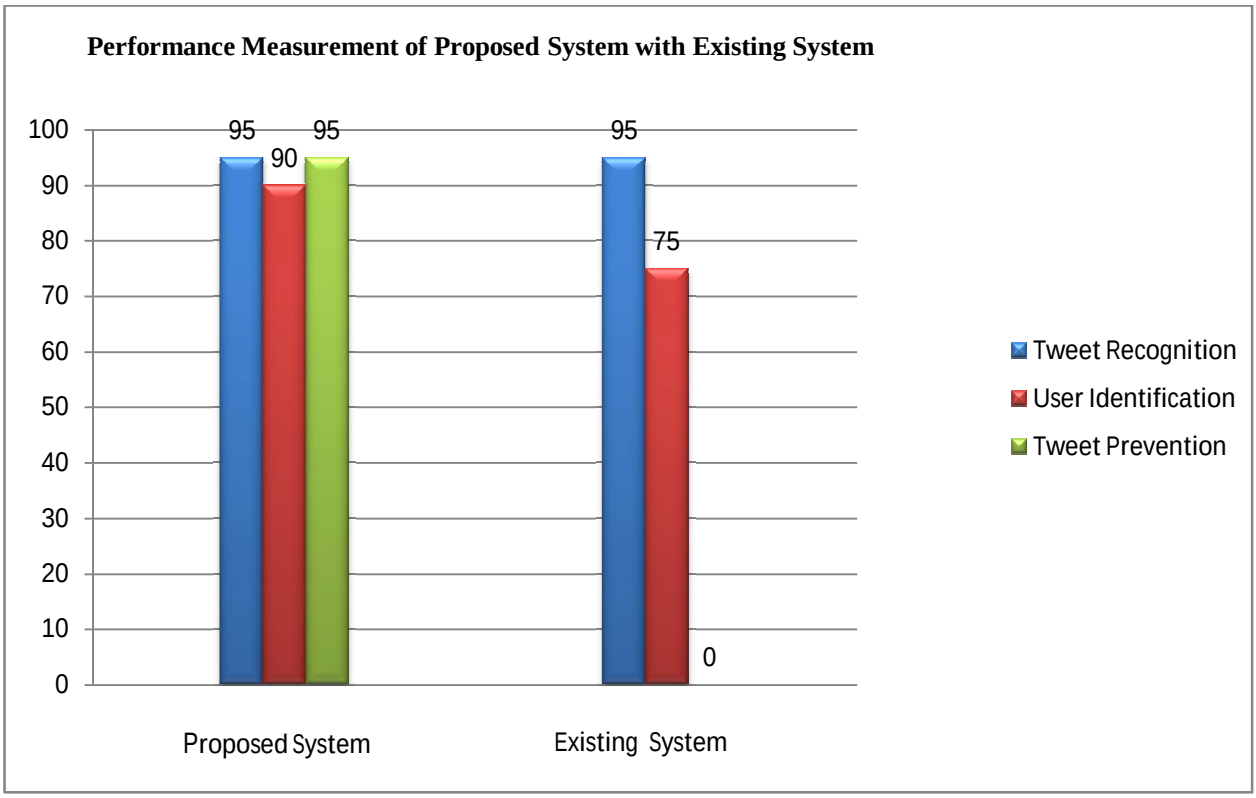

Figure 2 : Graph of performance measurement of proposed system with existing system.

V. CONCLUSION

In this paper, we are using two algorithms. First is String searching algorithm is used for trying to find a place where one or several strings are found within a larger string or text. Second is Bubble sort algorithm is used for fastest search. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g., named entity recognition. This system enables us to develop advanced social medium which provides a way to restrict from writing fake and abuses tweets. Stop to make issues on social media and prevent the tweet. Just like as shown in below.

REFERENCES

[1]. Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, Members, "Tweet Segmentation and Its Application to Named Entity Recognition" in IEEE February 2015.
[2]. Chetan Chavan, Prof. RanjeetsinghSuryawanshi, "Tweet Segmentation andNamed Entity Recognition", IJSART –Volume: 2, Issue: 1, JANUARY 2016.
[3]. Anuja A.Thete, Prof. J. S. Karnewar, "A Review Paper on Tweet segmentation and its Application to Named Entity Recognition", Volume: 4, Issue: 1, January 2016.
[4]. Akshay Shinde, Sachin Yelmar, "International Journal of Innovative Research in Computer and Communication Engineering", Vol. 4, Issue 4, April 2016.
[5]. Vijay Choure, Kavita Mahajan, "Twitter Segment NER-Tweet Segmentation Using Named Entity Recognition", Volume 2 Issue2, Apr 2016.
[6]. Rashmi Rachh, Vanaja H Kulkarni, "Segmentation of Tweets for Multilingual Named Entity Recognition",Vol-2, Issue-ISSN: 2454-1362, Sept-2016.
[7].Deniz Karatay, Pinar Karagoz, "User InterestModeling in Twitter with Named Entity Recognition",Vol-1395,
[8]. April 2015. Alan Ritter, Sam Clark, "Named EntityRecognition in Tweets : An Experimental Study", June2015.
[9]. C. Li, J. Weng, Q. H Y.Yao, A. Datta, A. Sun, and B.-S Lee,"Twiner: Named entity recognition in targeted twitter stream," in SIGIR,2012, pp. 721–730.
[10]. X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, " Exacting social events for tweets using a factor graph", in AAAI Volume No. 2 , 2012.