IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Detection on sarcasm using machine learning classifiers and rule based approach

To cite this article: K. Sentamilselvan *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1055** 012105

View the article online for updates and enhancements.

Detection on sarcasm using machine learning classifiers and rule based approach

K.Sentamilselvan[1*], P.Suresh [2], G K Kamalam [3], S.Mahendran[4], D.Aneri [5]

[1*]Department of Information Technology, Kongu Engineering College, Anna University, India
[2]Department of Information Technology, Kongu Engineering College, Anna University, India
[3]Department of Information Technology, Kongu Engineering College, Anna University, India
[4]Department of Electronics and Communication Engineering, EBET Group of Institutions, Anna University, India
[5]Department of Information Technology, Kongu Engineering College, Anna University, India

Email: *ksentamilselvan@gmail.com

Abstract. Sentiment Analysis has been mainly used to understand the judgment of the text. It has been undergoing major provocation and irony detection is considered as one among the most provocations in it. Irony is the unusual way of narrating an information which disagrees the concept which leads to uncertainty. One primary task included by most developers is data preprocessing which includes many techniques like lemmatization, tokenization and stemming. Many researches are done under irony detection which includes many feature extraction techniques. Machine learning classifiers used for these researches are Support Vector Machine (SVM), linear regression, Naïve Bayes, Random Forest and many more. Results of these research works includes accuracy, precision, recall, F-score which can be used to predict the best suited model. In this paper various methodology used in irony text detection for Sentiment Analysis is discussed.

Keywords − Naïve Bayes, SVM, Sentiment Analysis, Irony detection

1. Introduction

Natural Language Processing (NLP) acts as an interface between computer and human languages. It has been widely used to make the machine understand and examine the data. Also used to examine the datasets and obtain the outcome. It can be used both for ordered and unordered data. The semantic order depends community factors, topical idioms, jargons and much more [13]. NLP seems to face some provocations in this area. Sentiment Analysis is considered to be the most dominant areas in NLP which examines the concept. Sentiment Analysis is used to examine the sentiments conveyed by narrator and obtain the point of view behind it. It is used to categories the contradiction of a report or a judgmental text. The Strength of the text is categorized by Sentiment Analysis [8]. Many examinations are performed using this concept. These logics are further used to decide and recover various degrees of opinion. The inspection concept is analyzed on many discrete units they are nothing but the words used in this report. It gives a brief understanding of narrator's point of view.

It has to be represented as opinion mining where it speculates on discrete units and debate the assessments. Various data pre-processing as well as categorization techniques are being used in this concept of Sentiment Analysis. Opinions in text represents a positive or a negative meaning. There are key provocations faced by Sentiment Analysis such as Units named identification, Analogy identification, irony detection etc., People nowadays convey their sentiments on many social platforms and sometimes convey their emotions in sarcasm. Sarcasm is considered to be the most leading provocations in Sentiment Analysis. Irony is an accidental manner of telling a message which is commonly a harsh declaration which is conveyed. Sarcasm can throwback the situation of doubt. Sarcasm can be expressed in speech as well as text [1]. Sarcasm can be delivered via many ways like a straight discussion, oration, text etc. In straight discussion, face expressions and body language provides the sign of irony [4]. Whereas in oration, irony can be understood by the change in voice tone. But in the text, it is very hard to find the irony where it may also be conveyed using a capital letter, wide range of use in exclamatory marks and emoticons etc.

2. Related Works

A message, analysis, response by the people has many features. Which rely on many characteristics like geographical area, ongoing article, popular facts, sexuality etc., [5].Many opposites are obtained using motive rules in Serbian Word Net ontology. Opposites and its pairs, Positive sentiment, arranged series of opinion labels, Sarcasm markers and Parts of speech labels are used [4].

Conjunctions in English mainly used to join the set of words when positive conjunction "and" is used when it gives a positive meaning and negative conjunction "but" is used when it gives an opposite meaning. A Network is raised with same meanings in word net [3].The basic eight emotions classified namely happy, irritated, hope, angry, expect, surprised, sad and normal, which are withdrawn from Social Cognition Engine Emotional Lexicon[2].

Irony can also be appeared as numbers. For example, its much supported to wake up at 5 in the morning. Where the example means accidentally that it isn't much supported waking up 5 in the morning [1]. Emoji's reflect the nature of the sentence. The intuition of sentimental words and emoji's are meant to be same when they appear in general. For intuition recognition, mainly two ways can be used which are corpora and dictionary-based approaches. In English usually to join two sentences we use Conjunctions [13]. In order to examine the sentiment Dialect words are commonly used. Categorization of text to their subjective and also to their objective forms is performed. Sentiment dialects is observed by using subjective sentences. Identification of Polarity is done using (IDF) weighted inverse document frequency [12]. A related scenario is the Part of Speech (POS) tag which may is accompanied with words in the report. Classifications of the feature are mostly to please a role with regard to the specification of classification expressive of irony [11]. The Twitter and Amazon datasets are commonly taught on punctuations, patterns enrich punctuation, enrich patterns, semi-supervised Identification algorithm. Star related ratings are also taken into account [10]. The twitter airline sentiment analysis by predicting the polarity in tweets, by making use of the common machine learning algorithm like Naïve Bayes, Logistic Regression together with the feature extraction techniques like count vectorizer, TF-IDF and word2Vec and found that logistic regression together with count vectorizer works well for the dataset.[14].

3. Proposed methodology

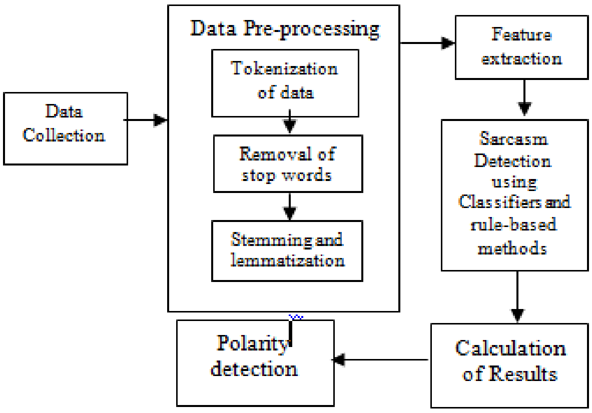

Figure 1. General workflow of Irony Detection in Sentiment Analysis

3.1 Data Preprocessing:

Pre-processing consists of three basic steps. First step is Tokenization. It is the process of tokenizing each word in the sentence. Second step is Stemming and lemmatization. It is the process of achieving the root form of the derived words in the sentence, the words are further converted into present tense format. Third Step is The Removal of stop words. It is the process of removing the unnecessary words in the document, the commonly considered stop words include (the, a, an, in etc.,) these are some words that make no sense in the sentence.

3.2 Tokenization:

It is the process of splintering an order of strings into chunks like words, clause, sign and further components called tokens. Tokens can also be discrete words, clause or full sentences. In Tokenization concept, few attributes such as punctuation marks are removed. The tokens itself act as an input for upcoming tasks such as parsing and text mining.

Example of Tokenization:

"Identifying the words"

From: flight hasn't arrive.

To: flight

has

n't

arrive

3.3 Removal of stop words:

Stop words in general are words that make no sense in the article, but act as a disturbance to it. Stop words like (the, a, an, in etc.) make no change in the article. So these stop words are removed for better performance.

Table 1. Example for Stop word removal

Sample text with stop words	**Text without stop words**
Flight has not arrived yet	Flight not arrived
Thanks for the response. We are hopeful!	Thanks response hope

3.4 Stemming and Lemmatization:

This concept is used to convert the words into their appropriate root forms so that they can be examined as an individual unit. In this process the words are then converted or changed into their appropriate present tense format.

Example for Stemming:

Table 2. Example for Stemming

Original word	**Stemmed word**
Flight	Fligh
Response	respons
Happy	Happi

Example for lemmatization:

Table 3. Example for Lemmetization

Original word	**Lemma word**
Troublesome	Trouble
Troubled	Trouble
Responsible	Response

4. Sarcasm detection classifiers

4.1 Decision Tree:

This algorithm comes under the supervised learning algorithms. It has been widely used to solve regression and categorization problems. One main reason for using the decision tree algorithm is the creation of the training example which is used to forecast class or value of the target variables by studying the rules and conditions of the decision tree algorithm which is deduced from the training example. This algorithm is very easy compared to other machine learning algorithms because it solves the problem by making use of trees i.e., tree representation, where each internal node agrees with the credit and each leaf node agrees with the labelling class

Algorithm:

```
Begin
        Place the good attribute of the dataset at the root of the tree
        Split the training set into subsets
        Each subset contains data with the same value for an attribute
        Repeat step 1 and step 2
        On each subset until you find leaf nodes in all the branches of the tree
End
```

4.2 Support Vector Machine:

Support Vector Machine algorithm is one of the Supervised Learning algorithms, that is used for both organization and Regression scenarios. Initially it was used for categorization problems in Machine Learning. The motive of the Support Vector Machine algorithm is to develop decision boundary that isolates n-dimensional space into categories in order that they can easily be placed to new data centre in the perfect class in the upcoming scenarios. Hyper plane is to be considered as the perfect decision boundary. Additional uses of this support vector machine algorithm includes text classification, image categorization and face detection etc.

Pros of SVM

SVM algorithm gives best accuracy and also works well with huge dimensional area. This algorithm generally make use of a subset of training points and finally uses limited memory space.

Cons of SVM

SVM have lot amount of training time that is why they are in practice not apt for big datasets. One more disadvantage is SVM algorithm do not work well with overlapping classes.

4.3 Random Forest:

Random forest is also considered as one of the supervised learning algorithms used in machine learning that is used in grouping and regression problems. Whereas the main usage is for the organization scenarios. We all already know that talking about a forest, it's full of trees and when there are more trees then it's considered to be the completely robust forest. Likewise, random forest algorithm in machine learning develops trees called as the decision trees out of the given examples and then finally predicts the best suited outcome. It generally dissolves the situation of over fitting by making the combination of the results of various decision trees. Random forest machine learning algorithm also undergo with large range of data products than one individual decision tree algorithm.

Algorithm:

```
Begin
    Select of random samples from a given dataset
    Constructs a decision tree for each sample
    Gets the prediction result from every decision tree
    Perform voting for every predicted result
    Select the most voted prediction result as the final prediction result
End
```

AdaBoost: This method is considered as the best meta-algorithm in machine learning concept. This is or else called as adaptive boosting. Researchers in [17] use adaboost.

Gradient Boost: One most common machine learning technique widely used for classification as well as regression problems is Gradient boost. Weak prediction models can be ensemble using this technique. It is used by researchers in [18].

5. Rule based methods

These are commonly used as a concept of irony identification which includes methods like Semantic, Syntactic etc.

5.1 Lexical Method:

Here the class, tasks and features are converted to lexical features like noun, verb and adjective respectively. To check and correct the misspelled words General English language dictionaries are being used which also is used to remove meaningless words [19].

5.2 Semantic Method:

The meaning of a language is basically considered as semantics. Natural language processing (NLP) can be used to know the way people think and communicate their views. Semantic matching is compared along with graph-based matching to give rise a score which is used to detect the level of sarcasm. Semantic processing is used for generating a meaningful review [25].

6. Results and discussion

Considering these four machine learning algorithms for the irony detection of the given dataset, it has been found as the Support vector machine provides more accurate results compared to the other there machine learning algorithms. In X axis it represents the various algorithms like SVM, Naïve Bayes. Decision Tree and Random Forest and Y axis represents accuracy of the each algorithms (Figure 2).

Results conforms that Support vector machine provides accuracy of 64 percentage for the SemEval2018-T3-train-taskA.txt given irony detection dataset.

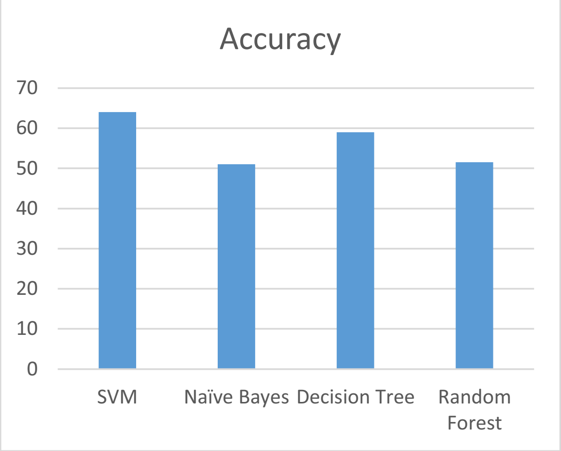

Figure 2. Graph based on Accuracy for SemEval2018-T3-train-taskA.txt given irony detection dataset

Table 4. Accuracy Table SemEval 2018-T3-train-taskA.txt given irony detection dataset

Algorithm	Accuracy
SVM	64%
Naïve Bayes	51%
Decision Tree	59%

Results conforms that random forest outperforms the support vector machine algorithm to the accuracy of 76% using the features of sentiment analysis while using the sarcasm-detection.txt dataset.

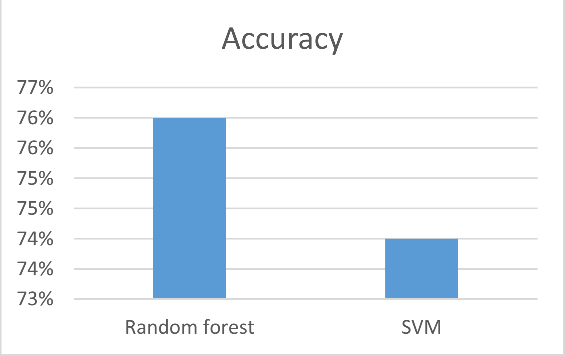

Figure 3. Graph based on Accuracy for sarcasm-detection.txt dataset

Table 5. Accuracy table for sarcasm-detection.txt dataset

Algorithm	Accuracy
Random forest	76%
SVM	74%

In X axis it represents the algorithms (SVM, Random Forest) and Y axis represents accuracy of the each algorithms (Figure 3).

7.Conclusion.

Several pre-processing techniques are used for the irony detection of the dataset. Developers worked on the categorization and results has been provided. A comparative study on these algorithms has been performed to identify which of these algorithms give the best results. Machine Learning algorithms used to detect the sarcasm here are Support Vector Machine, Naïve Bayes and Decision Tree for the SemEval 2018-T3-train-taskA.txt dataset and found Support vector machine algorithm to be the best suited for the particular dataset. Algorithms like Random Forest and SVM are used for the dataset Sarcasm Detection.txt and found Random Forest algorithm to give best results with the accuracy of 76%. Future work can be proceeded by using some more machine learning classifiers and compare the results to obtain the accuracy of the best suited algorithm.

References

[1]. Mondher Bouazizi and Tomoaki Otsuki, “A Pattern-Based Approach for Sarcasm Detection on Twitter,” IEEE Access Volume 4, 2016. pp. 5477- 5488

[2]. Pyae Phyo Thu and Than Nwe Aung. “Effective Analysis of Emotion-Based Satire Detection Model on Various Machine Learning Algorithms,” in Proc. IEEE 6th Global Conference on Consumer Electronics, 2017

[3]. Shuigui Huang, Wenwen Han, Xirong Que and Wendong Wang, “Polarity Identification of Sentiment Words based on Emoticons,” International Conference on Computational Intelligence and Security, 2017pp 134–138.

[4]. Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, Ranka Stanković, “Using Lexical Resources for Irony and Sarcasm Classification,” Proceedings of the 8th Balkan Conference in Informatics.

[5]. Anandkumar D. Dave and Prof. Nikita P. Desai, “A Comprehensive Study of Classification Techniques for Sarcasm Detection on Textual Data,” in Proc. International Conference on Electrical, Electronics, and Optimization Techniques, 2016, pp. 1985-1991.

[6]. Dmitry Davidov, Oren Tsur and Ari Rappoport, “Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon,” in Proc. Fourteenth Conference on Computational Natural Language Learning, pp. 107-116.

[7]. Michael Sejr Schlichtkrull, “Learning Affective Projections for Emoticons on Twitter,” in Proc. International Conference on Cognitive Infocommunications, 2015, pp. 539-543.

[8]. Ms. Payal Yadav and Prof. Dhatri Pandya, “SentiReview: Sentiment Analysis based on Text and Emoticons,” International Conference on Innovative Mechanisms for Industry Applications, 2017, pp 467- 472.

[9]. Setra Genyang Wicana, Taha Yasin İbisoglu and Uraz Yavanoglu, “A Review on Sarcasm Detection from Machine-Learning Perspective,” in Proc. International conference on Semantic Computing, 2017, pp. 469-476.

[10]. M. Boia, B. Faltings, C.-C. Musat, and P. Pu,A :) Is worth a thousand words: How people attach sentiment to emoticons and words in tweets," in Proc. Int. Conf. Soc. Comput., Sep. 2013, pp. 345_350.

[11]. Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, RaynerAlfred and Frans Coenen. “Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts,” in Proc. International Conference on Information Technology, 2017, pp. 703-709.

[12]. A. Joshi, P. Bhattacharyya, and M. J. Carman. (Feb. 2016),"Automatic sarcasm detection: A survey" Available: https://arxiv.org/abs/1602.03426

[13]. Shuigui Huang, Wenwen Han, Xirong Que and Wendong Wang, “Polarity Identification of Sentiment Words based on Emoticons,” International Conference on Computational Intelligence and Security, 2017. pp 134–138.

[14]. K. Sentamilselvan, D. Aneri, A. C. Athithiya and P. Kani Kumar International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9, Issue-3, February 2020

[15]. Ankita, Nabizath Saleena, “An ensemble classification system for twitter sentimental analysis”, International Conference On Computational System for Twitter Sentiment Analysis (ICCIDS), Procedia Computer Science, pp.937-946, 2018.

[16]. Bala Durga Dharmavarapu, Jayang Bayana “Sarcasm detection in twitter using sentimental analysis”, International Journal of Recent Technology and Engineering (IJRTE), volume 8, issue -1, 2018.

[17]. Anukarsh G Prasad; Sanjana S, Skanda M Bhat, B S Harish. “Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data” in Proc. International Conference on Knowledge Engineering and Applications, 2017, pp. 1-5

[18]. Pyae Phyo Thu and Than Nwe Aung. “Effective Analysis of Emotion-Based Satire Detection Model on Various Machine Learning Algorithms,” in Proc. IEEE 6th Global Conference on Consumer Electronics, 2017

[19]. Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, Rayner Alfred and Frans Coenen. “Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts,” in Proc. International Conference on Information Technology, 2017, pp. 703-709.

[20]. K. Sentamilselvan, A. Nirmal, P. Nivethitha, S. Subalakshmi,"An Improving ICU Patient Mortality Prediction Using Artificial Neural

Networks", International Journal of Advanced Science and Technology 29 (5), 6079 - 6089, 2020

[21]. Sentamilselvan. K, Vinoth Kumar. S, Jeevananthan. A. “Preparing Data Sets by Using Horizontal Aggregations in SQL for Data Mining Analysis”. i-manager's Journal on Information Technology (JIT), Vol. 4, No. 3, PP.33-41, September – November 2015

[22]. K.Sentamilselvan, Dr.G.K.Kamalam, "Potential Finish Time and Min-mean Algorithm for allocating Meta-Tasks on distributed Computational Grid", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-4, pp.10580-10586, November 2019

[23]. Dr.G.K.Kamalam, K.Sentamilselvan, "Limit Value Task Scheduling (LVTS): an Efficient Task Scheduling Algorithm for Distributed Computing Environment", International Journal of Recent Technology and Engineering (IJRTE),ISSN: 2277-3878, Volume-8, Issue-4, pp.10457-10462, November 2019

[24]. PCD Kalaivaani, R. Thangarajan ,Enhancing the Classification Accuracy in Sentiment Analysis using Joint Sentiment Topic Detection with Naive Bayes Classifier,Asian Journal of Research in Social Sciences and Humanities,6,12,105-116,2016,Asian Research Consortium

[25]. K. R. Prasanna Kumar, M. Pranesh, T. G. Rakhul Raahje, P. Vignesh, K. Kousalya, and K. Logeswaran, ‘Factual product recommendation system by eliminating fake reviews using machine learning techniques’, Int. J. Adv. Sci. Technol, vol. 29, no. 7 Special Issue, pp. 2729–2735, Apr. 2020.