JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Glove Based Sarcasm Detection On Social Media

[1]Dr. Saroja T.V, [2]Aniket Shetty, [3]Sarika Shimpi, [4] Sohil Sheikh,

[1]Professor, [2] [3] [4] Student,

[1]Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering,

Dombivli, Maharashtra, India.

Abstract : Sarcasm is typically a statement that is intentionally the exact opposite of the literal meaning of the statement. For example, " I love waiting on line " or " I hate receiving thoughtful gifts. " . In each case the communicator or speaker obviously means the opposite of the actual statement . Sarcasm detection is of great importance in understanding people's true sentiments and opinions. However, sarcasm detection is also a very difficult task, as it's largely dependent on context, prior knowledge and the tone in which the sentence was spoken or written. The most common errors nowadays are False negatives sarcasm. Sarcastic tweets not being detected by the model, most probably because they are very specific to a certain situation or culture and they require a high level of world knowledge that DL models don't have. The purpose and objective of our project is to use a model based on machine learning algorithm which will be able to detect sarcasm from tweets, comments , headlines of the social media. GloVe based model which is an unsupervised machine learning algorithm will help to detect sarcasm with a very good accuracy rate. GloVe algorithm is a category of unsupervised learning algorithm . GloVe is a very popular algorithm for getting the vector representation for words. These vector representation of words through GloVe is achieved by mapping words into a meaningful space where the distance between the words is related to semantic similarity. Now once the words are converted into sequence of vectors will pass these sequences to the Recurrent Neural Network. The category of recurrent neural network used in the proposed system is LSTM(long short term memory). The main reason to select LSTM model is because the LSTM itself a type of RNN and recurrent neural networks are very good at learning sequences. The final evaluation will be detection whether the following text is sarcastic or not.

***IndexTerms* –** Sarcasm, GloVe , Long-Short Term Memory, Word2Vec, Sentiment analysis. NLP

I. INTRODUCTION

Sarcasm, a sharp and ironic utterance designed to cut or to cause pain, is often used to express strong emotions, such as contempt, mockery or bitterness. Sarcasm detection is of great importance in understanding people's true sentiments and opinions. Global Vectors for Word Representation (GloVe) has gained popularity for its ability to capture semantic relationships between words based on co-occurrence statistics in large text corpora. By leveraging GloVe embeddings, we can potentially uncover subtle semantic cues that distinguish sarcastic utterances from non-sarcastic ones. Our project aims to develop a novel approach to sarcasm detection by harnessing the power of GloVe embeddings. We hypothesize that sarcasm manifests itself not only through the explicit words used but also through the underlying semantic context captured by word embeddings. By training machine learning models on these embeddings, we intend to capture the intricate interplay of words and their contextual meanings, thereby enhancing the robustness and accuracy of sarcasm detection.

II. LITERATURE SURVEY

Sunusi Kabir, Adamu Habu ,Badamasi Imam , Abdullahi Madaki[1] the authors have proposed SVM and an ensemble method with Random Forest, Naïve Bayes, K-Nearest Neighbour to conduct experiment to detect sarcasm and sarcasm type on social media text from twitter and Facebook. The final result shows that ensemble method has better accuracy than SVM.

Hafsa Ahmad, Wasif Akbar Naeem Aslam, Azka and Mohsin[2] the project is based on TF-IDF for feature selection, along with other methods for feature engineering and machine learning algorithms like KNN, Random forest, and Naïve Bayes. In this proposed model unique ML models are used to attain more accurate results. Sarcasm sentiment tweets are used for taunting, insulting, or to make fun of someone.

Parmar K., Limbasiya N., Dhamecha M[3] it considered the potential of live tweets, using a hybrid approach, in processing lexical and hyperbole features, as well as improving the overall performance result. The proposed algorithm was called Feature-based Composite Approach (FBCA). Two lexical and hyperbole features of composite and map reduce were used to reduce the execution time.

Ren Y., Ji D., Ren H.[4] The authors explored the use of neural network models for classifying the sarcastic tweets using Model-Key (local). This technique depends on the convolutional neural network using two self-developed context augmented neural models for the sarcasm detection task. Three labels (negative, sarcastic, and positive) were applied and the result showed that the proposed context-augmented neural models can successfully decode sarcastic clues from contextual information and provide a relative improvement in the detection performance (63.28%).

Tay Y., Tuan L. A., Hui S. C., Su J.[5] proposed an attention-based neural model to explicitly model contrast and incongruity. The proposed technique called Multi dimensional Intra-Attention Recurrent Network (MIARN) was designed based on the intuition of compositional learning through leveraging intra-sentence relationships. The result of the MIARN achieved high performance of 86.47%.

Poria S., Cambria E., Hazarika D., Vij P.[6] developed several models based on a pretrained convolutional neural network (CNN) to extract sentiment, emotion, and personality features for the purpose of sarcasm detection. The developed algorithm combines CNN and SVM (CNN SVM).

Prasad A. G., Sanjana S., Bhat S. M., Harish B. S.[7] the developers used the Gradient Boosting algorithms to classify sarcastic tweets. Gradient boosting was used to optimize the cost function which aids in the classification of the data set by continuously assessing the features and modifies the classifier to avoid wrong classification.

Amir S., Wallace B. C., Lyu H., Silva P. C. M. J.[8] introduced a deep neural network technique to accomplish an automated sarcasm detection task. They proposed a method to automatically perform a learning process and then exploit user embeddings with lexical signals to recognize sarcasm. The authors labeled the data as sarcastic and non-sarcastic when Content and User Embedding Convolutional Neural Network (CUE-CNN) algorithm was used. The performance result was 87%, particularly due to the efficiency of CUE-CNN in inducing vector lexical representations using a convolutional layer.

2.1 SUMMARY OF LITERATURE SURVEY

In summary , detecting the sarcasm classes is an essential matter in text classification and thus has many implications. From studying the Literature review there are various machine learning algorithms that were used to classify the sarcastic statement in various social media platforms. Different machine learning algorithms based on natural learning processing such as Support vector machine, Term Frequency-Inverse Document Frequency (TF-IDF), Context and User embedding CNN, Gradient Boosting and Context augmented convolutional neural networks were used for sarcasm detection each algorithms having its own pros and cons which is briefly discussed in the literature survey. This shows that there are various factors affecting the sarcasm detection in each algorithms which results in precision and accuracy score of the proposed system.

III. PROBLEM STATEMENT AND OBJECTIVES

Given the volume of social data being generated , identifying and understanding sarcastic remarks and comments is becoming increasingly important for business, researches and individuals alike. Detecting sarcasm on social media can be a challenging task because the lack of non-verbal cues such as facial expressions and tone of voice can make it difficult to distinguish between sincere comments and sarcastic ones. Sarcasm can take on many different forms, form ironic statements to satirical humor , which can make it challenging to develop a universal approach for detecting it. Sarcasm detection system can have a very useful application in social media because social media is full of comments, opinions , tweets, etc. These algorithms use natural language processing techniques to analyse the text of social media posts , comments or headlines and identify the patterns and cues that suggest sarcasm. The most common errors nowadays are False negatives sarcasm. Sarcastic tweets not being detected by the model, most probably because they are very specific to a certain situation or culture and they require a high level of world knowledge that DL models don't have.

The most effective sarcasm is the one tailored specifically to the person, situation and relationship between the speakers. 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 conversation. Complimenting someone in a very formal way is a common way of being sarcastic. One of the biggest challenges of DL seems to be adding world knowledge to models, which would speed up training and help tremendously with generalization and bias. Thus we have used GloVe based model along with LSTM to detect sarcasm. The main objective of using this model is to provide maximum accuracy rate.

3.1 OBJECTIVES

As from the literature survey and problem definition , sarcasm is difficult to detect. The purpose and objective of our project is to use a model based on machine learning algorithm which will be able to detect sarcasm from tweets, comments , headlines of the social media. GloVe based model which is an unsupervised machine learning algorithm will help to detect sarcasm with a very good accuracy rate..

IV. PROPOSED SYSTEM

Sarcasm detection is of great importance in understanding people’s true sentiments and opinions. However, sarcasm detection is also a very difficult task, as it’s largely dependent on context, prior knowledge and the tone in which the sentence was spoken or written. The proposed system is based on GloVe algorithm and LSTM model which is an unsupervised learning algorithm. This will help the model to learn from itself even when the proposed model is introduced to a whole new testing data and obtain the same accurate result. So here the main aim is to create a model based on RNN that is LSTM model using Glove algorithm. The final evaluation will be detection whether the following text is sarcastic or not.

The proposed system is compared with another algorithm named word2vec algorithm which is also a unsupervised algorithm and the reason behind the selection of word2vec algorithm for comparison is that both the algorithms Glove and Word2vec are unsupervised algorithms and these algorithms would be useful when it is introduced to new data.

4.1 ALGORITHM

i. Glove Algorithm

GloVe algorithm stands for Global Vectors for Word Representation. GloVe algorithm is a category of unsupervised learning algorithm . GloVe is a very popular algorithm for getting the vector representation for words. These vector representation of words through GloVe is achieved by mapping words into a meaningful space where the distance between the words is related to semantic similarity. In short mapping words which are co related to each other or having similar meanings. The proposed system is also based on semantic analysis so it is important to map words into similar semantic similarity in order to detect the sarcasm. So here GloVe algorithm plays an important role and also it is an unsupervised learning algorithm which helps the system to learn from itself and detect sarcasm even when it is introduced to new testing data.

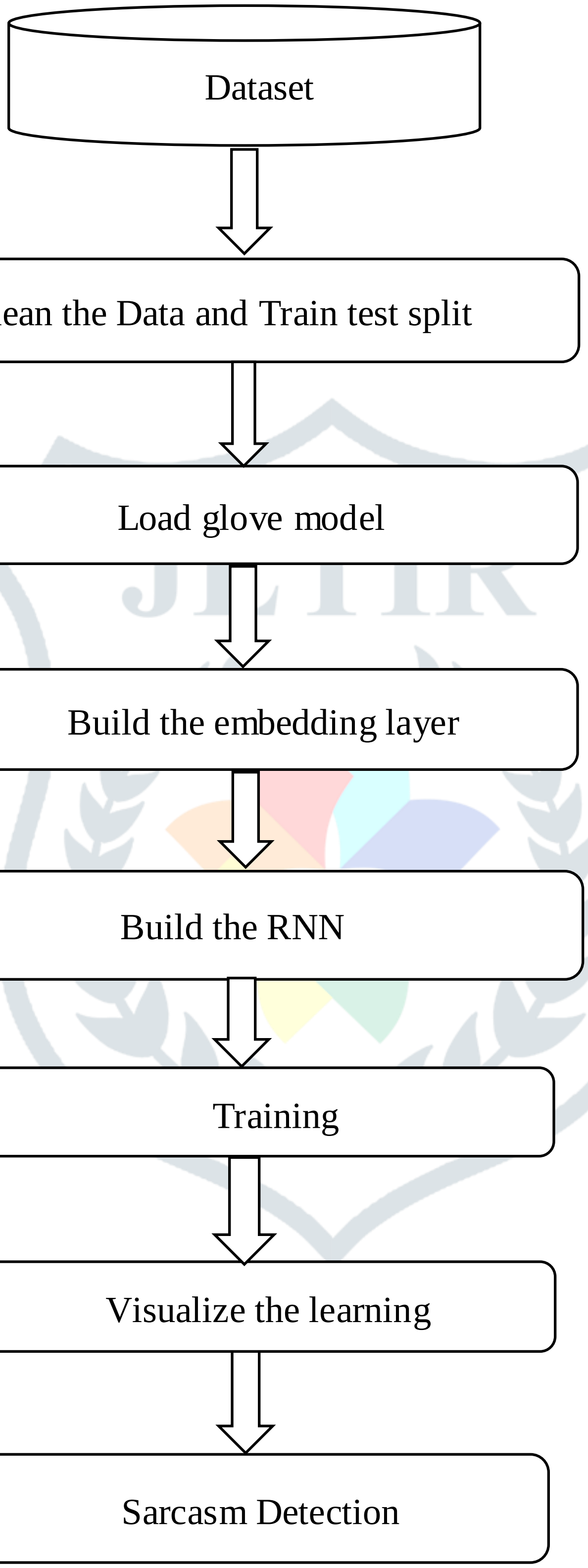
ii. .Long Short Term Memory(LSTM)

LSTM, which stands for Long Short-Term Memory, is a type of recurrent neural network (RNN) architecture specifically designed to handle the challenges of learning long-term dependencies in sequential data. Long Short-Term Memory is variety of recurrent neural network that are capable of learning long term dependencies, especially in sequence prediction problems. The main reason to select LSTM model is because the LSTM itself a type of RNN and recurrent neural networks are very good at learning sequences. This model will be helpful in producing the predictions. LSTM networks have demonstrated remarkable success in various sequential data tasks, including natural language processing, time series analysis, and speech recognition, owing to their ability to capture and retain long-range dependencies. They are widely used in a diverse range of applications, from language translation to sentiment analysis, and continue to be a cornerstone of modern deep learning architectures.

iii.Word2vec Algorithm

The Word2Vec algorithm is also very popular technique in natural language processing (NLP) for learning distributed representations of words in a continuous vector space. Word2Vec algorithm is also comes under unsupervised learning algorithm where the model will be learning from itself even when it is provided a new testing data. It aims to capture semantic similarities between words by representing them as dense vectors, where similar words are located close to each other in the vector space. This is the reason that Word2Vec algorithm is used to compare with the proposed system algorithm because both algorithms are unsupervised learning algorithms and also these algorithms are used to find the semantic similarity from the words which will directly help in detecting sarcasm. Word2Vec embeddings have proven to be useful in various NLP tasks, such as sentiment analysis, machine translation, and named entity recognition, due to their ability to capture semantic relationships between words. They are often used as pre-trained word representations in downstream NLP models.

V. PROCESS DESIGN



VI. METHODOLOGY

i. Dataset : First get the dataset from the Kaggle dataset and upload it in the system. Total two dataset in the form of .json file is uploaded and both the datasets are concatenated.

ii. Cleaning the Data : Second step is the processing and cleaning of the data. Cleaning and data processing is implemented because it will be difficult for the machine learning algorithm to do further classifications. Cleaning the data includes removing all the links, flags and emojis from the following tweets, comments and headlines. Converting all the upper cases into lower cases because machine learning algorithms works well if it is provided smaller case letters. Remove abbreviations and punctuations and symbols too. Also remove all the stop-words from the English language. Examples of stop-words are "to", " the ", " a" etc. After all the data cleaning the system will be left with a proper cleaned list of words.

iii. Train-Test split : In train-test split, the validation split will be equal to 0.2 which means 20% of the data will be reserved for testing purpose and the remaining 80% of the data will be used for training purpose.

iv. Loading the GLoVe model and building the Embedding layer : Now using GLoVe model the cleaned data in the form of words will be converted into vectors containing fixed dimensions. So it will result into sequence of vectors rather than sentences. This is achieved by mapping words into meaningful space where the distance between words is related to semantic similarity. This is nothing but the GloVe model.

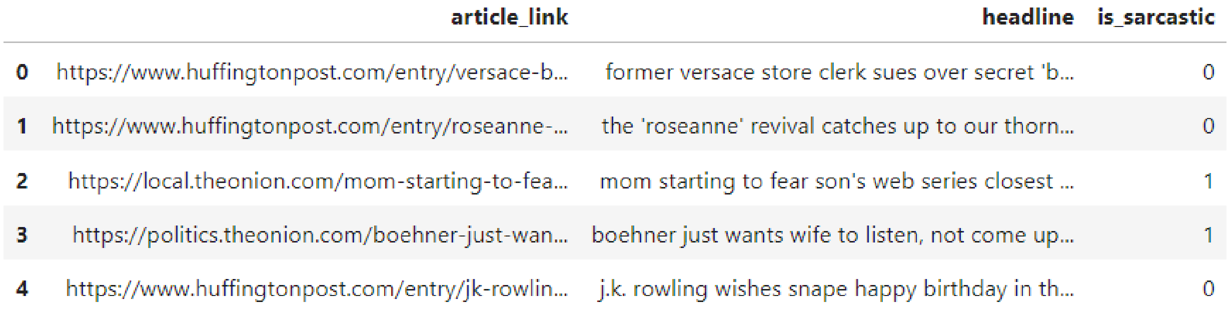
v. Building the Recurrent Neural Network (RNN) : Now once the words are converted into sequence of vectors will pass these sequences to the Recurrent Neural Network. The category of recurrent neural network used in the proposed system is LSTM(long short term memory). The main reason to select LSTM model is because the LSTM itself a type of RNN and recurrent neural networks are very good at learning sequences. This model will be helpful in producing the predictions.

vi. Training and visualizing : The training doesn't take long time. It will carry out total of 25 epochs throughout the training process. An epoch is the number of passes a training dataset takes around an algorithm . Here epoch will be useful to get the validation loss and accuracy . Once the validation loss and accuracy are obtained we visualize the trainings which will compare both training loss and validation loss and will also compare training accuracy and validation accuracy.

vii. Sarcasm Detection : A function will be made and used to check whether the model that is created is able to detect sarcasm or not. In this function we just need the pass the sentence as a parameter and check whether the following sentence is sarcastic or not.

6.1 DATASET DETAILS

The majority of earlier research on sarcasm detection relied on social media datasets gathered under hashtag supervision, although these datasets are noisy in terms of labels and language. Additionally, a lot of tweets or comments are answers to other tweets and comments thus having access to contextual tweets is necessary to detect sarcasm in them. Most of the time, raw data is not complete and it cannot be sent for processing(applying models). Here, preprocessing the dataset makes it suitable to apply analysis on. This is an extremely important phase as the final results are completely dependent on the quality of the data supplied to the model. However great the implementation or design of the model is, the dataset is going to be the distinguishing factor between obtaining excellent results or not. For the proposed model , total two datasets which contains social media tweets, headlines and comments are used. Both the datasets is available on Kaggle which is somehow related to sarcasm. It is not necessary that all the tweets or comments belongs to sarcasm class they may or may not belong to sarcasm class.

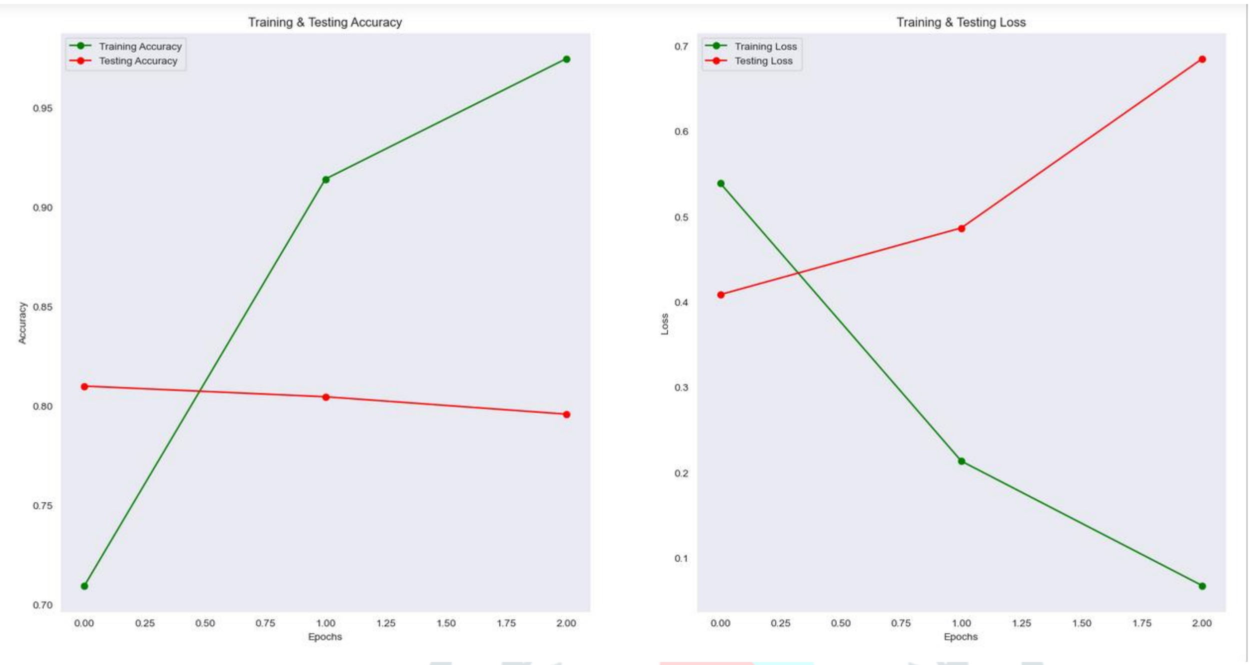


	article_link	headline	is_sarcastic
0	https://www.huffingtonpost.com/entry/versace-b...	former versace store clerk sues over secret 'b...	0
1	https://www.huffingtonpost.com/entry/roseanne-...	the 'roseanne' revival catches up to our thorn...	0
2	https://local.theonion.com/mom-starting-to-fea...	mom starting to fear son's web series closest ...	1
3	https://politics.theonion.com/boehner-just-wan...	boehner just wants wife to listen, not come up...	1
4	https://www.huffingtonpost.com/entry/jk-rowlin...	j.k. rowling wishes snape happy birthday in th...	0

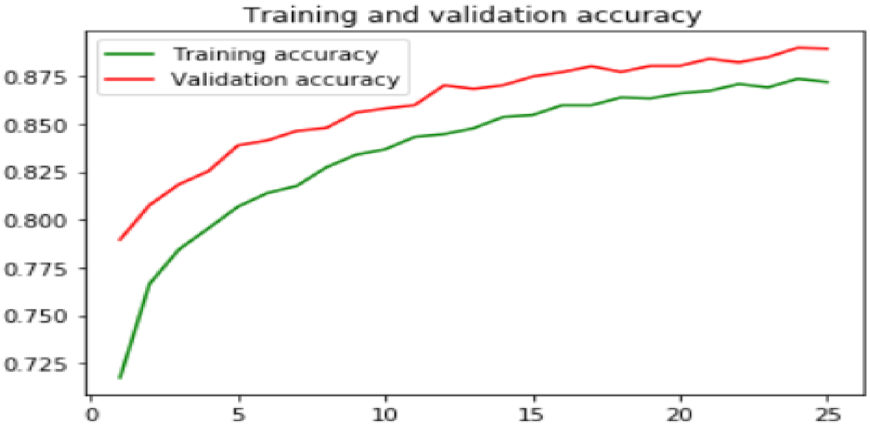
fig 6.1.1 Dataset

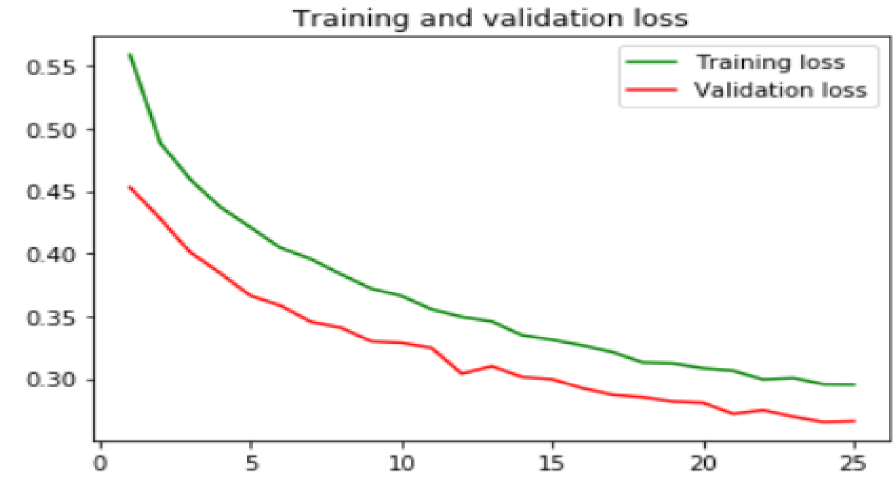
VII. RESULTS

7.1 Training and Validation Accuracy/Loss of Word2Vec Algorithm



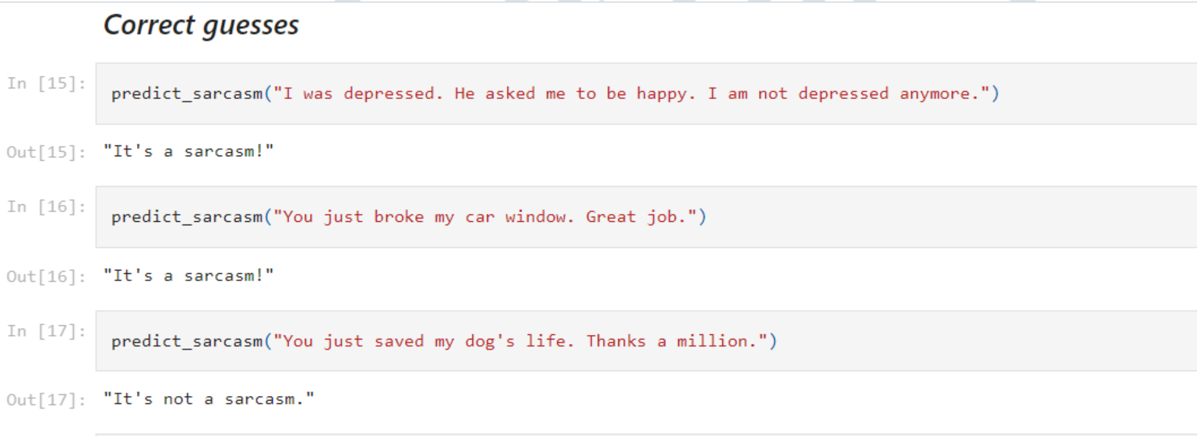
7.2 Training and Validation Accuracy/Loss of GloVe Algorithm





On comparison of both the unsupervised algorithms i.e GloVe and Word2Vec algorithms the accuracy achieved through Word2Vec algorithm is 79.58 whereas the accuracy achieved through the proposed system GloVe algorithm is 84.. The proposed system has successfully achieved more accuracy than the existing algorithm .

A detection function is implemented to find out whether the particular comment/sentence/headline is sarcasm or not.



Correct guesses

```
In [15]:  predict_sarcasm("I was depressed. He asked me to be happy. I am not depressed anymore.")

Out[15]:  "It's a sarcasm!"

In [16]:  predict_sarcasm("You just broke my car window. Great job.")

Out[16]:  "It's a sarcasm!"

In [17]:  predict_sarcasm("You just saved my dog's life. Thanks a million.")

Out[17]:  "It's not a sarcasm."
```

VIII. CONCLUSION AND FUTURE WORKS

In conclusion, the sarcasm detection project represents a significant advancement in natural language processing, with potential applications across various domains including sentiment analysis, social media monitoring, and customer feedback analysis. Through the utilization of state-of-the-art machine learning techniques, such as deep learning models or ensemble methods, coupled with feature engineering and fine-tuning, the project has demonstrated promising results in accurately identifying sarcastic utterances from text data. However, challenges persist, particularly in the realm of context-dependent sarcasm and subtle linguistic nuances. Future iterations of the project could benefit from incorporating multi-modal data sources, such as audio or visual cues, to enhance the robustness of sarcasm detection algorithms. Additionally, exploring techniques for capturing and incorporating socio-cultural context could further improve the accuracy and generalization capabilities of the model.

Leveraging the GloVe algorithm for sarcasm detection presents promising avenues for enhancing natural language understanding. Through embedding words into vector representations that capture semantic relationships, GloVe facilitates the identification of sarcastic utterances by discerning nuanced contextual cues. However, while GloVe provides a robust foundation, the detection of sarcasm remains a challenging task due to its multifaceted nature and context-dependent variations. Also the proposed system was successful in achieving more accuracy than the compared algorithm.

Future research should focus on refining algorithms, incorporating contextual information, and exploring multi-modal approaches to further improve the accuracy and reliability of sarcasm detection systems. Despite these challenges, the utilization of GloVe offers valuable insights and opportunities for advancing the field of sarcasm detection, ultimately contributing to more sophisticated and nuanced natural language processing capabilities.

IX. REFERENCES

Sunusi Kabir, Adamu Habu ,Badamasi Imam , Abdullahi Madaki (2023), “Sentiment analysis of sarcasm detection in social media”.

Hafsa Ahmad, Wasif Akbar Naeem Aslam, Azka and Mohsin (2023), “ TF-IDF Feature Extraction based Sarcasm Detection on Social media ”.

Parmar K., Limbasiya N., Dhamecha M. (2018, February). Feature based composite approach for sarcasm detection using MapReduce. In 2018 second international conference on computing methodologies and communication (ICCMC) (pp. 587–591). Institute of Electrical and Electronics Engineers.

Ren Y., Ji D., Ren H. (2018). Context-augmented convolutional neural networks for twitter sarcasm detection. Neurocomputing, 308, 1–7.

Tay Y., Tuan L. A., Hui S. C., Su J. (2018). Reasoning with sarcasm by reading in between. https://arxiv.org/abs/1805.02856

Poria S., Cambria E., Hazarika D., Vij P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. https://arxiv.org/abs/1610.08815.

Prasad A. G., Sanjana S., Bhat S. M., Harish B. S. (2017, October). Sentiment analysis for sarcasm detection on streaming short text data. In Proceedings of the 2nd International Conference on Knowledge Engineering and Applications (ICKEA) (pp. 1–5). Institute of Electrical and Electronics Engineers.

Amir S., Wallace B. C., Lyu H., Silva P. C. M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. https://arxiv.org/abs/1607.00976.