Search optimization of named entities from twitter streams

Search optimization of named entities from twitter streams

Mohammed Fazeel K, Simama Hassan Mottur , Jasmine Norman and Mangayarkarasi R

School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.

E-mail: jasmine@vit.ac.in

Abstract. With Enormous number of tweets, People often face difficulty to get exact information about those tweets. One of the approach followed for getting information about those tweets via Google .There is not any accuracy tool developed for search optimization and as well as getting information about those tweets. So, this system contains the search optimization and functionalities for getting information about those tweets. Another problem faced here are the tweets that contains grammatical errors, misspellings, non-standard abbreviations, and meaningless capitalization. So, these problems can be eliminated by the use of this tool. Lot of time can be saved and as well as by the use of efficient search optimization each information about those particular tweets can be obtained.

1.Introduction

Twitter is one of the most used social media by celebrities, business tycoons all over the world. Twitter is used by most of the people in the world, a lot of information's, posts, news and interactions are made by celebrities, business tycoons. The administration quickly increases world-wide prominence. Twitter tweet's are limited to only 140 characters strings. Celebrities, business tycoons and fans are being interacted using tweets personally by their respective user names. Twitter's tweets are trending these days as tweets of celebrities and business tycoons are shown in the news now days. Twitter's tweets post by using #tags. If the users of the twitter tweet about the several topics and issues consistently, the tweets become trending all over the world. So that, everyone knows about the most discussed topic across the globe.

2.Literature Survey

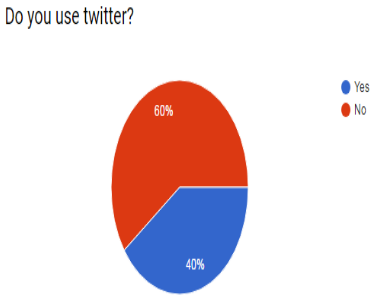


Figure. 1 Survey result of first Question

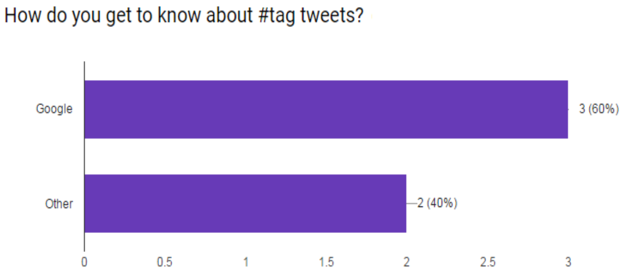


Figure. 2 Survey result of Second Question

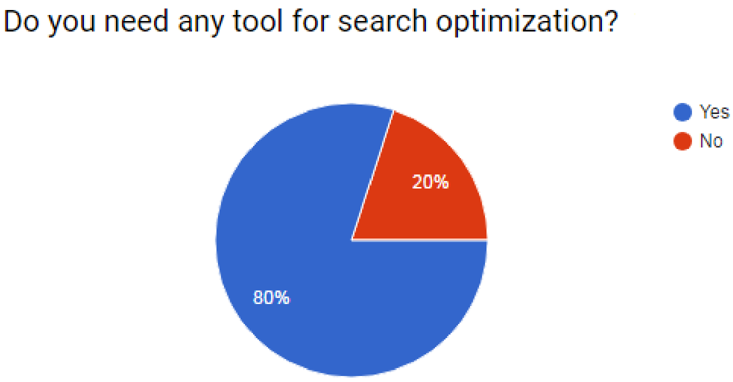


Figure. 3 Survey result of third Question

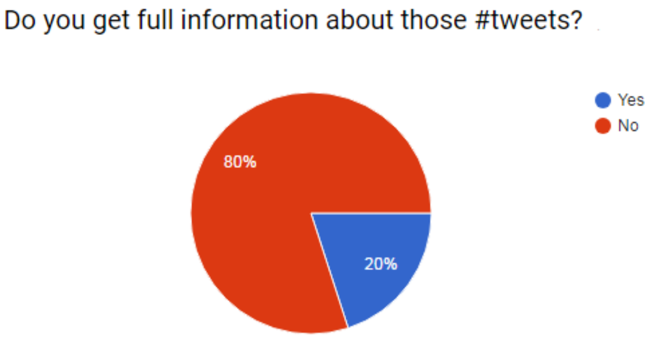


Figure. 4 Survey result of fouth Question

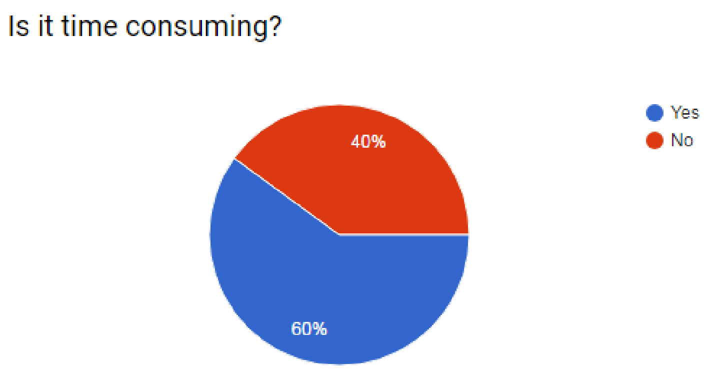


Figure. 5 Survey result of fifth Question

From the referred papers gain the information of normalization of social media verbal[4] ,conversion of grammatical mistake text to proper English[2] ,able to get the knowledge of removal of confusion from context, recognition of named entity from a context[6] and way of twitter normalization with the concept of 1 to n[5] ,ways of classifying tweets[1]. In the Existing System , there are several tools where normal text is disambiguated its accuracy is around 90% because it contains without any grammatical errors, but for tweets but its accuracy lies between 40-50% because tweets contain Accents, Abbreviation, Misspelled words and extra special characters which reduces the accuracy. Search Optimization is not efficient.

3.Proposed System

The proposed system consists of Search optimization, .In which tweets areconverted into information.By, which a lot of tweets accuracy can be improved. And detailed Information can be acquired with.in a short span of time. Including that, filtration of grammatical errors, Misspelled words, non- standard abbreviations and meaningless capitalizations are done. A lot can time be saved.

Methodologies

In the initial part tweets are retrieved and then normalization of tweets are done.Take each entity link it with Wikipedia to disambiguate, from there we could get the entire Details about the given nouns in the tweets ,by following this method we are saving the Time and providing an effective search operation for the twitter Streams.
It contains four phases: Data collection, pre-processing of tweets and normalization Valuation

3.1Data Collection

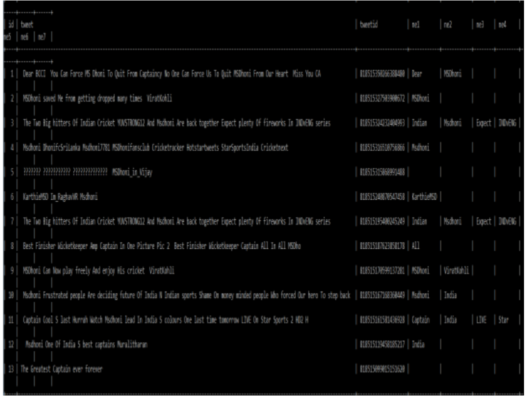
- Create an api in Twitter.
- Generate Tokens
- Make Use of Twitter package to fetch the tweets.
- Retrieve tweets from Twitter database.
- Store in local database for future Usage and Remove Extra Fields.

3.2Preprocessing of Tweets And Normalization

[4]By Using the Regular Expression we negate the unnormalised tweets
removal of urls
removal of non-alphanumeric characters
removal of hash tags
removal of retweet tags

3.3 Use Of NTLK

With the help of the NTLK package (Python), the normalized tweets are preprocessed
To extract the name entities.
[2]The extracted Entities are stored in a Separate table



```
+----+------+-----+
| id | tweet                                                                                                   | tweetid            | ne1       | ne2       | ne3    | ne4    |
ne5 | ne6 | ne7 |
+----+------+-----+
| 1 | Dear BCCI  You Can Force MS Dhoni To Quit From Captaincy No One Can Force Us To Quit MSDhoni From Our Heart  Miss You CA | 818515350266388480 | Dear      | MSDhoni   |        |        |
    |     |     |
| 2 | MSDhoni saved Me from getting dropped many times  ViratKohli                                        | 818515327503900672 | MSDhoni   |           |        |        |
    |     |     |
| 3 | The Two Big hitters Of Indian Cricket YUVSTRONG12 And Msdhoni Are back together Expect plenty Of fireworks In INDvENG series | 818515324332404993 | Indian    | Msdhoni   | Expect | INDvENG |
    |     |     |
| 4 | Msdhoni DhonifcSriLanka Msdhoni7781 MSDhonifansclub Cricketracker Hotstartweets StarSportsIndia Cricketnext | 818515316510756866 | Msdhoni   |           |        |        |
    |     |     |
| 5 | ??????? ??????????? ?????????????  MSDhoni_in_Vijay                                                  | 818515315868991488 |           |           |        |        |
    |     |     |
| 6 | KarthieMSD Im_RaghavNR Msdhoni                                                                       | 818515240870547458 | KarthieMSD |          |        |        |
    |     |     |
| 7 | The Two Big hitters Of Indian Cricket YUVSTRONG12 And Msdhoni Are back together Expect plenty Of fireworks In INDvENG series | 818515195400245249 | Indian    | Msdhoni   | Expect | INDvENG |
    |     |     |
| 8 | Best Finisher Wicketkeeper Amp Captain In One Picture Pic 2  Best Finisher Wicketkeeper Captain All In All MSDho | 818515187623858178 | All       |           |        |        |
    |     |     |
| 9 | MSDhoni Can Now play freely And enjoy His cricket  ViratKohli                                       | 818515170599137281 | MSDhoni   | ViratKohli |       |        |
    |     |     |
| 10 | Msdhoni Frustrated people Are deciding future Of India N Indian sports Shame On money minded people Who forced Our hero To step back | 818515167168360449 | Msdhoni   | India     |        |        |
    |     |     |
| 11 | Captain Cool S last Hurrah Watch Msdhoni lead In India S colours One last time tomorrow LIVE On Star Sports 2 HD2 H | 818515161581436928 | Captain   | India     | LIVE   | Star   |
    |     |     |
| 12 |  Msdhoni One Of India S best captains Muralitharan                                                  | 818515119458185217 | India     |           |        |        |
    |     |     |
| 13 | The Greatest Captain ever forever                                                                    | 818515099015151620 |           |           |        |        |
    |     |     |
+----+------+-----+
```

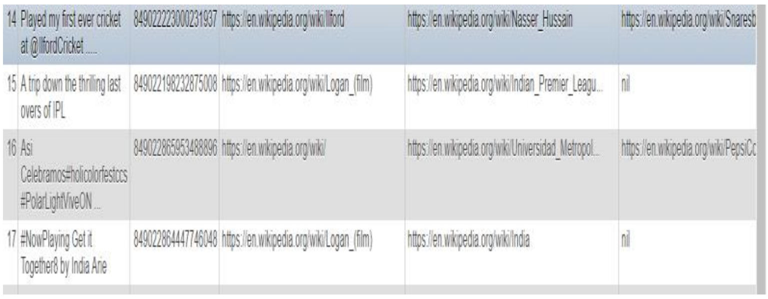
Figure. 6 NTLK

3.4 Valuation

- Take the named Entities.

3.5 Linking

- Take each entity
- Link it with Wikipedia to disambiguate.



14	Played my first ever cricket at @IlfordCricket	849022223000231937	https://en.wikipedia.org/wiki/Ilford	https://en.wikipedia.org/wiki/Nasser_Hussain	https://en.wikipedia.org/wiki/Snaresb
15	A trip down the thrilling last overs of IPL	849022198232875008	https://en.wikipedia.org/wiki/Logan_(film)	https://en.wikipedia.org/wiki/Indian_Premier_Leagu...	nil
16	Asi Celebramos#holicolorfestccs #PolarLightViveON ...	849022865953488896	https://en.wikipedia.org/wiki/	https://en.wikipedia.org/wiki/Universidad_Metropol...	https://en.wikipedia.org/wiki/PepsiCc
17	#NowPlaying Get it Together8 by India Arie	849022864447746048	https://en.wikipedia.org/wiki/Logan_(film)	https://en.wikipedia.org/wiki/India	nil

Figure. 7 Linking

4.Conclusion

As the data's are very large in twitter doing search optimization plays a major role in order to improve the accuracy rate. By the increase of accuracy rate, Detailed Information can be obtained. This can have a lot of impact in the years to come. Users don't need to go through Search engines every time. By using search optimization tool, accuracy of tweets information can be improved. As well as. People will be able to access the quality factors like user-friendly, Reliability.   .

References

[1] David Harrison Building a taxonomy of tweet types and automatically classifying tweets in to these types

[2] Neelmay Desaia and Narvekarb 2015 Normalization of Noisy Text Data *International Conference on Advanced Computing Technologies and Applications ICACTA*

[3] DrlimlianTze 2015 A Practical Introduction to natural language Processing

[4] Eleanor Clarka and kenji Arakia 2014 Text Normalization in social Media:Progress,problems and Applications for a preprocessing System of casual English *Pacific Association for computational Linguistics (PACLING)*

[5] Yafeng ren , Jiayuan Deng and Donghong 2016 Twitter Normalization via 1 to N Recovering *Springer International Publishing AG*

[6] Rizzo G and Trancy R NERD:A framework for evaluating Named Entity Recognition Tools in the web of data ISWC’11

[7] Gurpreet Singh Khanuja and Sachin Yadav 2013 Normalisation of SMS Text IIT Kanpur Conference of NLP 2013

[8] Catherine Kobus 2008 Normalizing SMS: are two metaphors better than one