Abusive Language Detection Against Immigrants and Women

[Introduction 3](#_Toc38537539)

[Project Aims 3](#_Toc38537540)

[Intended Audience 3](#_Toc38537541)

[Project Scope 3](#_Toc38537542)

[Approach 3](#_Toc38537543)

[Summary of Outcomes 4](#_Toc38537544)

[Background 5](#_Toc38537545)

[Project Context 5](#_Toc38537546)

[Identified Problem 5](#_Toc38537547)

[Likely Stakeholders 6](#_Toc38537548)

[Theory Associated with Problem Area 6](#_Toc38537549)

[Constraints on Approach 8](#_Toc38537550)

[Existing Solutions 8](#_Toc38537551)

[Methods and Tools 9](#_Toc38537552)

[Specification and Design 10](#_Toc38537553)

[Approach to Solving the Problem 10](#_Toc38537554)

[Implementation 12](#_Toc38537555)

[Supplementary Materials 13](#_Toc38537556)

Introduction

# Project Aims

The aim of this project is to design an effective system for detection of abusive language towards immigrants and women. The project is based on a similar project run online[[1]](#footnote-1), and as such I will be comparing my work and results to the results of other participants of the competition. As a result of there being other people working with the same data set that I have used for this project, a suitable goal of the project was to achieve the highest accuracy possible and compare this accuracy to other participant of the competition. This would be done through a thorough understanding of natural language processing, and experimentation with different forms of classifiers.

# Intended Audience

The intended audience of this project is anyone interested in natural language processing, but also anyone interested in the practical application of offensive speech detection online. This is especially beneficial in the current world, given the rise of hate speech online on popular platforms such as Twitter[[2]](#footnote-2). While this project was developed and tested on a specific set of tweets, with some modification it could be used to classify tweets in real time. Therefore, I would argue that the techniques and models explored in this project would be beneficial to anyone looking at implementing an automated hate speech detector into their website. Equally, this project would be beneficial to law enforcement agencies and governments looking to curb the rise in online hate speech.

# Project Scope

The scope of this project adapted over time, due to the fact that at the start of the project the libraries and concepts I was going to be working with were largely unknown to me. Therefore, it was hard to estimate the amount of time it would take me to complete pieces of work.

# Approach

The way that I approached carrying out the project was through regular meetings with my supervisor, coupled with a Kanban approach to working. This meant that I was focussed on small sections of deliverable work, that I could then discuss with my supervisor and plan out what to work on next. This also meant that the scope was able to remain flexible throughout the project, and I was able to pivot quickly in the event that circumstances changed, which they did at some parts of the project.

The way in which I ran the experiments that lead to my outcomes changed throughout my project as the data available to me changed, and as such can be split into two methods. Initially I did not have access to the test set that others who attempted this problem had, and as such I combined the two data sets I did have (train set and dev set), and then split this one set into multiple splits using a method known as “k-fold cross validation”. I was able to get access to the third data set during my project, and once I had this, I approached the project in the following way, which is considered best practice:

* Implement a method of converting the corpus into features.
* Select a classifier model.
* Provide the model with the train set and features.
* Predict whether the tweets in the dev set were offensive or not.
* Compare these predictions to actual results.
* Record results.
* Experiment with the classifier parameters and features.
* Repeat

[CHANGE THIS TO DIAGRAM?]

Once I had done this to a point where I was happy with each of the different

methods I had implemented, I then generated predictions for the test set, and compared these predictions to the actual results. These scores were then used for my final results.

# Summary of Outcomes

In terms of the important outcomes from the project, there were a number of interesting outcomes, as follow:

* Word embeddings based on Word2Vec were the most effective way of determining features for the models to use.
* SVC was the most effective model at classifying tweets from the corpus accurately.
* Your classifiers use word embeddings, but it’s not that word embeddings are for determining features for classifiers to use, they are general-purpose, semantically rich numerical representations of words.

[REWRITE THIS]

Background

# Project Context

The wider context of the problem of detection of abusive language towards women and immigrants is clear, as the rise of social media platforms and the anonymity that they provide has precipitated an increasing amount of online abuse, with one survey suggesting 18% of teenagers experienced abusive language while communicating online[[3]](#footnote-3). Given the frequency of abusive language use online, I would argue that there is a clear need for accurate automated detection methods of said language.

The main aim of this project is focussed around research and analysis into how best to approach the design and implementation of a program for detection of abusive language towards women and immigrants. I have chosen these two groups specifically due to the large amount of abuse they receive online[[4]](#footnote-4). In particular, immigrants have faced a rise in hate speech due to the refugee crisis[[5]](#footnote-5), and women have historically always suffered from hate speech[[6]](#footnote-6). The main project focus of detection of abusive language splits into a lot of sub problems, but the main 3 are feature extraction, model selection, and experiment setup. It is through working through these 3 problems that I was able to come up with varying different sets of results with varying accuracy.

Another aim is to produce a piece of software that given a tweet via the command line, will be able to accurately predict if the tweet is offensive or not towards the target groups. This was successfully developed using the research outcomes I found from the main aim of the project.

# Identified Problem

As identified in my initial plan, the pervasive nature of abuse towards women and immigrants’ online shows that this is a problem that needs solving. Although the focus of this project was experimentation and analysis, the outcomes and findings are useful in determining the best approach to developing an accurate automated method for detection of abuse towards women and immigrants. Using the findings of the project, I was able to implement a script that takes a single tweet as an input and returns a JSON blob containing an offensive/not offensive prediction for the tweet. This could be quickly added to an online platform to aid in hate speech detection.

# Likely Stakeholders

In terms of likely stakeholders in the problem area, those that would benefit from developing and implementing an automated system for the detection of abuse language would be social media sites and law enforcement agencies. There is a vested interest among social media site owners to stop and remove abusive language on their platforms, in order to make their platform more welcoming and nicer to use, as end users will stop using the site if they are receiving abuse. Equally, some of the abusive language that could be posted online could constitute hate speech, which is a crime in the United Kingdom[[7]](#footnote-7) as well as many other nations, and therefore it would be beneficial to law enforcement agencies if they could automatically detect the abusive language. This project would improve the first two steps in the hate crime prosecution process, namely reporting and investigation of said crime. By automatically detecting hate speech, platforms would be able to report the comments, police would be able to investigate said comments and report them to the Crown Prosecution Service as necessary. Prosecution could then take place as normal, but the previous two steps would be enacted faster and with less manual intervention, for example in the form of human moderators.

# Theory Associated with Problem Area

The focus of this project is around machine learning, of which the three main areas are models, features, and experiment setup. The majority of the work on this project involved experimenting with these three areas in different formats and combinations to try and achieve the best accuracy results possible.

Model selection involves making use of different classifier models for more accurate hate speech detection. There are two main differences in the models that I used, prediction accuracy and program run time. I made use of four different classifiers that I chose for varying reasons that I will elaborate on below.

The first model I used was a Decision Tree[[8]](#footnote-8). This model works by looking for a common word in the corpus that can split the data into offensive/not offensive tweets, and then repeating this step again with a new word. It is possible to set a maximum number of leaves for the tree, which will improve run time but potentially lower the accuracy of the model. I chose this model to begin with as it was a simple model to understand and implement, as it requires minimal data preparation to implement correctly.

The second model that I implemented was the Random Forest model[[9]](#footnote-9). This model is “a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting”. This was a logical follow on from using the Decision Tree classifier, due to the similarity in the way that they work.

The third model I implemented was Logistic Regression[[10]](#footnote-10). This model involves using a Logistic Function[[11]](#footnote-11) in order to model probability and generate a prediction. I chose this model due to the belief that it would be a good balance between slightly longer run time and improved prediction accuracy.

The fourth and final model that I implemented was Support Vector Classification[[12]](#footnote-12). This model works by plotting a line through the corpus matrix in order to makes its classifications. This model was chosen in the belief that it would provide the best accuracy, albeit in exchange for the longest run time.

Feature extraction encompasses the tokenization of the corpus, and the subsequent extraction of features that will be used to train the selected model. I ended up using three forms of feature extraction, a Count Vectorizer, TF/IDF extraction, and a Word2Vec model.

The Count Vectorizer was by far the least effective form of feature extraction, as it went through the corpus and selected the 50 most common words excluding English “stop words” such as “and, if, it, to” for example. Although this was not very effective in terms of accuracy, it was a quick and simple form of feature extraction to implement and was a good starting point for the project.

The Gensim Word2Vec model is a more effective method of feature extraction that involves grouping similar words together to produce word embeddings, that can then be tokenized and used as features to train a classifier model. I was given access to a Word2Vec model that was trained on a large corpus that produced a very effective set of word embeddings that I then tokenized and used to produce highly accurate results.

Term frequency–inverse document frequency or TF/IDF is the final form of feature extraction I implemented. This vectorizer works by putting weightings on features to determine how important they are to documents in the corpus. I made use of parameters such as “n grams” to specifically point the vectorizer to look for common phrases as well as words based on characters.

In terms of experiment setup, I went through varying setups involving differing forms of feature extraction and models together, with varying parameters. I made sure to record and export all of these results during the course of my work in order to catalogue the variation in results.

One of the major changes in experimental setup throughout my work on this project was the change from using “k-fold cross validation”[[13]](#footnote-13) to the best practice train/dev/test split. The best practice for work on machine learning algorithms involves training your model on the train data set, and then testing it on the dev data set, before implementing changes to the model or feature extraction to improve results. Only once you are satisfied with your model and feature extraction should you perform the final test on the test data set, to get your final results. Unfortunately, I did not have access to the test data set until halfway through the project, so I combined the train and dev data sets and implemented “k-fold cross validation”. This performed the train/dev/test splits for me, and I then combined the results to create an average.

# Constraints on Approach

As stated in my introduction, I was originally constrained from approaching the project in what is considered the best practice method for developing a models to classify offensive tweets due to not having access to the test data set that is used to generate final results, and ultimately compare these results to other teams who have worked on the same problem. This constraint was solved halfway through the project, and I was able to work in the best practice approach for the remainder of the project.

Due to this constraint, I did have to adopt an alternative approach for the initial period of my work on this project, involving using “k-fold cross validation” splits to alleviate the issue created by not having a test set. This proved to be a viable approach to the project, as the results and outcomes generated during this period of the project were essentially similar to when I approached the project with best practice in mind. The trends methods I found to get the best results were the same and could be adopted when I was given access to the test set.

# Existing Solutions

There were a large number of varying methods and research presented by others participating in the competition this project is based on. They range from machine learning methods similar to the work that I have to done, to deep learning. For example, there was use of Support Vector Machines, similar to what I have used. Although there was use of both machine learning and deep learning techniques, deep learning was by far the most popular method of approaching the problem with 70% of teams making use of deep learning[[14]](#footnote-14). Further to this, the most successful teams to participate made use of deep learning, with the highest ranked team to not use deep learning placing 6th.

As well as use of varying models and feature extraction, there was also a wide range of different methods of data pre-processing used. These methods ranged from making use of additional training material such as Word2Vec models and or external data sets, to using sentiment analysis models for prediction, or even using offensive word lists[[15]](#footnote-15). Some teams used twitter specific tokenizers, for example NTLK TweetTokenizer. Teams also made efforts to normalise URL’s, hashtags, and elongated words to improve results. There were also teams that converted emoji’s to text to be able to analyse them effectively.

# Methods and Tools

My project was developed using Python 3.7.6[[16]](#footnote-16), and in particular I made use of 3 specific libraries to complete the project: Sci-Kit Learn, Gensim, and Numpy. In terms of what makes up my program, each of my models is made up of a form of classifier (For example, a decision tree of Logistic Regression), and a form of feature extraction that helps to train the model with features that are deemed as important markers for if a tweet is offensive or not.

Sci-Kit Learn[[17]](#footnote-17) is a library that I used to give me access to a wide number of different classifier models that I could train my corpus on, and use to generate predictions for if a tweet is offensive or not. As well as this, at once stage I also used this library to extract features from the corpus as a further method of training the classifier model.

Gensim[[18]](#footnote-18) is a library that allowed me to implement a form of neural networks into the project, using a model of feature extraction called Word2Vec. Word2Vec works by creating word embeddings for each word in the corpus, and then looking at what words are “near” that word, to determine links. I implemented two forms of feature extraction in my project, with one Word2Vec model that was trained on my data set, and another Word2Vec model that was trained on a much larger data set of tweets. Both of these models were used in my work, and both produced varying results. I also used Word2Vec to analyse what words were “nearest” to certain words, to see what words are commonly used along with what could typically be considered as abusive language.

Numpy[[19]](#footnote-19) is a library that extends the operations that Python can perform on data structures, particularly arrays and matrices. This is useful for my project given that Sci-Kit Learn makes extensive use of sparse matrices to represent the predictions output by the classifier models, and by using Numpy it is much easier to record these results.

Specification and Design

# Approach to Solving the Problem

The exploratory nature of this project meant that my approach to the project changed over the duration of the work. As mentioned previously, there were changes in the experiment setup, methods of feature extraction, and classifier models. I have illustrated the way in which the program works in a flow chart that can be found in the supplementary materials under Figure 1. The reasoning behind why there was so much of an evolution in all areas of the project over the duration, is due to my extremely limited experience with the tooling and concepts used in the project prior to beginning work. As a result of this, a lot of the concepts and tooling I initially used were not best practice or did not provide very accurate predictions. The final iteration of the project, as shown in Figure 1, implements all of the concepts and tooling that I found to produce the most accurate predictions, as measured by precision, recall, and f-1 score.

As I have shown in Figure 1, the data flow of the program is relatively simple and remains the same despite variations in experimental setup, feature extraction, and model selection. However, Figure 1 does abstract a lot of the workings going on underneath the simple step descriptions I have included. For example, the step “Tweets are loaded into the program” abstracts the need to load the tweets and their corresponding labels into lists in such a way that Sci-Kit Learn can make use of the corpus.

Another flaw in Figure 1 is that it does not sure the evolution that this project went through over the course of my work on it. For example, although I was always comparing the predictions to the actual labels to determine the accuracy of the system, the way in which I calculated this accuracy did vary. In the final iteration of the program I had one set of predictions to compare to the actual tweet labels, but in earlier iterations I had 5 different sets of predictions. This was due to my use of “k-fold cross validation”, the use of which and use case I have previously explained. As a result of this, I had to average these 5 results together to get an overall result, thus complicating the step of result generation.

The way in which I recorded my results was to include a function that first noted down the parameters of the experiment to the results file, and then copied down the classification report containing metrics such as precision, recall, and f-1 score of the experiment. All of this information was recorded in a CSV, for two main reasons. First, because CSV files are to manipulate with Python, and second because they provide a neat and simple way of displaying my results in a tabulated format. As a result of recording my results automatically every time an experiment was run, I had an extensive catalogue of results that I have been able to go over and analyse, allowing me to accurately present a summary of outcomes as to what the most effective methods for detection of abusive language are.

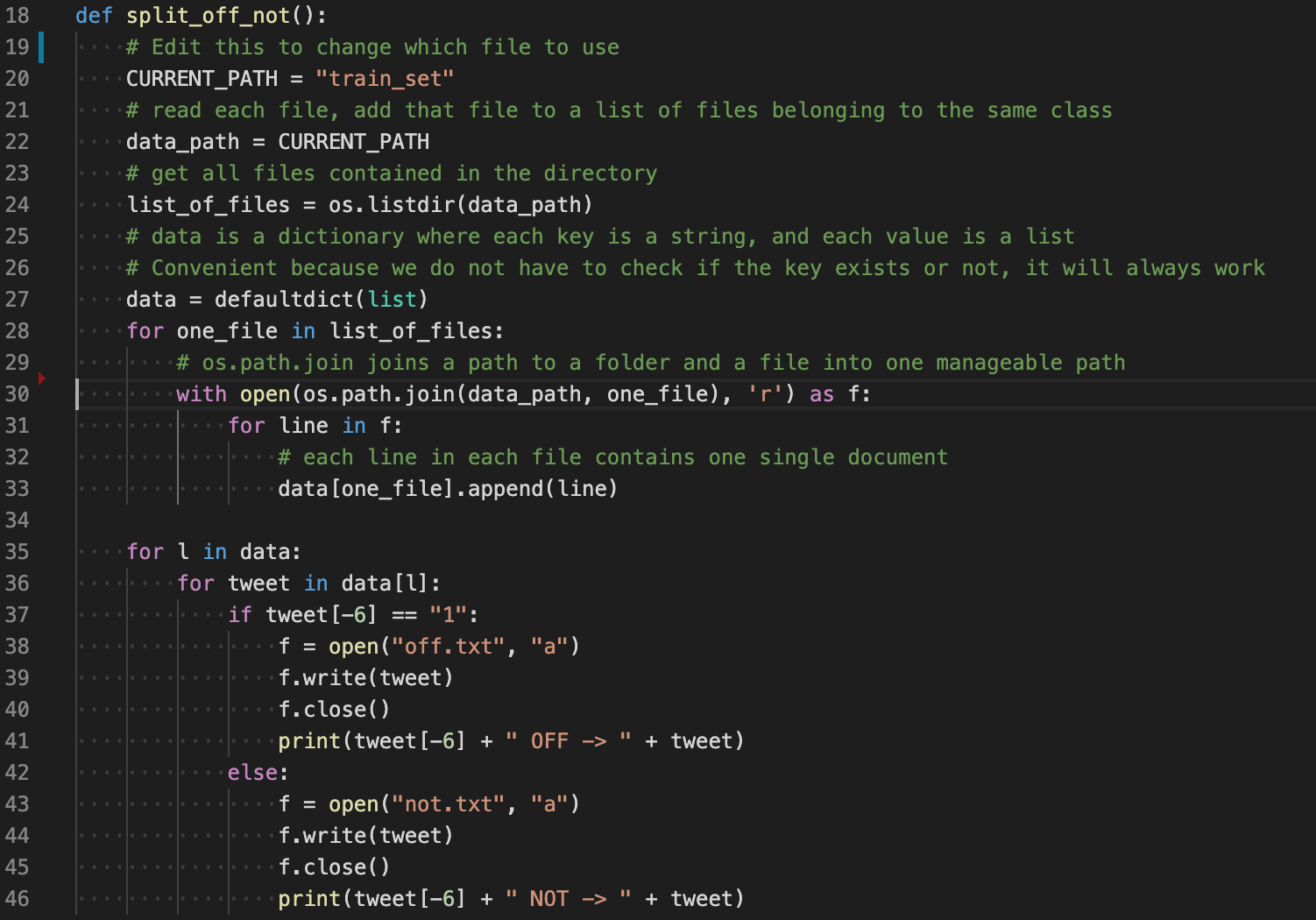
Another key step that is abstracted in Figure 1 under the “tweets are loaded into the program” step is data pre-processing. This step involves normalising the tweets so as to improve the overall accuracy of the results produced by the experiment. I performed varying levels of data pre-processing throughout the project as I learned more about natural language processing. In the final state of the project, the most effect form of pre-processing I used was to allow for the use of “n grams” to specifically look for common phrases instead of words. This vastly improved the accuracy of my results.

I have already previously touched on the variations in feature extraction and model selection, and I will go into depth about just how I implemented these varying methods that make up the core of the experimental setup.

Implementation

# Data Handling

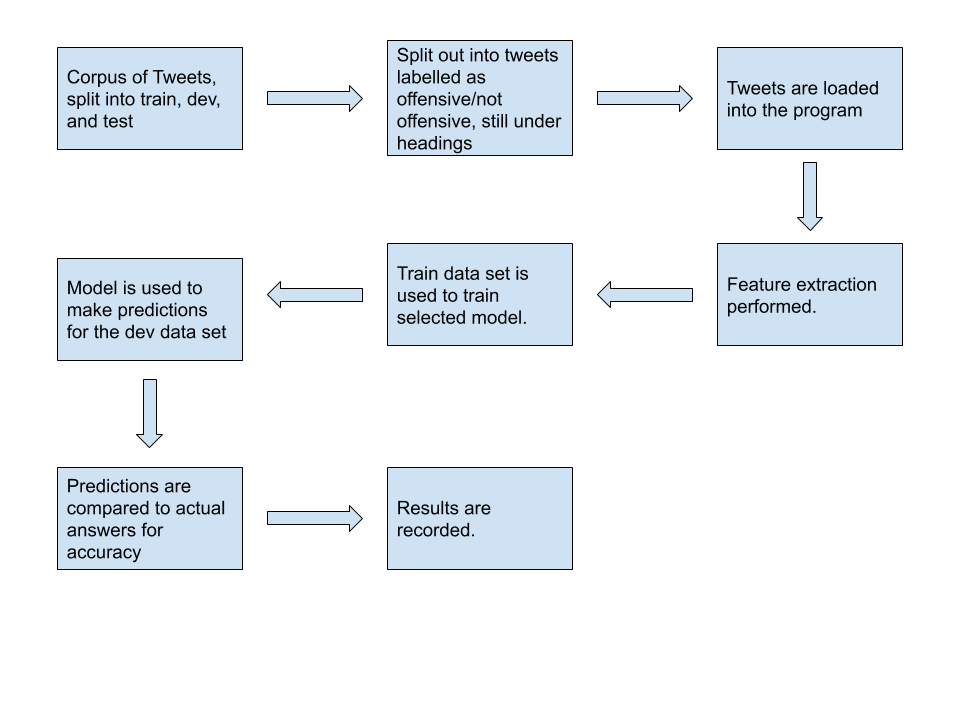
The first part of the project that I worked on was the handling of the data sets, as the libraries that I made use of require data to be formatted in a specific way before you can make use of the corpus to train the classifier models, or extract features from them. To do this in Python, I made use of the “OS”[[20]](#footnote-20) library as my method of file handling to load the tweets from the text files I had them stored in. I used this library due to the fact that it is built into Python, and as such is an effective method of file handling.

 Once the text files have been loaded, the next problem that I had to solve was the splitting of the training data, in this case the train data set, into separate offensive and not offensive text files, based on their labels within the original train set text file. The reason that this had to be done is to facilitate the creation of a Python list of tweets with corresponding labels that Sci-Kit Learn can use to train a classifier model, as well as compare predictions to actual tweet labels to determine the accuracy of said model. The screenshot below shows the function that I wrote to perform this task, along with comments explaining what the loading lines do. The step of splitting the tweets into their separate offensive/not offensive text file only needs to be done once, as the files are saved.

Following on from splitting the tweets out, I needed to load the separate offensive/not offensive files into the program. The way in which I did this essentially involved recycling the loading code shown in the “split\_off\_not” function shown above, as they essentially need to perform the same role of loading the tweets into a list, and because of this I will not show the “load\_file” function that performs this task.

Supplementary Materials

Figure 1



1. [https://competitions.codalab.org/competitions/20011](https://competitions.codalab.org/competitions/20011#results) [↑](#footnote-ref-1)
2. Banks, J., 2010. Regulating hate speech online. *International Review of Law, Computers & Technology*, *24*(3), pp.233-239. [↑](#footnote-ref-2)
3. Kawate, S. and Patil, K., 2017. Analysis of foul language usage in social media text conversation. *International Journal of Social Media and Interactive Learning Environments*, *5*(3), pp.227-251. [↑](#footnote-ref-3)
4. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P. and Sanguinetti, M., 2019, June. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 54-63). [↑](#footnote-ref-4)
5. Bosco, C., Patti, V., Bogetti, M., Conoscenti, M., Ruffo, G.F., Schifanella, R. and Stranisci, M., 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *SYMPOSIUM III. SOCIAL INTERACTIONS IN COMPLEX INTELLIGENT SYSTEMS (SICIS) at AISB 2017* (pp. 79-84). AISB. [↑](#footnote-ref-5)
6. Cresti, M., Martino, V. and Rosola, M., Kate Manne. Down Girl. The Logic of Misogyny, Oxford University Press, 2017, pp. XXIV, 338. *APhEx*. [↑](#footnote-ref-6)
7. <https://www.cps.gov.uk/hate-crime> [↑](#footnote-ref-7)
8. <https://scikit-learn.org/stable/modules/tree.html> [↑](#footnote-ref-8)
9. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [↑](#footnote-ref-9)
10. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> [↑](#footnote-ref-10)
11. <https://en.wikipedia.org/wiki/Logistic_function> [↑](#footnote-ref-11)
12. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> [↑](#footnote-ref-12)
13. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html> [↑](#footnote-ref-13)
14. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R., 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*. [↑](#footnote-ref-14)
15. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R., 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*. [↑](#footnote-ref-15)
16. <https://www.python.org/> [↑](#footnote-ref-16)
17. <https://scikit-learn.org/> [↑](#footnote-ref-17)
18. <https://radimrehurek.com/gensim/> [↑](#footnote-ref-18)
19. <https://numpy.org/> [↑](#footnote-ref-19)
20. <https://docs.python.org/3/library/os.html> [↑](#footnote-ref-20)