PSYCHOLOGY AND EDUCATION (2021) ISSN: 0033-3077 Volume: 58(3): Pages: 1227-1232
Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

A Robust Pre-processing model for Textual Data Analysis using PCL-COVID-19 Datasets

Mrs.U.Hemamalini[1*], Dr.S.Perumal[2]

[1] Research Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai.

[2] Professor, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai.

*hemababu2501@gmail.com

ABSTRACT

Global access to internet and flexibility enrolled with it attracted many peoples utilizing the benefit of its usage in many ways. Peoples started utilizing the services of internet with flexible and user friendly options. Peoples have their own freedom on conveying their opinion on products that they were experienced in the form of reviews. These online reviews act as one of the game changing factor for many products that would reach the customers or creates a great impact on particular incident, speech towards the society etc. Analysing those textual data in a specific approach is mandatory nowadays to improve the online image of the organization. The commonly used tweets, reviews impact the decision making angle of the users widely. The user feedbacks written by the consumers normally looks as a unstructured text data, hence to extract the real impacted emotion of the given comments, these reviews are required to be cleaned up. Pre-processing of original unstructured data is act as an important step of Data Mining. The main objective of this research paper to examine several pre-processing and feature selection techniques along with feature representation methods. The paper is focused on discussing various opinion mining levels available and the learning techniques adopted for sentiment analysis (SA). The raw data from the twitter is cleaned in many ways before handling the data for classification. The current paper formulates the Sample data from PANCEALAB website on COVID-19 discussions.

Keywords

Sentiment analysis, Machine learning, supervised learning techniques, Data analytics, opinion extraction, data optimization.

Introduction

Extracting the users's original emotion of the particular product or service is called Sentiment analysis (SA) called assessment mining, which is an examination that break down individuals' conclusions, opinions, assessments, evaluation, perspectives, and real emotion towards products or services, for example, products, operative systems, gathering, common, issues, festivals, political issues, themes. Most associations, managers, organizations, clients recognize these audits as a significant piece of their dynamic which in the long run bolsters business just as site support. The customers can investigate the survey of recorded clients for improvement of procurement choice and site advancement. Shockingly, these audits are in immense sum and unstructured. Consequently, it is dreary work to peruse and break down the surveys physically. In this way, the notion examination work gets famous since it is equipped for breaking down large number of audits and presents the yield to the client in a straightforward and reasonable way. Figure 2 speaks to the Textual Data Task Framework.

The proposed work considers the PANCEALAB master dataset on COVID-19 discussions. The tweeter dataset comprises of attributes such as hash tags, count of tweets, count of hash tags that are frequently used, apparent words, discussion key words and impacted words etc. the tweeter messages are the short but impulsive form of information that grab the attention of the social media users on various incidents. The real emotion and truth present in the textual expressions are mandatory nowadays to evaluate the fact behind the scenes. These tweeter data are publicly available information for further research works. Handling the raw data with best suited learning model and formulate the actual depiction is done through proper preprocessing techniques.

Literature Review

[1] Clarified the effect of pre-preparing. Tweets considered are loaded with images, unidentified words, and contraction. URLs, accentuation, client specifies. The words are pre-processed and analysed with the help of stop words, since it holds the key factor for keeping the slang of the written statement.

[2] The journal discussed about the sentimental analysis of movie reviews in which different language slangs are engaged. These slangs are pre-processed and sentiment dimensionality is formulated using support vector machine

modelling. The system achieves good accuracy since it handles the static reviews of the movie.

[3] The author utilizes the tweeter dataset that consists of 140 sentiments, where the author initiated the bags of words concept. The input data also holds the speech information and hence the features of the speech data are clearly extracted. Speech recognition is one of the complex part of artificial intelligence, hence the proposed work carried out with three algorithms like support vector, regression model and naïve Bayes model etc.

[4] Here the author compares four different classifiers to focus on detailed feature extraction process. The classifiers are tunes to remove the errors in feature extraction and that can be validated through accuracy calculations

[5] The author focused on deriving a robust classifier model that cascaded with support vector, j48 and naïve Bayes model. The unique points at the input test data is normalized and assessment is made

In [7], author addressed the formulation of tweeter reviews using bigrams. The speculated content provided by the author holds the real emotion on positive and negative between the reviews. These frequently used reviews are helpful to analyse the real emotion present with the written content instead of misconception of the system generated emotion etc.

[8] Most of the evaluations of the author related to the AI recommended hash tags. The addressed hash tags are consistent and frequently used in the most of the supporting reviews provided by the reviewers. Support vector and naïve bayes model is hybrid here. The selection of classifier is given an options to get better results with the processed data input.

[9] Evaluated a study on tweeter sentiments. Their journal focused on manipulating the basic knowledge on natural language processing, text mining and opinion mining etc. the analysed the various levels of sentiment analysis(SA) and some of the feature extraction techniques to grab the tweeter sentiments.

[10] Presented a detailed study on sentiment analysis (SA) through various algorithms involved in prediction. Their approach is based on transfer learning models, emotion of the textual data, reviewing the contribution and actual trustworthy sentiment present in it. They might have analysed large set of research journals to formulate the algorithmically better one for the prediction of sentiments.

[11] Evaluated a paper on analysing the micro blogging approach on sentiments finder. The hash tags, lexicon features and real emotions of the tweets are extracted. The tweet data are converted into ASCII to eliminate the commonly used words in the raw data. The proposed steps also remove the empty tweets and abbreviation incompletes.

[12] Authors presented the learning model that contains the supervised learning approach using support vector machine (SVM). The author explored the critical reviews and their impact on movie analysis is made. These comments are passed through various formulated process to extract the real emotion and the performance of the same is measured through accuracy of the system and precision in which the process executed.

SYSTEM DESIGN

A.Components of opinion Mining

Mining the data as a record level in which the entire record of information is being considered as a single component and fetched for analysis. This type of record level mining enables faster data processing to reduce time.

Mining of data as a sentence level is the next level of opinion mining technique in which the unique components present with the sentences are extracted through processing the set of comments alone. The frame by frame processing enable the system work even more robust and reliable.

.

Mining of data in certain perspective of unique meaning and enable them to express the real meaning of the written comments. The view points are extracted with certain keywords too. Such type of mining technique also called as point mining.

B. Guided Learning Model

At some point the system needs a prediction model that keeps the array of labelling information to improve the accuracy of the system. These supervised systems tend to provide high accuracy and static in nature.

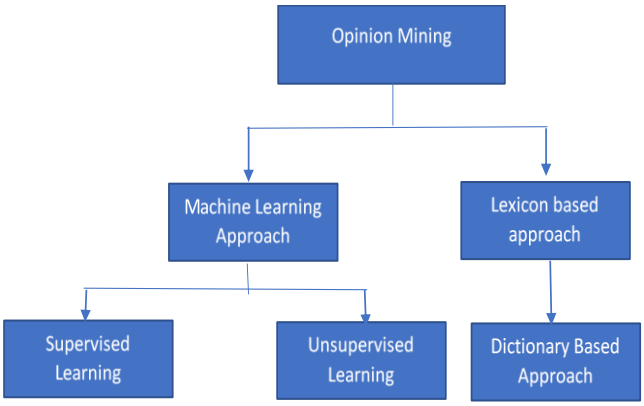

Figure 1: OPINION MINING TECHNIQUES

Supervised-learning approaches are used to evaluate the correlations between the each data and dependent features of the each attributes. Based on the relative features of the inputs, the database information is compared and mapped. These relationships are neurons that act as a comparison factor in learning models. The prediction is purely based on amount of relative information analysed by the neurons.

C. Unguided learning Model

As the name implies the unsupervised learning is the method of handling the data without and labels of the data. They don't need any supervising information to learn the data. The input acts as a vector and the feedback from the part of the output act as the learning feature. Unsupervised learning approach is almost like the learning model of human brain on learning new techniques. Based on trial and error approach the hidden layers get trained by itself. Unsupervised approach is not likely to be applicable for classification of inputs since the output data is not labelled. The unsupervised model enables the data to split up into groups based on the similarity in the data features.

Clustering algorithms, association rules are the examples of unsupervised learning approaches.

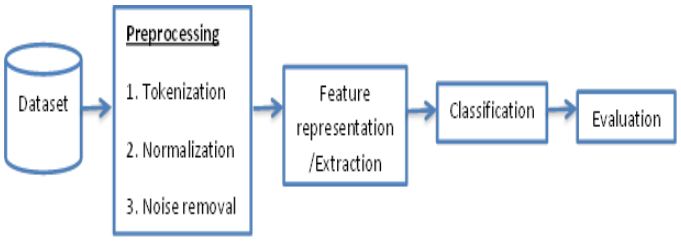

Figure 2. Textual data task framework

A. Data Acquisition

SA has discovered its applications in different fields that are presently assisting undertakings with assessing and gain from their clients accurately. Sentiment investigation is progressively being utilized for web-based media observing, brand checking, and criticism of the client, client support, and statistical surveying. Estimation investigation utilizes Natural language preparing strategies and calculations that are either rule-based, cross breed, or depend on AI procedures to take in information from datasets. The information required in supposition examination ought to be specific and is needed in huge amounts. The most testing part about the estimation examination preparing measure isn't restricted in discovering information in enormous sums rather it finds the significant dataset. These informational collections should approaches expanded region of conclusion examination operations and specified cases. The absolute most mainstream datasets for slant analysis.MDB Movie Reviews Dataset, Sentiment140, Twitter US Airline Sentiment, Paper Reviews Data Set, Sentiment Lexicons For 81 Languages, Lexicoder Sentiment Dictionary, pin-Rank Review Dataset.

B. Data Preprocessing

Information preprocessing is changing the information into a fundamental structure that make it simple to work. Information preprocessing is a significant method for handling the exhibition of Data Mining calculations. Fig 2 determines the overall portrayal of text preprocessing. Tokenization is the process of dealing the longer test patterns into couple of smaller text sequence and the operation is probably called as tokens making or tokenization of text in sentiment evaluation.

As we consider the raw text reviews as input, it obviously carries the information at prefix and suffix to provide valuable meaning to it. These infixed data does not contain specific meaning. The occupation of prepositions and articles that helpful for forming the sentence and meaning are just filtered out here with the process called stemming-process.

One of the most important step with the information mining process is keeping the meaning unchanged. The criteria need the exact word that is more sensitive to the sentence grammar and holder for meaning of the sentence. These sensitive part of the word is normally referred as Lemma (the sensitive word). Few different words with similar meanings are filtered with the help of Lemmatization-process with the sentiment evaluation model.

Some of the principle steps to be followed with the information mining is given below.

Formulate the text by removing the junk words and infinite numbers.

In case of any extra empty spaces, that should be removed.

Wrongly formulated characters in the sentences are filtered out and stop words are analyzed for slang formation.

The sentences are clearly checked to ensure the unchanged meaning of the reviews.

The commonly repeated words are removed to normalize the sentence to reach meaningless.

Clamor evacuation proceeds with the replacement undertakings of the structure. Tokenization and Normalization were by and large pertinent as-is to almost any content lump. The crude dataset comprises of endless qualities; spaces are eliminated totally in this progression.

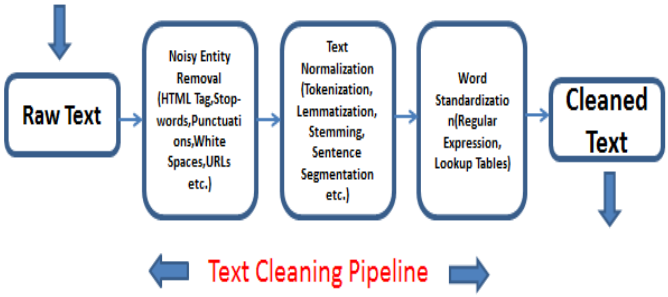

Figure 3. Generalized view of preprocessing

PSEUDOCODE

GET INPUT String = Text_Input;
- *Store x=String(Text_Input)*
- *String_split(x) = Y*

Loop =1:numberofElements(Y)
Check(not a char)
If TRUE Skip
If FALSE then Store to Y
End Loop;
- *Check (' ') and (inf)*
- *Remove prepositions*
- *ASSIGN temp_weigtage(Char)*

Map Words;

PSYCHOLOGY AND EDUCATION (2021) ISSN: 0033-3077 Volume: 58(3): Pages: 1227-1232
Article Received: 13th September, 2020; Article Revised: 25th January, 2021; Article Accepted: 12th February, 2021

RESULTS AND DISCUSSIONS

TABLE I. PREPROCESSING STEPS

Steps	Processed Text	Data description
1	"Even in the current situation we need to stay positive to motivate the medical support team and help……* public"	Raw Data
2	Even the current situation need stay positive motivate the medical support team and help……* public	Prepositions removed
3	Even current situation need stay positive motivate medical support team help ……*public	common words removed
4	Even current situation need stay positive motivate medical support team help public	junk removed
5	current situation need stay positive motivate medical support team help public	processed data
6	Stay, positive, motivate, support, help, team	Positive Words Detected

TABLE II. OVERVIEW OF PROCESSED DATASET

SlNo	Processed Data	Tweet Counts	Hashtags Used	Weigtage
1	trying times share	4249	coronavirus	5600
2	italy trying times	4229	nan	4573
3	stand italy trying	4228	covid	4439
4	times share support	4223	19	4334
5	share support italian	4217	covid19	4321
6	support italian friends	4215	people	4254
7	italian friends colleagues	4196	trump	4252
8	tests positive coronavirus	1966	via	4246
9	href rel nofollow	1959	us	4231
10	true href rel	1913	virus	4225
11	coronavirus covid 19	1766	amp	4217
12	rel nofollow twitter	1703	cases	4207
13	covid 19 pandemic	1476	get	3298
14	covid 19 cases	1336	New	3258

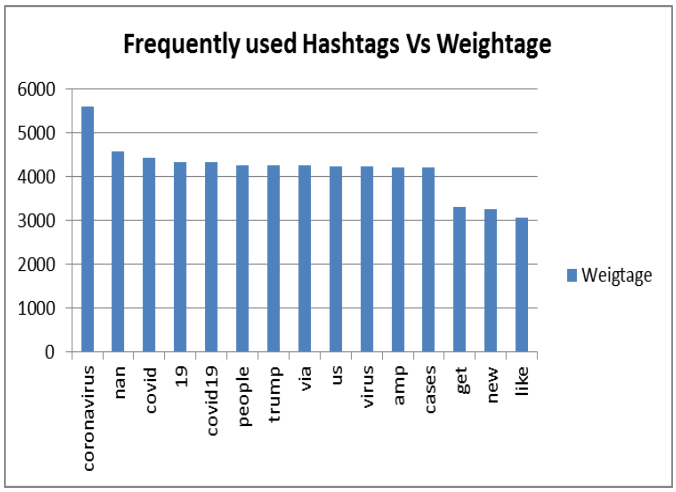

Figure 4. Frequently used hash tags vs weigtage of the data used

Table I. shows the preprocessing steps invlved in single tweet handling. It clearly depicts the stages of processed text from the raw dataset to the detection of few positive sentiment words through programatic approaches. The initial processing removes the unused words and prepositions. Further the text is checked for spces and infinite values. Common words that are not considered among the sentiment analysis is removed from the processed data. Finally the preprocessed data is compared with few postive words from the bags of word datasheet and assigned with random weigts for further proceedings.

Conclusions

Textual processing becomes more attractive field of study in which research focuses on extracting the original emotions present in the text, trust worthy sentiments present in the text data and weightage of the words used through bags of words approach. The proposed work presented preprocessing models that perform the steps clearly formulated here to remove the unrelated words from the raw text data. The approach is based on previous literatures discussed [1][2][3] etc. the proposed system comprises steps that remove the unassociated data from the tweets. The preprocessing approach also compares the filtered data with the positive words bag created for comparison. The system assigns a random weightage to each word that correlate with the positive words present in the database. The proposed model is simulated and the PANCEALAB dataset is categorized as sown in TABLE I. the system is further

improvised by using the supervised learning algorithms to evaluate the exact emotion present in the tweets.

References

[1] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca
Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011

[2] Shi, Y., Xi, Y., Wolcott, P., Tian, Y., Li, J., Berg, D., Chen, Z., Herrera-Viedma, E., Kou, G., Lee, H., Peng, Y., Yu, L. (eds.): Proceedings of the First International Conference on Information Technology and Quantitative Management, ITQM 2013, Dushu Lake Hotel, Sushou, China, 16–18 May 2013, Procedia Computer Science, vol. 17. Elsevier (2013)

[3] Fouad M.M., Gharib T.F., Mashat A.S. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In: Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA 2018. Advances in Intelligent Systems and Computing, vol 723. Springer, Cham

[4] Prusa, Joseph D., Taghi M. Khoshgoftaar, and David J. Dittman. "Impact of Feature Selection Techniques for Tweet Sentiment Classification." In FLAIRS Conference, pp. 299-304. 2015.

[5] Angulakshmi, G., and Dr R. Manicka Chezian. "Three level feature extraction for sentiment classification." International Journal of Innovative Research in Computer and Communication Engineering 2, no. 8 (2014): 5501-5507.

[6] 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

[7] 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.

[8] Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association forComputational Linguistics: Short Papers-Volume 2 (pp. 90- 94). Association for Computational Linguistics.

[9] Kishori K. Pawar, R. R. Deshmukh, "Twitter Sentiment Analysis: A Review", International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 9 ISSN 2229-5518, pg.957-964.

[10] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal (2014) 5, pg.1093–1113.

[11] Kishori K. Pawar, R. R. Deshmukh, "Twitter Sentiment Analysis: A
Review", International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 9 ISSN 2229-5518, pg.957-964.

[12] Harnani Mat Zin, Norwati Mustapha, Masrah Azrifah Azmi Murad, and Nurfadhlina Mohd Sharef, "The effects of pre-processing strategies in entiment analysis of online movie reviews" AIP Conference Proceedings 1891, 020089 (2017); https://doi.org/10.1063/1.5005422
Published Online: 03 October 2017.

[13] T. Nikil Prakash, Dr. A. Aloysius ," Data Preprocessing In Sentiment Analysis Using Twitter Data" International Educational Applied Research Journal (IEARJ) Volume 03, Issue 07, July 2019 E-ISSN: 2456-6713.

[14] Emma Haddi, Xiaohui Liu, Yong Shi, "The Role of Text Pre-processing in Sentiment Analysis" Information Technology and Quantitative Management (ITQM2013) Procedia Computer Science 17 (2013) 26 – 32.

[15] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta,
Pratyush Ahuja, "The Impact of Features Extraction on the Sentiment Analysis", Procedia Computer Science,Volume 152,2019,Pages 341-

348,ISSN
18770509,https://doi.org/10.1016/j.procs.2019.05.008.

[16] Swati Redhu, Sangeet Srivastava, Barkha Bansal, Gaurav Gupta, "Sentiment Analysis Using Text Mining: A Review", International Journal on Data Science and Technology. Vol. 4, No. 2, 2018, pp. 49-53. doi: 10.11648/j.ijdst.20180402.12

[17] Henrique Siqueira and Flavia Barros, "A Feature Extraction Process for Sentiment Analysis of Opinions on Services" Centro de Inform_atica (CIn) - Universidade Federal de Pernambuco (UFPE) Recife - PE – Brazil.

[18] Hung, Lai & Alfred, Rayner & Hijazi, Mohd. (2015). "A Review on Feature Selection Methods for Sentiment Analysis". Advanced Science Letters. 21. 2952-2956. 10.1166/asl.2015.6475.

[19] T, Nikil & Amalanathan, Aloysius. (2019). "Data Preprocessing In Sentiment Analysis Using Twitter Data". 3. 89-92.

[20] A., Vishal, and S.S. Sonawane. "Sentiment Analysis of Twitter Data: A Survey of Techniques." International Journal of Computer Applications
139.11 (2016): 5–15. Crossref. Web.

[21] Ahuja, Ravinder & Chug, Aakarsha & Kohli, Shruti & Ahuja, Pratyush. (2019). The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science. 152. 341-348. 10.1016/j.procs.2019.05.008.