iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47151

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Term Frequency Based Approach for Binary Classifications on Short Sentences

Kaustubh Keshav

Abstract: ***Great strides have been taken in advancing the domain of Natural Language Processing(NLP) over the past few years. The introduction of attention mechanism in neural networks has led to the creation very complex and efficient architectures like the BERT and GPT. These architectures have been well established as the first choice for most problem statements but, there application in certain simpler use cases generally undermines their real potential. The use case I explore in this research paper is the Natural Language Processing with Disaster Tweets. This is a binary classification task for picking out tweets that refer to a disaster. In this research paper I report the performance of Term Frequency based algorithm without leveraging any vector embedding techniques. Since I am making use of an open source dataset from Kaggle, I have received a leaderboard score of 0.78. The scoring metric used for the leaderboard is the F1 metric and the best score stands at 0.85. The approach leverages the ability of a inverse document frequency(IDF) metric to readily eliminate "stop word" from a corpus and uses a modified term frequency metric<put definition reference> to classify the sentences. The said term frequency metric has been designed in a way to eliminate any impact of stop words that might have escaped from the IDF filter. This paper presents the detailed description of all the algorithms and metrics utilized before coming to a final conclusion.***

Keywords: Tf-idf; Natural language processing; Intent detection; Stop word removal.

I. INTRODUCTION

Sentence classification is one of the most famous algorithmsstudied in the domain of NLP. There is a glut of algorithms, each with its own merits and demerits, available to a practitioner to choose from and build impactful products and tools. In the light of social media gaining massive acceptance all over the globe, it becomes imperative to build systems that are able to categorize social media posts. One such application is to classify the posts of twitter into two classes i. Talking about disaster and ii. Not taking about disaster. For the remainder of this paper I will refer to i. as positive class and ii. as negative class. With the ubiquity of social media users, the task described above can be used by a lot of organizations to help people in an emergency. This made me really interested in the problem. While designing come up with a solution, it made intuitive sense to make use of one the state-of-the-art language models to tackle the problem but, lately there has a been a growing interest in creating lightweight-models that are commensurate with performance to these complex model. The idea implemented in this paper has to do with term frequency. I introduce a metric that is defined for every word in training data. This metric can be intuitively thought of as a vote cast by a word of sentence toward one of the two classes. We define this metric in detail later in this paper. The said metric can also be used to filter out stop words from a text.

II. TERM FREQUENCY RATIO

Term Frequency Ratio is the metric we use to calculate the affinity of a word towards a particular document. We define this metric for a binary classification task. We call the positive sentences with P and negative sentences with N. It is defined as follows:
Term frequency is an ubiquitous text statistic, which measures the importance of a term to a particular document. Term frequency(TF) is defined as below:

$$TF(t, d) = f\,t, d \qquad (2)$$

here

$$f\,t, d \qquad (3)$$

the frequency of term t in document d.

A. Distribution of TFR

The Term Frequency Ratio lies in the range [0, |D'|], where D' denotes the negative document.

B. Weaknesses Of TFR

TFR metric mentioned in the text is not standardized. The maximum value of the TFR for a token in a document D equals the |D'| and hence in situations where the numberof words in individual documents in not the same, TFR might develop a bias towards the smaller documents.

III. DESCRIPTION OF DATA

For testing the algorithm proposed in this paper we make useof a publicly available dataset on Kaggle. The data is a part of Natural Language Processing with Disaster Tweets competition. The goal of this competition is to classify tweets into 50 pertaining to disaster or not pertaining to disaster.

Tweet Class	Average Length
Disaster	22
Non Disaster	24

Table 1: Average Number Of Words In Tweets

$$T(t, d) = f\,t, d \quad (1)$$

IV. SOLUTION

The algorithm performs the following steps to generate pre- dictions for the tweets.

A. Generate TFR for Positive and Negative Sentences

As a first step we create two mappings one for each class of the tweets. The mapping contains the TFR ratio for each word occurring in the tweets of corresponding classes. Once the TFR for both positive and negative classes are generated, we take sum of TFR for each class as the vote towards a particular class. The class with the higher vote is then considered the predicted class.

B. Model Selection

The hyper-parameter for this algorithm is the IDF threshold to reject words as stop words. A simple grid search was per-formed on the IDF ranging from 0.1 to 0.99 and algorithms performance was tracked on a validation set.

The maximum IDF value chosen was 0.57 which gave re-sults presented below. The grid search was performed in two stages for better speed. First a search was performed on a exponential scale and a region of interest was selected. A linear search was performed in the region of interest to get the final value of theIDF.

C. Results

Using this approach we are able to obtain the performance mentioned below.

Data Split	F1 Score	Accuracy
Train	0.836	0.864
Test	0.737	0.789
Public Test	NA	0.777

Data Split	F1 Score	Accuracy
Train	0.83	0.86
Test	0.74	0.79
Private Test	NA	0.78

Table 2: Model Performance On Various Data Splits

The algorithms is able to achieve a F1 score of 0.77 on the public test set which when compared to the 99 percentile score of 0.85 presents the merits of using a simplistic approach for intent detection could save a lot of computation time while providing a competent performance.

V. CONCLUSION

Through the experiments presented in this paper we are able to motivate the use of general statistical algorithm for building baseline models for Intent Detection Tasks of NLP. A sophisticated model should be able to outperform such approaches with a significant margin to prove merit. In general for quick experimentation purposes a simplistic or rather naive statistical model can come close in terms of performance to a much more nuanced solution.

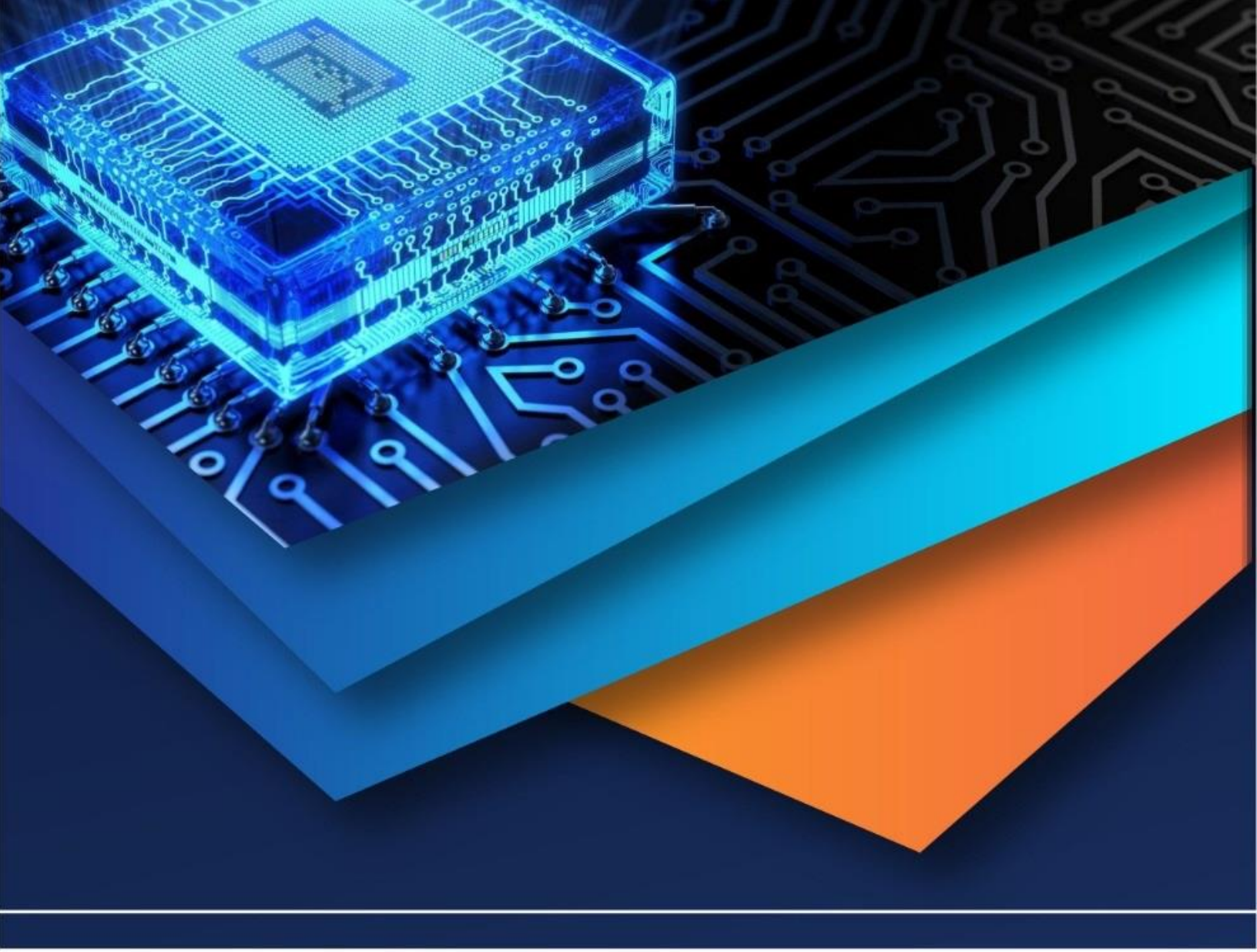
VI. ACKNOWLEDGEMENTS

I would like to thank kaggle.com for hosting the competition and keeping the leaderboard live for validating the results generated in this research paper.

VII. DATA AVAILABILITY

All data used in this experiment is available herehttps://www.kaggle.com/competitions/nlp-getting-started/data .

VIII. COPYRIGHT NOTICE



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY