Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2301.09

2023, 42(1) 88-94

Comparative analysis of TF-IDF and loglikelihood method for keywords extraction of twitter data

Muhammad Adeel Abid [a], Muhammad Faheem Mushtaq [b,*], Urooj Akram [b], Mateen Ahmed Abbasi [a], Furqan Rustam [a]

[a] *Faculty of Information Technology, Khwaja Fareed University of Engineering and Information Technology, 64200 Rahim Yar Khan Pakistan*

[b] *Department of Artificial Intelligence, The Islamia University of Bahawalpur, 63100 Bahawalpur Pakistan*

[*] Corresponding author: Muhammad Faheem Mushtaq, Email: faheem.mushtaq@iub.edu.pk

Received: 20 October 2020, Accepted: 15 December 2022, Published: 01 January 2023

K E Y W O R D S

Twitter

Social Media

Classification

Loglikelihood Methods

Term Frequency-Inverse Document Frequency

A B S T R A C T

Twitter has become the foremost standard of social media in today's world. Over 335 million users are online monthly, and near about 80% are accessing it through their mobiles. Further, Twitter is now supporting 35+ which enhance its usage too much. It facilitates people having different languages. Near about 21% of the total users are from US and 79% of total users are outside of US. A tweet is restricted to a hundred and forty characters; hence it contains such information which is more concise and much valuable. Due to its usage, it is estimated that five hundred million tweets are sent per day by different categories of people including teacher, students, celebrities, officers, musician, etc. So, there is a huge amount of data that is increasing on a daily basis that need to be categorized. The important key feature is to find the keywords in the huge data that is helpful for identifying a twitter for classification. For this purpose, Term Frequency-Inverse Document Frequency (TF-IDF) and Loglikelihood methods are chosen for keywords extracted from the music field and perform a comparative analysis on both results. In the end, relevance is performed from 5 users so that finally we can take a decision to make assumption on the basis of experiments that which method is best. This analysis is much valuable because it gives a more accurate estimation which method's results are more reliable.

1. Introduction

Twitter is the most widely used a social network that connects people with different ideas, opinion, and taste. Organizations, people, groups, schools, colleges, universities, celebrities, musicians begin a twitter account to share views, ideas, events, and news related to them and share it with the social network [1]. Twitter is also used by different classes i.e. showbiz personalities, doctors, singers, researchers, sportsmen, politicians, singers, etc [2]. Twitter consists of only 140 characters that make it most interesting because its length is short and contains more valuable contents. Now, as the number of mobile phone users is increased so there are large numbers of users that accessed it through mobile. The Twitter official also admit that near about 80% of accounts are accessed through mobile.

Most of the users used their Twitter accounts on frequently basis that Twitter officials told that near about 500 million tweets are sent in a day. Twitter is also making a large influence in business because it is widely used to make a strong relationship partners and customers. Twitter is additionally a platform between audience, viewers, media and TV organizations. Fact that were recorded in March 2019 which illustrates that the 335 million users are active in a single month [3]. Near about 80% of users use their mobile phones to access the twitter account. The popularity of Twitter is estimated from this fact that near about 500 million tweets are recorded to send in a single day. Furthermore, the Tweeter supports 35+ languages that contain additional feature. The most interesting fact is that only 21% of users are from the US, the remaining 79% are from outside the US.

When a large number of populations of the world keeps going on posting data on Twitter on a daily basis. So, we have a huge amount of data [4] related to news, user's views, foods, sports events and even have a bird's eye view on trending things [5]. People related to different areas of life post tweets on a daily basis and there is a bulk of data [6] that would be needed to categorize interm of interest or advertisement purpose [7, 8]. To do this desired people exactly want to extract keywords so they can take benefit from it [9,10]. In this paper, the Term Frequency-Inverse Document Frequency (TF-IDF) [11] and Loglikelihood [12] methods are used for keywords extraction. This research essentially focused on the analytical analysis of both methods and evaluates [13] which method is more suitable for keywords extractions.

This rest of the paper is illustrated as follows: Section II describes some related work which incorporates the techniques associated with the extraction of keywords. Section III explains the gathering of data from Twitter and describes TF-IDF and Loglikelihood methodology. Section IV discusses the results and discussion of planned methodology. Section V presents the conclusion and future work of this analysis.

2. Related Work

In the decade, micro blogging website became famous and get more attention to the peoples. Most of the people use it on daily basis as well as multiple times in a day. Twitter have near about 335 million users. Due to its daily usage of twitter, a large volumn of data is produced. It attracts many researcher to check, investigate and apply multiple methods in order to analyze the data. Twitter is considered as a large source of data [14], [15]. The first tweet are extracted with the search that was made in First Story Detection (FSD) technique [16]. Different researchers uses local sentive hashing for the first time to make FSD [17]. The aim of this approach is minimize the number of comparisons the find the nearest neighbor among tweets. Recently, the authors performed a detection on streaming data with an application on twitter by using LSH [18]. This research proposes a new reduction strategy which focus that query is only compared with previous few points. Most of the researchers uses TF-IDF in finding match in twitter [18]–[21]. In 2012, Vogiatizis [22] works on improving the speed and scalability of FSD using distributed approach which can process a large amount of tweets and requires very high power of computation in real time. Word2vec is the efficient method that deals with the continuous vector representation of words in unsupervised computation [23]. In 2014, Leand Mikolov [24] introduces a new model called paragraph vector that is based on Word2vec model to create an algorithm for fixed length text representation from variable length text. In 2013 and 2014, BM25 algorithm is used for the improvement of results to be more accurate in [25], [26]. Latest search in 2016 also focus on enhancement of its effectiveness for Word2vec [27]. A new method for summarization of news that is based on 3-nearest clustering which is most effective than a baseline that uses dissimilarity of individual document from its nearest neighbor [28]. Most of the approaches based on TF-IDF that convert each tweet as vector and after that make it LSH. TF-IDF is most simple approach that gives more accurate result. Besides TF-IDF there is another method named "loglikelihood" that is also useful in extracting keywords from a corpus by comparing it with a general corpus.

The term "likelihood" refers to a mathematical function that was proposed by Ronald [29]. It introduces a new term that is known as likelihood interval But later it said to as a method of maximum likelihood. The author mentioned the concept that likelihood should not be mixed with probability. He told that likelihood is treated as a sort of probability. The first result that were obtained is different measures of rational belief is about different cases. It can be express in terms of probability to understand the details of population. In simple words, it can be express in terms of likelihood if population is known. The statistical likelihood method is invented in the response of earlier form of reasoning that are known as inverse probability [30].

3. Experimental Setup

Twitter APIs are used to collect tweets data only for the music, then the TF-IDF and Loglikelihood method are used to extract keywords and compare the results.

3.1 Data Collection

Twitter API is used for the purpose of collection of tweets data and downloaded. A Java Program [31] is used to collect the tweet online through twitter API [32] and insert into a database that built-in MYSQL. The main focus on the collection is Music area tweets. As far as quantitative details of data collected are concerned. Total tweets that are collected are 691604 occupying a size of 288mb on Hard Disk. 353783 tweets are in the English Language. For our Music area, only 21000 tweets are selected.

3.1.1 TF-IDF method

The Term Frequency-Inverse Document Frequency (TF-IDF) defines the terms occurrence with higher frequency and are minimum in a document. With the combination of both techniques, it becomes a strong method of identifying keywords that are of great value. For the computation of TF-IDF [11] from TFij, the Term Frequency (TF) of *i* terms from the document that are found in domain *j* and Inverse Document Frequency (IDF) is calculated. This method focuses on ranking keywords according to the frequency or occurrence. The calculation of TF-IDF from TF*ij*, the frequency of term *i* in documents *j* is described in Eq. 1.

$$TF_{ij} = \frac{n_{ij}}{\sum_k nkj} \tag{1}$$

The IDF_i (Inverse document frequency) is obtained by calculating the total number of documents divide by document in which that term occurs. IDF gives more importance to that term that is found rare in the given domain. The calculation illustrated using Eq. 2.

$$IDF_i = \log\left(\frac{|D|}{|\{d : t_i \in d\}|}\right) \tag{2}$$

Where |D| is considered a set of all domains and t_i is considered as the term.

Finally, the Term Frequency-Inverse Document Frequency (TF-IDF) is obtained by multiply term frequency (TF) with corresponding Inverse document frequency (IDF).

$$TF - IDF_i = TF_{ij} \times IDF_i \tag{3}$$

3.1.2 LogLikelihood method

This method describes the comparison of domain corpora with domain-independent data. This method is used to determine keywords in the corpora which differentiate one from another [12]. This method is also significant for extract keywords of the given document with the respect to domain-independent data. The LogLikelihood method is simple and easy to apply as shown in Eq. 4.

$$G = 2 \times \left[\left[freq_{domain} \times \log\left[\frac{freq_{domain}}{freq\_Expected_{domain}} \right] \right] + \left[freq_{general} \log\left[\frac{freq_{general}}{freq\_Expected_{general}} \right] \right] \right] \tag{4}$$

where $freq_{domain}$ and $freq_{general}$ are considered to be the actual frequencies in the domain corpus and in the reference corpus. $freq\_Expected_{domain}$ and $freq\_Expected_{general}$ are expected frequencies in the domain corpus and reference or another corpus. The $freq\_Expected_{domain}$ and $freq\_Expected_{general}$ are calculated using Eq. 5 and 6.

$$freq\_Expected_{domain} = size_{domain} \times \frac{freq_{domain} + freq_{general}}{size_{domain} + size_{general}} \tag{5}$$

$$freq\_Expected_{general} = size_{general} \times \frac{freq_{domain} + freq_{general}}{size_{domain} + size_{general}} \tag{6}$$

Furthermore, Eq. 4 is not able to distinguish between the corpus of two domains. Eq. 4 consider being worked as symmetrical for both corpuses. So, it necessary to correct this situation with another Eq. 7, i.e. their relative frequency is much larger in the domain corpus as compared to the reference corpus. If this condition is false, then we will certainly discard the word as a possible term: in our case, we multiply its weight by -1.

$$\frac{freq_{domain}}{size_{domain}} > \frac{freq_{general}}{size_{general}} \tag{7}$$

4. Results and Discussion

Term Frequency-Inverse Document Frequency (TF-IDF) and Loglikelihood method are applied for the extraction of keywords. Following are the result that we obtained from both methods.

TF-IDF Result
5sos , Added , album , am , amazing , art , artist , Ashton5SOS , ass , awesome , band , bands , beautiful , Boy , Brothers , Buy , Calum5SOS , cant , check , Chocolate , dance , day , days , de , dont , Download , eating , Facebook , favorite , feel ,

Festival , Follow , Format , free , ft , gonna , guys , Happy , hear , Heard , help , hiphop , History , hot , im , iTunes , Ive , Life , liked , listen , Listening , live , lol , looking , love , Luke5SOS , lyrics , makes , Michael5SOS , movie , MP3 , music , name , night , NowPlaying , official , oh , omg , people , photos , pizza , Play , playing , playlist , Please , Posted , Radio , real , release , Rock , shit , SING , singing , single , song , songs , Summer , Thats , time , Top , tweet , ur , via , video , World , youre , YouTube, Youve

Loglikelihood Result

5sos , Added , album , art , artist , Artists , Ashton5SOS , ass , band , bandlads , bands , beatles , bld1 , Brothers , Calum5SOS , chances , check , Cimorelliband , Concert , Cover , dance , DaniCim please , dj , Download , Ed , edsheeran , Facebook , fav , favorite , Favourite , feat , Festival , Fivemember , Follow , Format , ft , FUNK , Guitar , guys , hear , hiphop , hot , indie , is perfect , iTunes , Jazz , KathCim , King , liked , listen , listened , Listening , live , lol , love , Luke5SOS , lyrics , Michael5SOS , MIXTAPE , movie , MP3 , Music From , music , NowPlaying , official , partied , photos , Pitbull , pitbull , pizza , playing , playlist , pop , Posted , Radio , rap , release , Releases , Rock , ROYALTY , SING , singer , singing , single , so much , song , songs , Spotify , TalentEverywhere , Teaser , Theme , tweet , Upcoming , ur , video , write , YouTube , Youve , Desire

Then the relevancy test is performed with 5 users to validate our result are shown in Fig. 1.

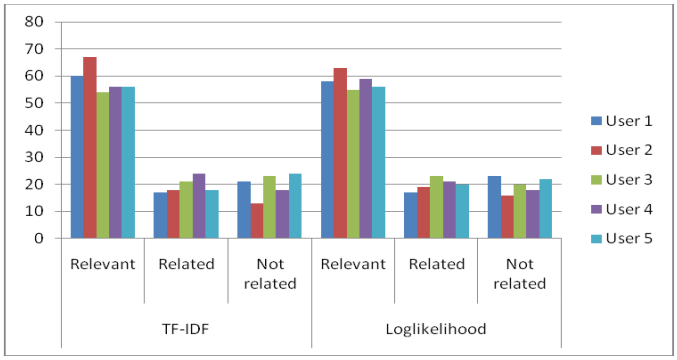

Fig. 1. Relevancy Test based on TF-IDF and Loglikelihood method's result

The result of the relevancy test on the keywords that are extracted using Term Frequency-Inverse Document Frequency (TF-IDF) and Loglikelihood method. It can be clearly seen that the result obtained from TF-IDF is slightly more relevant than the result obtained from the Loglikelihood method. If both methods are used for keywords extraction, then we find out common between them that shows the result is much concise, and accuracy would be better than the result obtained by applying a single method. Common words for music domain that are obtained from Term Frequency-Inverse Document Frequency (TFIDF) and Loglikelihood method are as follows:

Common Words for Music

5sos , Added , album , art , artist , Ashton5SOS , ass , band , bands , Brothers , Calum5SOS , check , dance , Download , Facebook , favorite , Festival , Follow , Format , ft , guys , hear , hiphop , hot , iTunes , liked , listen , Listening , live , lol , love , Luke5SOS , lyrics , Michael5SOS , movie , MP3 , music , NowPlaying , official , photos , pizza , playing , playlist , Posted , Radio , release , Rock , SING , singing , single , song , songs , tweet , ur , video , YouTube , Youve ,

The common keywords are shown above that are obtained after comparison of keywords extracted from TF-IDF and Loglikelihood method. These keywords are much more reliable than the result obtained from both the techniques separately.

If the relevancy test is applied to common keywords extract, then the following graph illustrates the statistics.

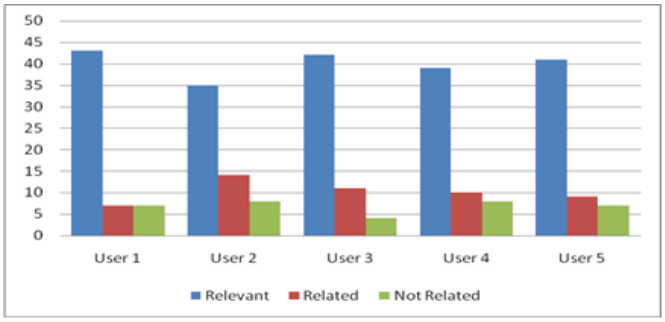

Fig. 2. Relevancy test for common keywords extracted from TF-IDF and Loglikelihood method

Fig. 2 shows the relevancy test performed between the 5 users against common keywords for TF-IDF and Loglikelihood method. Furthermore, the Noun and Verbs are identified for keywords that are extracted from both the TF-IDF and Loglikelihood method.

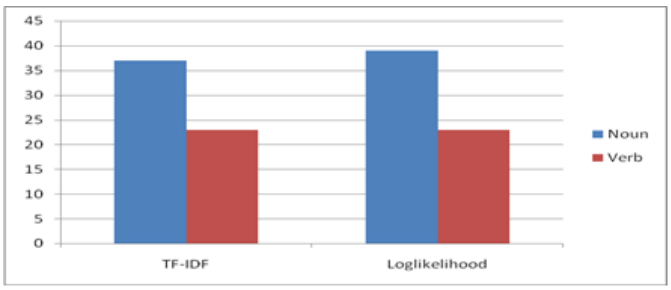

Fig. 3. Nouns and Verb in keywords extracted from the TF-IDF and Loglikelihood method

Fig. 3 shows the number of nouns and verb found in keywords against each domain. Sentiment words [33] for keywords are also identified by using TF-IDF and Loglikelihood method from sentiwordnet shown in Fig. 4.

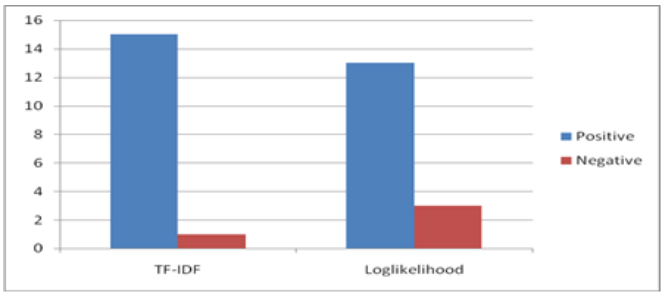

Fig. 4. Keywords as positive or negative of each domain

Fig. 4 shows the keywords as positive and negative for both TF-IDF and Loglikelihood method. TF-IDF shows more accuracy then loglikelihood method which concluded that TF-IDF is better. Accuracy of TF-IDF is further boosted with the addition of some other techniques.

5. Conclusion and Future Work

Twitter is the most famous social networking site and thought to be a reliable network on the internet nowadays. People of different categories used it on a daily basis. Due to the large volume of data, it is important to cluster the tweet data based on keywords. This paper focused on comparative analysis of keywords extraction methods namely Term Frequency-Inverse Document Frequency (TF-IDF) and Loglikelihood. After applying the relevancy test, it is concluded that TF-IDF is the most efficient method because most of the results verified in relevancy test. The keywords are a valuable asset because it is used to identify some patterns and can be used for clustering of the huge volume of data. Thus, it is beneficial for advertisement purpose, trends and set business policies for the future.

Furthermore, some other techniques may be applied with TF-IDF that helps to more refine the result. By applying further techniques will further improve the resulting quality effectively that is much better.

6. Acknowledgement

The authors would like to thank the Faculty of Information Technology, KFUEIT, Rahim Yar Khan and Islamia University of Bahawalpur, Pakistan for providing a support and research oriented environment.

7. References

[1] M. Asghar, M. F. Mushtaq, H. Asmat, M. M. S. Missen, T. A. Khan, and S. Ullah, “Finding correlation between content based features and the popularity of a celebrity on twitter”, International Journal of Computer Science Issues, vol. 11, no. 4, pp. 177–181, 2014.

[2] Twitter, “About twitter”, 2019. [Online]. Available: https://about.twitter.com/en_gb.html.

[3] C. Petrov, “Twitter statistics”, 2019. [Online]. Available: https://techjury.net/ stats-about/twitter/.

[4] R. Schroeder, S. Everton, and R. Shepherd, “Mining twitter data from the Arab spring”, Combating Terrorism Exchange, vol. 2, no. 4, pp. 54–64, 2012.

[5] Caiv, Yichuan and Y. Chen, "Mining influential bloggers: from general to domain specific", In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 447-454. Springer, Berlin, Heidelberg, 2009.

[6] Byrd, Kenny, A. Mansurov, and O. Baysal, "Mining twitter data for influenza detection and surveillance", In Proceedings of the International Workshop on Software Engineering in Healthcare Systems, ACM, pp. 43-49. 2016.

[7] Hickman, Louis, K. Saha, M. D. Choudhury, and L. Tay, "Automated tracking of components of job satisfaction via text mining of twitter data", In ML Symposium, SIOP, 2019.

[8] W. Fan, and A. Bifet, "Mining big data: current status, and forecast to the future", Association for Computing Machinery (ACM) SIGKDD Explorations Newsletter, vol. 14, no. 2, pp. 1-5, 2013.

[9] M. Ali, and A. I. Wagan, "An analysis of Sindhi annotated corpus using supervised machine learning methods", Mehran University Research Journal of Engineering and Technology, vol. 38, no. 1, pp. 185-196, 2019.

[10] J. Asmussen, "Automatic detection of new domain-specific words, using document

classification and frequency profiling", In Proceedings of the Corpus Linguistics conference, 2005.

[11] S. N. Kim, T. Baldwin and M. Kan, "An unsupervised approach to domain-specific term extraction", In Australasian Language Technology Association Workshop, pp. 94-98, 2009.

[12] A. Gelbukh, G. Sidorov, E. Lavin-Villa, and L. Chanona-Hernandez, "Automatic term extraction using log-likelihood based comparison with general reference corpus", In International conference on application of natural language to information systems, Springer, Berlin, Heidelberg, pp. 248-255, 2010.

[13] Q. Javaid, F. Memon, S. Talpur, M. Arif, and M. D. Awan, "Mining Frequent Item Sets in Asynchronous Transactional Data Streams over Time Sensitive Sliding Windows Model", Mehran University Research Journal of Engineering and Technology, vol. 35, no. 4, pp. 625-644, 2016.

[14] G. Luo, C. Tang, and P. S. Yu, "Resource-adaptive real-time new event detection", In proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 497-508, 2007.

[15] J. Allan, V. Lavrenko, D. Malin, and R. Swan, "Detections, bounds, and timelines: UMass and TDT-3", In Proceedings of topic detection and tracking workshop, pp. 167-174, 2000.

[16] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection", In proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 28–36, 1998.

[17] P. Indyk, and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality", In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604-613, 1998.

[18] S. Petrović, M. Osborne, and V. Lavrenko. "Streaming first story detection with application to twitter", In Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguistics, pp. 181-189, 2010.

[19] J. Benhardus, and J. Kalita. "Streaming trend detection in twitter", International Journal of Web Based Communities, vol. 9, no. 1, pp. 122-139, 2013.

[20] G. M. Huddar, M. M Ramannavar, and N. S. Sidnal, "Scalable distributed first story detection using storm for Twitter data", In 2014 International Conference on Advances in Engineering & Technology Research, pp. 1-5, 2014.

[21] J. Allan, V. Lavrenko, and H. Jin, "First story detection in TDT is hard", Proceedings of the ninth international conference on Information and knowledge management ACM, pp. 374-381, 2000.

[22] M. Vogiatzis, "Using Storm for Real-Time First Story Detection (Master's thesis, University of Edinburgh)," [Online]. Available: https://micvog.com/2013/09/08/storm, 2012.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.

[24] S. Moran, R. McCreadie, C. Macdonald, and I. Ounis, "Enhancing first story detection using word embeddings", Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 821-824, 2016.

[25] F. Godin, B. Vandersmissen, W. De Neve, and R. V. d. Walle, "Multimedia lab @ ACL W-NUT NER shared task: named entity recognition for twitter microposts using distributed word representations", In Proceedings of the workshop on noisy user-generated text, pp. 146-153, 2015.

[26] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, "Efficient online novelty detection in news streams", In International conference on web information systems engineering, Springer, pp. 57-71, 2013.

[27] I. Brigadir, D. Greene, and P. Cunningham. "Adaptive representations for tracking breaking news on twitter", arXiv preprint arXiv:1403.2923, 2014.

[28] J. B. Vuurens, and A. P. Vries, "First story detection using multiple nearest neighbors", In proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 845-848, 2016.

[29] Hal and Anders, "On the history of maximum likelihood in relation to inverse probability and least squares", Statistical Science, vol. 14, no. 2, pp. 214-222, 1999.

[30] Fienberg and E. Stephen, "Introduction to RA fisher on inverse probability and likelihood", Statistical Science, vol. 12, no. 3, pp. 161, 1997.

[31] ywwbill, “Github”, 2019. [Online]. Available: “https://github.com/ywwbill/TwitterCrawler.”

[32] K. Safari, and S. Sanner, "Optimizing search API queries for twitter topic classifiers using a maximum set coverage approach", arXiv preprint arXiv:1904.10403, 2019.

[33] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. "Sentiment analysis of twitter data", Proceedings of the Workshop on Language in Social Media, pp. 30-38, 2011.

[34] M. A. Abid, M. F. Mushtaq, U. Akram, B. Mughal, M. Ahmad, and M. Imran, “Recommending domain specific keywords for twitter”, In International Conference on Soft Computing and Data Mining, pp. 253-263, 2020.

