ABSTRACT

Sarcasm is an important part of communication. Sarcasm is expressed in words, facial expressions and can be noticed in the intonation of voice. So in this digital age , sarcastic comment are passed everyday via tweets, comment sections of different social media outlets , and even news headlines. Newspapers often seem to use sarcasm in their headlines to grab the reader's attention. Some of these headlines can be misunderstood and taken to mean something different than the original intentions. However, more often than , not the readers find it difficult to detect the irony within the headlines thus, getting the incorrect idea about the actual news and further passing on their understanding to their friends and colleagues .This leads to a need to detect sarcasm especially in the news and on social media. The sarcasm detection of text has its challenges because text lacks intonations and deflection of voices that occur when a sarcastic statement is formed vocally by a person's.

This project focuses on the effect of the different encoding methods used in the text to extract the feature of machine learning models. A deep learning model is used and the results are compared. Prior to the feature extraction, pre-processing data techniques such as tokenization, training set and testing set, and embedding concepts and punctuation are used by researchers in any work that involves textual analysis. These pre-processing methods are widely used and accepted and used in this project. Different methods of extracting features such as Count Vectorizer and word embedding were used in this project. Random Forest , and Logistic Regression were the machine learning algorithms used in this project.

Keywords: Sarcasm Detection, Convolutional Neural Networks (CNN), Random forest, Logistic regression, deep learning.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Over the past decade, there has been an increase in the use of internet services disseminate information. These texts have distinctive features that are explored through the use of Natural Language Processing techniques in machine learning for knowledge and understanding. One of the most important things features the presence of sarcasm in the text. Sarcasm is a complex communication process that allows speakers to express rich ideas in an ambiguous way. There is a lot of ridicule in texts that are distributed online and on social media [1]. This includes news headlines and other media. Mocking is traditionally defined as the process of intentional misuse of words in order to convey a specific meaning (often the opposite of what is being said) [1]. For example, if someone says ‘How lucky I am? I caught a corona virus! "It is obvious that although the words used are constructive, the particular meaning is wrong, making the speech sarcastic. Sarcasm is characterised by the utilization of irony that reflects a negative meaning.

No one has ever used sarcasm in online comments or in the headlines. People can usually see this kind of sarcasm in person, but not all sarcastic words are so clear in the text. Computers strive to tell the difference in any way. We ask ourselves, can the language of the sarcastic language itself tell the computer that the word is sarcastic? Can a computer read these rules, or is the context involved? In our project, we will be taking an advanced route. Our goal is to build a sarcasm detector using machine learning through neural networks. The entries in our project contain a series of articles (newspaper articles) labeled as sarcastic or non-sarcastic. The Onion aims to produce current sarcastic versions of events, and we collected all the news headlines from the News Articles (funny). We collected real and non-sarcastic news headlines in HuffPost. This dataset is provided in Kaggle.

This project addresses the most problem of NLP known for Sarcasm Detection employing a combination of models supported Convolutional Neural Networks (CNNs). The discovery of humor is vital in other areas like affective computing and Sentimental analysis because such an expressions can change sentence variability.

OBJECTIVES AND IMPORTANCE OF SARCASM DETECTIION IN NEWS HEADLINES

Although sarcasm is well-known and widely used, it is a challenge not only for computers but also for people can detect them quickly. Some people find it difficult to identify and understand the use of sarcasm [2]. Because of this fact, the presence of sarcasm, if not detected and calculated can interfere with machine learning activities such as sentimental analysis and opinion mining. This makes the sarcasm detection is an important task. Receiving automatic jokes can be seen as text classification problem. Text data is one of the easiest forms of data. The machine learning algorithms are unable to process non-numerical data and as a result the need for the feature extraction arises. It is important to take useful information from any type of data, especially the data which is in unstructured forms such as text data. Feature extract and appropriate representation of the text to achieve the purposes of classification is an important factor that affects the accuracy of the classification.

The Project consists of mainly three parts. The first task is pre processing data and second one is classification models and the third one is model elevation.

The first is pre-processing the dataset which contains of a data set, tokenizer and sequence, training set and testing set, word emdeddings , tokenizing and padding the splitted data.

The second part is classifiers used in this project are random forest, and Logistic regression. The main classifier is Convolutional Neural Networks (CNNs) which is used in the deep learning algorithm. By using these models ,we can get the accuracies for the sarcasm detection appropriately.

The third part consists of the performance elevation and results of the projects that involves combining all the models and provides the best result.

The importance of sarcasm detection can be in understanding the real feelings or sentiments and opinions of people. The use of sarcasm detection can benefit many areas of NLP applications, including marketing research, opinion mining and information classification. Acquiring complex ideas and detection of ironic opinions can help companies and governments develop products and services. Reliable identification of sarcastic and sarcasm in the text can improve the effectiveness of natural language processing techniques.

SCOPE

We have successfully implemented and completed all three objectives in sarcasm detection where we have imported the necessary modules in the implementation , loading data into arrays and tokenizing the data and padding the data etc., and using the machine learning and deep learning techniques ,we have checked the accuracies using the random forest , nd logistic regression and finally we have used the Convolutional Neural Networks (CNNs) classifier and found out that it is the best classifier to detect sarcasm in our project. The results can be seen in the last section.

This report will explain the details of the literature used in this project (the list of references used and the review of some key reference journals/papers) and how the proposed model is different and better from the pre-existing ones, the limitations of the pre-existing and the proposed model and the future scope of how the proposed model can be improved.

We will explain the implementation of this project with diagrammatic explanation, the results derived after the implementation as well as the performance evaluation of the sarcasm detection in news headlines.

The scope of this project is to detect sarcasm in news headlines , where people like all of us sometimes doesn't understand the sarcastic meaning in the news paper article. So this leads to misunderstandings between us. So by this sarcasm detection , opinions of the people can help companies and governments develop products and services in their domain. When the newspaper headlines containing sarcasm sometime leads to conflicts between Politian's and the government . By using this detection , there will be less chances of conflicts and the government can also understand the peoples opinion on their political parties , so that they will know where to improve themselves.

MOTIVATION

Why sarcasm detection? Currently , the need to understand the sarcastic meaning in text is very less and people don’t usally understand the sarcasm in text very quicly. There are so many ways of understanding sarcasm very easily such as like in pictures and pitch of voice by a human or vocally by a person can be understood very easily. The sarcasm is the use of words that can be a meaning opposite to the one you actually intend to pass on. It has the ability to flip the sentiment of the sentence this makes sarcasm detection an important part of the sentiment analysis. This detection is used to detect whether the newspaper headline is sarcastic or non sarcastic. Passing on sarcastic comments effects the persons feeling or sentimental emotions of the humans.

We are motivated to take this project because there are so many projects like Twitter sarcasm detection , social media sarcasm detection and images sarcasm detections and emoji’s sarcasm detection and etc., these are done by many people. So we wanted to take this Sarcasm detection in news paper headlines which are done by very few people , so we wanted to experiment this project.

CHAPTER 2

LITERATURE REVIEW

The literature review on this project is done by gathering the work done by the researchers and authors.

PREVIOUS WORK

The detection of sarcasm is very important in the context of sentimental analysis and also in opinion mining. Different machine learning algorithms were used in this problem. Some researchers have already used the Naive Bayes Classifier and Support Vector machines for data analysis in social media in Indonesia [5]. Classifiers used in the work [5] described and did well in the job and appreciated it the decision to use Naive Bayes and SVM in this analysis [7]. They used vector support machines with Tf-IDF and Bag-Of-Words. The work of Davidov et.al [6] provides details on how to detect sarcasm using semi supervised methods on two different data sets, one containing a review of Amazon products and one with tweets collected on Twitter. These researchers focuses on features such as punctuation, syntax of sentences, hashtags used etc. The proposed system has more than 75% accuracy [6]. The work performed [14] by the reasearcher shows the use of Support Vector Machines and in order to detect ridicule from the media news headlines. The proper method of extracting the element used in this work results in points of accuracy are about 80%. Some researchers have used deep learning to the problem strategies [2]. Collecting data for the purpose of identifying sarcasm is even more challenging to read. This is especially so social media data case. This study uses some of the methods described in previous works with better accuracy results and other steps such as well.

Previous work has focused on almost Twitter data and sentimental analysis, but such data sets are noisy in terms of labels and language. In addition, many tweets are a response to some tweets and finding sarcasm and this need to be found contextual tweets. Therefore, we try to differentiate and find sarcasm in different ways. To overcome the noise-related restrictions on Twitter data sets and taking into account the fact that headline data has a similar structure compared to tweets, this discrimination News Database collected on two news websites.