SARCASM DETECTION BY MACHINE LEARNING USING TWITTER DATA

Athira K Menon,Neethu P,Dr.G.Kiruthiga

[1] *Student, Department of Computer science, IES College of Engineering Thrissur, Kerala, India*

[2] *Assistant Professor, Department of Computer science, IES College of Engineering Thrissur, KeralaIndia*

[3] *Associate Professor, Department of Computer science, IES College of EngineeringThrissurKerala, India*

ABSTRACT

Sarcasm is a sophisticated form of irony widely used in social networks and micro blogging websites. It is usually used to convey implicit information within the message a person transmits. Sarcasm might be used for different purposes, such as criticism or mockery. However, it is hard even for humans to recognize. Therefore, recognizing sarcastic statements can be very useful to improve automatic sentiment analysis of data collected from micro blogging websites or social networks. Sentiment analysis refers to the identification and aggregation of attitudes and opinions expressed by internet users towards a specific topic. For the detection of Sarcasm in plain text we are going to use Machine Learning Classification Methods. By detecting Sarcasm from social media's like twitter we can identify irrelevant and sarcastic opinions of people. These opinions can be used as reviews for the effective business decisions.

Keyword: *Sarcasm, Stemming, Unigrams, Classification*

1.INTRODUCTION

Sarcasm, commonly defined as An ironical taunt used to express contempt, is a challenging NLP problem due to its highly figurative nature. Interjections, punctuation, and sentimental shifts have been considered as major indicators of sarcasm. When such lexical cues are present in sentences, sarcasm detection can achieve high accuracy. The use of sarcasm also relies on context, which involves the presumption of common sense and background knowledge ofan event. When it comes to detecting sarcasm in a discussion forum, it may not only be required to understand the context of previous comments but also the necessary background knowledge about the topic of discussion. Sarcastic tweets not being detected by the model, most probably because they are specific to a certain situation. Also the sarcastic tweets written in a very polite way are undetected. Sometimes people use politeness as a way of being sarcastic, highly formal words that don't match the casual conversations. Complementing someone in a very formal way is a common way of being sarcastic

1.1 PROBLEM STATEMENT

Proposing a hybrid approach of both content and context driven modelling in online social media discussions to detect sarcasm.

1.2 OBJECTIVE

To detect sarcasm by a content driven modelling in Twitter using machine learning algorithms. As we know that sarcasm detection is a very narrow research field in Natural Language Processing, a special case of sentimental analysis where instead of detecting a sentiment in the whole spectrum, the focus is on sarcasm.

1.3 OUTLINE

The project is divided mainly into four phases as data collection, data preprocessing, feature extraction and classification.The whole report starts with a small description about the theoretical background of the proposed approaches are included in Chapter 2. Chapter 3 contains the literature survey which helped to improve the systems performance. Chapter 4 describes the methodology of the project. The implementation details, hardware and software requirements are described in Chapter 5. Chapter 6 presents the results. Performance analysis is discussed in Chapter 7. Conclusion and future work are included .

2. DESIGN DESCRIPTION

.The design of the proposed system is divided mainly into the phases Data Collection, Data preprocessing, Feature extraction and Classification. In the first phase, raw data is collected. The data set in its raw form is merged into a single file for the ease of the next phases.In the second, the data cleaning and preparation work is done. It is manipulated in such a form that is suitable for further analysis and processing. Features are extracted in the third phase and classify the tweet comments using the machine learning algorithms in the fourth phase. After classification a representational graph is build up for the users of the system to view the model comparison of the classification algorithms. These are the modules involved in the design of the system. The system architecture is shown in Figure 3.1

3.METHODOLOGY

In this section provides a methodological description of the proposed approach, the twitter data is collected and the aim is to classify the sarcastic tweets. tThis model primarly consists of five major components (I) data collection (ii)data preprocessing (iii)feature extraction (iv)classification and(v)data visualization.The aim is also to classify the tweets using Naive Bayes classifier, Support vector machine (SVM), Random Forest and Decision Trees (DT) and to differentiate between the accuracy, precision, recall and F-score of Naive Bayes classifier, Support vector machine (SVM), Random Forest and Decision Trees (DT).Each of the approaches is explained in the section below.The figure 3.1 shows the overall system architecture of the proposed approach which is shown below.

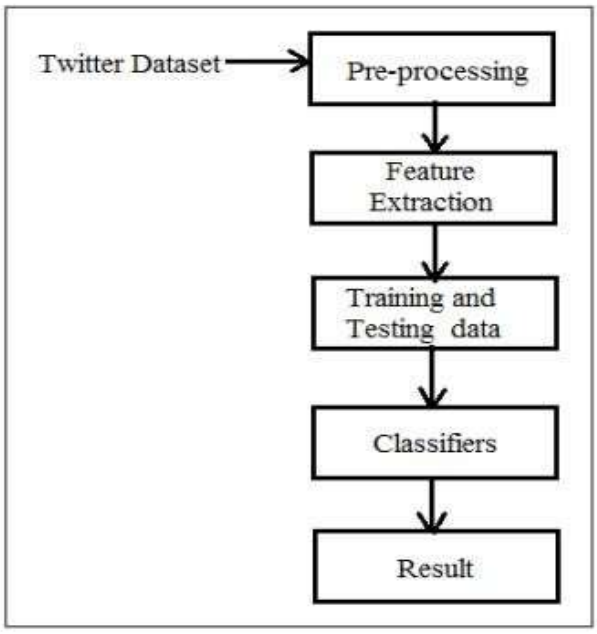

Figure 3.1: System Architecture

3.1 DATA COLLECTION

Collection is the first stage of the cycle, and is very crucial, since the quality of data collected will impact heavily on the output. The collection process needs to ensure that the data gathered are both defined and accurate, so that subsequent decisions based on the findings are valid. This stage provides both the baseline from which to measure, and a target on what to improve. Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed. A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate and that subsequent decisions based on arguments embodied in the findings are valid. The process provides both a baseline from which to measure and in certain cases an indication of what to improve. More data especially data from more diverse sources enables finding better correlations, building better models and finding more actionable insights. While individual records are often useless, having every record available for analysis can provide real value.

3.2 DATA PREOPROCESSING

In order to prepare our corpora for use, it first had to be sanitized. The preprocessing aims to minimize the vocabulary of terms used in the tweets. This involved a certain amount of preprocessing steps which involved.

• Tokenizing, stemming, and filtering out non English tweets. This process is known as cleaning useless and meaningless words from each tweet.

• Filtering out duplicate html tags and hyperlinks removed because html tags and hyperlinks expresses no meaning.

• Hash tagged sarcasm and sarcastic were filtered out in order to not influence our models due to their presence. All other hash tags were kept in place.

Each and every tweet of corpora preprocessed by following previous steps. All tweets were tokenized and stemmed. Hyperlinks, non-English tweets were filtered out.

3.3 FEATURE EXTRACTION

Feature extraction has a huge role in determining the outcome of any machine learning task. The quality of classification, both qualitatively and quantitatively, depends on the features selected. This section, at a high level, focuses on extracting the features from tweets that can be categorized into various types, namely, punctuational, sentiment intensity, frequency related and structural features. There are three main approaches exist to extract features i.e. Bag of Words, TF-IDF and Word2Vec.

3.4 CLASSIFICATION

While looking to classify the data, we chose four methodologies in which to test features:

• Naive Bayes (NB)

• Support vector machine (SVM)

• Decision Trees (DT)

• Random Forest (RF)

3.5 DATA VISUALIZATION

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. It is the process of conveying information in a way that can be quickly and easily digested by the viewer. A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

4.IMPLEMENTATION

The task is to predict whether a tweet is sarcastic or not. This is a typical supervised learning task where given a text string, we have to categorize the text string into predefined categories. To solve this problem, follow the typical machine learning pipeline. First import the required libraries and the dataset then do exploratory data analysis to see if we can find any trends in the dataset. Next, perform the text preprocessing to convert textual data to numeric data that can be used by a machine learning algorithm. Finally, use machine learning algorithms to train and test models.Software requirements used in this proposed system are Python(2.7 or 3.3 - 3.6), NLTK, NumPy, Pandas Matplotlib,Scikit-learn and CountVectorizer.The sytem configuration is Dual core processor 2GB RAM , 500GB Memory.

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming. Many other paradigms are supported via extensions, including design by contract and logic programming. Python uses dynamic typing and a combination of reference counting and a cycle detecting garbage collector for memory

management. It also features dynamic name resolution, which binds method and variable names during program execution.

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.NLTK is available for Windows, Mac OSX, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrames and allows you to store and manipulate tabular data in rows of observations and columns of variables.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientifific libraries NumPy and SciPy.

Scikit-learns CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the preprocessing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

5.CONCLUSION AND FUTURE WORK

The sarcasm detection system is to detect sarcasm by a content driven modelling in Twitter using machine learning algorithms. As we know that sarcasm detection is a very narrow research field in Natural Language Processing, a special case of sentimental analysis where instead of detecting a sentiment in the whole spectrum, the focus is on sarcasm.The paper conferred illustrations containing the methods, datasets and performance values.This clearly indicates that incorporating features that describe the psychological and behavioural aspects of the user goes a long way in helping the process of automatic identification of sarcasm. The classifier is working as intended, and have successfully examined sarcastic and non-sarcastic tweets using the set of features. We observed the best results in Random Forest classifications. Some best working classification methods for the detection of sarcasm are analysed and compared till now. Better classification method for sarcasm detection with images and audio shall be observed.

6. REFERENCES

[1]. Reference 1 Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann and Rada Mihalcea *"CASCADE: Contextual Sarcasm Detection in Online Discussion Forums"*,2018
[2] Reference 2 Rajeswari K and ShanthiBala P, *"SARCASM DETECTION USING MACHINE LEARNING TECHNIQUES"*,2018

[3] Reference 3 Shubhodip Saha, Jainath Yadav and Prabhat Ranjan. *Proposed Approach for Sarcasm Detection in Twitter*,2017
[4] Reference 4 D.V.Nagarjana Devi, Dr.T.V.Rajanikanth, Dr.V.V.S.S.S. Balaram, *"Sarcasm Detection in Plain Text Using Machine Learning"*,2018
[5] Reference 5 Aditys Joshi,Pushpak Bhattacharya and Mark J. Carman, *"Automatic Sarcasm Detection: A Survey"*,2017
[6] Reference 6 Karthik Sundararajan and Anandhakumar Palanisamy, *"Multi-Rule Based En semble Feature Selection Model for Sarcasm Type Detection in Twitter"*,2019