# SOCIAL NETWORK SARCASTIC DATA ON CONTEXTUAL TEXT INVOLVING IN SENTIMENT ANALYSIS ON TWITTER CORPUS OF BIG DATA APPLICATION

**N. Karthikeyan**
*Department of Computer Science, Srimad Andavan Arts and Science College, India*

*Abstract*

***Information retrieval in context based has become one of the impecuniousness in the lively information trends. Commonly information retrieval on context based is slowly important on sarcastic functions and now days should be very tough to data heterogeneity and not easy on modes of data circulate function. This paper presents a context based information retrieval sarcastic system to perform real time retrieval of data. This model operates on the semantic comparative of the data, rather than the content similarity. Hence this technique exhibits better and efficient retrieval levels providing high adequate approach. The retrieved data's are pre-processed and feature vectors are created from the small instance of ranking methods. Polarity matching is used to filter sentimentally correlated results and result based output ranking is performed to perform further elimination of proper circumstances of the results. Experiments conducted using the proposed model exhibits very high true retrieval rates, along with high precision and recall levels exhibiting the competence of the proposed work on the sarcastic text information retrieval is considered to be effective.***

*Keywords:*

***Context Based Data, Regression, Polarity, Semantic, Contextualization***

## 1. INTRODUCTION

Context based information retrieval is most popular and value based aspect on high corpus data in social media industry. In social media having much more sentiment comments as we observed as a sentiment polarity of manners should be important [3]. Information retrieval is the process of find out the related with right content to the context from a huge corpus at accomplish by some time stages at the limit [4]. In omniscient existing in the query and the document large data base also tend to an increase in the intricacy of the retrieval process. Problems existing in the initial steps of data storage, when the memory was not sufficient to hold the data. When the challenge of information retrieval process should be more constrain the method need to avail as easiest way in any shape of methods [5].

Several sub-operations are required for an effective information retrieval system. These include, effective content matching, effective polarity identification of text, effective analysis of polarity to identify sentiment related to the text and the magnitude of the sentiment levels contained in the text, and finally the overall analysis of these components to identify the most related content to be retrieved for the current query.

Effective content matching is the first phase of the process. This performs a rough comparison of the query and the content to be retrieved. The first phase roughly filters out most of the unrelated content from the major repository. This acts as a major filtering phase. The second level proceeds with identifying the polarity levels of the content. Polarity is the process of identifying the level of positivity or negativity contained in the text. This is usually identified using a sentiment repository. The significant terms are identified and for every significant term identified, its polarity is identified from the repository. These polarity levels define the sentiment corresponding to each word or token. However, it is also necessary to identify the polarity levels of the entire text. This corresponds to identifying the sentiment related to the text.

Sentiment analysis is the process of identifying the sentiment related to the document as a whole. The polarities that have been individually identified are aggregated to form the sentiment levels of the document as a whole. These levels are either identified as fuzzy levels; positive or negative, or as magnitude levels of the positive or negative intensities of the document. These effectively define the sentiment levels of the document. These levels can be compared with the sentiment levels associated with the query to identify the contextual relationship of the content with the query. Documents exhibiting high contextual similarities are filtered for the user.

Effective contextual based analysis hence requires effective operation of multiple modules together. This paper presents an effective model for contextual information retrieval. The model presents an effective feature extraction phase that analyses the input content to provide the most significant features contained in the text. These features are then used to identify the sentiment associated with the document. Sentiment identification is performed using the Naïve Bayes Classifier (NBC). Naïve Bayes is a probabilistic model that operates based on the assumption of attribute independence. This hence makes Naïve Bayes powerful on domains like sentiment analysis. The results from NBC is used to determine the contextual relationship of the text with the query. Based on the contextual relationship identified, the results are ordered and filtered and provided to the user. Experiments were performed with standard benchmark datasets and the results show that the proposed model exhibits high performances, depicting the high efficiency of the NBC model.

## 2. RELATED WORKS OF THE CONTEXT BASED INFORMATION RETRIEVAL

Information retrieval is a big destruction process of polarity analysis of social media networks like a Twitter, Facebook etc. with heterogeneous data. Context based information retrieval systems are on put on growth the effective due to the append in the amount of data available on live streams. Multiple combinations of contributions in this domain has been evidence and talk about more of the most projects and the latest researches in this domain. A semantic based information retrieval trace on graph theory was proposed in [7]. This method of functions based on graph-of-concepts, rather than the accordance of the logic-concepts. A graphical represents is created by reviews by the relationship among concepts and named entities. This also presents a similarity identification module to correlate graphs and found circumstances with information. Another context based information retrieval approach is specifically designed for retrieving information from the on net was proposed in [8]. This technique functions by identifying the gravity of the term during the retrieval process, making the process rely towards the contextual nature of the outputs rather than the term frequency. A context based text process files relevance assessment system for rich of information retrieval was proposed in [9]. This technique proposes an extend technique for logic by representation and document relevance of validate.

The vector space model and statistical language model were applied as major components for analysis. The model was created to enhance the atmosphere decision making process. A collective learning approach for effective geographical information retrieval was introduced in [10]. This technique operates on query contextualization to address the problem about the data assortments of the past entries. Context based information retrieval systems are on the high level owing to the increase in the amount of data available online. Multiple contributions in this domain has been focused and discusses some of the most important and the most recent researches in this domain. A semantic based information retrieval system based on graph theory was proposed in [11]. This technique operates based on graph-of-concepts, rather than the conventional bag-of-concepts. A graph is created by considering the relationship among concepts and named entities. This also presents a similarity identification module to correlate graphs and identify appropriate information. Another context based information retrieval technique specifically designed for retrieving information from the web was proposed in [13]. This technique operates by identifying the significance of the term during the retrieval process, making the process rely towards the contextual nature of the results rather than the term frequency. A context based document relevance assessment system for effective information retrieval was proposed in [12]. This technique proposes an enhanced technique for concept representation and document relevance recognition. The vector space model and statistical language model were used as the major components for analysis. The model was constructed to enhance the environmental decision making process. A collaborative learning approach for effective geographical information retrieval operates on query contextualization to address the problem of data heterogeneity.

The compare the prediction between the different type of versions with SentiWordNet and its features are also neat explained along with the research calculations of such a lexical resource in various automated text classification and sentiment polarity analysis. They have also describes of the algorithm for automatic WordNet annotations and how it specific applied as classifiers text into positive, negative and neutral elements.

A hybrid approach is developed for sentiment analysis based on rule based classification, supervised learning and machine learning. They have utilized that to movie comments as a reviews and product relevance feedback as a reviews and reported effective classification of sentiment polarity. Though the results are comparatively good the hybridization increases the computational complexity of the approach to a greater extent.

Sentiment analysis is considered on positive sentiment polarity and negative sentiment polarity of the emotional values with text methods upon the single hand and independent of related index with brief functions. Naive Bayes, maximum entropy classification, and support feature vector machines have been used for sentiment analysis by them and they have also reported that machine learning working on the datasets are good than human baseline when it comes to sentiment polarity

## 3. METHODOLOGY

Machine learning algorithms are commonly use that could learn from data and improve from periodical trains of the data. The function that maps the input to the output, learning the back end structure in not identify the correct ranking of the data. A class label is produced for an occurrences of the matrix from the training data operational and kept in a memory. Initially it does not take any impact of trained datasets. In the paper work presents Naive Bayes Classifier working very effective way approach on the basis related to Bayes theorem. Naive Bayes is a probability method in machine learning algorithm based on the Bayes theorem, used in a wide variety of classification tasks. The NBC is used (Fig.1) because it assumes the features that go into the model is independent of each other. That is changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm. The algorithm is unique approach of the data.

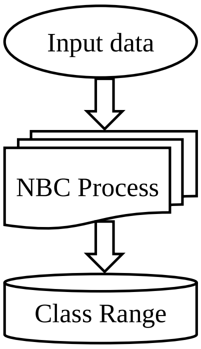

Fig.1. NBC Process

### 3.1 CONTEXT BASED INFORMATION RETRIEVAL IN SARCASTIC TEXT USING NBC METHODS

Context based retrieval of information has essentially important specific need for the active of current life industry in information retrieval systems [5]. However, analyzing the text based on the context rather than the content can provide effective of mutual relationship of connection between the tasks and hence better semantically towards the content [6]. Naive Bayes model is simple to build and particularly useful for data sets. In simplicity, Naive Bayes is known to outperform even highly involving with classification methods. The major reason for this requirement is that analyzing the content might not necessarily provide the required results. This leads to inappropriate content presented to the user. The proposed information retrieval architecture is presented in information Retrieval Models in the Context of Retrieval Tasks by machine learning process.

## 4. RESULTS AND DISCUSSIONS

In Machine Learning consists a set of input values as in axis of *x* that are used to find the output values of the datasets *y*. A relationship exists between the input variables and the output variable. The aim of ML is to quantify this relationship. We collected the sentiment word and workout the scores as a sarcastic and put on to trained of classifying with the method of regression method by NBC [1] [2].

Trained data: In Linear Regression, the re the values between the input variables *x* cons and output variable *y* pons is expressed as an equation of the form $y = a + bx$. The output results of linear regression is to represents out the values of coefficients *a* and *b*. In this paper obstruct the values *a* and *b* is the slope of the line. Graph shows the plotted *x* and *y* values for a dataset. The goal is to fit a line that is nearest to most of the points. This would reduce the error of classifying the SVM values *x* between the *y* value of a data point and the line.

The NBC is a way of going from $P(X|Y)$, known from the training dataset with the classifier's data in calculations to find $P(Y|X)$ [17]. In this work *A* positive polarity and *B* negative polarity in the above formula, with the feature *X* and response *Y*. For observations in test or scoring data, the *X* would be known while *Y* is unknown. And for each row of the test dataset, you want to compute the probability of *Y* given the *X* has already happened.

The result is probability of each class of *Y* and let the highest win. The Bayes Rule provides the formula for the probability of *Y* given *X*. But, in real-world problems, you typically have multiple *X* variables. This paper presents a system that solves this problem of context base applications out the required data with reduced ambiguity using big data techniques by NBC techniques. Generally, the regression is explained as a line in the form of $y = a + bx$.

Experimental results obtained are used to formulate the confusion matrix. Confusion matrix is composed of the number of true positives, false positives, true negatives and false negative values obtained from the proposed model. All the other metrics are obtained from these values. The other performance metrics used for analysis are True Positive Rate (*TPR*), False Positive Rate (*FPR*), Precision, Recall and F-Measure. The process of identifying the performance metrics from the confusion matrix is given below.

$$TPR = TP/(TP+FN)$$

$$FPR = FP/(FP+TN)$$

$$Recall = TP/(TP+FN)$$

$$Precision = TP/(TP+FP)$$

$$F\text{-}Measure = (2*Precision*Recall)/(Precision+Recall)$$

The performance of the proposed model in terms of Recall, Precision, Accuracy and F-Measure is shown in the Fig.2. The metrics shown in figure are aggregate metrics and hence represent the overall performance of the model. It could be seen that the recall level of 88%, precision of 100%, accuracy of 92% and F-Measure of 93% are shown by the model. This shows the high efficiency of the model.

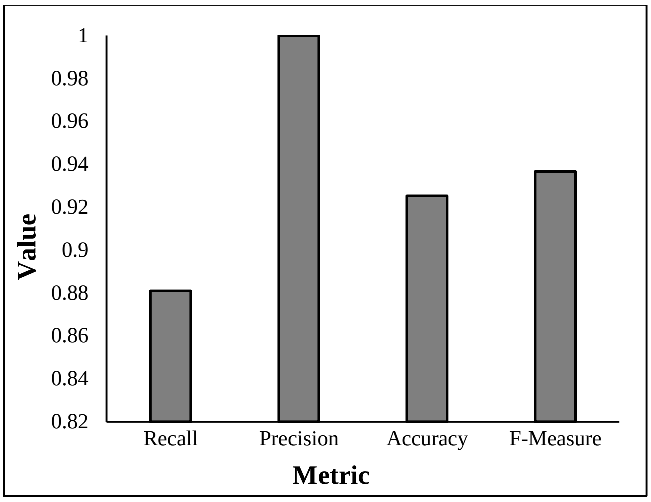

Fig.2. Aggregate Performance Measures

The ROC curve of the model is shown in Fig.3. ROC curve is plotted with FPR in *x*-axis and TPR in *y*-axis. It shows the prediction level of a model. The proposed model shows high TPR and low FPR showing high efficiency.

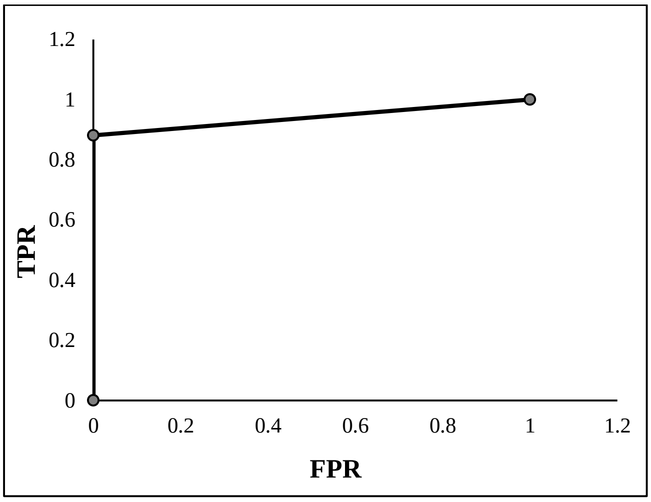

Fig.3. ROC Curve

A tabulated form of the results is shown in the below table.

Table.1. Performance Metrics

| Metric | Value |
|---|---|
| FPR | 0 |
| TPR | 0.881 |
| Recall | 0.881 |
| Precision | 1 |
| Accuracy | 0.9254 |
| F-Measure | 0.9367 |

## 5. CONCLUSIONS

Naive Bayes Classifier is good technical approach of the big data industry in live applications and predict for future behavior based on past results. The bias should value and understanding current sentiment actions in observational data's either if it is rest of movie functions. A comparative research on these classifiers was done to know which classifiers confirm with narrow results. The randomly effect on performance in Gradient developed classifier with features namely positive sentiment polarity, positive sentiment words, parts-of speech tagging, irony markers and ordered sequence of sentiment tags has achieved highest accuracy. The proposed model was designed with Naïve Bayes classifier to identify the sentiment levels pertaining to the documents. The NBC model was observed to exhibit high performance on the benchmark data exhibiting the high performing nature, hence promising high accuracy in retrieving information based on context.

## REFERENCES

[1] R. Shobana, K.S. Shalini, S. Leelavathy and V. Sridevi, "De-Duplication Of Data In Cloud", *International Journal of Chemical Sciences*, Vol. 14, No. 4, pp. 2933-2938, 2016.

[2] N. Kaaniche and M. Laurent, "A Secure Client-Side De-Duplication Scheme in Cloud Storage Environments", *Proceedings of IEEE International Conference on New Technologies, Mobility and Security*, pp. 1-7, 2014.

[3] J. Stanek, A. Sorniotti, E. Androulaki and L. Kencl, "A Secure Data De-Duplication Scheme for Cloud Storage", *Proceedings of International Conference on Financial Cryptography and Data Security*, pp. 99-118, 2014.

[4] K. Akhila, A. Ganesh and C. Sunitha, "A Study on De-Duplication Techniques over Encrypted Data", *Procedia Computer Science*, Vol. 87, pp. 38-43, 2016.

[5] B. Harish and K. Harshitha, "Data De-duplication In Cloud", *International Journal of Pure and Applied Mathematics*, Vol. 115, No. 8, pp. 353-358, 2017.

[6] M.P.D. Thakar and D.G. Harkut, "Hybrid Model for Authorized De-Duplication in Cloud", *International Journal of Emerging Trends and Technology in Computer Science*, Vol. 4, No. 1, pp. 147-151, 2015.

[7] F. Shieh, M.G. Arani and M. Shamsi, "De-Duplication Approaches in Cloud Computing Environment: A Survey", *International Journal of Computer Applications*, Vol. 120, No. 13, 2015.

[8] P. Puzio, R. Molva, M. Onen and S. Loureiro, "Perfect Dedup: Secure Data Deduplication", *Proceedings of International Conference on Data Privacy Management, and Security Assurance*, pp. 150-166, 2015.

[9] P. Priyadharsini, P. Dhamodran. And M.S. Kavitha, "A Survey On De-Duplication In Cloud Computing", *International Journal of Computer Science and Mobile Computing*, Vol. 3, No. 11, pp. 149-155, 2014.

[10] G.U. Devi and G. Supriya, "Encryption of Big Data in Cloud using De-duplication Technique", *Research Journal of Pharmaceutical Biological and Chemical Sciences*, Vol. 8, No. 3, pp. 1103-1108, 2017.

[11] D. Harnik, B. Pinkas and A. Shulman-Peleg, "Side Channels in Cloud Services: De-Duplication in Cloud Storage", *IEEE Security and Privacy*, Vol. 8, No. 6, pp. 40-47, 2010.

[12] S. Bharat and B.R. Mandre, "A Secured and Authorized Data De-Duplication in the Hybrid Cloud with Public Auditing", *International Journal of Computer Applications*, Vol. 120, No. 16, pp. 1-8, 2015.

[13] J. Hur, D. Koo, Y. Shin and K. Kang, "Secure Data De-Duplication with Dynamic Ownership Management in Cloud Storage", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 11, pp. 3113-3125, 2016.