International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2011-2017 www.ijaem.net ISSN: 2395-5252

Bias Detection and Neutralization of Wikipedia Articles

[1] Harsha Bang, [2]Komal Kamble, [3] Snehal Brahmane, [4]Piyush Muthal,

Dept of Computer Engineering ,MESCOE, Pune.
Dept of Computer Engineering ,MESCOE, Pune.
Dept of Computer Engineering ,MESCOE, Pune.
Dept of Computer Engineering ,MESCOE, Pune.

Submitted: 05-06-2021 Revised: 18-06-2021 Accepted: 20-06-2021

ABSTRACT: Biased means the perceptual strong instinct or a judgment on something.Bias articles are introduced by unique offensive words or phrases during a declaration and must be excluded in order to form the objective declaration. Wikipedia focuses on the policy of the Rational Point of View that implies that Wikipedia data must be neutral.Due to sizable amount of articles on Wikipedia and operating guidelines with voluntary basis of Wikipedia editors, the quality assurance and Wikipedia guidelines can not be always followed.There has been many researchers to developed different techniques ideas to get rid of subjectivity during this paper we are that specialize in those techniques and briefly study about them.

KEYWORDS: Neural Network , Word Embedding, Random Forest, Naive Bayes, NLP , Linguistic Feature Extraction.

I. INTRODUCTION

Since Wikipedia is primarily created by users, it is assumed that the expression of viewpoint is expected. Wikipedia follows a Rational Point of View policy according to which documents should, to the extent possible, be objective, appropriate and bias-free.Wikipedia's policy document advises editors to avoid presenting as an opinion unproblematic facts and, on the other hand, to avoid stating personal views or disputed statements as facts, historical narrative, from the very first field study to the present day[1].

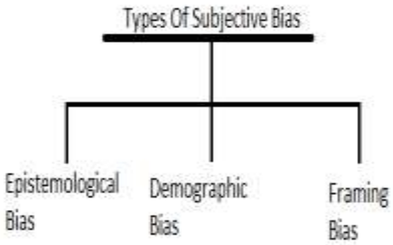
Understanding objective and subjective terminology and the distinction between the two allows us to make informed decisions on data on the concept of subjectivity and the need to neutralize essential and important documents. Subjective is a perception of something.Whereas Objective refers to something which is not influenced nor interpreted by others opinion.This prejudice is introduced by provocative terms and phrases in natural language, putting doubt on evidence and assuming the truth.[2].Every language plays a very significant role to balance our communication and represent our point of view by expressing our thoughts,sharing ideas with others.It mostly depends on our personal experiences and our own perspective.Unbiased language is an important part of balanced and fair representation in writing.Biased language can discriminate between the opinions,demean or offend the society which won't be acceptable in news,Wikipedia articles according to the guidelines.We should make sure that the review or the content does not demean or offend anybody's thoughts or opinions. Usually ,it is influenced by emotions or opinions. Statements and terminology in collaborative contexts or environments where objective language is required (e.g. Wikipedia, news media) should be equally interpreted by the parties concerned and neutrally articulated. Through the presence of offensive terms or phrases, or comments that may be wrong or one-sided, biased language is adopted, thereby breaching such consensus[3].The vocabulary used in Wikipedia should be impartial [4]. The processing of natural language sets such as word embedding increases the accuracy of predictive models. The word embeddings often include and amplify biases such as stereotypes and prejudice[5] present in data.

TYPES OF SUBJECTIVE BIAS

As shown in Fig1.there are 3 types of subjective sentences:

Epistemological Bias: Linguistic features that concentrate subtly on the credibility of the proposal. Example: I Piyush Goel said that all Indian villages were electrified and tended to favor liberal viewpoints. This can be written as 'Piyush Goel said that all Indian villages were electrified and tended to favor liberal viewpoints[6].

Demographic Bias: Presumptions about highly relevant gender or other demographic categories.Example: i) In general, an IT consultant spends his career mired in darkness.It can be corrected as '**IT consultants** often spend their careers mired in darkness.'[6]



Framing Bias:Framing bias happens when individuals make those choices based on the way evidence is framed, as opposed to only the facts themselves.That is the same fact presented with different ways can lead to different judgements. In simple words it is the interpretation of facts by one's subjective opinion.[1]

Other forms of bias, such as selection bias and bias based on particular subjects, such as gender bias or cultural bias, are also addressed in some studies. Open views were not found biased by Hube C and Fetahu B. For instance, the argument that I think this movie is very bad is not, by definition, biased because the author makes it clear that it is her own opinion.

Even though there are more types of subjectivity bias ,the categories in which the subjectivity can be included ,need to be studied and understood. These categories can be separated or identified by analyzing the pattern of editing the articles.This can be done by using WNC corpus. As Wikipedia follows the NPOV corpus, Reid Pryzant Alt studied the WNC where they ignore some edits to maximize the precision,like [1].

- The edits where more sentences were changed.
- The one in which nouns are more than half of the words in the sentence.
- The edits in which spelling or grammatical errors are present.
- Edits that includes references or hyperlinks.
- The one in which the paragraph has symbolic elements, such as tables or punctuation

Wiki Neutrality Corpus (WNC) :

The Wiki Neutrality Corpus consists of pre and post-neutralization aligned paragraphs by English Wikipedia editors[1]. The corpus was extracted from Wikipedia edits that were designed to ensure a rational point of view for writings.WNC is the first parallel corpus targeting biased and neutralized language.In WNC they categorize the types of subjective bias in sentences.It helps to understand the characteristics of the subjective bias on Wikipedia.In history, politics, philosophy, sports, and language categories, subjectively biased editions are most frequent. They are least common in the categories of meteorology, science, landforms, broadcasting, and arts[1]. This indicates that there is a connection between the matter of a text and the revelation of bias. We can work on categorical bias neutralization by using the data provided in WNC.Since the NPOV corpus version is primarily designed to eliminate the subjectivity in which logistic regression and linguistic features, including factive verbs, hedges, and subjective intensifiers have been used to detect bias-inducing words[2]. They have used some methods that include multi-word edits by detecting sentence-level bias.

Word Embedding:

A word embedding is a learned text representation in which a similar representation is provided to words that have the same meaning.This approach to the representation of words and documents can be considered one of the main advancements of deep learning on difficult issues related to the processing of natural language. Word embedding is a commonly-used collection of natural language processing techniques that describe words to real number vectors.. These vectors are used to raise the standards of relational and qualitative models[5].The use of word embeddings to form a model which is highly precise to perform text generation, translation, classification and regression is advancing without taking into account the effect of their inherent biases [5]. BERT (Bidirectional Encoder Representations from Transformers) models pre-trained with big dataset of sentences. There has been significant work on detecting subjectivity using text-classification models ranging from linguistic-feature-based models to fine tuned pre-trained word embeddings such as BERT[4]. It is a powerful tool for NLP, such as calculating similarity measures within the sentences.

Linguistic Features:

The linguistic characteristics of the various form of bias categories differ according to the context of the article. These characteristics can be captured on the idea of the cues encountered during the training of datasets.The context must be analyzed, as biases can be rely on the context, especially epistemological bias, as they depend on the fact of the proposal[6]. There is a sociolinguistic theory in which language and linguistic structure are the medium for the function of a specific social group. Language represent the group's parameters and other characteristics (i.e., ideology, economical, cultural). This generally leads to public consensus on the usage of terminology on a specific topic and the meaning of particular phrases and words[3].In their work, Rohit Raj and Rahul Agarwal listed the kinds of features used in the logistic regression model along with their significance domain. The count for it is 36,787.It includes previous revisions of the documents and the calculation of the ratio between the number of times the word was altered by term of neutralization and its frequency of occurrence. The aim of this feature was to remove framing bias. For linguistic theory, understanding linguistic bias is essential; equally critical for computational linguistics is the computerized detection of biases. The edits related to the NPOV tags allow us to identify the text in its biased (before) and objective (after) form, letting us understand the linguistic realization, as Wikipedia keeps the overall history of revised words

II. RELATED WORK

Christoph Hube and Besnik Fetahu explained [3] the RNN based approach for classifying statements that contained biased language.They focused on the case of biased phrasing,that is statements in which words and phrases were inflammatory or partial. The representation of words in a phrase was an important prerequisite for the effective implementation of RNN models in their assignment. The three main phrase representations have been differentiated.They have distinguished three main sentence representations that were Word Representation ,POS Tags LIWC Word Functions.The RNN models were superior in performance when compared to feature-based models, and were able to capture the important words and phrases that introduced bias in their statement. They were able to predict the bias with a very high precision of 91.7%.

Reid Pryzant Et al. [1] For this neutralization mission, they had proposed a pair of sequence-to-sequence algorithms. Both strategies exploit autoencoders and token-weighted loss function denoising.The algorithm had splitted their problem into (1) identification (2) editing, specifically identifying problematic words using BERT-based detector and a novel join embedding in which the detector could modify an editors' hidden states. This paradigm encouraged an important human-in-the-loop approach to understand the bias and modeling generative language.Second, it was easy to train and use, but the CONCURRENT method was not transparent.BERT encoder was used as part of the generation process to define subjectivity. Also they used LSTM-based editing. They pretrained model for each stage and then combined them into a one system. But the scope was limited to single-word edits, which only focused a quarter of the edits in their data, and was probably for the simple bias sentences[1].

Oriestis Papakyriakopoulos Et al. [5] explained a new technique[5] for gender language bias detection and which was used to analyze biases in Wikipedia-trained embeddings and political social media data. The results were divided into three sections in this paper, firstly, they presented findings on Wikipedia bias and social media word embeddings.Second, they studied how the bias was distributed and how to minimize it. When used in sexism detection models, they also demonstrated the efficacy of biased word embedding. The assessed bias in word embedding was further diffused in the changed sentiment classifiers in the last. They had trained one classifier for each embedding dataset, with accuracy of around 85%.The drawbacks were not found there because the semantic quality of words has always been related to a society's sociopolitical ties and reliance on the existence of the input data on word embedding[5].

Desislava Aleksandrova Et. al.[8] proposed a multilingual method for extracting Wikipedia sentences and used them to create corpus in Bulgarian, French and English.. The hypothesis was that having similar examples in both bias and unbiased classes would help to identify discriminatory words targeted by NPOV-related edits. As this method did not rely on language-specific features other than the NPOV tag list and a stop word list, it was easily applied to Wikipedia archives in other languages[8].

Marta Recasens Et al.explained [7] Actual bias and bias-driven edits extracted from Wikipedia. For each word that initially appeared in the NPOV sentences of the training set, trained a logistic regression model on a feature vector, with bias-inducing words as a positive class, and all the other words as a negative class. At the time of the test, a set of sentences was given to the model and, for each of them,the words were placed as per their probability of being biased.

Rohit Raj and Rahul Agarwal [6] explained the importance of detecting biases in various articles .Since unbiased language was very important to be followed for sources such as news articles, Wikipedia articles. As for the Wikipedia policy of neutral point of view i.e. (NPOV) that suggested that articles should be declared impartial. This paper helped to detect the biased sentences in the articles.They have written in python3, all the scratch modules that were trained and many models such as linear regression, SVM, CRF using the sklearn library. It was modelled as a sequence labeling problem where each word was categorized in O(Unbiased) or B(Biased) in a phrase, then they used a CRF for the labeling task[6].

Tanvi Dadu Et. Al. [2]the implementation of BERT-based models to the task of subjective language detection was being explained. Numerous BERT-based models have been studied, including BERT, Ro BERT, AL BERT, along with their native classifiers with their base and broad specifications.. They have also provided an ensemble model that used multiple ensemble techniques to give predictions.Their proposed model exceeded the baselines by 5.6 percent of the F1 score and 5.95 percent of the Accuracy. FastText, BiLSTM, BERT were its baseline models used in the projects. They also integrated multi-word edits by detecting bias at the sentence level[2].

Hube C. and Fetahu B [4] explained that Wikipedia had a set of editing guidelines and policies for the demographic groups and interests of editors,and to achieve the quality of the information provided [4]. In the paper,Wikipedia statements, they addressed had quality problems that dealt with language bias that were in violation of points(i)avoid stating opinions as facts and(ii)prefer non judgemental language.They have structured lexicon of words using word representation techniques such as word2wec which was found to be effective in disclosing words for a particular word that were similar to or used in a similar context. Using two steps(i)Seed word extraction where high-density bias words were extracted from the list(ii)Bias Word Extraction, which extracted words from the list of seed words from which word embeddings were computed using word2Vec and skip-gram model.

III. METHODOLOGY

The proposed methodology for Bias Detection and Neutralization and Wikipedia articles through machine learning is depicted in the following project overview. The steps that are involved in the process of evaluation are described in detail below.

Step 1: User input and Dataset Preprocessing – For the purpose of management and organization of the construction waste the proposed system takes the user input. User details are stored using SQLite.The input for the system is analyse and preprocessed.

Lexicons are stored in the backend to analyse and extract linguistic features of the given input.Data preprocessing includes removal of null values, to neglect stop words, neglect special characters etc.

Step 2: Developing Baseline Modules – The preprocessed dataset obtained in the previous step is utilized as an input in this step. These models includes analysis of data and with respect to that the task selection for the particular module is decide.Input is in the form of training and testing dataset.There are two tasks defined in the project.Task 1 does the detection of subjective and objective words in the sentence and pass the output to other models as per the requirement.Task 2 includes calculation of subjective score and suggesting of alternate words with minimum/no subjectivity.

Step 3: Linguistic Features selection – In this linguistic features for the given input is generated.This output is then pass on to the respective baseline models to detect the subjective words.

Step 4: Neural Network and TextBlob - The input for this model is vector representation of sentence done by word embedding of the give user input data.If the sentence is objective then it is represented as it is to the user.If the sentence is subjective then it will pass on to the task 2 models as input.

Step 5: Modules for task 2 - There are three different modules are developed for task 2 to find out subjective score and alternate suggestion of words.The average of these along with WordNet database by NLTK corpus is considered as output of task 2.This output is then passed to Bias Detection.

Step 6:Bias Detection - This module generates the flow of the system.The output given to the user at the front end is passed from this module at the backend.Calculation of average of values given by task 2 modules are done here.The results are either as it is the user input in case of objective sentences or appropriate sentences generated by system as per requirements without changing meaning of the sentences.

IV. SYSTEM DESIGN

Tools and Technologies used:

The proposed methodology for Bias Detection and Neutralization of wikipedia articles is developed on the JupyterLab using the Jupyter Notebook. The implementation machine is a laptop equipped with a standard configuration such as the Intel Core i3 processor assisted with 4GB of RAM and 500GB of storage. The SQL lite database server is in charge of the database capabilities for the realization of the presented technique for construction management.

Requirements:

Software:

Operating System: WindowsXP/7/10
Coding Language : Python 3
IDE: Jupyter Notebook,SQL lite
\ Tool:JuptyerLab

Hardware :

Processor :2.2 GHz ,For Fast Processing
Hard Disk :200 GB ,For Fast Processing
RAM: 4 GB ,For Fast Processing
Monitor, Keyboard and UPS : 1 Quantity

Database:

A database is a system intended to organize, store, and retrieve large amounts of data easily. It consists of an organized collection of data for one or more uses, typically in digital form. One way of classifying databases involves the type of their contents, for example: bibliographic, document-text, statistical. Digital databases are managed using database management systems, which store database contents, allowing data creation and maintenance, and search and other access. Database architecture consists of three levels like External, Conceptual and Internal. The external level defines how users understand the organization of the data. A single database can have any number of views at the external level. The internal level defines how the data is physically stored and processed by the computing system. Internal architecture is concerned with cost, performance, scalability and other operational matters. The conceptual is a level of indirection between internal and external. It provides a common view of the database that is uncomplicated by details of how the data is stored or managed, and that can unify the various external views into a coherent whole.
Database: CSV database

System Architecture:

Proposed system, we evaluate the bias sentences on our dataset by using machine learning algorithm.Before classification, a classifier that contains the knowledge structure should be trained with the prelabeled bias and unbiased sentences. After the classification model gains the knowledge structure of the training data, it can be used to predict the bias and unbiased sentences. The whole process consists of two steps: 1) learning and 2) classifying.

First, features of sentences will be extracted and formatted as a vector. The class labels (biased or unbiased) could be get via some other approaches (like manual inspection).

Features and class label will be combined as one instance for training. One training sentences can then be represented by a pair containing one feature vector and the expected result,and the training set is the vector.The training set is the input of machine learning algorithm, the classification model will be built after training process.

DIVISION OF PROBLEM STATEMENT:
1. Task 1 : Classify sentence into subjective or objective class
2. Task 2 Part A : For subjective sentences obtained from Task1, Identify words that are introducing subjectivity in the sentence.
3. Task 2 Part B : Suggest alternative words for subjective words so that sentence will have no or subtle subjectivity but have same/similar meaning as the original one

APPROACH:
INPUT: Representation of words in source sentence:
1. Word2Vec or Glove embedding
2. BERT embedding (If possible) Representation of source sentence:

Average of embeddings of the words in the sentence.Concatenation of embeddings of the words in the sentence by limiting length of the sentence. (To decide length threshold, we may need to analyse data to check length of sentences.

BASELINE MODELS:
Baseline models for Task1 :
(Input: Average embedding representation of sentence)
1. Random Forest
2. Naive Bayes
3. Logistic Regression
4. SVM
5. Neural Network

Baseline model for Task 2

Part A: (Input: Subjective sentences identified in Task 1)

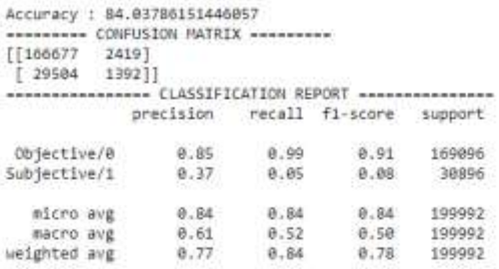
1.Identify subjective words by checking if any word is present in the lexicon collection of lexicons

2.Identification of subjective words by using sentiment analyser like Stanford's coreNLP tool, Textblob, Affinn, etc. and explore if there are any functions which can give us subjective/ sentiment score for words in the sentence or give subjective words in the sentence.

3.Identification of subjective words by using WordNet to find sentiment of words in the sentence

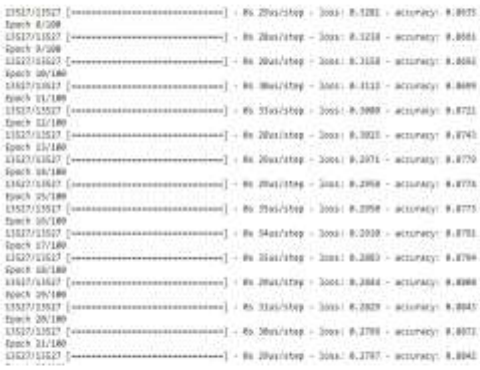
4.Explore more ways/tools to identify subjective words in the sentence. Also, check if it is possible to provide subjectivity score for words if multiple subjective words are found in one sentence for above models.

V. RESULTS



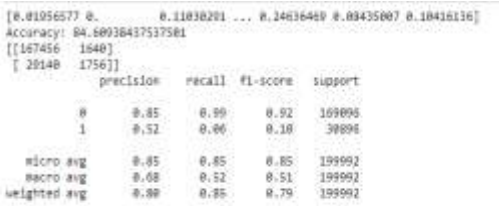
```
Accuracy : 84.03786151446057
--------- CONFUSION MATRIX ---------
[[166677   2419]
 [ 29504   1392]]
---------------- CLASSIFICATION REPORT ---------------
              precision    recall  f1-score   support

 Objective/0       0.85      0.99      0.91    169096
Subjective/1       0.37      0.05      0.08     30896

   micro avg       0.84      0.84      0.84    199992
   macro avg       0.61      0.52      0.50    199992
weighted avg       0.77      0.84      0.78    199992
```

Fig. 3 Accuracy For LR model



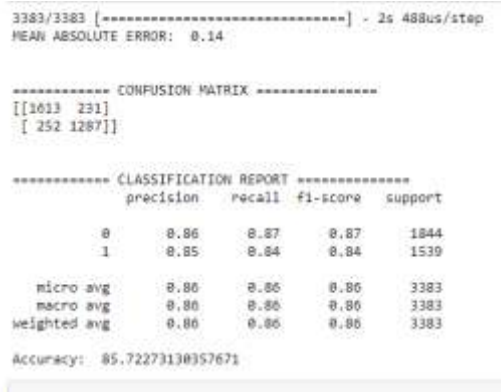
```
13517/13527 [==============================] - 0s 29us/step - loss: 0.3281 - accuracy: 0.8635
Epoch 8/100
13517/13527 [==============================] - 0s 28us/step - loss: 0.3218 - accuracy: 0.8661
Epoch 9/100
13527/13527 [==============================] - 0s 20us/step - loss: 0.3158 - accuracy: 0.8681
Epoch 10/100
13517/13517 [==============================] - 0s 30us/step - loss: 0.3113 - accuracy: 0.8699
Epoch 11/100
13517/13527 [==============================] - 0s 33us/step - loss: 0.3080 - accuracy: 0.8721
Epoch 12/100
13527/13527 [==============================] - 0s 28us/step - loss: 0.3021 - accuracy: 0.8743
Epoch 13/100
13527/13527 [==============================] - 0s 26us/step - loss: 0.2971 - accuracy: 0.8770
Epoch 14/100
13527/13527 [==============================] - 0s 29us/step - loss: 0.2950 - accuracy: 0.8774
Epoch 15/100
13517/13517 [==============================] - 0s 35us/step - loss: 0.2950 - accuracy: 0.8775
Epoch 16/100
13527/13527 [==============================] - 0s 54us/step - loss: 0.2930 - accuracy: 0.8781
Epoch 17/100
13527/13527 [==============================] - 0s 31us/step - loss: 0.2883 - accuracy: 0.8794
Epoch 18/100
13517/13527 [==============================] - 0s 29us/step - loss: 0.2844 - accuracy: 0.8808
Epoch 19/100
13527/13517 [==============================] - 0s 31us/step - loss: 0.2829 - accuracy: 0.8843
Epoch 20/100
13517/13517 [==============================] - 0s 30us/step - loss: 0.2788 - accuracy: 0.8872
Epoch 21/100
13527/13527 [==============================] - 0s 29us/step - loss: 0.2787 - accuracy: 0.8841
```

Fig. 4 Output For NN model



```
[0.01956577 0.         0.11030201 ... 0.24636469 0.03435007 0.10416136]
Accuracy: 84.60938437537501
[[167456   1640]
 [ 29140   1756]]
              precision    recall  f1-score   support

           0       0.85      0.99      0.92    169096
           1       0.52      0.06      0.10     30896

   micro avg       0.85      0.85      0.85    199992
   macro avg       0.68      0.52      0.51    199992
weighted avg       0.80      0.85      0.79    199992
```

Fig. 5 Accuracy For RF model



```
3383/3383 [==============================] - 2s 488us/step
MEAN ABSOLUTE ERROR:  0.14


============ CONFUSION MATRIX ===============
[[1613  231]
 [ 252 1287]]


============ CLASSIFICATION REPORT ==============
              precision    recall  f1-score   support

           0       0.86      0.87      0.87      1844
           1       0.85      0.84      0.84      1539

   micro avg       0.86      0.86      0.86      3383
   macro avg       0.86      0.86      0.86      3383
weighted avg       0.86      0.86      0.86      3383

Accuracy:  85.72273130357671
```

Fig. 6 Accuracy of NN model

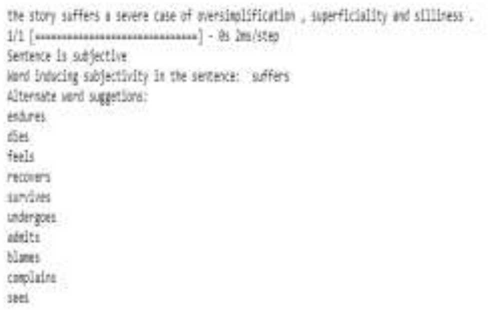
VI. CONCLUSION

This Project implements Bias detection and neutralization of Wikipedia articles.The project models were developed to avoid opinionated thoughts and save editors time . Many points relevant to Wikipedia posts are well known after much study on identifying bias words. Articles such as Wikipedia should be portrayed equally, proportionately, and without any prejudice to the extent possible.

The consolidation of existing quantitative results and carry out comparative analysis is done in first phase of project development.We have successfully studied previous research papers. We have found that the dataset they have been through has a lack of quality and variability and this directly affects the accuracy of the model.According to that the new proposed model is developed which runs more effectively.Classification of sentences done successfully in task 1 with great accuracy as well as suggestion and replacement of alternative words done with negligible errors in task 2.

VII. FUTURE SCOPE

For the Future research approach, the proposed methodology can be enhanced further through the transformation into an API for easier integration. The methodology can also be converted into a mobile application for ease of use.The scope of the langauge can be increased as more knowledge and understanding of the project will grow day by day.



```
the story suffers a severe case of oversimplification , superficiality and silliness .
1/1 [==============================] - 0s 2ms/step
Sentence is subjective
Word inducing subjectivity in the sentence:  suffers
Alternate word suggetions:
endures
dies
feels
recovers
survives
undergoes
admits
blames
complains
sees
```

Fig. 7 Output of the System

REFERENCE

[1]. Reid Pryzant, Richard Diehl Martinez, Nathan Dass,Sadao Kurohashi,Dan Jurafsky,Diyi Yang,"Automatically Neutralizing Subjective Bias in Text. "34th AAAI Conference on Artificial Intelligence (2020)

[2]. Tanvi Dadu(NSIT Delhi) , Kartikey Pant(IIIT Hyderabad)Radhika Mamidi(IIIT Hyderabad) "Towards Detection of Subjective Bias using Contextualized Word Embeddings."(February 2020)

[3]. Hube, C.and Fetahu B. "Neural based statement classification for biased language". In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 195–203. ACM.(2019).

[4]. Hube C. and Fetahu B. "Detecting biased statements in Wikipedia" In The Web Conference, 1779–1786. International World Wide Web Conferences Steering Committee(2018).

[5]. Orestis Papakyriakopoulos ,Simon Hegelich ,Juan Carlos, "Bias in Word Embeddings ". In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). (2020).

[6]. Rohit Raj and Rahul Agarwal, "Bias Detection" CSE Publications IIT Delhi 2018.

[7]. Marta Recasens ,Cristian Danescu-Niculescu-Mizil and Dan Jurafsky, "Linguistic Models for Analyzing and Detecting Biased Language".Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2013).

[8]. Desislava Aleksandrova , François Lareau and Pierre-Andre Menard , "Multilingual Sentence-Level Bias Detection in Wikipedia"Conference: Recent Advances in Natural Language Processing At: Varna, Bulgaria(september 2019).