# A machine learning algorithm for binary classification of gender bias in text

TANBIR SARKAR KAMAL SONI APRATIM BHATTACHARYA BODDEPALLI NAVJOTHY

## Abstract

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions. Technological developments such as artificial intelligence can strengthen social prejudices prevailing in society, regardless of the developer's intention. Therefore, researchers should be aware of the ethical issues that may arise from a developed product/solution. In this project, we were asked to experiment with a text dataset, and to explore how machine learning algorithms can be used to find the patterns in data. We were expected to classify the data to some meaningful insights using a common data mining and machine learning library. To test the effect, we perform gender classification using logistic regression technique of machine learning where the data is classified as a set of 0 and 1.

**Keywords**: Machine Learning, Pattern Recognition, Binary classification, Bias in text, Supervised learning, Logistic regression.

## 1.1.0 Introduction

Detecting bias is becoming gradually important based on its relevance in many fields, ranging from evaluating publications to understanding political perspectives. The progress of technology day by day and its inclusion in many areas of our lives has brought new social problems. Bias is difficult to detect and evaluate because it is usually implicit. People can develop discrimination toward or against an individual, an ethnic group, and gender identity. Therefore, researchers should be aware of this ethical issue. In the area of sentiment analysis, algorithms are not unbiased. They give more accurate results when trained on female-authored data, which implies that they over-represent females' viewpoints in a gender-mixed collection. Considering the risk of increasing data bias, word embeddings have been analysed first. It's shown that these word representations reflect social tendencies that exist in the data used to train them. Gender bias is the preference of one gender over another or approaching one gender prejudicially. In NLP literature, different methods have been used to reveal gender bias. Gender bias is also evaluated in classification problems.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labelled data. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman's terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naive Bayes Classifier, Nearest neighbour, Bayes Net, Majority Classifier etc., and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem.

In this project, we try and find patterns in a dataset having pairs of text and determine if they are similar or dissimilar based on some form of gender bias present in the text.

### 1.2.0 Problem Definition and Algorithm

Textual similarity is an important problem in NLP. The broad goal of textual similarity is to measure the extent of similarity between a given pair of text fragment (sentence/paragraph etc) based on a specific aspect/criterion (such as topic, sentiment, etc).

In this problem, a set of pairs of text is given as dataset. The end goal is to determine if they are similar or dissimilar based on some form of gender bias present in the text. The forms of the gender bias could be: i) firstness: where a gender (male/female) is always mentioned first, ii) stereotyping of a particular gender, iii) subordination: where the text reflects a gender is subordinate compared to other. This is binary classification problem which can be solved with various machine learning methods for classification of data from a given dataset. The desired output will help make a classification to make a segregation and come to a decision making. Binary Classification refers to predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc.

### 1.2.i Task Definition

Training data format: The training data will contain a set of text pairs (p1 p2) along with their labels (0 or 1), where 0 indicates p1 and p2 can be both biased or both unbiased, similarly, 1 indicates if one is biased but the other is unbiased. There are two files for training data. The first file (name: text-and-id) contains the text (2nd column) and its unique id (1$^{st}$ column) in each line, while the second file (name: pairs-label-training) contains the ids of the two text fragments and the corresponding label (0/1) in each line.

Test data: Test data contains a set of text pairs, and you have to produce the 0/1 tag for each pair in the test data.

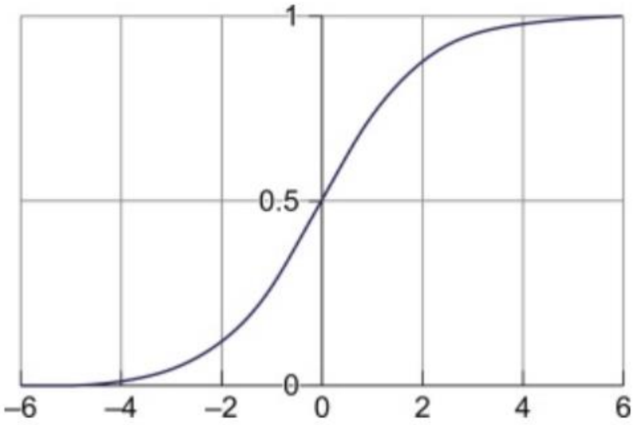
Desired output: Given a text pair (p1 p2), your goal would be to mark this as 0 or 1. The meaning of 0 and 1 is the same as in the training data.

### 1.2.ii Algorithm Definition

To address this issue, a logistic regression technique is used. Logistic regression is one of the most popular binary classification algorithms. Given a set of feature samples, the goal of logistic regression is to output a value between 0 and 1 on a linear combination of inputs that can be interpreted as the probability that each sample belongs to a particular class. Logistic regression maps the continuous output of traditional linear regression (-∞, ∞) to probabilities (0, 1). This transform is also symmetric, so reversing the sign of the linear output reverses the original probabilities. The formula for the logistic function is:

$$F(x)= \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

When we plot the above equation, we get S shape curve like below.



An important point in the figure above is that the output along the vertical axis is always between 0 and 1, regardless of the value of x you use in the logistic or sigmoid function. Classify a label as class 1 or the positive class if the sigmoid function result is greater than 0.5. If less than 0.5, it can be classified as negative class or 0. Logistic regression uses iterative optimization algorithms such as gradient descent or stochastic techniques such as maximum likelihood to obtain the "best" S-curve.

Let's understand the math behind the sigmoid function. As a rule, logistic regression values should be between 0 and 1. The constraint that it cannot exceed a value of 1 result in an 'S' curve on the chart. This is an easy way to identify sigmoid or logistic functions. Regarding logistic regression, the concept used is the threshold. Thresholds help define the probability of 0 or 1. For example, values above the threshold tend to be 1, and values below the threshold tend to be 0. Unlike generative classifier there’s a discriminative classifier that aims to learn the probability distribution over all classes. Logistic regression is an example of a discriminative classifier and is commonly used in text classification, as a baseline in research, and as an MVP in real-world industry scenarios.

## Related Work

We all check our emails every day, sometimes many times. A useful feature of most email service providers is the ability to automatically separate spam emails from regular emails. This is a use case for a common NLP task known as text classification and is the focus of this chapter. Text classification is the task of assigning one or more categories to a given text from a larger set of possible categories. In the email spam identifier example, he has two categories, spam and non-spam, and each incoming email is assigned to one of these categories. This task of classifying text based on some properties has wide applications in various fields such as: B. social media, e-commerce, healthcare, legal, marketing, etc. The purpose and application of text classification may vary from domain to domain, but the underlying abstract problem remains the same. The immutability of this core problem and its application across a wide range of fields has made text classification the most widespread NLP task in industry and one of the best-studied NLP tasks in academia. This chapter discusses the usefulness of text classification and how to create a text classifier for your use case, along with practical tips for real-world scenarios.

Gender influences myriad aspects of NLP, including corpora, tasks, algorithms, and systems. For example, statistical gender biases include word embeddings (including multilingual embeddings), cross-referencing, part-of-speech and dependency analysis, unigram language modelling, appropriate turn-taking classification, relationship extraction, and objectionable content. Affects various downstream tasks. identification and machine translation. We also note that the translations were written by speakers who were older and more masculine than the originals. Machine learning has been found to reinforce gender bias in training corpora, especially dialogue texts. Although many of the articles cited above suggest ways to mitigate the undesirable effects of gender on texts, it is especially important to modify training distributions to compensate for statistical gender imbalances in counterfactual data. depends on gain. Also relevant is the introduction of a parallel style corpus that demonstrates the advantages of style transfer across binary his genders. This work provides a clean and fresh way to understand gender bias that extends to dialogue use cases by separately examining author gender contributions to human-generated data. Most relevant to this work is the proposal of a framework for modelling the practical aspects of many social prejudices in texts. These works focus on complementary aspects of the overarching goal of making NLP safe and inclusive for all but differ in some respects. Here we look specifically at statistical gender bias in human- or model-generated text to give such a complex phenomenon the focused and nuanced attention it deserves. It takes a different perspective and aims to characterize the broader landscape of negative stereotypes in social media copy. It reveals similarities between different types of socially harmful content. It is an approach to Also, they consider practical dimensions that differ from ours:

While they target the negative stereotypical implications of common sense with their arguably harmless remarks, the conversation examines practical dimensions that directly correspond to roles (i.e., subject, recipient, and content creator). increase. Therefore, we believe that the two frameworks are fully compatible.

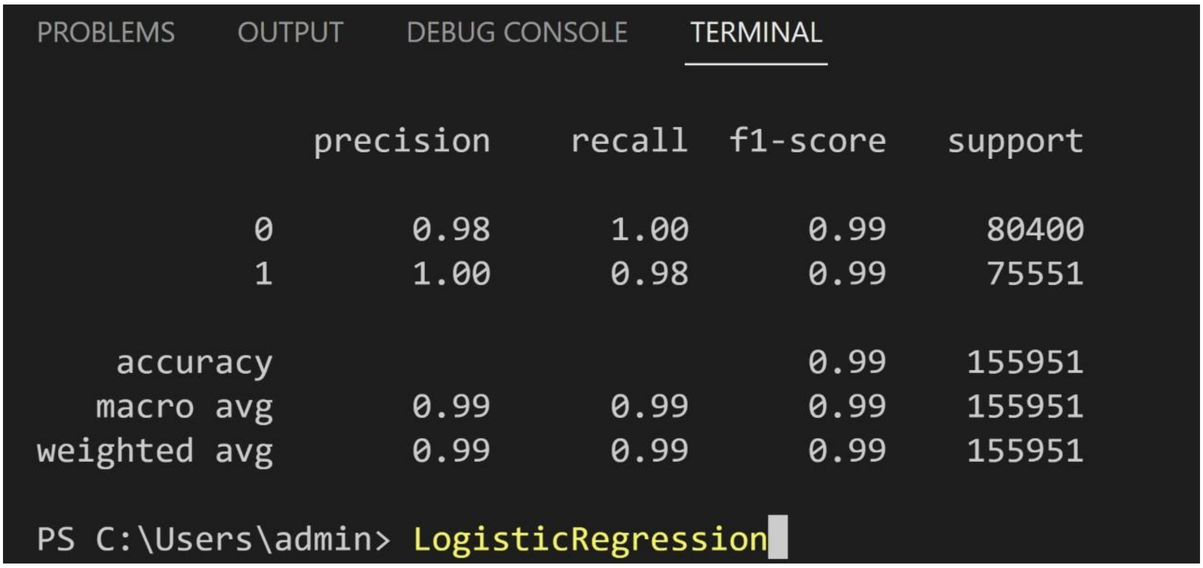
## Experimental Evaluation

### Methodology

**Building the model:** You should normalize your data and shift the mean to the origin. This is because the nature of the logistic equation requires exact results. Next, create a method that helps you make predictions that return probabilities. After that, you can proceed to train the model. The model is trained for 96.795 %, and partial derivatives are computed, and weights are updated these 96.795 %. Although they target the negative stereotypical implications of common sense with their arguably harmless remarks, the conversation has practical aspects that directly correspond to roles (i.e., subject, recipient, and creator of content). Explore. Therefore, we believe that the two frameworks are fully compatible.

**Training the model:** As mentioned earlier, the forecast equation returns probabilities. Since this is a classification task, we need to convert it to a binary value. To do this, we need to select a threshold. For this example, choose a threshold of 0.5. This means that all predicted values above 0.5 are treated as 1 and all below 0.5 are treated as 0. You can also calculate accuracy by looking at the number of correct predictions and dividing it by the total number of test cases.

## Results



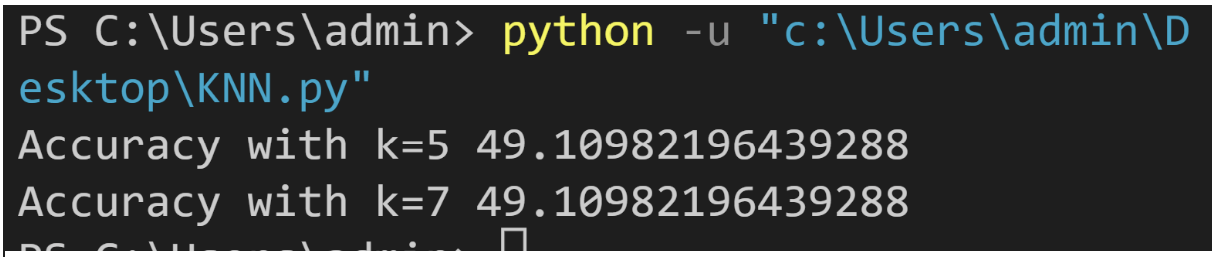
```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     80400
           1       1.00      0.98      0.99     75551

    accuracy                           0.99    155951
   macro avg       0.99      0.99      0.99    155951
weighted avg       0.99      0.99      0.99    155951

PS C:\Users\admin> LogisticRegression
```

*Figure 1 Accuracy score of Logistic regression*



```
PS C:\Users\admin> python -u "c:\Users\admin\D
esktop\KNN.py"
Accuracy with k=5 49.10982196439288
Accuracy with k=7 49.10982196439288
PS C:\Users\admin>
```

*Figure 2 Accuracy score of KNN model*

From the accuracy score it has been observed that from logistic regression model 99% of accuracy has been achieved, whereas from KNN model it much lesser, only 49.1%. From the testing it has been clear that greater accuracy can be achieved from a logistic regression model from a binary classification of this kind.

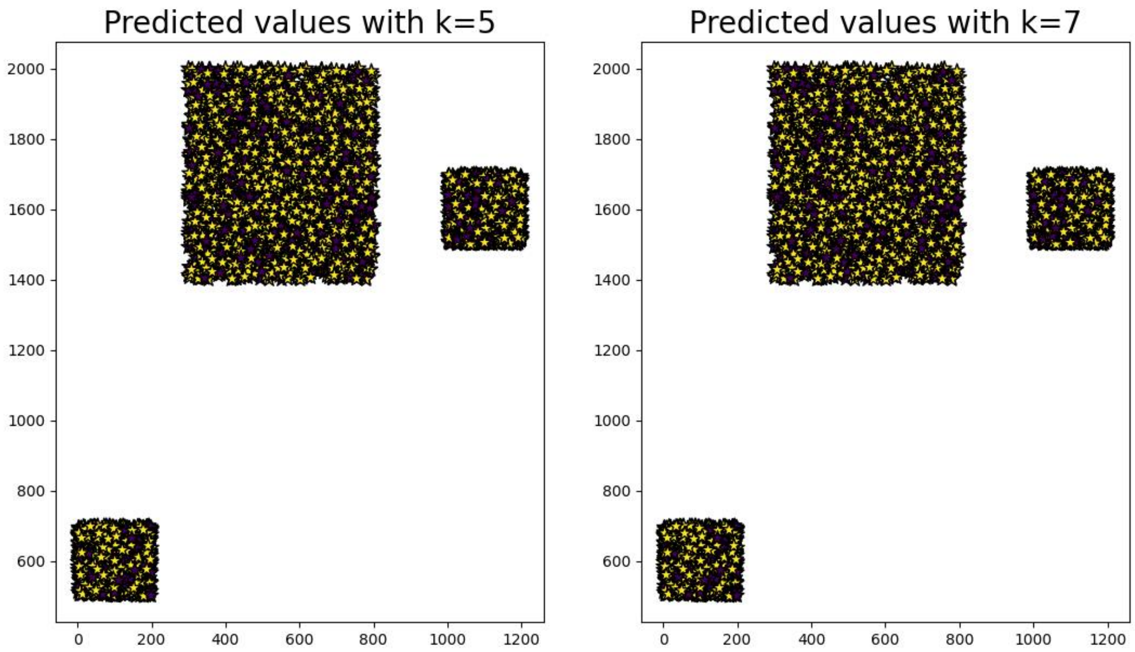


*Figure 4 data visualization of KNN cluster*

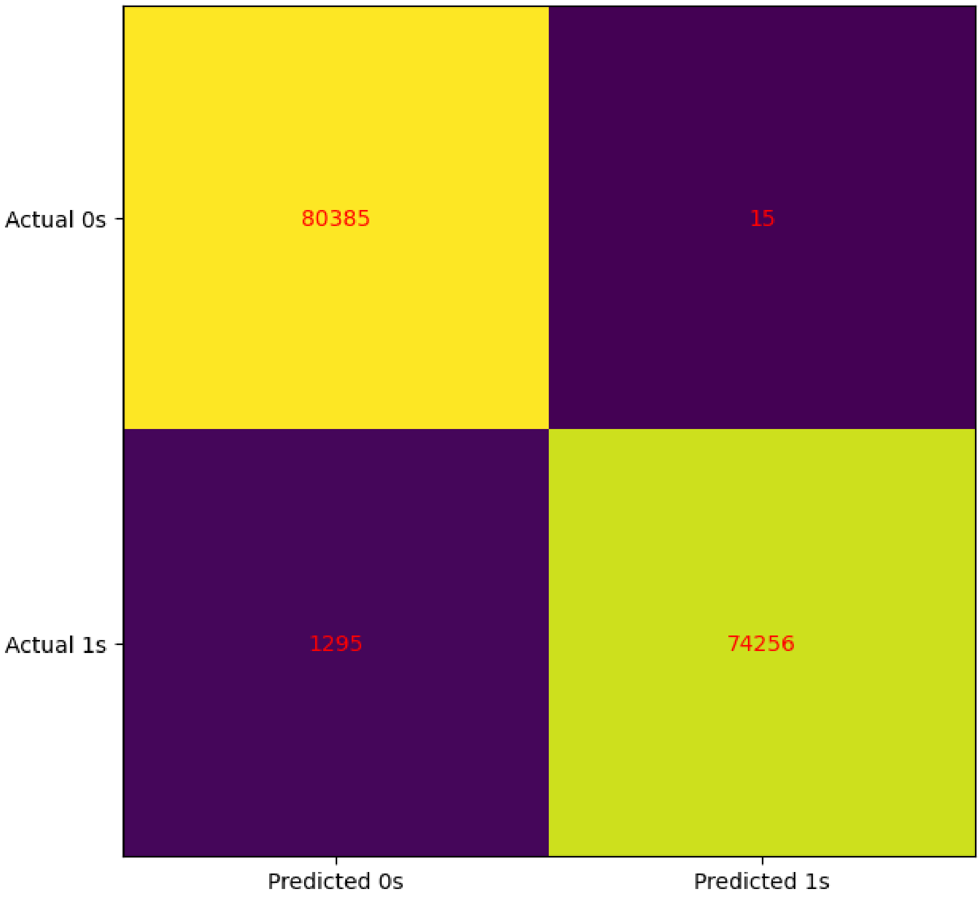


*Figure 3 Confusion matrix from logistic regression*

## Discussion

Logistic regression is a widely used technique because it is very efficient and does not require many computational resources. Logistic regression works more efficiently if you remove variables that have little or no relationship to the output variable. Feature engineering is therefore a key factor in performing logistic regression. Logistic regression is very useful for classification tasks. However, it is not one of the most powerful algorithms. It easily outperforms other more complex algorithms. Anyway, it is light and easy to handle. However, due to its simplicity, it can be used as a good baseline to compare the performance of other, more complex algorithms. The test data was also passed to KNN (K Nearest Neighbour) to see if the logistic regression application technique provided higher accuracy. From our tests, we found that the logistic regression application method yielded higher values compared to the KNN model. There are several other machine learning algorithms that can be used to develop classification models.

## Conclusion

There are numerous machine learning techniques for binary classification. In a nutshell, logistic regression is used for classification problems when the output or dependent variable is dichotomous or categorical. There are some assumptions to keep in mind while implementing logistic regressions, such as the different types of logistic regression and the different types of independent variables and the training data available.

The methodology provided by this study is primarily in the ability to recognize and perform binary classification within digital text documents. Text classification has proven to be an excellent method for studying gender bias in digital texts for a variety of purposes. B. Offensive language detection, male or female detection, text-to-speech, etc. A neutral data set is gender neutral. Besides creating manuals investigate if there are names and other gender words in the corpus feature importance can be examined by examining which features have the highest and lowest coefficients in logistic regression.

## Bibliography

1. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.
2. Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 25–32.
3. Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts, Hong Kong, China. Association for Computational Linguistics.
4. Christine Charyton and Glenn E Snelbecker. 2007. Engineers' and musicians' choices of self-descriptive adjectives as potential indicators of creativity by gender and domain. Psychology of Aesthetics, creativity, and the arts, 1(2):91.
5. Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. Digital Investigation, 8(1):78–88.
6. Marta R Costa-jussa. 2019. An analysis of gender bias ` studies in natural language processing. Nature Machine Intelligence, pages 1–2.

## Annexure

GitHub - icekamal786/Data_Analytics: OPEN IIT