iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5089

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Survey Paper on Algorithms used for Sentiment Analysis

Meghashree K M[1], Radhika S[2], Shilpashree A[3], Soujanya Dinni[4], Mrs. Monika P[5]
[1, 2, 3, 4, 5]*Department of CSE, Dayananda Sagar College of Engineering*

Abstract: Analysis of human sentiments is a trending topic in this era. The motive of this survey is to determine the emotional state of people or a person by using data mining concepts. Here, the process analyses the state based on keywords provided by the user. Feeling, opinion, emotion or attitude are the Sentiments. By computational identification of a text and processing it, Sentiment analysis can be carried out to find whether the writer's feelings in a specific context is positive, negative, or neutral. So implementing NLP and to categorize, classify or statistically view the report of analysis from a group of text is what is considered as "sentiment analysis".
The motive of this paper is to perform a thorough survey on sentiment analysis algorithms that use machine learning concepts. The topics henceforth explored include various algorithms, quality measurement metrics, feature selection and training methods. Our paper also focuses on the three kinds of machine learning algorithms for Sentiment Analysis- Supervised, Unsupervised and Semi-Supervised Machine Learning Algorithms.

I. INTRODUCTION

Nowadays there are many users on social media. All the users discuss, chat, recommend and share their views about a product, place, public issue through the internet. The Internet has become an integral part of our life . Most people seek social validation before making their final, firm decision.

In order to help people with others' views on a product or a movie or a novel, processing and analysing the data(opinions/feedback) and processing it to categorise into various categories is required. To fulfil this requirement Sentimental analysis is a significant and trending field of research among researchers. This deals with collection of information and processing The gathered data is analysed to identify the emotion of the text provided by the user .There are many applications for this area. Few of them are- product recommendations, where the user gets a recommendation on what to buy based on his /her social media circle. To solve the problem of how to classify the sentiment various approaches have been reviewed, among which we chose to work with machine learning techniques of supervised learning, unsupervised learning and semi supervised approach.

II. SUPERVISED LEARNING

Problem is based on sentiment analysis. Bo Pang, et al. Compared Supervised machine learning techniques with human generated baselines and they employed supervised machine learning techniques will not perform well on sentiment when compared to traditional based categorization. Finding out the factors which make the problem for sentiment classification using supervised learning methods like Naive Bayes Classification and SVM. Also, concluded that the review is negative or positive. [1] Machine learning algorithms are better when compared to baselines generated by humans and also, SVM technique is best when compared to Naive Bayes based on their performances.

Sentiment analysis can be performed using various techniques providing variant results. [3] The comparative study on how different classification models behave on different feature selection techniques for unigrams and higher n-gram data. Sentiment analysis is performed stepwise on Movie review datasets using Naïve Bayes and SVM for classification. First step is preprocessing a dataset where tokenization, pruning, filtering of tokens, and stemming are performed. Various feature selection techniques implemented in combination with classification models like Binary Term Occurrence, Term Occurrence, TF-IDF and Term Frequency. Gautami Tripathi et al. found that when linear SVM, TF-IDF is used they gave the best accuracy of 84.75% and for Naïve Bayes Term Occurrence provided a maximum accuracy of 70% for Unigrams. The study was extended for other n-gram models such as bigrams, trigrams and four-grams and reached the conclusion that bigrams provide higher accuracy, precision when compared to others while recall stays the same. The research can be extended for dynamic data implementing hybrid techniques and also improving the results of Naive Bayes algorithm.

[4] Performs Sentiment Analysis on the Movie Reviews. 2000 movie reviews is provided as a dataset having multi dimensional features. Support Vector Machine(SVM) is used as the classifier model for classifying dataset into positive and negative sentiments. Since different feature selection techniques give different performance over classification, this paper performs various feature selection techniques such as Chi-Square, Mutual Information, TFIDF and Information gain and chooses the one that selects best features for efficient classification. A relative study to lower unigram features is also performed on shorter feature length to enhance accuracy. Collecting dataset, Preprocessing, Feature extraction and selection, Classification are performed stepwise in this process. Shahana P.Ha et al. concluded that the best method to distillate sentiment is Unigram. Unigram with stemming and without stop words gives better accuracy than with stop words. In the future, to perform extended research in this field, ensemble feature selection methods could be useful.

Nurulhuda Zainuddin et al. describes how Sentiment Analysis is performed upon benchmark datasets and classifies them into positive and negative datasets. Support Vector Machine is the supervised machine learning technique used for this purpose. Adopting different feature selection methods for classification, impacts on accuracy of classification significantly. [5] Experiments upon this fact and implements Term Occurrence(TO),TFIDF and Binary Occurrence (BO) over different n-grams and states which among them is best.

The whole process of Sentiment Analysis using SVM is implemented in various steps. Preprocessing is the first step where stemming, removal of stop words, tokenization, lowercase conversion are performed. Second step is Feature Extraction where the input data is transfigured into a group of features that work efficiently on the classifier model. Third step is Feature selection and the fourth step is Classification using SVM.

Last step is to calculate effectiveness measures, the one used here is Confusion Matrix. The paper concludes that unigrams perform best among other n-gram models and TFIDF, BO play the significant role in selecting best features appropriate for classification. It also shows using Chi-square for feature selection enhances the accuracy of classification.

In [10] the representation of some aspects of natural language meanings uses vector space models(VSM). It uses a distributional hypothesis which states that words in similar situations have the same meanings. These types use the Helmholtz principle and a supervised algorithm SVM. This algorithm can be used for a large set of unlabeled datasets. Berna Altınel et al. used a large amount of dataset given the number of unlabeled datasets is less. There are two approaches in semantic classification. The first approach applied in this is used for small datasets, the second approach is considerably good at reducing noise in large datasets.

III.UNSUPERVISED LEARNING

Sergio Canuto et al. Fetched meta features clustering like KNN from short messages of sentiment analysis. Among giving k near neighbors of sentiment distribution short test documents and given a short test document, the information is derived. Distribution neighbor and their distance of x is found in [6] Unsupervised lexical based methods give the polarity of these neighbors. Meta level features gopal et' s meta feature and canuto et al' s meta level feature are used. Finally Sergio Canuto et al. concluded by ranking short messages with queries.

Milagros Fernandez-Gavilanes et al. Proposed a novel approach to predict the sentiment in online tweets based on an unsupervised method that holds a variety of NLP techniques and sentiment features like lexicons. To increase accuracy of the lexicons, a semi automatic polarity expansion algorithm is used. [7]

Used unsupervised method to provide a sentiment on tweets and also semi automatic polarity expansion algorithm to increase the accuracy of lexicons. They concluded that supervised techniques have disadvantages like taking more time for training an algorithm and being highly dependent on the quality and quantity of features. Also, supervised technique requires human judgement for labeling the classes.

The twitter tweets can be analyzed by clustering-based methods to observe the views and sentiments of users . But twitter dataset is subjective in nature, therefore clustering methods like metaheuristic performs better than other methods of sentiment analysis. Avinash Chandra Pandey et al. proposed metaheuristic technique which is built on K-means , cuckoo search(CSK). Cluster heads from twitter dataset using csk method are efficient. In this paper,[11] efficiency of CSK technique is better when compared with outcomes of optimised cuckoo search, particle swarm optimization, two n-grams technique and so on. The efficient cluster from the sentimental data is found.

The random cluster formation in k means is improved by cuckoo search technique. Accuracy is better than existing methods but needs to be improved further.

IV. OTHERS

Antonio Reyes et al. analyzed domain languages like humor, irony inorder to find key terms for self processing on twitter. [2] The model based on textual features is evaluated on two dimensions like relevance and Representativeness. Assumptions are satisfied by finding out metrics like classification, accuracy score, precision, recall score and F-measure. Concluded that there is no single feature which is distinctly humorous and ironic, together they provide useful inventory for figurative devices at textual level. Humor and irony are no more single features and they together form an useful discovery for figurative devices at textual level.

[8] Sentiment sorting and auditing can be preceded by varieties of skills and by considering it's superiority constraint of the apparatus . Apparatus that are considered for the different kinds of process of sentiment auditing also have a great part. Steps for sentiment auditing are: Gathering data, text composing, attitude noticing, attitude sorting. Sentiment sorting can be done by machine learning, linguistic based, and amalgam. Machine learning auditing focuses on forecasting which is the contradiction of sentiment that forms on prime data and trail data. In linguistics it doesn't need any anterior information for drawing information. Based on the machine learning process, most used skills are Bayesian webwork. It is anticipation auditing that represents the connection of attributes in common strategy. Alessia D'Andrea et al. focused on the skills and tools that are used for sentiment auditing and sorting.

Sarcasm is a way of conveying the negative feelings in positive words. But in real life it's along with a gesture of sarcasm . In [9] text or twitter this gesture cannot be found, which results in a normal positive comment. Because of these difficulties it's been an important area of research for researchers. It proposes a framework which is based on hadoop, this framework process and predicts the sarcasm in the text. S.K. Bharti et al. concluded that this method outperforms all other present methods. 66% of processing time is reduced by this framework. TCUF, TCTDF ,LDC these algorithms need to be deployed in this hadoop framework for future enhancement.

Under-sampling (a typical resampling approach)is to deal with the imbalanced problem in providing data for sentiment classification. Novel semi-supervised learning method based on random subspace generation which generates various subspaces in the iteration process [12] dynamically to provide enough variation among the involved classifiers. Undersampling, Random Subspace Generation(RSG), Co-training techniques are used in semi-supervised learning for imbalanced sentiment classification. Shoushan Li et al. concluded that CoTraining-Dynamic robustly outperforms under-sampling and the two baselines (CoTraining-static and CoTraining-undersampling) in all the domains.

V. CONCLUSION

In the current generation, the rapid growth of technology is affecting human lives significantly. People buying the products by entering the shops physically is getting replaced by people logging into the various online shopping websites. Judgement of a product being good or bad by the own experience of an individual is replaced by experience and views of other people over the product. The ratings on the product, number of reviews , number of positive and negative reviews, polarity of opinions towards positive or negative have become the major criteria for product sale. More reviews and its polarity towards positivity, more reliable it is for a new buyer. Thus, collecting tons of data(reviews/opinion/feedback), analysing and processing it to categorise it into positive, negative or neutral becomes the necessity in this online shopping pool.

Sentiment analysis is the current hot topic trying to provide a solution for this. Our Literature survey has presented several approaches for sentiment analysis by applying various algorithms, predominantly machine learning algorithms(supervised, unsupervised and semi-supervised). Different algorithms in combination with different feature selection techniques have also been applied to extract the best feature to perform classification and identification of polarity. It is observed that unigrams give better performance than any n-grams model. The results also show implementing feature selection using chi square will improve classification accuracy. Supervised machine learning algorithms provide higher accuracy and performance.

REFERENCES

[1] Bo Pang , Lillian Lee, Shivakumar Vaithyanathan ."Thumbs up? Sentiment Classification using Machine Learning Techniques". The Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, DOI:10.3115/1118693.1118704.

[2] Antonio Reyes , Paolo Rosso , Davide Buscaldi. "From humor recognition to irony detection: The figurative language of social media". Institut de Recherche en Informatique de Toulouse (IRIT), Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse, France, April 2012, DOI: 10.1016/j.datak.2012.02.005 .

[3] Gautami Tripathi , Naganna .S. " Feature selection and classification approach for sentiment analysis".Machine Learning and Applications: An International Journal (MLAIJ), June 2015, DOI: 10.5121/mlaij.2015.2201 .

[4] Shahana P.Ha , Bini Oman. " Evaluation of Features on Sentimental Analysis". International Conference on Information and Communication Technologies (ICICT), Elsevier, April 2015, DOI: 10.1016/j.procs.2015.02.088 .

[5] Nurulhuda Zainuddin , Ali Selamat. "Sentiment Analysis Using Support Vector Machine". IEEE International Conference on Computer, Communication, and Control Technology (I4CT), Sept 2014,DOI: 10.1109/I4CT.2014.6914200 .

[6] Sérgio Canuto, Marcos André Gonçalves , Fabrício Benevenuto. "Exploiting New Sentiment-Based Meta-level Features for Effective Sentiment Analysis". WSDM'16: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, February 2016, DOI: https://doi.org/10.1145/2835776.2835821 .

[7] Milagros Fernandez-Gavilanes, Tamara Alvarez-Lopez, Jonathan Juncal-Martõnez, Enrique Costa-Montenegro , Francisco Javier Gonzalez-Castano. "Un-supervised method for Sentiment Analysis in online texts", Expert Systems With Applications, October 2016, DOI: https://doi.org/10.1016/j.eswa.2016.03.031.

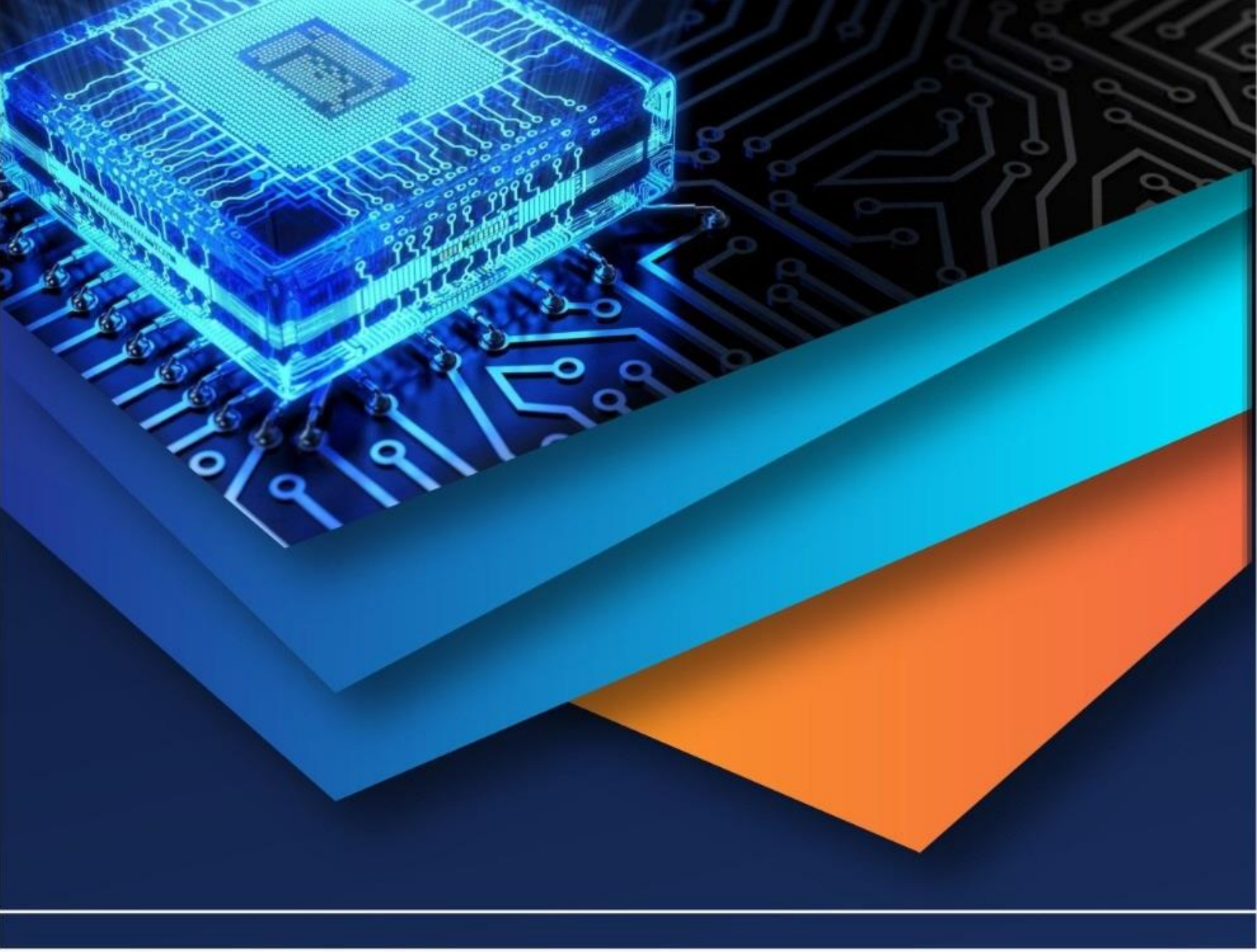
[8] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni , Tiziana Guzzo. "Approaches, Tools and Applications for Sentiment Analysis Implementation". International Journal of Computer Applications, September 2015, DOI:10.5120/ijca2015905866.

[9] S.K. Bharti, B. Vachha, R.K. Pradhan, K.S. Babu , S.K. Jena. "Sarcastic Sentiment Detection in Tweets Streamed in Real time: A Big Data Approach" Digital Communications and Networks, August 2016, DOI: https://doi.org/10.1016/j.dcan.2016.06.002 .

[10] Berna Altınel , Murat Can Ganiz. "A new hybrid semi-supervised algorithm for text classification with class-based semantics". Knowledge-Based Systems, Elsevier, September 2016, DOI: https://doi.org/10.1016/j.knosys.2016.06.021.

[11] Avinash Chandra Pandey , Dharmveer Singh Rajpoot , Mukesh Saraswat. "Twitter sentiment analysis using hybrid cuckoo search method". Information Processing and Management, Elsevier, July 2017, DOI: https://doi.org/10.1016/j.ipm.2017.02.004.

[12] Shoushan Li, Zhongqing Wang, Sophia Yat Mei Lee. "Semi-Supervised Learning for Imbalanced Sentiment Classification", Proceedings of the 22nd International Joint Conference on Artificial Intelligence, January 2011, DOI: 10.5591/978-1-57735-516-8/IJCAI11-306 .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY