International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

A Proposed Approach for Sentiment Analysis and Sarcasm Detection on Textual Data

Himani Khullar, Amritpal Singh

Abstract: ***Sarcasm in straightforward dialect infers the utilization of incongruity to taunt or pass on hatred. It is portrayed as the unexpected and humorous mind. It changes the extremity of a clearly positive or negative articulation to inverse of it. Sarcasm is a rich method for the passing on message in understood way which makes difficult to recognize it. The objective of this paper is to detect sarcasm with high accuracy. In this paper bagged gradient boosting is proposed with particle swarm optimization as feature selection. It is compared with other classifiers such as random forest, gradient boosting, bagged gradient boosting. The emoji and acronym dictionary mapping is done, part of speech labelling is introduced. Hashtags and stop words are recognized and removed. Particle swarm optimization is used to remove noisy data.***

Keywords: Bagging,Gradient Boosting,Particle Swarm Optimization, Sarcasm

I. INTRODUCTION

Social Network has developed and increased in the previous couple of years. It is a kind of mode of communication that can be utilized by any individual lives in any area of the world. With such all inclusiveness and high speed information, sentimental analysis on such systems has been most focused on research subjects in Natural Language Processing(NLP) in the previous decade. The fundamental point of sentimental analysis is to identify the extremity of the content. The web has drastically changed the manner in which individuals express their perspectives and sentiments [1]. Emotions can be positive, negative as well as neutral. For identification of emotions via communication through face is an easy task rather than communication through text. Number of data which are in textual form present on internet so that an individual can find them easily via internet. Reputation can be easily tracked through internet for any organization which includes people opinion that helps media to make opinion regarding any organization's product Corservice.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Himani Khullar*,M.Tech, Lovely Professional University, Punjab, India.
Amritpal Singh, Assistant Professor Department of Computer Science and Engineering, Lovely Professional University, Punjab.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

A. SentimentalAnalysis

Sentimental Analysis (SA) is a data mining technique. SA is a procedure of examining of emotional tone beyond a processing of words that is utilized for grabbing an appreciation of a perspective, suppositions and sentiments conveyed byperson. Sentiment analysis is an investigation of individuals' appearance, feelings, and feelings to groupregardless of whether it is certain, negative or neural. It very well may be connected on different types of information, for example, content, emojis, pictures, sound, and video. Individuals post

their perspectives, feelings and feelings on long range interpersonal communication sites with wide utilization of emojis. Twitter clients generally utilize hashtags and smileys for underlining their perspectives behind any thought.

B. Sarcasm

arcasmisspeechorwritingwhichactuallymeanstheopposite of what it seems to say. It is a rich method for the passing on message in understood way which makes difficult to recognize it. In Sentiment Analysis (SA) one of the challenges is Sarcasm detection. These days, various online social media sites like Facebook, Twitter have empowered customers for communicating their emotions in the lingo alongside name in their messages.

C. Challenges of Sarcasm DetectionApproach

Sarcasm sentences explicit the negative conclusion about an objective utilizing positive words[8].

I do not like being loved by someone!!

I love to work on weekends!!

some of challenges are saidunderneath:

a) It might be utilized in an indirect way, and have the type of incongruity which makes it hard to comprehend the sentiments of the individual.

b) The snide tweets communicates negative estimation utilizing positive words. In this way the classifier would mistakenly assign sentiments to these tweets. "I love attending lectures on end of theweek.

c) There is wide usage of slang words, abbreviations, smileys, special symbols and unstructured data which makes it difficult to identifysentiments.

II. LITERATUREREVIEW

González-Ibánez et al. (2011) Author utilize the dependable corpus to contrast snide articulations in Twitter with articulations which explicit positive and negative states of mind except sarcasm. Author research the effect of lexical for machine learning adequacy to recognize sarcasm expressionsand they think about the execution of machine learning systems or human judges on this errand [2].

Luonando E et al. (2013) Author proposed two extra highlights for identifying sarcasm later a typical assessment examination is directed. Highlights are the hatred data and the quantity of interposition words. They likewise utilized deciphered SentiWordNet for the slant characterization. Every one of the orders was directed for machine learning calculations. The trial outcomes demonstrated the extra highlights are very powerful in the sarcasm detection[3].

Rajadesingan A et al. (2015) Their paper intended to address thetroublesome errand of sarcasm recognition on Twitter by utilizing social qualities characteristic for clients communicating sarcasm. They distinguish such qualities utilizing the client's past tweets. They utilize hypotheses from social and mental examinations to develop a conduct demonstrating system tuned for distinguishing sarcasm. They show proficiency in distinguishing mocking tweets[4].

Joshi A et al. (2015) present a computational framework that saddles setting incoherency as a reason for sarcasm identification. Their factual sarcasm classifiers consolidate two sorts of confusion highlights: implicit and explicit. They demonstrate the advantage of disjointedness highlights for two content structures - tweets also, dialog gathering posts[5].

Ravi K et al. (2015) paper exhibits a thorough study on SA, which depicts sees introduced by more than one hundred articles distributed in the most recent decade with respect to important assignments, methodologies, and uses of supposition examination.[6].

Anandkumar D. Et al. (2016) In their examination they have attempted to distinguish diverse supervised classification systems chiefly utilized for sarcasm identification and their highlights. Also they have investigated the effects of the arrangement methods, on printed information accessible in different dialects on survey related destinations, online networking locales and small scale blogging destinations.. They likewise completed primer analysis to recognize sarcasm sentences in "Hindi" dialect. [7].

Goel L et al. (2016) invest a great deal of their energy in the web what's more, a prevailing piece of which is spent on searching diverse casual networks. In this undertaking, OM because of social swarming is found utilizing revolutionary algorithm of swarm knowledge. The development of assessment in a discussion about an item or occasion is contemplated here as it gives an elective method for dissecting the nearness of estimation in online correspondences. [8].

JainT et al. (2017) utilizes two gathering based methodologies – random forest classifier and voted ensemble classifier. Not at all like current ways to deal with sarcasm identification which depend on existing corpus of positive and negative opinions for preparing the classifiers, they utilize a seeding calculation to create preparing corpus. The proposed demonstrate additionally utilizes a sober minded classifier to distinguish emoji based sarcasm [9].

Yadav P et al. (2017) concentrated more on how emojis assume an imperative job in estimation examination. Different variables that influence opinion examination are talked about quickly in this paper. Additionally different issues like sarcasm recognition, multilingualism, taking care of acronyms and slang dialect, lexical variety and dynamic word reference dealing with are examined[10].

Dharwal P et al. (2017) paper concentrate around different sarcasm examining procedures utilized for sifting of sarcasm explanations from content and the utilization of Automatic sarcasm identification in the arrangement of tweets and item audit writings [11].

Cambria E et al. (2017) Albeit most works approach it as a straightforward order issue, notion investigation is really a bag inquire about issue that requires handling numerous NLP assignments. Author address the composite idea of the issue by means of a three-layer structure enlivened by the "jumping NLP bends" worldview. Specifically, they contend that there are 15 NLP issues that should be comprehended to accomplish human-like execution in SA [12].

III. METHODOLOGY

The proposed methodology has three major stages: (A) Data preprocessing,(B) Optimization, (C) Sarcasm detection

A. Data Preprocessing

It is the most crucial step of the algorithm. It includes five steps.

- *Hashtag recognition andreplacement*

Hashtags are words or articulations gone before by a hash sign (#), used through electronicsystems administration media locales and applications, especially Twitter, to recognize messages about a specific point.The hashtags in the dataset are abundance.

- *Emoji dictionarymapping*

The noticeable example of using emojis in tweets can't be ignored as it passes on an impressive proportion of weightage for course of action. The emojis in a tweet are perceived and after that mapped with the physically created emoji word reference introduced in this paper. The word reference contains the notable emojis named as positive or negative. Comparable names supplant the emojis in a tweet in the midst of this movement. In this way every one of the emojis are supplanted by the imprintsmoreover.

- *Word parsing and tokenization*

In this stage, every client survey parts into expressions of any normal handling dialect. Tokenization is the way toward breaking input content into little ordering components – tokens.

- *Parts of speech labelling(POS)*

POS labeling incorporates the utilization of modifiers at that point there is a probability that the tweet is depicting something with a lot of acclaim that will give us the clue about it beingsarcasm.

- *Removal of stopwords*

Stop words are the words that contain little information so ought to have been ousted. As by ousting them, execution increases.

- *Stemming andlemmatization*

The objective of both stemming and lemmatization is to decrease inflectional structures and some of the time derivationally related types of a word to a typical base frame.

For instance, "studies", "studying" as dependent on the root word "study".

B. Optimization

The optimization is done using particle swarm optimization (PSO). PSO is a computational strategy that enhances an issue by iteratively endeavoring to enhance an applicant arrangement with respect to a given proportion of value.PSO is utilized for highlightdetermination. It is for separating immaterial or excess highlights from our dataset. The guideline of PSO calculation can be portrayed as pursued. An answer of the ideal issue is considered to a molecule in multi-measurement space and every molecule has a comparing versatile esteem which is chosen by complaint work; flying at a specific speed in the looking space, a molecule can change its speed and position by following its present neighborhood best esteem and the populaces' worldwide ideal esteem with the goal that the ideal arrangement of the issue might be found. In this, an occurrence in dataset is viewed as a molecule while include characteristics establish of multi-measurement space aside from choice properties. basic coordinating coefficient, Jaccard coefficient and soon.

C. Sarcasm Detection

After applying particle swarm optimization, we get optimized data, that is, noisy data is being removed. To detect sarcasm with high accuracy and good performance, ensemble technique bagged gradient boosting is used. Gradient boosting is used as a base learner in bagging. The proposed model is compared with other classifiers such as bagging, gradient boosting, random forest.

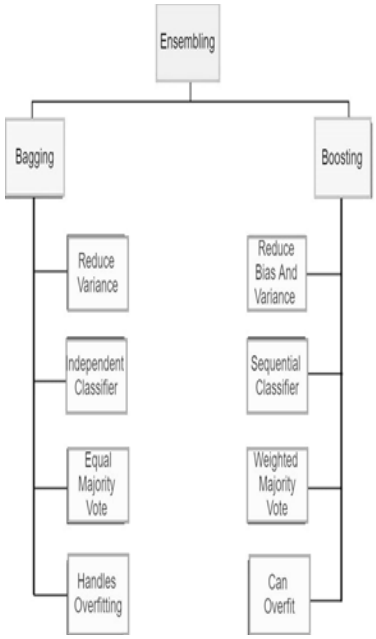

Figure 1: Ensembling methods

Ensemble method is a collection of two or more learners/classifiers. The primary driver of blunder in learning are because of noise, bias and variance. Ensemble limits thesecomponents. Random Forest classifier works by building multiple decision tress and obtaining class label.Bagging (Bootstrap Aggregation) is utilized when objective is to decrease the variance of a decision tree. Here thought is to make a few subsets of information from preparing test picked haphazardly with substitution. Presently, every gathering of subset information is utilized to prepare their choice trees. Thus, we end up with a group of various models. Normal of the considerable number of expectations from various trees are utilized which is more powerful than a solitary decision tree. Gradient boosting generates learners during the learning process. It builds first learner to predict the values/labels of samples, and calculates the loss (the difference between the outcome of the first learner and the real value). It then builds a second learner to predict the loss after the first step. The step continues to learn the third, fourth and so on, until a certain threshold.The reason we utilize ensembles is that various indicators attempting to foresee same target variable will play out a superior job than any single indicator alone. Boosting calculations are seen as more grounded than bagging on noise free data; regardless, bootstrap aggregation is essentially more robust than boosting in uproarious settings. Consequently, in this work, we manufactured an ensemble technique bagging using gradient boosting as a base learner.

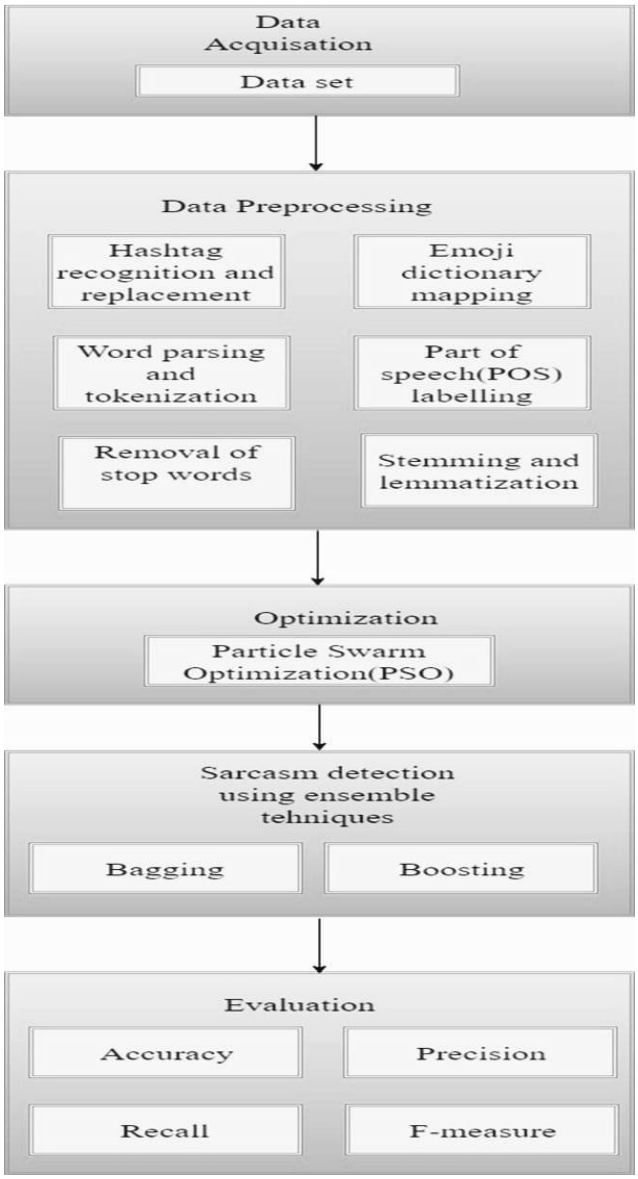

Figure 2: Workflow of proposed methodology

IV. EXPERIMENTAL SETUP

A. Dataset

Twitter produces a huge number of tweets every hour. It is a troublesome errand to acquire an ongoing dataset forpreparing or even a subject explicit dataset as it will prompt either too numerous or too less highlights for order. In this paper, we have considered a dataset which is manually arranged, the tweets in the dataset are physically named sarcastic and non-sarcastic dependent on human instinct which has organized an exact dataset for preparing. The physically ordered dataset is oneof the presentation in this paper. The dataset contains an accumulation of 1000 tweets. The dataset taken is that of around 1000 pre-characterized tweets.

B. Experimentation

The experiment is done using weka tool. Weka contains an accumulation of visualization tools and algorithms for information analysis and prescient demonstrating, together with graphical UIs for simple access to these capacities. Weka supports several standard data mining tasks, morespecifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. The simulation has been done in Java Net Beans. The setup was used to classify data as sarcastic or non-sarcastic. The results are presented below.

Table 1: Results

	Gradient Boosting	Random Forest	Bagged Gradient Boosting	Bagged Gradient Boosting with PSO
Accuracy	90.8909%	90.8184%	91.4915%	91.7918%
Precision	0.910	0.909	0.917	0.919
Recall	0.909	0.908	0.915	0.918
F-measure	0.909	0.908	0.915	0.918
TP Rate	0.909	0.908	0.915	0.918
FP Rate	0.093	0.094	0.088	0.084

Table 1 shows the results obtained by using different classifiers. The best accuracy is shown by bagged gradient boosting with PSO. In this case the correctly classified instances are 917 and incorrectly classified instance are 82.

C. Discussion

The paper presents comparative results of the four classifiers utilized. The accuracy of the outcomes endure a hit because of the measure of information considered for preparing. The dataset contains 1000 physically arranged tweets. It can be noticed that optimization followed by classification improves the accuracy. The best performing algorithm is bagged gradient boosting with PSO as feature selection. PSO optimizes the data and thus helped in reducing the time by removing noise and improved the accuracy. Bagged gradient boosting has reduced bias and variance. The goal of any supervised machine learning algorithm is to accomplish low noise, low bias and low variance, which can be achieved by using PSO followed by bagging and boosting. Alsoit will enhance the performance andaccuracy.

V. CONCLUSION

Sarcasm detection is a really fascinating subject. The paper presented illustrations containing the dataset, approaches and performance values.

A method for improving the existent sarcasmdetection calculations by includingbetter pre-processing techniques and optimization is introduced to improve the performance. Ensemble techniques bagging and boosting are combined which contribute to the improvement of accuracy. In the future work, the dataset can be expanded to get better results.

REFERENCES

1. Manuel K, Indukuri KV, Krishna PR. Analyzing internet slang for sentiment mining. In2010 Second Vaagdevi International Conference on Information Technology for Real World Problems 2010 Dec 9 (pp. 9-11). IEEE.
2. González-Ibánez R, Muresan S, Wacholder N. Identifying sarcasm in Twitter: a closer look. InProceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2 2011 Jun 19 (pp. 581-586). Association for ComputationalLinguistics.
3. Lunando E, Purwarianti A. Indonesian social media sentiment analysis with sarcasm detection. InAdvanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on 2013 Sep 28 (pp. 195- 198).IEEE..
4. Rajadesingan A, Zafarani R, Liu H. Sarcasm detection on twitter: A behavioral modeling approach. InProceedings of the Eighth ACM International Conference on Web Search and Data Mining 2015 Feb 2 (pp. 97-106).ACM.
5. Joshi A, Sharma V, Bhattacharyya P. Harnessing context incongruity for sarcasm detection. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) 2015 (Vol. 2, pp.757-762).
6. RaviK,RaviV.A surveyonopinionminingandsentimentanalysis:tasks, approaches and applications. Knowledge-Based Systems. 2015 Nov 1;89:14- 46.
7. Dave AD, Desai NP. A comprehensive study of classification techniques for sarcasm detection on textual data. InElectrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on 2016 Mar 3 (pp. 1985-1991).IEEE.
8. Goel L, Prakash A. Sentiment Analysis of Online Communities Using Swarm Intelligence Algorithms. InComputational Intelligence and Communication Networks (CICN), 2016 8th International Conference on 2016 Dec 23 (pp. 330-335).IEEE.
9. Jain T, Agrawal N, Goyal G, Aggrawal N. Sarcasm detection of tweets: A comparative study. InContemporary Computing (IC3), 2017 Tenth International Conference on 2017 Aug 10 (pp. 1-6).IEEE.
10. Yadav P, Pandya D. SentiReview: sentiment analysis based on text and emoticons. InInnovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on 2017 Feb 21 (pp. 467-472).IEEE..
11. Dharwal P, Choudhury T, Mittal R, Kumar P. Automatic sarcasm detection using featureselection.
12. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intelligent Systems. 2017Nov;32(6):74-80.

AUTHORS PROFILE

Himani Khullar: An M.Tech student at the Lovely Professional University, Punjab. A merit student with 94% in 12th and 92% in 10th. Current cgpa is 9.5.

Amritpal Singh: AnM.Tech graduate and currently an Assistant Professor in Computer Science and Engineering at the Lovely Professional University, Punjab. Currently pursuing Ph.D. from Lovely Professional University aiming to find solution of Unit Commitment problem in power systems.

Publications: **Amritpal Singh,** Nitin Umesh, "Implementing log-based security in data warehouse" (International Journal of Advanced Computer Research Vol 3 (1) March 2013)

Amritpal Singh, Implementation model for access control using log-based security: A practicalapproach" (Conference paper: Computer Engineering and Applications, March 2015)

Gaurav Kumar Pandey, **Amritpal Singh**, "Energy conservation and efficient data collection in WSN-ME: A survey" (Indian Journal of Science and Technology Vol 8(17) August 2015),

Amritpal Singh, Sushil Kumar, "Differential evolution: an overview" (Conference Paper: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, March 2016) and in the book: Advances in Intelligent Systems and Computing, March 2016)

Gaurav Kumar Pandey, **Amritpal Singh**, "Recent Advancements in Energy Efficient Routing in Wireless Sensor Networks (WSN): A survey" (Conference paper and book: Advances in Intelligent Systems and Computing, March 2016)

Gaurav Kumar Pandey, **Amritpal Singh**, "Residual Energy based One-Hop Data Gathering in Wireless Sensor Networks" (International Journal of Computer Science and Information Security (IJCSIS),Vol 14 (4), April 2016)