# INTRODUCTION

Media bias meaning distorted news coverage, which strongly yet passively polarises the public opinion of the current events. In linguistics, manual techniques are employed. Whereas, the Natural Language Processing and Machine Learning techniques revolutionize the process into exponentially performing systems[1].

It is essential for the public to have reliable, factual, and unbiased news to have a well-informed society with a healthy understanding of the events. Media outlets in the modern world have a crucial part in the popular public opinion. However, editorial slant results in bias which is reflected in news articles and TV reports as well. Various factors influence the bias which may be the source of income or just the ideological stand of the media outlet or of their readers or viewers. From a Linguistic point of view, the formation of editorial slant constitutes a set of frames, which comprise words in a certain combination that seek to “promote a particular interpretation” of a story [2]. The literature determines various techniques through which news articles can indicate bias. Journalists deliberately choose certain sources, and from these sources,

they collect data that is distorted and then report events with a biased political stance.

The domination of corporations in the media has substantially escalated in the past decades[6]. Thus, there is a greater chance for media bias.

The media bias identification, and their analysis, are grabbing the attention of those in Machine Learning, how Machine Learning techniques can help achieve the goal of recognizing and mitigating political slant and media bias by eventually providing a factual and neutral point of view news which can shape a healthy society with a well-formed opinion[3].

# 1.2 Media Bias:

Media bias must have the intention and sustained meaning, it must represent a deliberate act and it must be systematic i.e. must not a single isolated incident[7].

To analyze bias it is essential to have two datasets. The first one can be various news articles in which we want to identify bias and second is employed for providing baseline which contains all events and these events are already classified as bias or not bias. After comparing the two datasets we can deduce whether the given source is biased or not.[8].

The media biases that were identified in the research paper were:[9]

1. Coverage
2. Gatekeeping
3. Statement

## 1.2.1 Coverage Bias:

Coverage bias is a form of bias where in time or space given in TV or news articles respectively to topics or entities, such as a person or country, are deliberately adjusted in a way to favour an entity. Making these topics to be perceived as more or less important for the audience.[2]

## 1.2.2 Gatekeeping Bias:

Gatekeeping bias, which is also known as selection bias or agenda bias. As the name suggests, it is when the media outlets select their headline and ignore events that are to be reported. Only a few selected events make it to the news.

## 1.2.3 Statement Bias:

Statement bias, also known as presentation bias, signifies how news articles choose to report on concepts or ideologies to create a perception. We can consider, in the US elections, an easily identifiable bias surfaces from media outlets as a result of editorial slant, in which the journalist’s stance for a presidential candidate affects the quantity and tone of a newspaper’s coverage being aggressive or defensive for the candidate and thus affecting his candidature. [3,5].

## 

Fig. 1.1: Identification of media Bias by grouping[5]

## 1.3 Motivations

It is essential for the public to have reliable, factual, and unbiased news to have a well-informed society with a healthy understanding of the events. With the increasing polarising in media reporting in India is was an inevitable need to make such a system.

## 1.4 Problem Statement

To Design and develop a system that can detect bias or political slant in various News articles using tools of Natural Language Processing.

## 1.5 Objectives and Scope

* Automate identification of news bias and analysis of news articles.
* Provide unrestricted access to unbiased information.
* Help shape correct public opinion on important issues.
* To identify the political bias of media houses using machine learning algorithms.
* To detect and visualize the occurrence of political bias in real-time.

# LITERATURE SURVEY

**Four Rules of numeric text analysis:**

(1) The Numeric models of language are incorrect—however few might be helpful.

The data generation process for media articles is a conundrum. Every random sentence has an intricate dependency structure, its meaning is affected extremely with the addition of a new frame of words, and the context of the sentence might extremely affect what it means.[3]

(2) Numeric techniques for text amplify resources and augment linguists.

Significant performance throughout the various significant problems is exhibited by Automated content analysis techniques. These techniques don't terminate the necessity for the cautious insight provided by researchers nor eliminate the need for analyzing texts. Indeed a deep understanding of the texts is one amongst the key benefits of the scientists in applying automated techniques.

(3) A standard technique for automated text analysis that performs consistently does not exist.

Distinct datasets and distinct research problems often cause distinct numbers of interests. Text models generally exhibit these properties. Occasionally, the objectives are to determine the frame of words that identify the difference in the language of classes.

(4) Verify, Verify, Verify.

Automated text analysis techniques can considerably decrease time and the costs of analyzing enormous amounts of political articles. When employed upon a problem, however, the output models might be deceiving or just incorrect. Hence, it is necessary for scholars to verify their application of automated text analysis. This verification might have many different forms, many of which we describe below. When classes are available in a dataset, researchers must exhibit that the supervised techniques are proficient enough to consistently perform like that of a person. Verification for unsupervised techniques is rarely easy to perform[4,1].

Using easier algorithms like the Naive Bias and Perceptron for the classification of the Bias resulted in the proof of the first principle. That some are indeed useful. Just by identifying the following words in an article to be related to the article being Conservative or Democratic gives an insight into the bias. The below tables show the given [6].

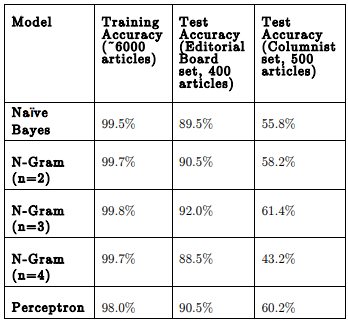
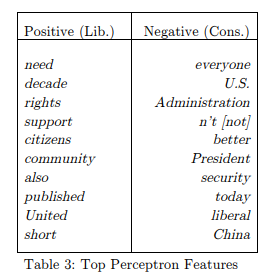
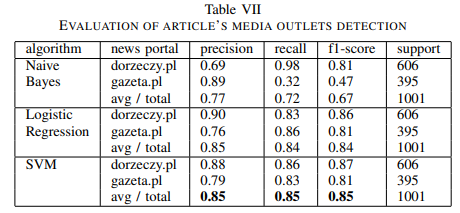
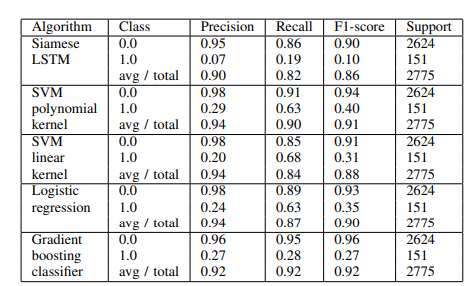


Fig 2.1: Evaluation and results of news articles[]

Above is the result of the experiment to identify if a pair of articles is similar or not. The data was split into test and train datasets as presented in. making a set of similar articles 320 for training and 151 for testing. And a set of not similar articles 7837 for training and 2624 for testing. Similar articles were labelled as ’1’ and not similar articles as ’0’. Then created the following features: cosine similarity on tf-idf vectors, number of similar keywords, normalized number of similar entities that is the number of similar entities/sum of entities in both articles[3].



EVALUATION OF ARTICLE’S MEDIA OUTLETS DETECTION [3]



EVALUATION OF ONE TO ONE ARTICLES PAIRS[3]

# SYSTEM DESIGN

1. Create a Dataset from the televised audio.
2. Add a few features from that analysis such as the keywords and tone.
3. Apply K-Means on the transcript and the generated features.
4. This generated model with highest silhouette score will have an optimal number of Ks .
5. On this generated model we can cluster the given televised audio into a cluster

|  |
| --- |
| 3.1 Block diagram |
|  |
|  |
| 3.2 Hardware and Software Requirements  3.2.1 Software Requirements  python3  pip  nltk  matplotlib |

3.2.2 Hardware Requirements

Processor: Intel Pentium

RAM: 2 GB

Storage: 10 GB

# IMPLEMENTATION DETAILS

|  |
| --- |
| 4.1 Algorithm and flowcharts (If applicable):  4.1.1 Newspaper Library (Python):  The Newspaper Library in python uses Natural language processing and gives us the summary as well as the keywords used in the news article that shall be passed to it. It uses web scrapping first to get the article just by using the URL then the article is parsed. This parsing and processing gives a summary of that article and provides us with the list of the keywords used in that article. Same can be done by providing it with a transcript. [11]  4.1.2 K Means Clustering:  The K Means Clustering is an unsupervised learning algorithm procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location creates different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as the barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more. [10,12]      (src please)  Fig 4.1 The K Means Clustering Flowchart |
| 4.2 Results:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | No. of K | Silhouette | Accuracy | F1-Score | Recall | |  |  |  |  |  | |  |  |  |  |  | |  |  |  |  |  | |
| 4.3 Discussion |

The reason to use an unsupervised learning algorithm is to get the bias as unbiased as possible as the human error has to be avoided in such a delicate topic. However, when one takes a look at the results it actually exhibits that this is representing closely the way of journalism rather than just the bias. The Bias is the slant but when studying the words being used in polarisation also differ by the media outlets and thus making the bias detection automation hard. By, the use of Sarcasm and asking rhetorical questions or questions that are real. Causing a greater conundrum but still demonstrating the similarity.[4,6]

# CONCLUSION & FUTURE SCOPE

The use of unsupervised learning gives a satisfactory output but the algorithm has substantially clustered the similarity in the style of journalism. This is similar to the finding as that the Polish researchers but the SVM when employed has better results but then risking the reliability on the dataset. However, the implementations of Decision trees are yet to be worked with but they seem to give an overfitted model.

Although the model has achieved remarkable success, it is currently working on the written media data. We aim at achieving even higher accuracy on the TV news channels by using the proposed model by voice analysis. At this present instance, there is no dataset of the Indian television news channels as the problem remains unquestioned in this current scenario of the news channels being biased.

In the course of study, it would be better to not only get reports but also analyse the debates being held. They attract prime time slots and have a relatively higher impact than that of regular reports. The reason is, there are experts representing different views and the role of the adjudicator and his stance on each ideology shapes the debate. The image processing of that of the reports analysing the expressions to overcome the rhetorical questions conundrum.