COMPARATIVE ANALYSIS OF PRE-PROCESSED SOCIAL MEDIA DATA GENERATION IN MODERN WORLD

[1]Nikitha kumari, [2]Dr.Sangeeta gupta
[1]Mtech, [2]Associate professor
[1]CSE,
[1]Vardhaman college of engineering and technology

Abstract: Numerous clients share their emotions via web-based networking by which it has become a most widely used a popular site online networking is one of the greatest stages where a monstrous number of messages are been posted each day which makes it a perfect hotspot for catching the feelings towards different inquisitive subjects, for example, items, similar interest, entertainments, games or celebrities, etc. This paper presents an analysis on existing works in the area of pre-processing based sentiment analysis of data generated based on social media.

Index Terms – pre-processing, Twitter.

INTRODUCTION

Microblogging administrations, for example, Twitter has separated a large number of clients to share their data between the general population and concentrate learning from the common data, [1] as they offer substantial volumes of continuous information, with around 200 a huge number of standard clients posting the tweets every day. And it has 320m month to month dynamic client, which can be gotten to through the site interface, SMS or versatile devices80% clients are dynamic through portable [2].

By collecting the tweets from twitter tweets are been unaliased in the form of positive, negative and neutral tweets which are known as the sentiment analysis. Where most of the research papers are based on the sentiment analysis. Supposition Classification methods can be generally partitioned into the AI approach, vocabulary based methodology and half breed approach [3]. The AI approach applies the popular ML calculations and utilizations semantic highlights. The Lexicon-put together Approach depends with respect to a opinion dictionary, an accumulation of known supposition terms. It is partitioned into word reference based methodology furthermore, corpus-based methodology which utilize measurable or semantic techniques to discover supposition extremity. The cross breed Approach consolidates the two methodologies. The exactness of a feeling examination depends on how well it concurs with human decisions. This can be estimated by utilizing accuracy and review. Sentiment analysis is another sort of content investigation which goes for deciding the supposition and subjectivity of analysts. With the developing fame of sites like Amazon.com, twitter and Epinion.com where individuals can express their supposition on various items and rate them, the web is packed with audits, remarks, and evaluations. It is therefore simple to discover abstract surveys on explicit items[4].

In the micro-blogging scale administrations, clients commit spelling errors and use feelings for communicating their perspectives and feelings [5]. To correct these spelling mistake and remove the emotions, punctuations etc the pre-processing techniques are been used.

This paper is been organised such as section2 consists of the privious works, section 3 consist of the comparision work of previous papers and section4 is the conclusion and future work.

2. Related work

The creators in [6] built up the close constant Twitter information warehousing by utilizing the NO-SQL database (Cassandra) and thought about its putting away and Querying execution by utilizing the strategy online dashboard to exhibit it in a continuous diagram, times are being noted. as per the outcome, Cassandra gives better execution to close constant information distribution center usage. Despite the fact that Cassandra for sure isn't the best when perusing generally little information yet quicker on huge information.

The authors in [7] had Compared SQL database and no SQL database Cassandra which handles a large amount of data with easy manage and low cost. Cassandra is an open source technology. they have compared the oracle and Cassandra by fetching the student information by writing the queries and explains how the Cassandra handles the large data efficient with almost 30-35% less than the oracle. by comparing the results, it shows that Cassandra taking less time when compared to Oracle and also working with records over 40000 which shows Cassandra is more efficient than the Oracle queries for large data set.

The Authors in [8] assess the execution of the five NoSQL group (Redis, MongoDB, Couch base, Cassandra, HBase) is been estimated by the YCSB (yippee cloud serving benchmark) instrument by appearing at equalization productivity and cost in NoSQL determination; correlation NoSQL in different perspectives, for example, working deferral, even scaling proficiency, and furthermore plan to make explore on the most proficient method to do improvement initially and after that think about and select NoSQL in a particular application. what's more, clarified the outcome got. By which the client can choose the required NoSQL.

The creators in [9] Cassandra which utilizes huge Table (of Google) for information stockpiling which has a large number of sections and a large number of columns, so it isn't productive to run the inquiry on such a gigantic table, in the wake of realizing that the aftereffect of the past question is adequate to answer our present inquiry. To take care of such issues of Cassandra database, they actualized another request language named as "setting Based Cassandra Query Language" which is inside mapped to Cassandra Query Language (CQL) so it has a comparative power as Cassandra however gives extra usefulness of questioning on consequence of past question. CBCQL is clarified with precedents. These give a system by which a client can solicit a grouping from related questions and gave the office of sparing a specific situation and reviewing it, so back tracking is additionally simple for the client for questioning. CBCQL has a similar control as Cassandra with extra usefulness since it worked well beyond Cassandra.

The creators in [10] Authors have built up an Android application 'Brilliant News' utilizing Apache Cassandra to store the gushing live news information. these applications are intended for the cutting-edge clients who wish to be most recent or inclining news with assistance of social media. These give access to disconnected every day drifting news in different classifications, for example national, specialized, sports, science and so on in a solitary swipe by utilizations Cassandra, a conveyed NoSQL database the board framework, as an information store for circulation of news related information crosswise over 4 hubs. Cassandra been utilized because of its capacity to oversee and control a lot of information and by the outcomes is been turned out to be an ideal database for putting away powerful, live and expansive dataset. Which performs proficient and shortcoming tolerant burden adjusting and dissemination of information among the hubs.

As per [11] in assumption examination, the execution of Sack of words some of the time stays restricted because of a few key lacks in taking care of the extremity move issue. Along these lines, to address this issue for slant grouping they proposed a model called double estimation investigation (DSA). They initially proposed a novel information development system by making an assessment switched survey for each preparing and test audit. On this premise, they proposed a double preparing calculation to utilize unique and turned around preparing surveys in sets for learning a notion classifier, what's more, a double forecast calculation to arrange the test audits by thinking about opposite sides of one audit. They likewise broadened the DSA structure from an extremity (positive-negative) arrangement to 3-class (positive - negative-impartial) the arrangement, by mulling over nonpartisan emotions. At long last, they built up a corpus-technique to develop a pseudo-antonym word reference.

Creators in [17] have proposed a system which can record information. backhanded access to official records of the association, for example, username secret word email client utilizing LDAP get to association it records the client's information who and what time of the tweet, and devotees will confide in chief to post on the official records. who simply enlisted via web-based networking media the board framework for the post is picked by the director level before it is been posted in authority account.

3. comparison of previous works

Table1 consists of the description about the previous research work which authors had done on their research papers which include details about the twitter data by taking the tweets as a data for collecting all the related information.

s.no	Research work	Algorithm used	Pre-processing technique used	attributes	Data source
1	Assumption investigation in Arabic the examination is finished by gathering the Arabic tweets at that point sifting it by the pre-preparing systems and estimated the exactness level by contrasting it with the three machine learning algorithms [12]	Machine learning algorithm such as: KNN, SVM, NB were used	Stemming, tokenization, filtering unwanted words.	Php and SQL, scripts	Tweets
2	they had proposed a technique utilizing Naïve Bayes, KNN furthermore, altered k implies bunching and found that it is more precise than Naïve Bayes and KNN strategies independently. furthermore, acquired a general grouping precision of 91% on the test set of 500 portable surveys. [13]	Machine learning algorithms such as KNN, NB, modified K-means+ NB and modified K-means+ NB+KNN	__-	Implementation in weka2 toolkit	Tweets
3	Extracted the tweets of Arvind Kejriwal and Kiran Bedi during the elections and conducted an experiment by applying pre-processing step and then analyses using LSW and showed by the result that Arvind Kejriwal's votes are higher than Kiran bedi. [14]	Lexicon sentiment word net and word net	Removing URL's, tags and negation handling	Word sense disambiguation	Tweets
4	Sentiment classification is been done airlines service analysis by tweets of airline services [15]	Machine learning algorithms	Tokenization, Stop words and lemmatization	python	Us airlines tweets
5	collected the data from many Indonesian official twitter accounts regarding football news by using LDA [16]	LDA	_		Tweets related to football

Table 1comparision of methods in previous work

In all these research paper most commonly, used techniques were machine learning and lexicon-based methods for the sentiment analysis of the real time data. Fig1 describes the two methods of sentiment analysis.

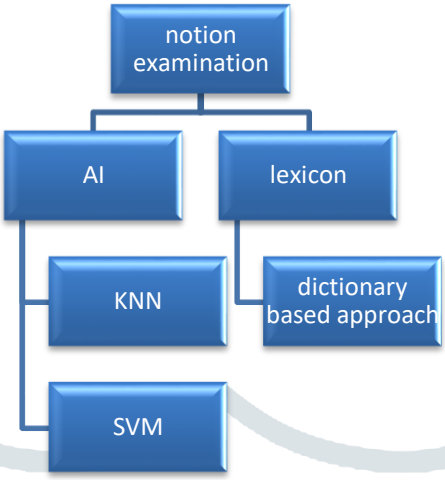

Figure 1 methods of sentiment analysis

4. Conclusion

This paper presents an analysis on the existing works in the area of pre-processing based notion examination of data generated based on social media. The literature considered in this paper used a variety of AI techniques and lexicon-based methods for the sentiment analysis of the real time data. This analysis is of much worth to implement the identified pitfalls in terms of the development of appropriate models for huge amount of social data being generated in modern world.

References

[1]. B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a #twitter," in Proc.49th Annu. Meeting. Assoc.Comput. Linguistics: Human Language Technol., 2011, pp. 368–378.
[2]. http:// twitter community.com.
[3]. Walaa Medhat a, Ahmed Hassan b, Hoda Korashy b, *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal science direct 2014.
[4]. Alessia, D., et al. "Approaches, Tools and Applications for Sentiment Analysis Implementation." International Journal of Computer Applications125.3 (2015).
[5]. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, Sentiment Analysis of Twitter Data, Department of Computer Science Columbia University New York, NY 10027 USA.
[6]. Muh.Rafif Murazza, Arif Nurwidyantoro: "Cassandra and SQL database comparison for near real-time twitter data warehouse "International seminar on intelligent technology and its application held in Lombok, Indonesia on 28-30 July 2016.
[7]. S. Anand, P. Singh, B.M. Sagar "Working with Cassandra database" Information and decision science, google scholar.
[8]. Enqing and Yushun fan: "Performance Comparison between Five NoSQL Databases" international conference on cloud computing and big data in Macau, china on 16-18 nov,2018.
[9]. Shivendra Kumar Pandey and Sudhakar: "Context Based Cassandra Query Language" international conference on computing, communication and networking technologies (ICCCNT) in Delhi, India on 3-5 July 2017.
[10]. Shubham Dhingra, Shreeya Sharma, Parmeet Kaur, Chetna Dabas: "Fault Tolerant Streaming of Live News using Multi node Cassandra" International Conference on Contemporary Computing (IC3) in Noida, India on 10-12 Aug 2017.
[11]. Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li. "Dual Sentiment Analysis:Considering Two Sides of One Review", IEEE Trans.on Knowledge and Data Engineering, 2015.

[12]. R.M. Duwairi, Raed marji, Narmeen Sha'ban, Sally Rushaidat, "sentiment analysis in Arabic tweets" 2014 5th International Conference on Information and Communication Systems (ICICS).
[13]. Onam Bharti Mrs. Monika Malhotra "sentiment analysis on twitter data" IJCSMC, Vol. 5, Issue. 6, June 2016, pg.601 – 609.
[14]. Rincy Jose and Varghese S Chooralil "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation" 2015 IEEE.
[15]. Bing Liu, *Sentiment Analysis and Opinion Mining* Morgan and Claypool Publishers, May 2012.
[16]. Ahmad Fathan Hidayatullah, Elang Cergas Pembrani, Wisnu Kurniawan, Gilang Akbar, Ridwan Pranata: "Twitter topic modelling on football news". International conference on computer and communication systems,2018 in Indonesia.
[17]. Dwiki Jatikusumo, Hanny Hikmayanti H., Feriadi, Wendi Usino: "Securing Official Account Twitter Using Social Media Management System" International Conference on Cyber Security, Cyber Warfare, and Digital Forensic in 2015.