International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Sarcastic Detection of Twitter Comments using Python

Nishan.A.H, Joy Winnie Wise.D.C, Malaiarasan.S, Gopala Krishnan.C

Abstract*: A sarcasm / joke is a language of expressing the feeling in an opposite manner. In case of aural or cinematographic information the sarcasm can be identified easily through the tonal stress, gestures and facial expressions. But the main challenge of sarcasm is sarcastic detection in textual information where there is an absence of expressions or tonal stress. Presently a days, the utilization of informal organizations are enormously expanded. Most of the people express their feelings in their posts and comments in sarcastic manner. Sarcasm creates inquisitiveness and courtesy towards it in sentimental analysis. Wistful Analysis is the procedure or investigation of examining the sentiments. In this project, I have chosen twitter comments for this sarcastic sentimental analysis which is commonly an opinion mining. The importance of the project is to increase the accuracy rate by feeding huge data set for training. The purpose of finding the sarcasm in social networks is to block the user who points particularly or attack any victim which is not considered as sarcasm.

***Keywords* : Blocking, gesture, sarcasm, sentimental analysis, twitter comments.**

I. INTRODUCTION

Sarcasm is a piece of human instinct and maybe a developmentally honorable element. It is the daily schedule of comments that without a doubt allude to something contrary to what the people say and made so as to miffed somebody's sentiments or to deride something in a crazy manner. The understanding the delicacy of this training needs second-request clarification of the storyteller's or creator's targets; various pieces of the mind must exertion together to get sarcasm. Sarcasm seems to work out the mind more than certifiable tributes do. Sarcasm has a deceptive quality: it's both humorous and implies. In this way, the specialists show interest in sarcasm recognition of internet based life content, particularly in tweets.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author

Nishan.A.H*, Department of Computer Science and Engineering, Francis **Xavier** Engineering College, Tirunelveli, India. Email:caaju196@gmail.com

Dr. Joy Winnie Wise.D.C, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, India. Email:joywinniewise@yahoo.com

Malaiarasan.S, Department of Computer Science and Engineering, **Francis** Xavier Engineering College, Tirunelveli, India. Email: malai.sweng@gmail.com

Dr. Gopala Krishnan.C, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, India. Email:skywarekrish@gmail.com

The quick development of tweets prompts basic in the investigation of information. It is otherwise called conclusion mining that infers the assessment of an individual or frame of mind of a speaker. Numerous specialists center their enthusiasm towards wistful investigation especially in the field of the informal community from the previous hardly any years. AI strategies and calculations clear another route for notion examination especially sarcasm discovery by giving a lot of calculations and methodology. The headways of innovations and new creations of gadgets, there have a fast development in producing information by every single gadget of the system. At the period of digitalized time and blasting of web utilization in the two urban areas and rustic regions, numerous individuals currently can compose and impart their insight as simple as taking treat from an infant. Web based life or miniaturized scale blogging like Face book or Twitter give the client numerous abilities to impart their insight. In a study directed by Princeton Survey Research Associates International done in 2014, from 81% of all out American grown-ups who utilize the web, 52% utilize at least two internet based life locales. It shows an extension of 10% from the previous year. In online life the client posts numerous things and for that post individuals will remark beneficial thing just as awful moreover. As a result of this a few people get influenced of awful remarks posting out in the open. To beat this issue twitter information is utilized for dissecting and recognizing whether the given remark is mocking or typical.

II. RELATED WORK

Satire exposes humanity's vices and foibles through the utilization of irony, wit, and sometimes sarcasm too. It is likewise oftentimes utilized in online networks. Acknowledgment of parody can help in numerous NLP applications like discourse framework and survey synopsis. They filtered online news articles as satirical or true news documents using SVM (Support Vector Machine) classification method combined with machine learning techniques. With ample training documents SVM tends to offer good classification results. For getting promising outcomes with SVM a comprehension of its working and approaches to impact its exactness is required. They also use various feature extraction strategies and conclude that TF-IDF-BNS feature extraction gives maximum accuracy for detection of satire in web content [1].Sarcasm analysis, being one of the toughest challenges in tongue processing (NLP), has become a hot topic of research lately . A lot of labor has already been wiped out the sector of sentiment analysis, but there are huge challenges still being faced in identification of sarcasm.

The property of sarcasm that creates it difficult to research and detect is that the gap between its literal and intended meaning. Distinguishing opinion in web based life like Face book, Twitter, online sites, and audits has become a significant undertaking as they impact each business.

In this part, four methodologies were proposed, to be specific parsing-based lexical age calculation, different preferences logical inconsistency, tweet negating all inclusive realities, and tweet repudiating impermanent actualities. The aim of the proposed methods is to extract text features like lexical, hyperbole, behavioral, and universal facts. Further, four machine learning classifiers, to be explicit support vector machine, Naive Bayes, maximum entropy, and decision tree, were sent. Finally, trained those classifiers using an extracted feature set to identify sarcasm in Twitter data. This work achieves an extensive precision improvement over existing systems [2].During the last decade majority of research has been carried out in the area of sentiment Analysis of textual data available on the web. Sentiment Analysis has its challenges, and one among them is Sarcasm. Classification of sarcastic sentences may be a difficult task thanks to representation variations within the textual form sentences. This can affect many tongue Processing based applications. Sarcasm is that the quite representation to convey the various sentiment than presented. They have tried to identify different supervised classification techniques mainly used for sarcasm detection and their features. Furthermore, for every method studied, our paper presents the analysis of knowledge set generation and have selection process used thereof. They carried out preliminary experiment to detect sarcastic sentences in “Hindi” language. They found that this simple model based on “bag-of-words” feature accurately classified 50% of sarcastic sentences. Thus, primary experiment has revealed the fact that simple Bag-of-Words are not sufficient for sarcasm detection [3].

Sarcasm presents a negative meaning with positive expressions and may be a non-literalistic expression. Sarcasm detection is a crucial task because it contributes on to the development of the accuracy of sentiment analysis tasks. First, analyzed sarcastic sentences in product reviews and classify the sentences into 8 classes by that specialize in evaluation expressions. Next, generated classification rules for each class and use them to extract sarcastic sentences. Our method consists of three stage; judgment processes supported rules for 8 classes, boosting rules and rejection rules. Then, compared the method with a baseline based on a simple rule. The experimental result shows the effectiveness of our method [4].Sarcasm presents a negative meaning with positive expressions and is a non-literalistic expression. Sarcasm detection is a crucial task because it contributes on to the development of the accuracy of sentiment analysis tasks. They proposed an extraction method of sarcastic sentences in product reviews. First, they analyzed sarcastic sentences in product reviews and classify the sentences into 8 classes by that specialize in evaluation expressions. Next, they generated classification rules for each class and use them to extract sarcastic sentences. Our method consists of three stage, judgment processes supported rules for 8 classes, boosting rules and rejection rules. They compared their method with a baseline based on a simple rule. The experimental result shows the effectiveness of our method [5]. Sarcasm may be a sort of language during which individual convey their message in an implicit way i.e. the opposite of what is implied. Sarcasm detection is that the task of predicting sarcasm in text. This is the crucial step in sentiment analysis thanks to inherently ambiguous nature of sarcasm. With this vagueness, sarcasm detection has consistently been a troublesome errand, in any event, for people. Therefore sarcasm detection has gained importance in many tongue processing applications [6].

Natural Language Processing (NLP) is one of the significant areas in Artificial Intelligence. It goes about as a stage between the PC and human languages. It works for structured and unstructured data. Sentimental Analysis is one of the important fields in Natural Language Processing (NLP) which deals with analyzing the context. Sentimental Analysis is the processes of analyzing the opinions expressed by the writer and determine the attitude towards the topic. It is wont to classify the polarity of a document or an opinioned text. Several analyses can be performed using sentimental analysis. It provides a quick understanding of the writer’s attitude. It is sometimes known as opinion mining where it speaks about a particular entity and pre-processing and classification techniques are used in sentimental analysis. Tweets are characterized into six types: abusive, comparison, passing judgment, religious or ethnic, sarcasm or joke, whataboutery. The main challenge in sentimental analysis is sarcasm detection. A sarcasm / joke is a language of expressing the feeling in an opposite manner. In case of aural or cinematographic information the sarcasm can be identified easily through the tonal stress, gestures and facial expressions. But the main challenge of sarcasm is sarcastic detection in textual information where there is an absence of expressions or tonal stress. Now a day, the usage of social networks is hugely increased. Most of the people express their feelings in their posts and comments in sarcastic manner. Sarcasm creates inquisitiveness and courtesy towards it in sentimental analysis. Sentimental Analysis is the process or study of analyzing the feelings. There are various applications of sarcastic text detection. It is used for letting the reviewer know the intent of the writer and the context in which it is said [7]. Opinion mining and sentiment analysis ask the identification and therefore the aggregation of attitudes or opinions expressed by internet users towards a specific topic. However, thanks to the limitation in terms of characters (i.e. 140 characters per tweet) and therefore the use of informal language, the state-of-the-art approaches of sentiment analysis present lower performances in Twitter than that once they are applied on longer texts. Moreover, presence of sarcasm makes the task even tougher .Sarcasm is when an individual conveys implicit information, usually the other of what's said, within the message he transmits. They proposed a method that makes use of a minimal set of features, yet, efficiently classifies tweets regardless of their topic. It also implies the importance of detecting sarcastic tweets automatically, and demonstrate how the accuracy of sentiment analysis are often enhanced knowing which tweets are sarcastic and which are not [8]. The nearness of sarcasm in content can hamper the exhibition of sentiment analysis. The test is to identify the presence of sarcasm in writings. This challenge is compounded when bilingual texts are considered, for instance using Malay social media data.

In this paper a feature extraction process is proposed to detect sarcasm using bilingual texts; more specifically public comments on economic related posts on Face book. Four categories of feature which will be extracted using tongue processing are considered; lexical, pragmatic, prosodic and syntactic. They also investigated the use of idiosyncratic feature to capture the peculiar and odd comments found in a text. To determine the effectiveness of the proposed process, a non-linear Support Vector Machine was wont to classify texts, in terms of the identified features, consistent with whether or not they included sarcastic content or not. The results obtained demonstrate that a mixture of syntactic, pragmatic and prosodic features produced the simplest performance with an F-measure score of 0.852 [9]. In general, the most challenging problem for the opinion mining task is sarcasm detection. To be ready to do this, many researchers tried to explore such properties in sarcasm like theories of sarcasm, syntactical properties, psycholinguistic of sarcasm, lexical feature, semantic properties, etc. Studies wiped out the last 15 years not only made progress in semantic features, but also show increasing amount of method of study employing a machine-learning approach to process data. Because of this reason, this paper will try to explain current mostly used method to detect sarcasm [10]. Sarcasm detection has been modeled as a binary document classification task, with rich features being defined manually over input documents. Traditional models employ discrete manual features to deal with the task, with much research effect being dedicated to the planning of effective feature templates. They investigated the utilization of neural network for tweet sarcasm detection, and compare the consequences of the continual automatic features with discrete manual features. In particular, they use a bi-directional gated recurrent neural network to capture syntactic and semantic information over tweets locally, and a pooling neural network to extract contextual features automatically from history tweets. Results show that neural features give improved accuracies for sarcasm detection, with different error distributions compared with discrete manual features [11].

III. PROPOSED WORK

Manual analysis of numerous sentences is costly. Therefore, generating rules automatically becomes necessary. The classifiers used here is Naive Bayes for predicting the accuracy rate higher. Here, we use 21 special features along with usual unigrams and bigrams for classification. These 21 features were divided in to 4 categories: Text expression-based features, Emotion-based feature, Familiarity-based feature and Contrast-based features.

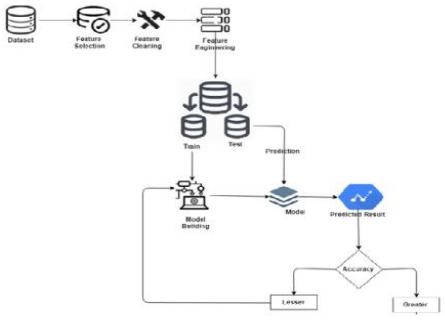

Fig: Architecture Diagram

A. Modules Implementation

Analyzing Data

The benefit of using these words based on their entropy score in the feature-set is that we were able to reduce uncertainty in the prediction outcome as these words have a different impact of frequency count in sarcasm and non-sarcasm tweets. After collecting 400,000 labeled tweets, we extract around 350,000 English tweets.

Feature Selection

Feature selection is based on which features will make an impact in our project and which feature we don't need to use. The features we need to use are extracted from the dataset and other features are left as it is. The feature can be multiple class as well as Single feature so we need to decide how our feature should come.

Preprocessing

Analyzing of the data helps in screaming of the data carefully which can rectify misleading results. Preprocessing is done in three major steps like: Feature Extraction, Feature Cleaning, Feature Engineering

Feature Extraction and Feature Engineering

The text must be parsed to eliminate words, called tokenization. Then the words need to be determined as integers or floating point value for use as input to a machine learning algorithm, called feature extraction.

This is the most important phase in the development of the system. Before applying feature extraction algorithms, the stemming of words was performed. Stemming is the process in which the words are shortened and normalized to their stem and their tenses are ignored. The root of the word is preserved for better efficiency of feature extraction and to reduce redundancy. This system takes into account the features developed from N-grams, sentiments, topics, pos-tags, capitalization, etc. The features from N grams are majorly unigrams and bigrams. Topics are basically words which have a high probability of appearing together. We extract the topics from the dataset and assign separate scores to them.

Feature cleaning

After the feature extraction we'll search for the any null parameters in my data's. If there is any null parameters there means we need to fill the parameters with the related content. In this process the steaming of the words are done. The similar words are considered as one and the model is being build.

Model Building

After the Preprocessing we need to build the model according to the features if our feature has labels means model can be built as Supervised algorithm if my feature doesn't have the labels means we need to use Unsupervised learning algorithms if there is combination means Semi supervised Ensemble Model should be used.

IV. METHODOLOGY

A. Classification

Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification issue, the objective variable (or yield), y, can take just discrete qualities for given arrangement of highlights (or data sources), X.

The goal of logistical regression is to properly predict the class of outcome for individual cases victimization the foremost ungenerous model.

A new model is created with all predictor variables to accomplish the goal that are useful for predicting the response variable. Logistic regression may be a mathematical model utilized in statistics to estimate the probability of a happening occurring having been given some previous data. Logistic regression works with binary information, where either the event happens (1) or the event does not happen (0).

Decision Tree

There are two categories of decision trees: Classification trees & Regression trees. The decision tree learning is the creation of a decision tree from class-labeled training tuples. A decision tree consists of nodes that form a tree structure; the top most node is called the root node. Each non-leaf node denotes a test on a characteristic; each division represents the result of a test, and each leaf node hold a class label. Leaf nodes represent classes that are return if reach as the final calculation by the model. As Zaki & Meira (2014) elaborate, given an occurrence with its features values, the model is able to classify the instance by traverse the decision tree. There are a few decision tree algorithms including: ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3) and CART (Classification and Regression Tree).

Support Vector Machine

Support vector machines (SVM) were first introduced by Vapnik, 1992 to solve binary classification problems, then they are comprehensive to nonlinear regression problems. SVMs are based on structural risk minimization dissimilar ANNs which is based on experiential risk minimization. They use a nonlinear mapping to change the input data into a multidimensional feature break. After this alteration the SVM finds the best overexcited plane within the feature space. The nonlinear mapping depends on what so call a most important part function.

B. Prediction

Metric Score

Metric score is after the model constructed can't implement in real world without testing. So we'll test the model using test set and we are having the original values. After the model's prediction it is compared with the original values and we can find the metric score called Accuracy.

Accuracy

Accuracy is that the most intuitive performance live and it's merely a magnitude relation of properly foretold observation to the entire observations. One might imagine that, if we've high accuracy then our model is best. Wherever values of false positive and false negatives area unit virtually same. Therefore, you have got to appear at different parameters to judge the performance of your model. For our model, we've got zero.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

Precision

Preciseness is that the magnitude relation of properly foretold positive observations to the entire foretold positive observations. High accuracy identifies with the low bogus positive rate. We have got zero.788 preciseness that is pretty sensible.

Precision = TP/TP+FP

Recall (Sensitivity)

Recall is that the magnitude relation of properly foretold positive observations to the all observations in actual category affirmative. We have got recall of zero.631 that is nice for this model as it's on top of zero.5.

Recall = TP/TP+FN

F1score

F1 Score is that the weighted average of preciseness and Recall. Therefore, this score takes each false positives and false negatives under consideration. Intuitively it's not as straightforward to know as accuracy, however F1 is sometimes a lot of helpful than accuracy, particularly if you have got AN uneven category distribution. Accuracy works best if false positives and false negatives have similar price. If the value of false positives and false negatives area unit terribly totally different, it's higher to appear at each preciseness and Recall. In our case, F1 score is 0.701.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

V. RESULT

As a result, the sarcastic content was classified as plain text, emotion, contrast and context as shown in a graph below.

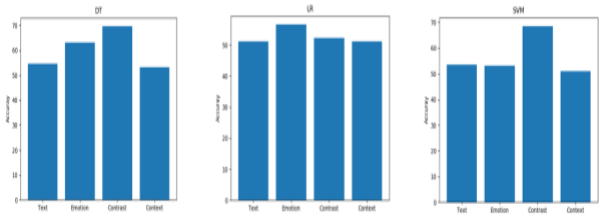

Fig: Sarcasm in various forms

The sarcastic text are detected with the accuracy ratio which was calculated from the formula (Accuracy = TP+TN/TP+FP+FN+TN) and those were tabulated below along with the classification method used to detect the sarcastic text.

Tab: Method versus Accuracy

Classification Method	Accuracy
Logistic Regression	0.70
Decision Tree	0.64
Support Vector Machine (SVM)	52.24

VI. CONCLUSION

The way of up the existent caustic remark detection algorithms by as well as higher pre-processing and text mining techniques like emoji and slang detection area unit given. For classifying tweets as sarcasm and no sarcasm there are various techniques used, however, the paper takes up a classification algorithm and suggests various improvements that directly contribute to the advance of accuracy. The project derived analytical views from a social media dataset and also filtered out or reverses analyzed sarcastic tweets to achieve a comprehensive accuracy in the classification of the info that's given. The model has been tested in time period and may capture live streaming tweets by filtering through hash tags so perform immediate classification.

For future work, the proposed model will be implemented in several other algorithms to improve the accuracy of the system. The accuracy of sarcastic detection can be improved by training the system with the algorithms which gives an accurate percentage.

REFERENCES

1. Ahmad, Tanvir, et al. "Satire detection from web documents using machine learning methods." 2014 International Conference on Soft Computing and Machine Intelligence. IEEE, 2014.
2. Bharti, Santosh Kumar, et al. "Sarcasm analysis on twitter data using machine learning approaches." Trends in Social Network Analysis. Springer, Cham, 2017. 51-76.
3. Dave, A. D., & Desai, N. P. (2016, March). A comprehensive study of classification techniques for sarcasm detection on textual data. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 1985-1991). IEEE.
4. Dave, Anandkumar D., and Nikita P. Desai. "A comprehensive study of classification techniques for sarcasm detection on textual data." 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE, 2016.
5. Hiai, Satoshi, and Kazutaka Shimada. "A sarcasm extraction method based on patterns of evaluation expressions." 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, 2016.
6. Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., & Carman, M. (2016). Are word embedding-based features useful for sarcasm detection? arXiv preprint arXiv:1610.00883.
7. Rajesh Basak,ShamikSural, NiloyGanguly, Soumya K. Ghosh "Online Public Shaming on Twitter: Detection, Analysis and Mitigation" IEEE Transactions on Computational Social Systems 2019.
8. Selvan, LokmanyathilakGovindanSankar, and Teng-Sheng Moh. "A framework for fast-feedback opinion mining on Twitter data streams." 2015 International Conference on Collaboration Technologies and Systems (CTS). IEEE, 2015.
9. Suhaimin, MohdSuhairiMd, et al. "Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts." 2017 8th International Conference on Information Technology (ICIT). IEEE, 2017.
10. Wicana, SetraGenyang, TahaYasinİbisoglu, and UrazYavanoglu. "A Review on sarcasm detection from machine-learning perspective." 2017 IEEE 11th International Conference on Semantic Computing (ICSC). IEEE, 2017.
11. Zhang, Meishan, Yue Zhang, and Guohong Fu. "Tweet sarcasm detection using deep neural network." Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers. 2016.

AUTHORS PROFILE

Nishan.A.H, received the B.Tech degree in Information Technology from Francis Xavier Engineering College, Affiliated to Anna University, Tirunelveli, Tamil Nadu, India in 2014. She is currently pursuing the M.E degree in Computer Science and Engineering with Francis Xavier Engineering College, An Autonomous Institution, Tirunelveli, Tamil Nadu, India.

Dr. Joy Winnie Wise.D.C, Principal, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India. She received her B.E., M.E., and Ph.D degrees in Computer Science and Engineering. Her Academic Research was Computer Architecture and Microprocessor & Microcontroller. Her Research Interests are Computer Network and Parallel Processing. She has fourty eight publications. She has attended twenty six seminars, fifteen conferences and seven workshop. She is an active member of CSI.

Malaiarasan.S, Assistant Professor, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India for the past ten years. He received his B.E. and M.E. degrees in Computer Science and Engineering during 2005 and 2013. His Research Interests are Image Processing and Data Mining. He has sixteen publications. He has attended two seminars, eleven conferences, five workshop and has delivered a lecture. He is an active member of CSI.

Dr. Gopala Krishnan.C, Professor, Department of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India. He received his B.E., M.E., and Ph.D degrees in Computer Science and Engineering. His Academic Research was Cloud Computing. His Research Interests are Mobile Communication and Cloud Computing. He has fifteen publications. He has attended three seminars, twenty conferences, nine workshop and has delivered a lecture. He is an active member of ISTE. He has applied for DST-SHRI Sponsored Research.