INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 1 | Issue: 1 | Year: 2019

Review on Text Mining Based on Social Media Comments using Big Data Analysis

P.J.V.G. PRAKASA RAO

Assistant Professor and Training & Placement Officer, Department of Computer Science and Engineering, Viswanadha Institute of Technology and Management, Visakhapatnam, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Abstract: *Text mining and conclusion investigation have gotten enormous consideration as of late, exceptionally as a result of the accessibility of tremendous information in the type of content accessible via web-based networking media, web-based business sites, web journals, and other comparative sources. This information is generally unstructured and contains clamor, in this way the assignment of picking up data is intricate and costly. There is a developing requirement for creating distinctive systems and models for productively preparing the writings and separating adept data. One approach to extricate data is Text mining and opinion examination, which include: information obtaining, information pre-preparing and standardization, highlight extraction and portrayal, naming, lastly the use of different Natural Language Processing (NLP) and machine learning calculations. This paper gives an outline of various techniques utilized in Text mining and notion investigation explaining all subtasks.*

Keywords: Sentiment Analysis, Supervised Learning, Unsupervised Learning, Text Mining, Feature Extraction, Feature Representation

1. Introduction

Sentiment analysis, otherwise called sentiment mining, generally, is the way toward evaluating the enthusiastic incentive in a progression of words or content, to pick up a comprehension of the frames of mind, feelings and feelings communicated. Notion investigation can be connected to different divisions, for example, web-based business, managing an account, mining web-based life sites like Facebook, Twitter, etc. Utilizing feeling investigation and content mining, associations can pick up customer understanding from the reaction about their items and administrations. This can be additionally used to think about clients' fulfillment with the administrations and if there should arise an occurrence of dissensions and issues, finding the conceivable explanations behind that. One of the utilization of opinion examination is proposal frameworks, for example, YouTube suggests based on customers likes, aversions, and remarks given by the client. In this paper, we broadly think about different Text mining and feeling examination strategies connected to various territories in the multilingual organization and from various assets [1].

A conclusion examination and Text mining system commonly incorporates following subtasks: obtaining content information, information cleaning and pre-handling, information standardization, change of content to machine lucid vectors, highlights determination, lastly applying NLP and machine learning calculations. In this paper, we present a writing audit on ongoing patterns in Text mining and slant investigation. For example, customer audit mining and application to the travel industry are the current fruitful applications [2]. Point displaying is effectively joined with estimation priors to create themes and slant classes at the same time. Emoticon and emoji slants are incorporated into a considerable lot of the investigations to enhance exactness of results, etc.

2. Related Work:

Text processing is effectively utilized in the analysis of sentiments about a particular topic or product to know the reviews or recommendations of various

public opinion. Research on text processing methods for sentiment analysis is continuously conducted to optimize the performance of text mining in obtaining the appropriate opinions as in studies that measure the effect of training data size using SVM and Naïve Bayes by forming two ensembles. The study states that the change in training set size does not significantly affect the level of classification accuracy using SVM or Naïve Bayes but by combining SVM and Naïve Bayes using AND-type fusion suggests increased accuracy and F-Score from SVM [3]. In addition, the development of research related techniques in managing data such as transaction data and customer interaction on social media is done to analyze various methods and analytical tools that can be applied in big data applications and support decision makers to gain insights based on data extraction from the dataset.

In Duwairi et. al mentioned that sentiment analysis determines the polarity of given text either using machine learning approach or using lexicon based approach. The classifiers applied on the datasets were Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN (k=10)) where SVM gave highest precision and KNN gave the highest recall. Also to test the data sets 10-fold cross validation was used. They demonstrated that the precision got by SVM i.e. 75.25 was the best precision and the recall got by KNN i.e. 69.04 was the best recall. Therefore, to get better classification results, bigger data sets were required and to label them crowd sourcing was considered followed by semi supervised learning [4].

The use of social media for the user has encouraged the increase of unlimited textual information so that there is a need to utilize textual data to be presented without reducing the value of the information. This can be done with text mining. Text mining is a text analysis where data sources are usually obtained from documents with the aim of searching for words that can represent the contents of a document so that interrelationships and inter-document classes can be analyzed [5]. It is used to know the pattern of issues and problems that occur in the community in real time so that it can be taken into consideration in preparing a more appropriate policy. Text mining can be done through classification (classifier) or just by looking at the frequency (word cloud) and followed by doing sentiment analysis.

In Kouloumpis et. al demonstrated the usefulness of linguistic features and existing lexical resources used in micro-blogging to detect the sentiments of twitter messages. From this paper the researchers concluded that microblogging features were more useful as compared to POS (Part-of-Speech) features and features from existing sentiment lexicon. They also concluded that if they include micro-blogging features then the training data will be of less benefit. Consists of a new method formed by combination of rule based classification, supervised learning and machine learning which showed the improvement in micro and macro averaged F1 [2]. To get better effect, Prabowo et. al considered semi-automatic approach. From this paper they concluded that hybrid classification was better than the classification by any individual classifier. They also concluded that reduction of rules will produce less effect on F1.

Mudinas et. al concluded that concept level sentiment analysis system (psenti) was better as compared to pure lexicon based system and pure learning based system due to more precision in polarity classification and well structured, readable results. On experimenting, they confirmed that hybrid approach was better than sent strength [4]. From their paper, they concluded that senti system obtained high precision than pure lexicon based system but near to pure learning based system. It also gave well structured, readable results and more resistance to writing style of text. They also concluded that sent system works better than sent strength. In short, the proposed hybrid approach was capable in combining a carefully designed lexicon and a powerful supervised learning algorithm.

This research proposed a text mining processing through SVM method with classification optimization with Feature Selection. Feature Selection is used to select the relevant feature of the dataset in order to get a better performance of SVM as a classifier [3]. Text mining aims to generate a classification on the sentiment about the problem of taxation based on data sources the public comments on Facebook and Twitter. In this study, the results of positive and negative sentiments are based on time period and the type of tax data namely service, website system, and tax news. For further research, information generated

from this text mining can be used as considerable of taxation and support services for future policies [6].

Complaints submitted by the public through Facebook and Twitter can be extracted into consideration in the evaluation of the quality of tax services. In addition, for further research, information in the form of public opinion results can be used as one of decision support for planning tax policies. The use of social media for the user has encouraged the increase of unlimited textual information so that there is a need to utilize textual data to be presented without reducing the value of the information [3] [4]. This can be done with text mining. Text mining is a text analysis where data sources are usually obtained from documents with the aim of searching for words that can represent the contents of a document so that interrelationships and inter-document classes can be analyzed. It is used to know the pattern of issues and problems that occur in the community in real time so that it can be taken into consideration in preparing a more appropriate policy. Text mining can be done through classification (classifier) or just by looking at the frequency (word cloud) and followed by doing sentiment analysis [7].

3. Conclusion and Future Work

The real utilization of content mining generally incorporates system mining, regular dialect administration, and data recuperation and data extraction. In this paper, we study a couple of agent works, for example, substance acknowledgment and connection extraction and data extraction. In this paper, we likewise talk about the notions of Spanish tweets, Arabic tweets, and a lot more dialects. For content mining and opinion investigation, the significant advances required are information obtaining, information transformation, highlight portrayal, include extraction, and distinctive machine learning calculations. We likewise broadly demonstrate the consequences of different regulated and unsupervised notion examination systems to effectively recognize the suppositions.

Future work incorporates a broad correlation of various content mining and assumption investigation approaches on various informational indexes procured from different assets and in numerous dialects. We will likewise move in the direction of finding the most computationally modest calculations for different undertakings and sub-errands. Different forecast applications will likewise be examined.

References

[1]. Arun, K., Srinagesh, A., & Ramesh, M. (2017). Twitter Sentiment Analysis on Demonetization tweets in India Using R language. International Journal of Computer Engineering in Research Trends, 4 (6), 252-258

[2]. Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A Survey of Arabic Text Mining. In Intelligent Natural Language Processing: Trends and Applications (pp. 417-431). Springer, Cham.

[3]. Elgendy N., Elragal A. 2014. Big Data Analytics: A Literature Review Paper. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557. Springer, Cham.

[4]. Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11 (538-541), 164.

[5].] Sulistiani, H., & Tjahyanto, A. Comparative Analysis of Feature Selection Method to Predict Customer Loyalty. Journal of Engineering, Vol. 3, No. 1, 2017 (eISSN:2337-8557).

[6]. Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3 (2), 143-157.

[7]. Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining (p. 5). ACM.

Author's Profile:

P.J.V.G. PRAKASA RAO working as Assistant Professor and Training & Placement Officer, Department of Computer Science and Engineering, Viswanadha Institute of Technology and Management, Visakhapatnam, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

He has 18 years good teaching experience with taught various subjects on good knowledge on Hadoop & Big Data, COA, DAA FLAT, and JAVA and along with CSE subjects. She had published 3 research papers in reputed International and national level conferences\Journals\Magazines. He attended 1 conference, 25 organized / attended workshops and

seminars. He is Participated and active member in academic, curriculum, Training & Placement and administrative works in various organizations.