DS-GA 1017: Responsible Data Science

Homework 2 Submission

Era Sarda
New York University
es6790@nyu.edu

March 2025

1 AI Ethics: Global Perspectives

Content Moderation in Social Media and AI, *By Serge Abiteboul*

Social media has brought a massive change to society, with platforms like Facebook (3.1 billion users), and TikTok (1.04 billion users) shaping how people communicate and consume information. No other technology has transformed society as significantly since the invention of the printing press. However, with great power comes great responsibility. While social media offers benefits such as the free flow of knowledge, increased awareness, removal of geographical boundaries, and entertainment, it also brings serious concerns, including fake news, incitement to violence, online harassment, and hate speech. According to a study cited by the *Times of India*, social media accounts for 85% of misinformation, making content moderation and regulation by relevant authorities a necessity.

Social media platforms function similarly to states, with a triptych of Territory (jurisdiction of use), Population (users), and Authority (platform executives). Relying solely on human moderation is impractical now due to the large volume and variety of data across languages and dialects. A news report suggested that "Facebook moderators in India are traumatized." Immediate intervention is often required, making ML-based content moderation another necessity.

Algorithms often outperform humans in detecting terrorism-related content, child exploitation, and hate speech. Harassment can also be detected more effectively using message graphs than a single message. However, they have limitations:

- Lack of judgment in non-text scenarios

- Privacy violations and uninformed data usage.
- Biases against protected groups due to unrepresentative training data.
- Cultural and legal differences in defining "inappropriate" content. Moreover, AI systems used for content recommendation may inadvertently propagate harmful content rather than eliminating it.

Users: These platforms rely on user data to fuel algorithms. Issues like data leakage, excessive tracking, and aggressive permissions access raise privacy concerns. If ML models exhibit biases against certain groups, users' well-being may suffer, or fuel communal tensions, as seen in misinformation-triggered violence.

Government: The propagation of fake news, hate speech, and online abuse can destabilize a state. Users are themselves biased, and often believe the information presented to them. For example, during the COVID-19 pandemic in Indore, false WhatsApp messages about vaccines incited violence against healthcare workers. However, social media also allows whistleblowers to expose corruption, making governance more transparent.

Regulators: Those responsible for overseeing social media—including governments, companies, or third parties—face challenges in striking a balance. If governments have full transparency, they could manipulate public opinion using AI-driven recommendations, especially during elections. If platforms self-regulate, they might prioritize profit over responsible moderation.

Others: Academics, think tanks, and human rights groups contribute to AI Safety research, legal actions, and audits but lack enforcement authority.

Transparency & Interpretability: Social media platforms use proprietary, black-box algorithms, making their operations opaque to users and external regulators. Users often consent to complex 'data enroachment' agreements without full comprehension. Personal information, passwords, and sensitive documents may be unknowingly exploited for training algorithms. Social media feeds are designed to maximize user engagement, often leading to addiction and mental health issues. Researchers and advocates also face black-box challenges, as the lack of transparency hinders independent audits and accountability. While governments could impose stricter regulations, the hidden nature of these algorithms makes comprehensive oversight difficult.

Company Incentive: Content moderation is cost, not a revenue stream. However, they are incentivized by user trust, legitimacy, legal compliance, and public perception. Regulatory noncompliance could lead to shutdowns, fines, or decreased ad auction prizes. Their revenue primarily comes from advertising, which requires a stable and moderated platform. Thus, platforms must balance responsible AI practices with profitability.

In conclusion, AI-driven moderation is essential for managing vast amounts of content but presents challenges in privacy, bias, and transparency. Achieving equilibrium between corporate interests, regulatory oversight, and public trust remains an ongoing challenge.

2 Data science lifecycle

- (a)

  White as well as Other, Female and Non-Binary groups, might be disadvantaged by Alex's imputation method. This is because the overall mean experience value used to replace missing data is lower than the average experience values within these groups. Consequently, the percentage change in their group means will be greater, as indicated in Figure 1.

  (However, any individual—regardless of gender or group—who has an experience level higher than the overall mean could also experience a negative impact. This is because replacing missing values with the mean would fail to capture their higher-than-average experience, potentially under-representing their qualifications.)

```
def f(mean, null, total):
  new = mean*(total-null)+6.12*null
  newmean=new/total
  ret=((newmean-mean)/mean)*100
  return ret
```

```
f(5.70,80,2001)
```

```
0.29458954733159726
```

```
f(5.66,63,1065)
```

```
0.48076444532922213
```

```
f(7.40,60,634)
```

```
-1.636968198482391
```

```
f(7.91,42,300)
```

```
-3.1681415929203474
```

Figure 1:

- (b) Replacing missing values (NULL) in the experience feature with the group's mean value for that feature in the dataset.

- (c) Pre-existing: Applicants from female and non-binary groups, who statistically have higher experience levels on average (as per the dataset), may see their qualifications diminished

because the overall mean is lower than their group-specific averages. This reinforces the disadvantage caused by pre-existing bias against them.

Emergent: The imputation method creates an emergent bias by favoring applicants closer to the mean experience level while devaluing those with higher-than-average experience. This devalues outliers, since the number of male applicants is much higher than female and non-binary applicants.

3 Shades of NULL

Simulating 50 percent missingness in every case.

MCAR: Missing Completely at Random
MCAR holds if data is missing due to administrative errors, unrelated to the any variables itself.

MAR: Missing at Random
If the disadvantaged group are more likely to withhold their law school performance, and this can be explained by observed covariates (i.e., race). Here, missingness depends only on observed features, not the missing values themselves.
condition=('race','Non-White')
cols = ['zfygpa', 'zgpa'] - First year GPA, Overall GPA after 4 years.

MNAR: Missing Not at Random
Consider a student whose academic performance depends by different law schools (i.e. different schools have different grading standards)—not captured in the data—rather than by race. Suppose students with lower academic performance are more likely to withhold this information. In this case, missingness is correlated with the missing value itself and can not be explained by observed covariates (i.e., race).
condition=(col,{'lt': mnar_df[col].median()})
If less than column median, then delete randomly 50 percent of such values.
cols = ['zfygpa', 'zgpa'] - First year GPA, Overall GPA after 4 years.

Comparison based on Fairness Metrics - FNR_Diff, FPR_Diff, Demo_Parity_Ratio, Eq_Odds_Ratio, Select_Rate_Diff

Depending on the context it may be better to focus on FNR diff (FNRD), Demo parity Ratio (DPR) and Selection rate diff (SRD). **Focus Metrics := [FNRD, DPR, SRD]**

In case of MCAR, for our focus metrics *Drop nulls (DN)* is although comparable, but more fair than median imputation (MI). However for FPRD and EOR, MI is fairer.

MAR: For the focus metrics, both DN and MI methods perform comparable. However, MI is fairer w.r.t. FPRD and EOR. Making *MI* better.

MNAR: For our focus metrics *Drop nulls (DN)* is much fair than median imputation (MI). However for FPRD and EOR, MI is slightly fairer.

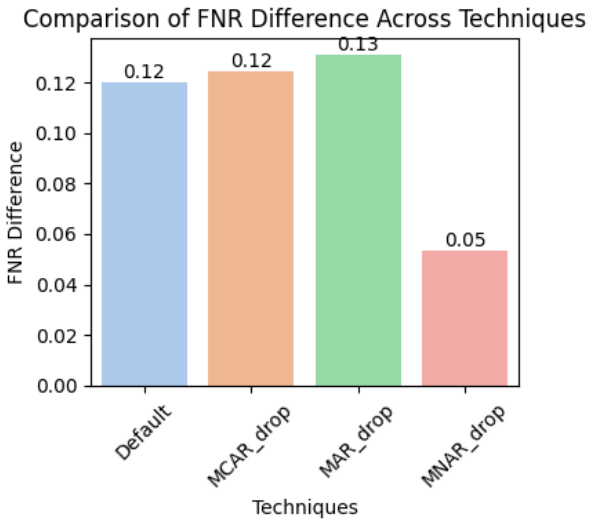

Figure 2: Drop Nulls

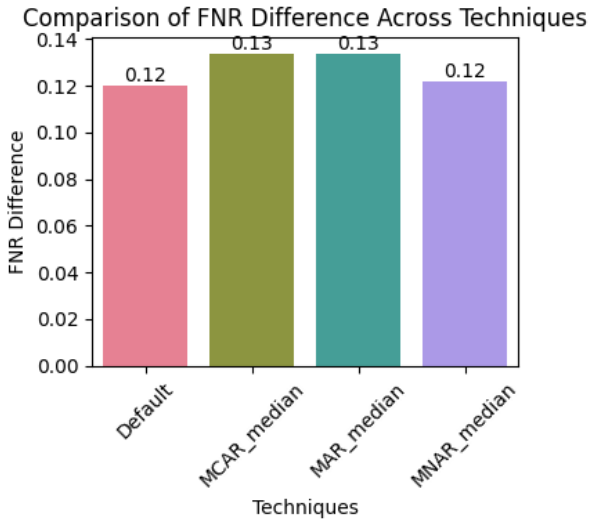

Figure 3: Median Imputation

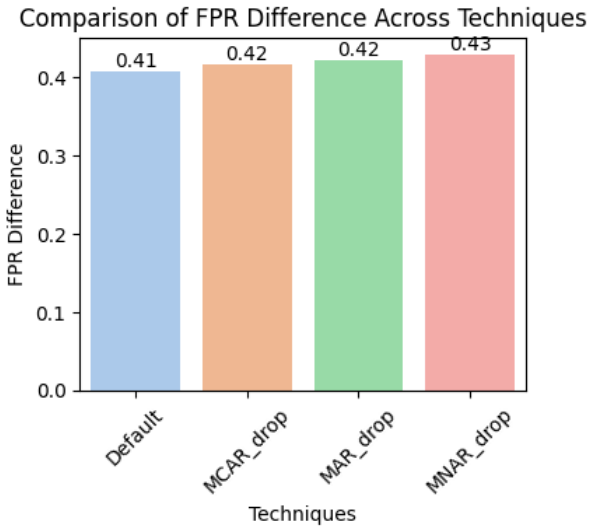

Figure 4: Drop Nulls

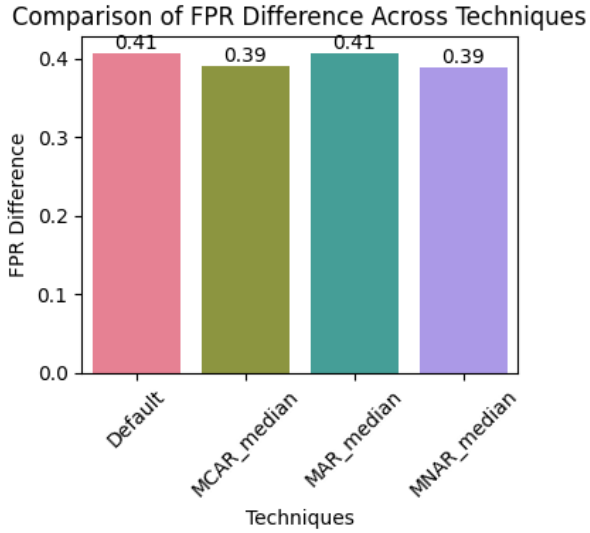

Figure 5: Median Imputation

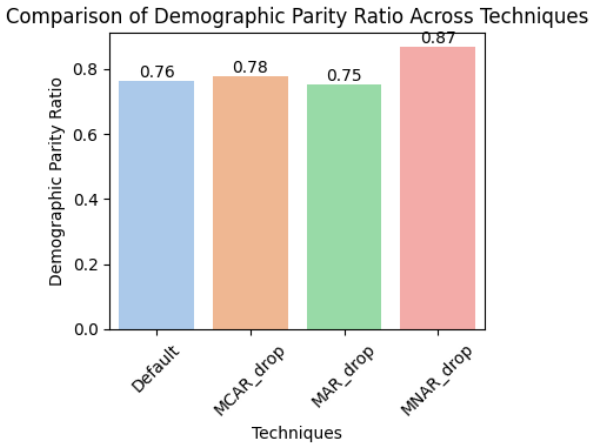

Figure 6: Drop Nulls

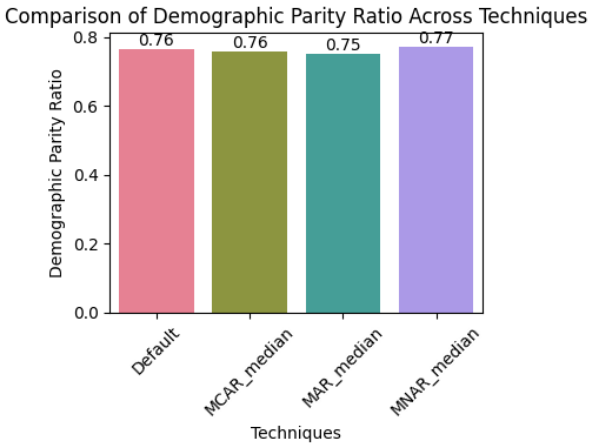

Figure 7: Median Imputation

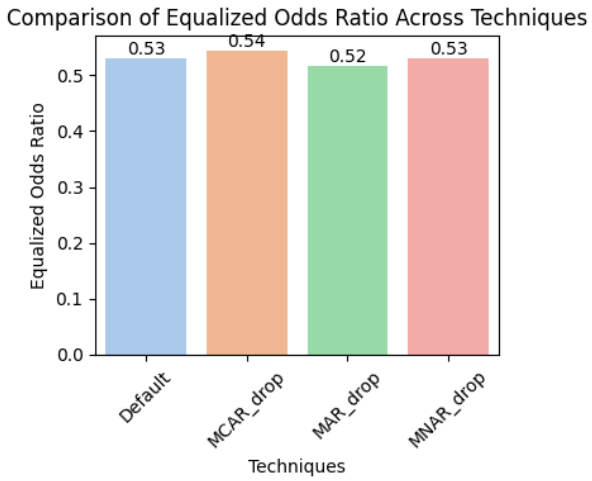

Figure 8: Drop Nulls

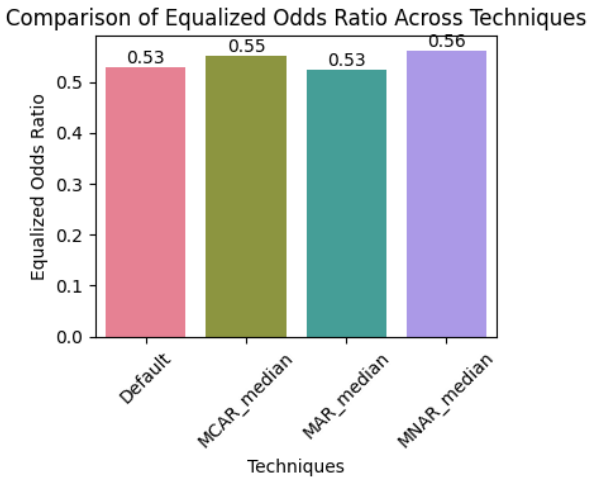

Figure 9: Median Imputation

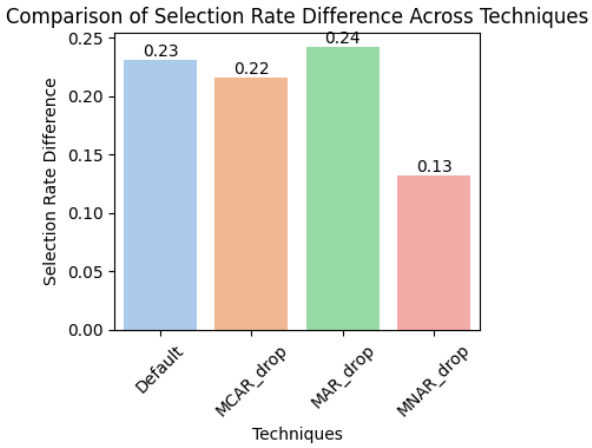

Figure 10: Drop Nulls

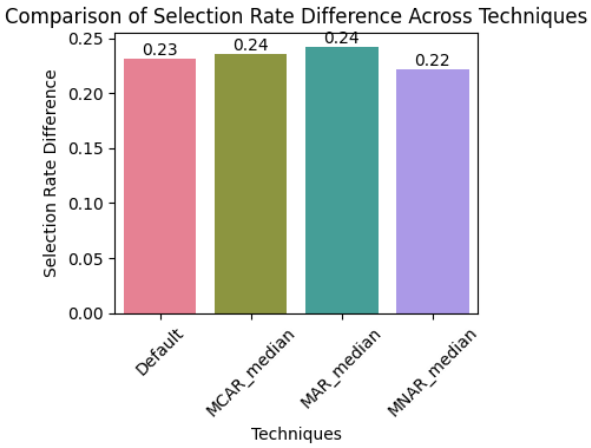

Figure 11: Median Imputation

Comparison by Group - White & Non-White - Accuracy, Precision, Recall, FNR, FPR

Since the difference in rates have already been considered, the focus here will be more on the absolute values of metrics for White and Non-White Population. **Focus Metrics = [Accuracy, Recall, FPR]**

MCAR: For white population, *MI* performs better in all metrics. For non-whites, *MI* performs slightly better for the focus metrics.

MAR: For white population, *MI* performs better in all metrics. For non-whites, altough MI and DN are very comparable, but *MI* performs slightly better for accuracy.

MNAR: For white population, *DN* performs better in all metrics except FPR. For white population, *DN* performs much better in all metrics except FPR.

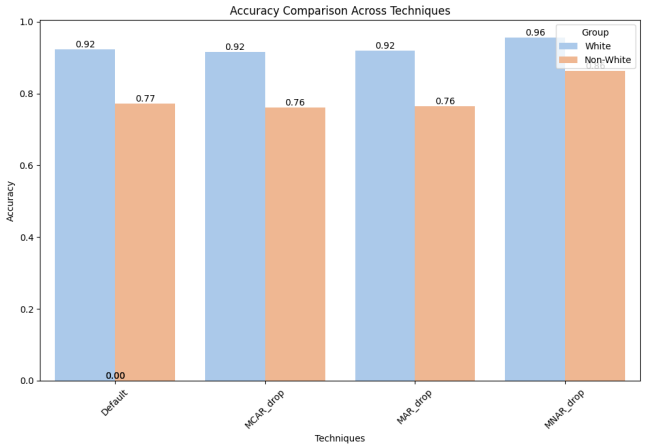

Figure 12: Drop Nulls

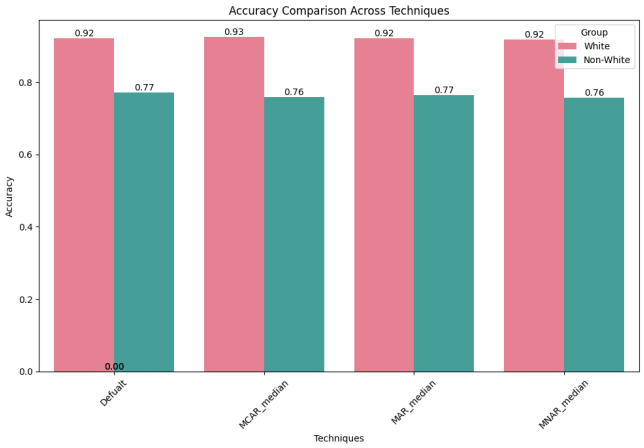

Figure 13: Median Imputation

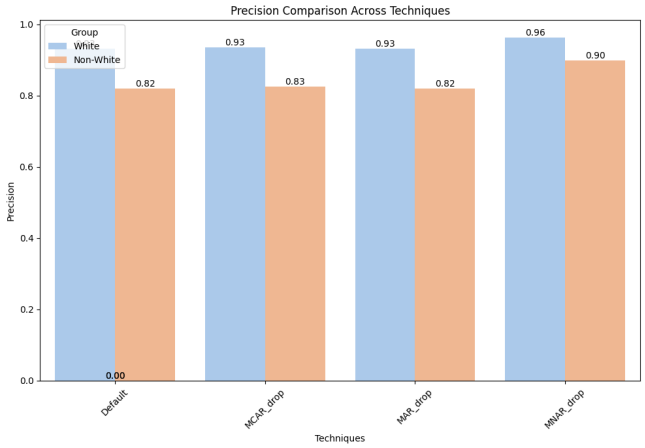

Figure 14: Drop Nulls

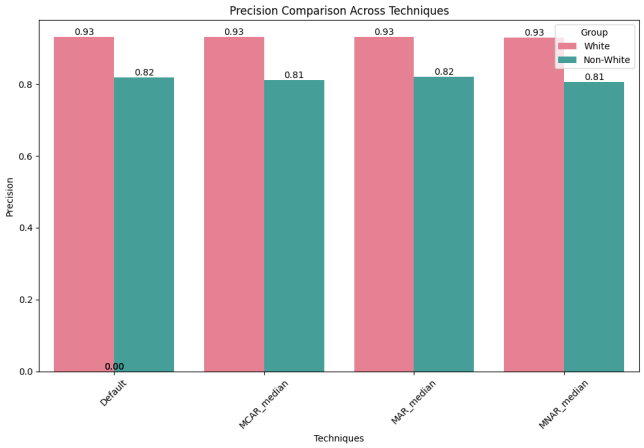

Figure 15: Median Imputation

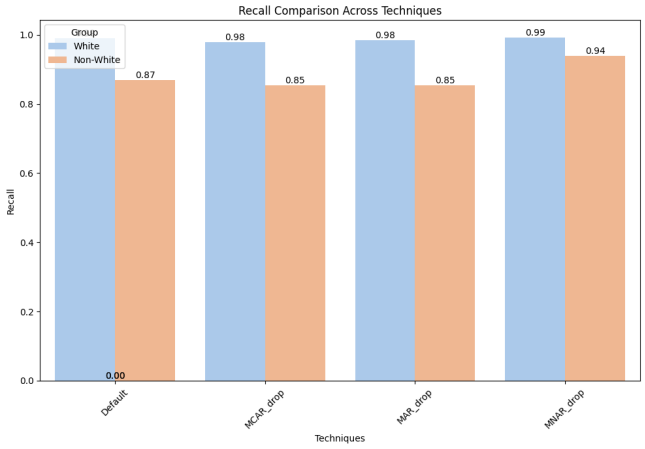

Figure 16: Drop Nulls

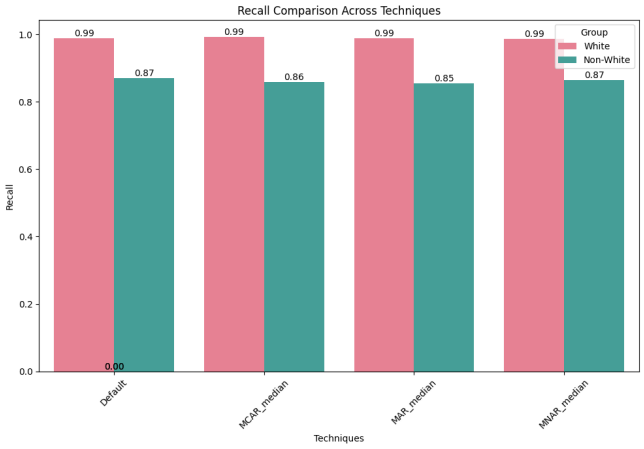

Figure 17: Median Imputation

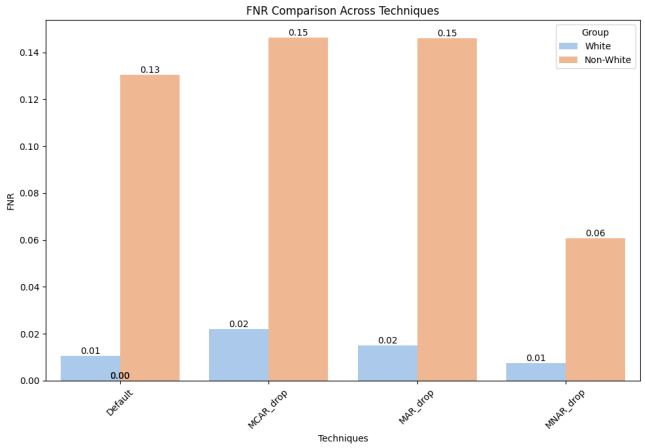

Figure 18: Drop Nulls

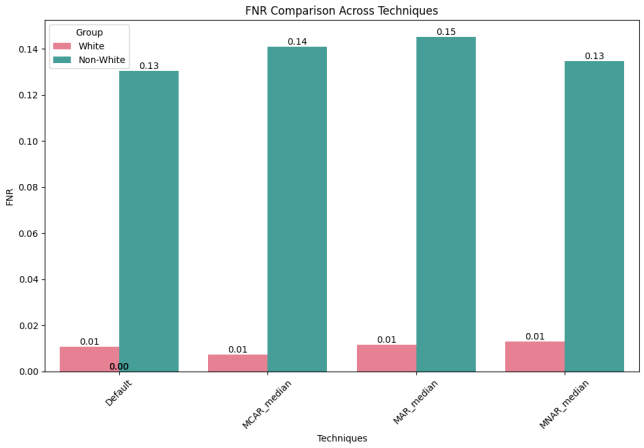

Figure 19: Median Imputation

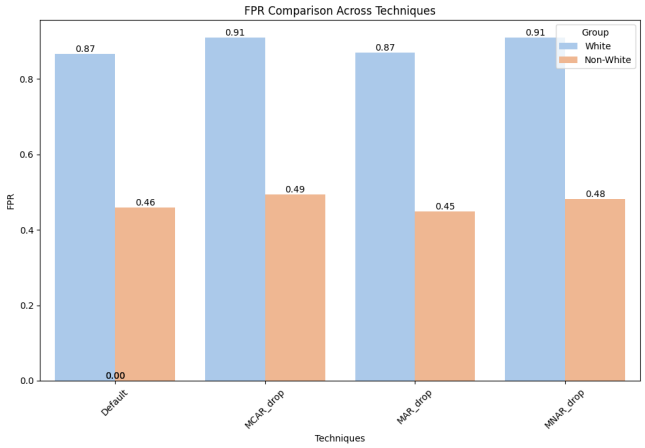

Figure 20: Drop Nulls

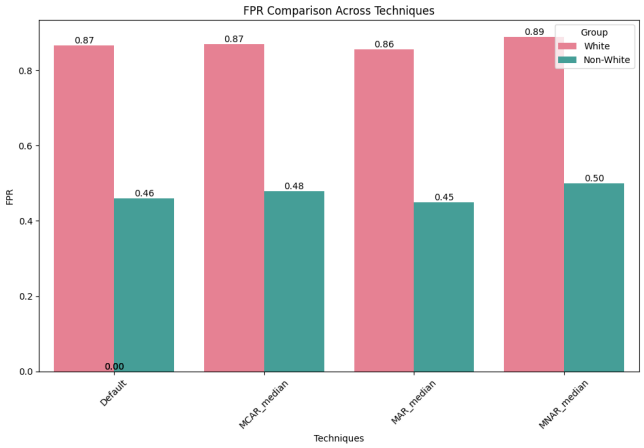

Figure 21: Median Imputation

Overall Comparison - Accuracy, Precision, Recall, FNR, FPR

pass_bar values = { '0' : 2293, '1' :18505} - imbalanced.

Accuracy: Useful if overall correctness is critical and/or the dataset is balanced (i.e., has roughly equal positive and negative samples).

Precision: When want to minimize false positives.

Recall & FNR: Prioritized if detecting all positives is crucial, even if some false positives occur.

False Positive Rate (FPR): Important when false alarms are costly or disruptive.

Therefore **Focus Metrics := [FPR, Recall, Accuracy]**

MCAR: For all the metrics *median imputation (MI)* performs better than Drop nulls (DN).

MAR: For the focus metrics and precision *DN* performs slightly better than Drop nulls (MI). However, FPR performance is better for MI.

MNAR: For the focus metrics and precision *DN* performs better than Drop nulls (MI). However, FPR performance is better for MI.

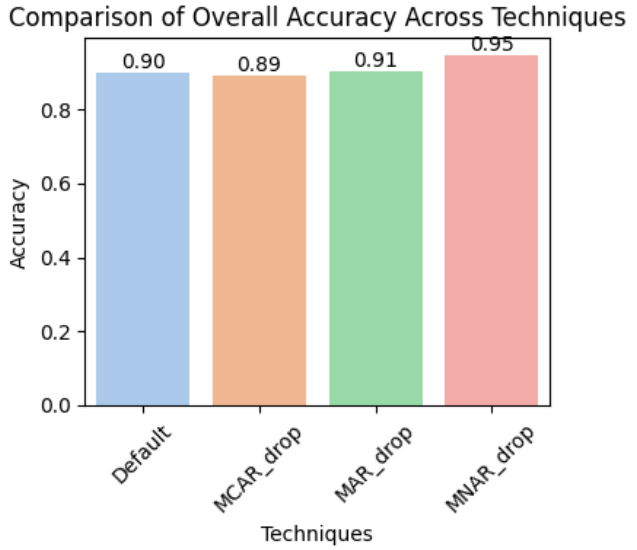

Figure 22: Drop Nulls

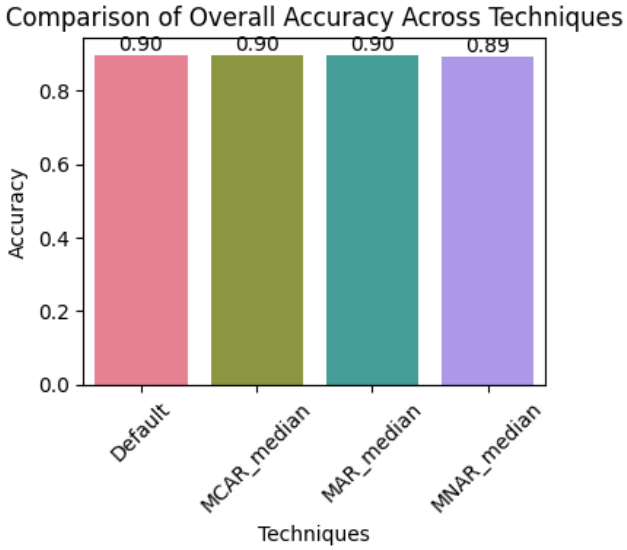

Figure 23: Median Imputation

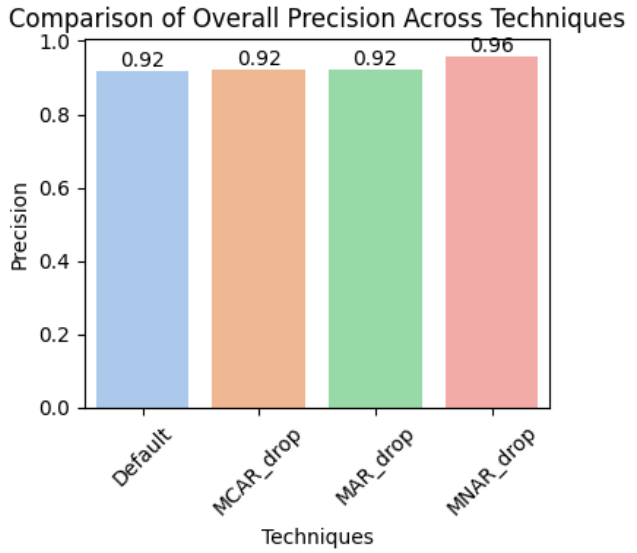

Figure 24: Drop Nulls

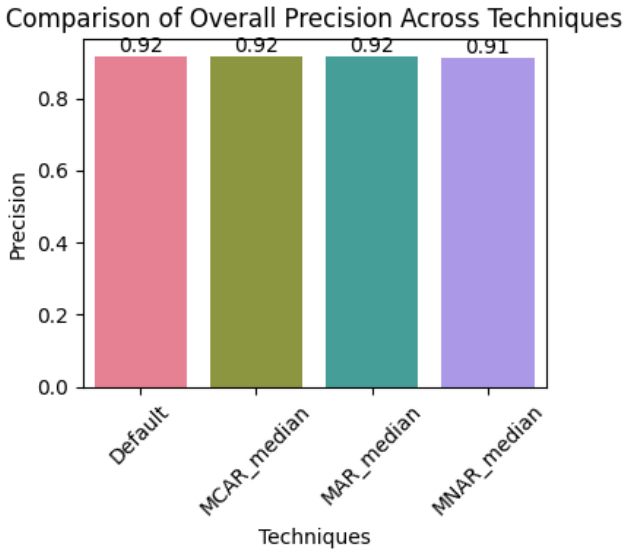

Figure 25: Median Imputation

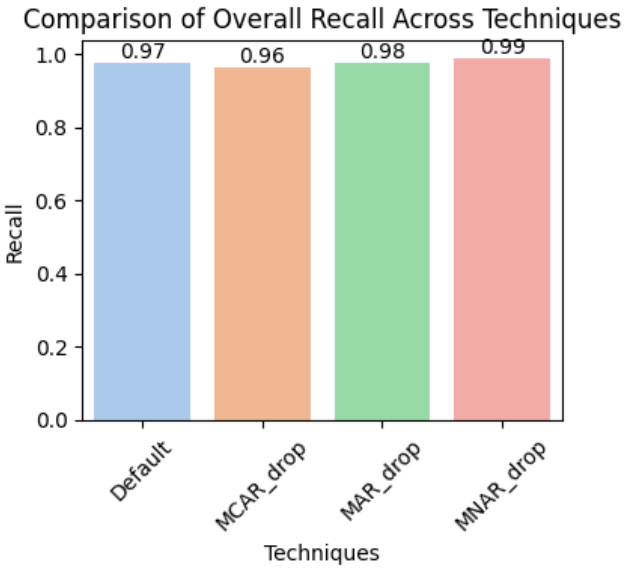

Figure 26: Drop Nulls

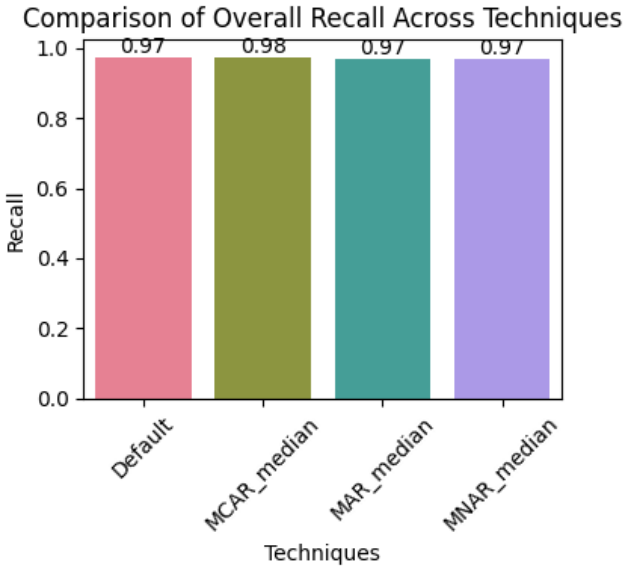

Figure 27: Median Imputation

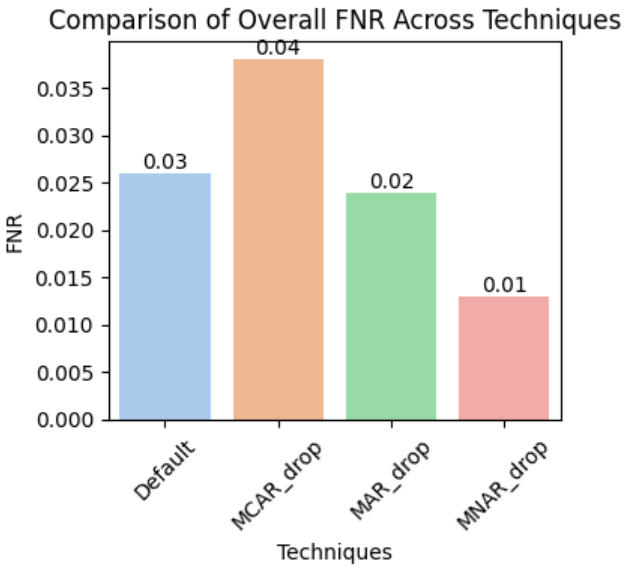

Figure 28: Drop Nulls

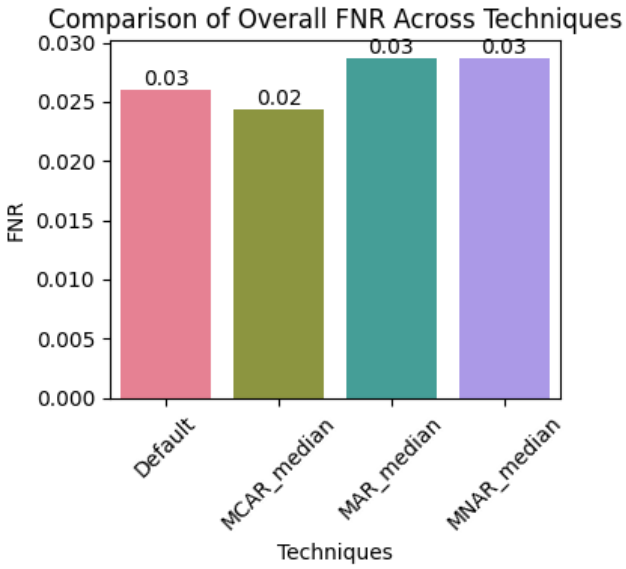

Figure 29: Median Imputation

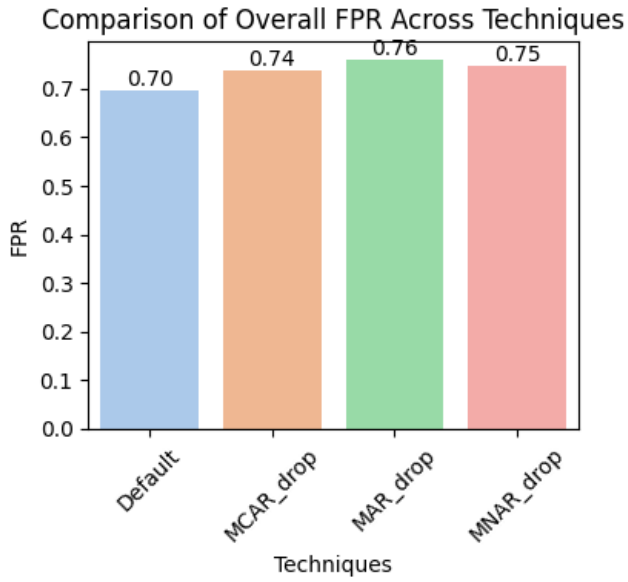

Figure 30: Drop Nulls

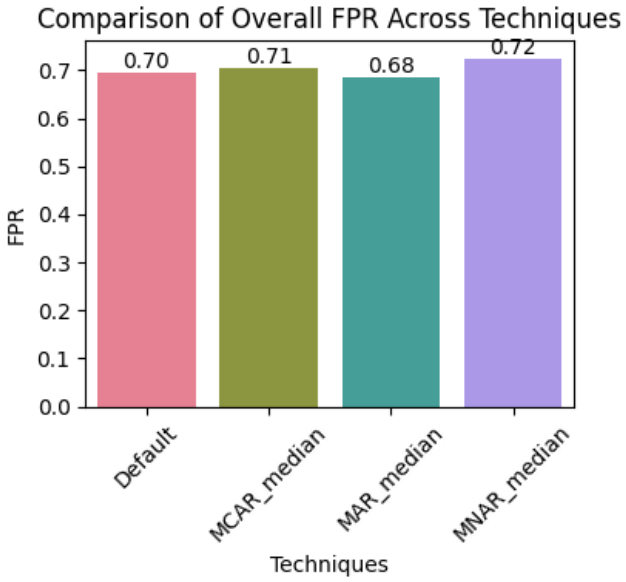

Figure 31: Median Imputation

Final Note:

- For MCAR, Drop Nulls performed fairer, but still was comparable to Median Imputation method. And for other comparisons - by group and overall, MI performed better. Making Median Imputation better method for MCAR.
- For MCAR, Drop Nulls performed very slightly better in focus metrics, but still was comparable to Median Imputation method. And for other comparisons - fairness and by group, MI performed better. Making Median Imputation better method for MAR too.
- For MNAR, Drop Nulls performed better in all comparisons, fairness, by group and overall.

Extra Credit Question:

Yes changing the parameters above causes slight variations in the fairness metric values, with changes ranging from 1-10 percent for each metric (accuracy, precision, recall, FNR, and FPR) for both White and Non-White groups. However, these changes do not significantly affect fairness. The model appears to be biased against the Non-White population, regardless of the seed or split used. Specifically, the model shows higher accuracy, precision, recall, and FPR for White individuals, while Non-White individuals experience higher FNR. This implies that adjusting these hyperparameters does not significantly reduce the model's bias. See figure 32.

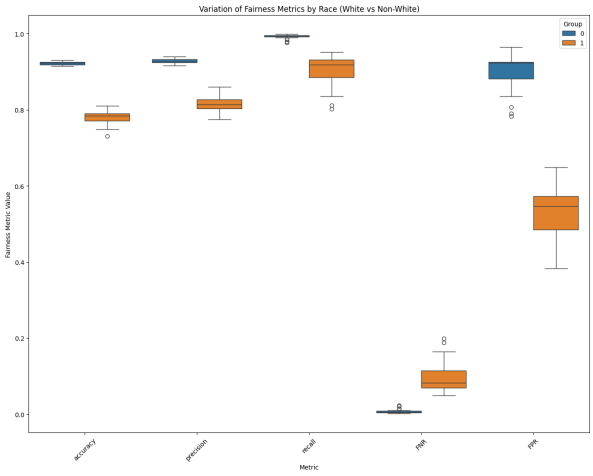

Figure 32: Varied Random seed and Train_Test Split

4 Explaining text classification with SHAP

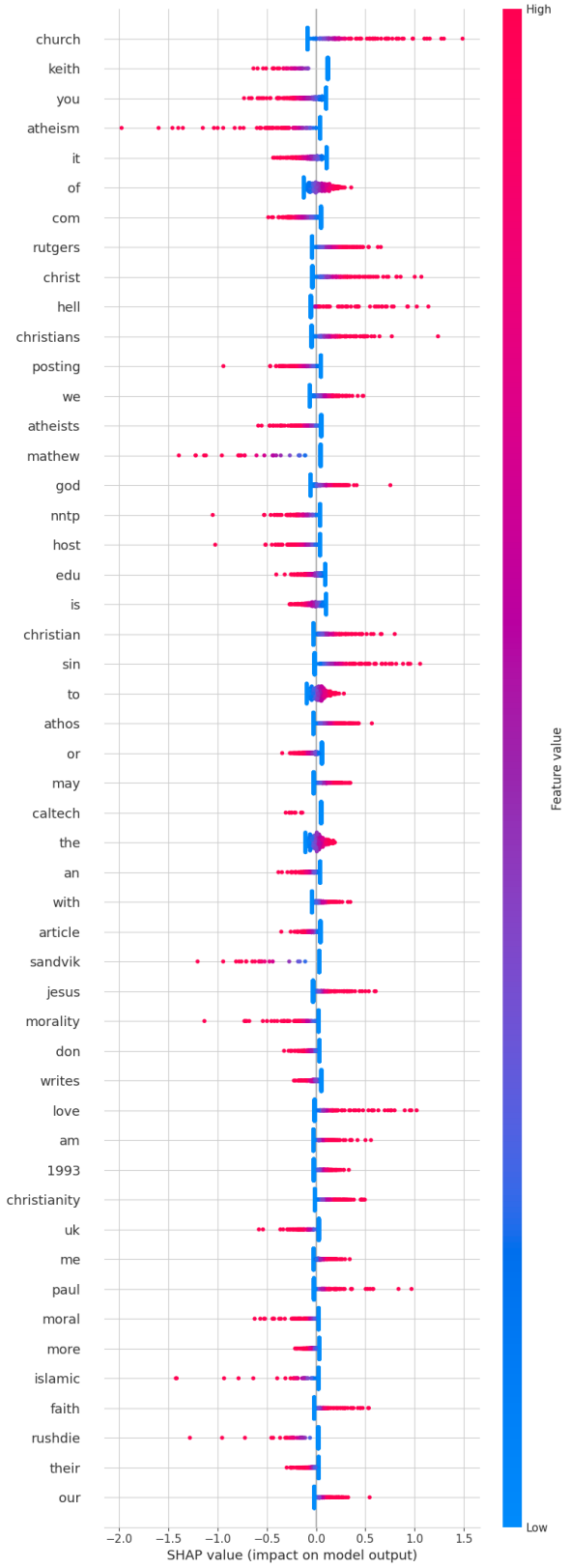

Figure 33: SHAP Values Summary Plot

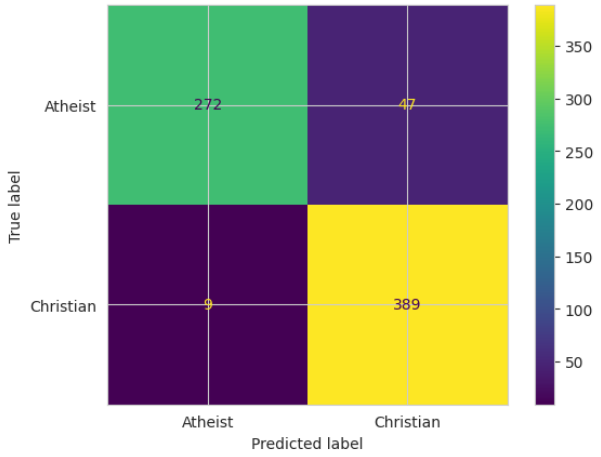

Figure 34: Confusion Matrix - 56 Incorrectly Classified

Visual explanations for 5 documents in the test set:

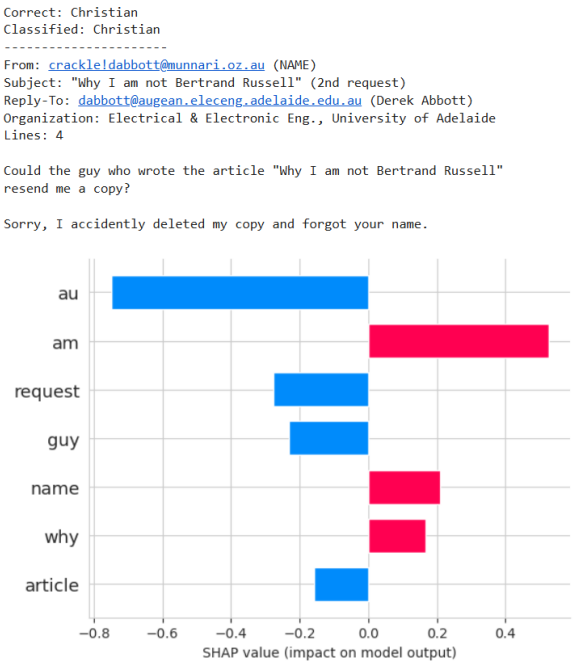

Figure 35: Correctly Classifiedidx-0

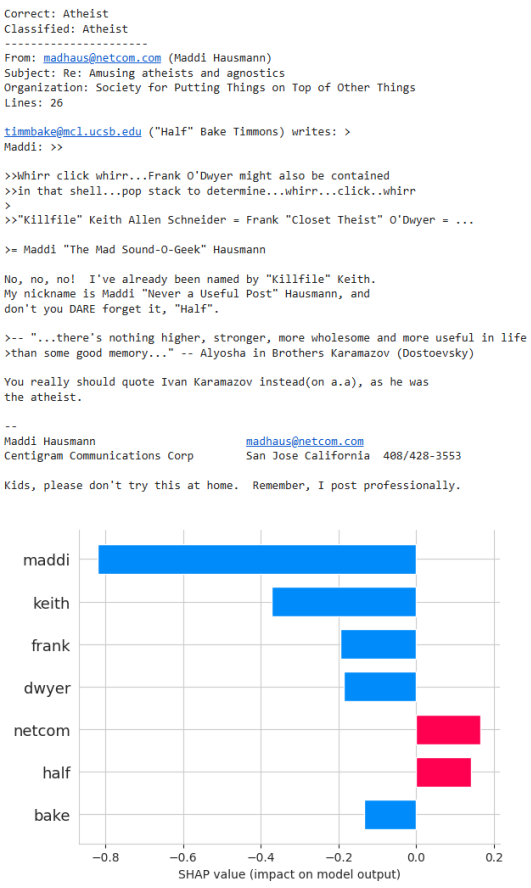

Figure 36: Correctly Classified -idx2

Correct: Christian
Classified: Atheist

From: tdarcos@access.digex.net (Paul Robinson)
Subject: Homosexuality is Immoral (non-religious argument)
Organization: Tansin A. Darcos & Company, Silver Spring, MD USA
Lines: 43

[This was crossposted to a zillion groups. I don't intend to
carry an entire discussion crossposted from alt.sex, particularly
one whose motivation seems to be having a fun argument. However
I thought readers might be interested to know about the
discussion there. --clh]

I intend to endeavor to make the argument that homosexuality is an
immoral practice or lifestyle or whatever you call it. I intend to
show that there is a basis for a rational declaration of this
statement. I intend to also show that such a declaration can be
made without there being a religious justification for morality,
in fact to show that such a standard can be made if one is an atheist.

Anyone who wants to join in on the fun in taking the other side,
i.e. that they can make the claim that homosexuality is not immoral,
or that, collaterally, it is a morally valid practice, is free to do
so. I think there are a lot of people who don't believe one can have
a rational based morality without having a religion attached to it.

This should be fun to try and figure this out, and I want to try and
expose (no pun intended) my ideas and see other people's and see where
their ideas are standing. As I'm not sure what groups would be interested
in this discussion, I will be posting an announcement of it to several,
and if someone thinks of appropriate groups, let me know.

If someone on here doesn't receive alt.sex, let me know and I'll make
an exception to my usual policy and set up a mailing list to automatically
distribute it in digest format to anyone who wants to receive it as I'll
use that as the main forum for this. By "exception to usual policy" is
that I normally charge for this, but for the duration the service will be
available at no charge to anyone who has an address reachable on Internet
or Bitnet.

I decided to start this dialog when I realized there was a much larger
audience on usenet / internet than on the smaller BBS networks.

To give the other side time to work up to a screaming anger, this will
begin on Monday, May 24, to give people who want to make the response
time to identify themselves. Anonymous postings are acceptable, since
some people may not wish to identify themselves. Also, if someone else
wants to get in on my side, they are free to do so.

This should be *much* more interesting than Abortion debates!

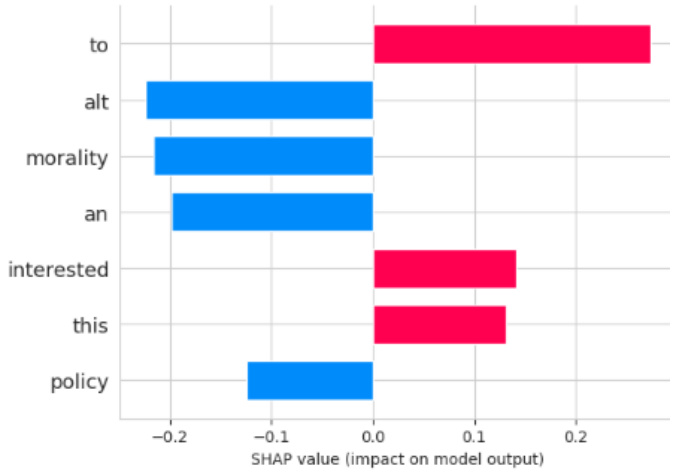

Figure 37: Incorrectly Classified -idx1

```
Correct: Atheist
Classified: Christian
----------------------
From: aaron@minster.york.ac.uk
Subject: Re: Gulf War / Selling Arms
Distribution: world
Organization: Department of Computer Science, University of York, England
Lines: 14

Mark McCullough (mccullou@snake10.cs.wisc.edu) wrote:
: I heard about the arms sale to Saudi Arabia.  Now, how is it such a grave
: mistake to sell Saudi Arabia weapons?  Or are you claiming that we shouldn't
: sell any weapons to other countries?  Straightforward answer please.

Saudi Arabia is an oppressive regime that has been recently interfering
in the politcs of newly renunified Yemen, including assasinations and
border incursions. It is entirely possible that they will soon invade.
Unluckily for Yemen it is not popular in the West as they managed to put
aside political differences during reunification and thus the West has
effectively lost one half (North?) as a client state.

                Aaron Turner
```

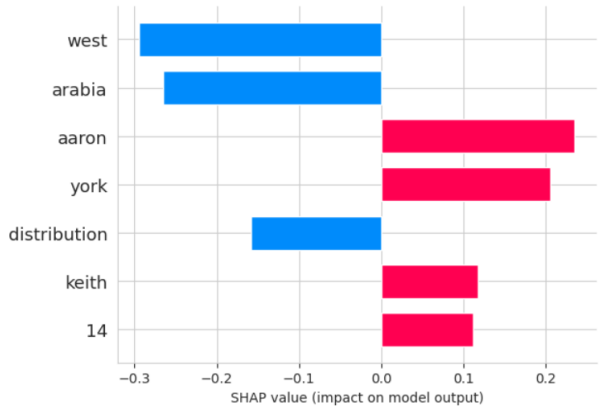

Figure 38: Incorrectly Classified -idx25

```
Correct: Christian
Classified: Atheist
----------------------
From: biz@soil.princeton.edu (Dave Bisignano)
Subject: Re: Why do people become atheists?
Reply-To: biz@soil.princeton.edu
Organization: Princeton University
Lines: 10

Ken,
Then what happens when you die?
Why are you here?
What is the purpose of Your life, do you think it's
just by chance you're in the family you are in and have the
friends you have?
Why do you think your searching?  To fill the void that
exists in your life.  Who do you think can fill that void

--Dave--
```

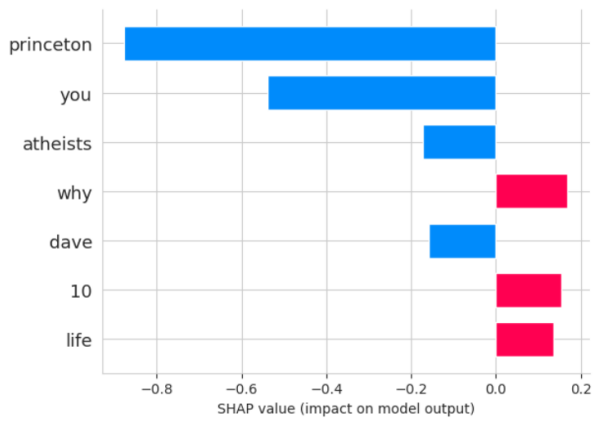

Figure 39: Incorrectly Classified -idx62

Accuarcy of Classifier: 0.9218967921896792
Number of misclassified documents: 56
Mean value of actual labels: 0.5550906555090656
For a baseline, about 50 percent of the test data imply Christian documents. Our model is pulling out meaningful structure by showing 92 percent accuracy!

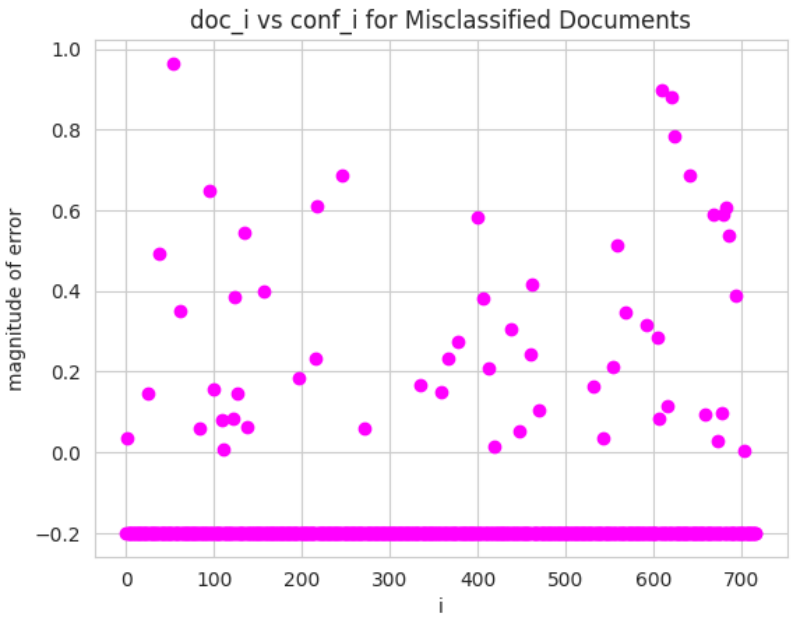

Figure 40: Doc_i vs Conf_i

-0.2 is proxy error for correctly classified documents.

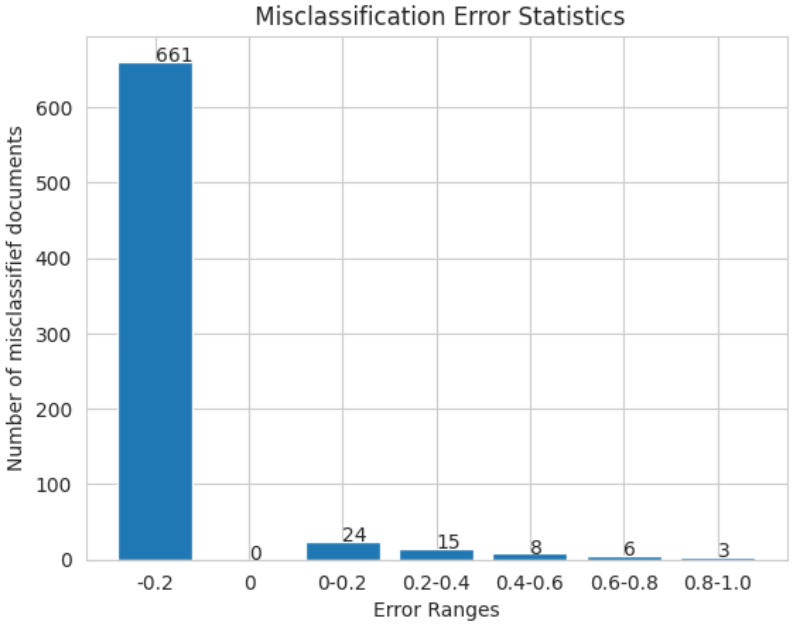

Figure 41: Error Stats

Misclassification Identification and Feature Selection:

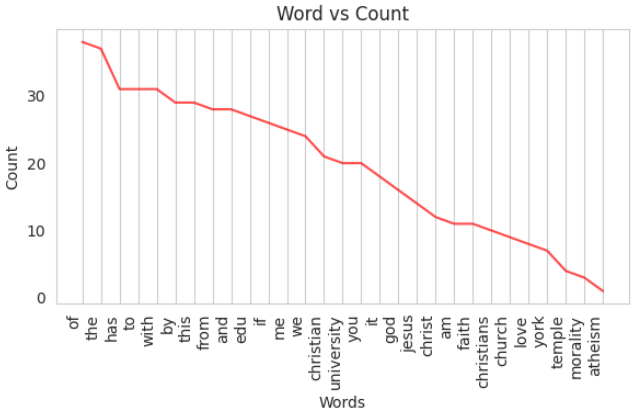

Figure 42: Overall

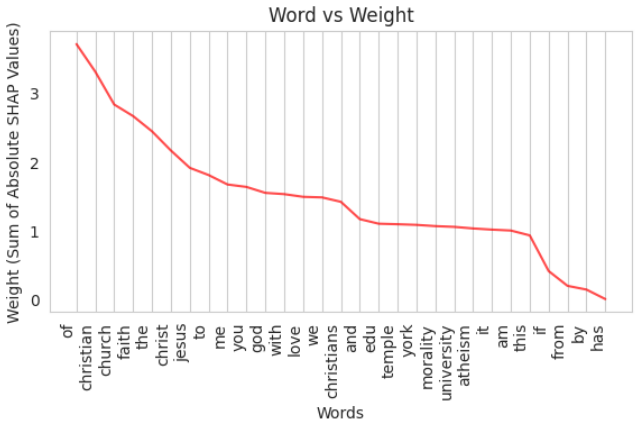

Figure 43: Overall

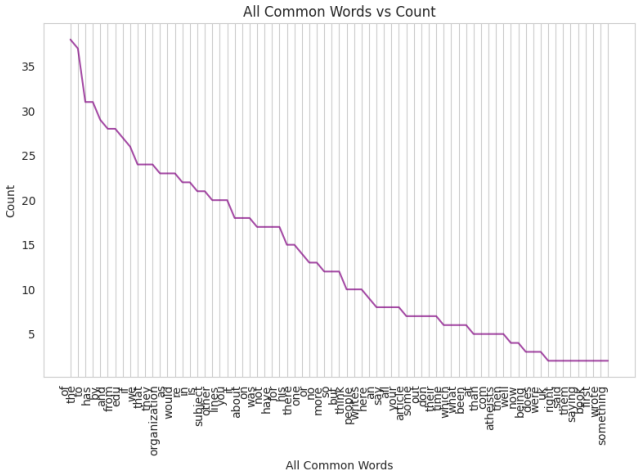

Figure 44: Overall

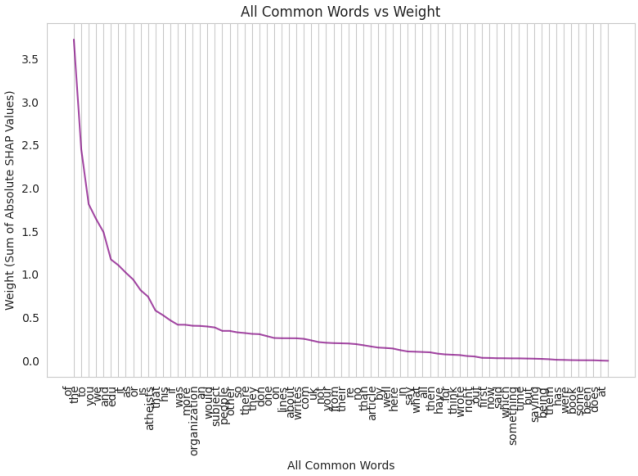

Figure 45: Overall

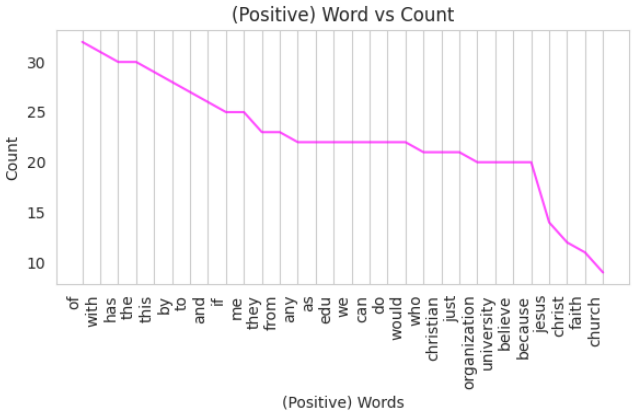

Figure 46: Toward Positive(Christian)

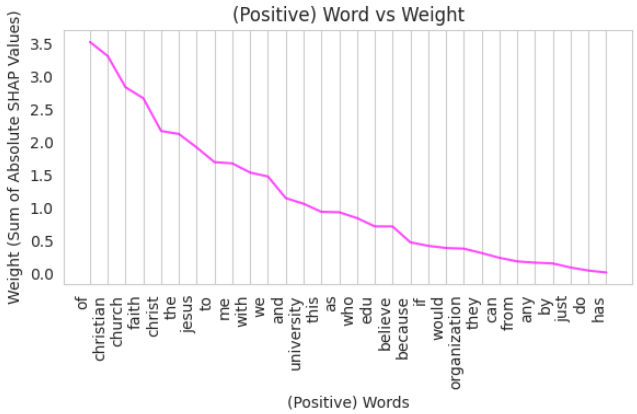

Figure 47: Toward Positive(Christian)

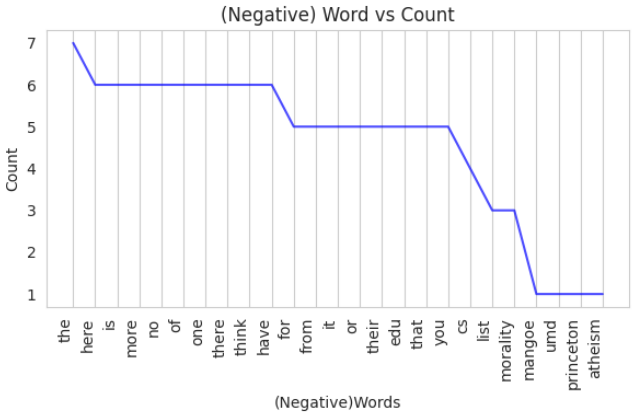

Figure 48: Toward Negative (Atheist)

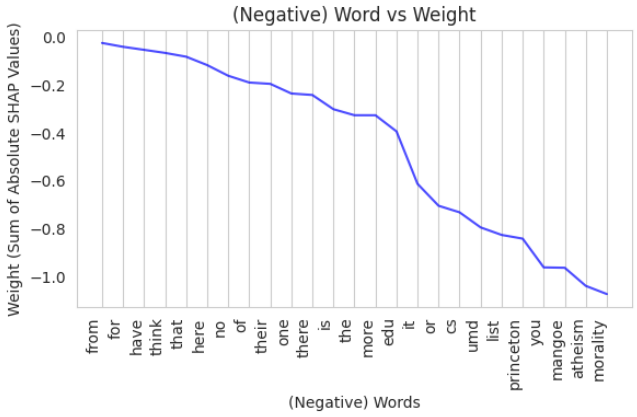

Figure 49: Toward Negative (Atheist)

Impact of Words in Misclassification:

The objective is to highlight impactful words that contributed the most to misclassifications across multiple documents.

By multiplying the count and the weight, we're essentially prioritizing words that have both high frequency (count_j) and significant contribution (weight_j) to the misclassifications.

This approach is for focusing on overall influence across misclassified documents.

An observation is that some common words, such as "of", "to", "the", "and", and "it", appear frequently in both the top 50 words Shapley summary plot and words with the highest weighted counts (overall/positive/negative). These words seem to be among the top contributors according to the Shapley values.

However, when considering the removal of these words, we must be cautious. Removing them could potentially reduce false positives (FP) or false negatives (FN), but we also need to ensure it doesn't drastically affect true negatives (TN) or true positives (TP), since they show high contribution in the summary plot.

Additionally, the Shapley values for words like "to", "of", and "the" are more evenly distributed around the '0.0' axis compared to other words. Based on this observation, we could hypothesize that at least one of these words might be a candidate for removal during feature selection. From a logical standpoint, these words likely shouldn't play a crucial role in distinguishing between categories like "atheist" and "Christian" in document classification.

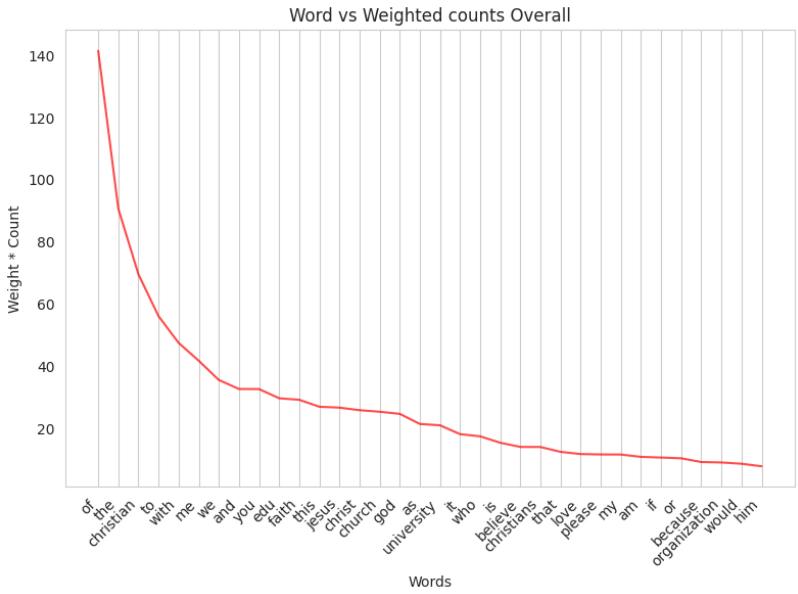

Figure 50: Overall Impact

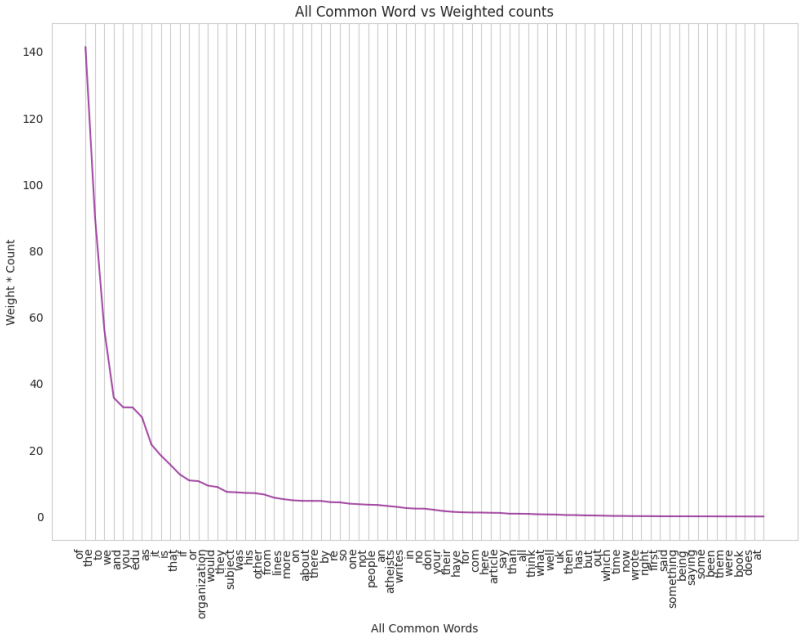

Figure 51: Overall Impact

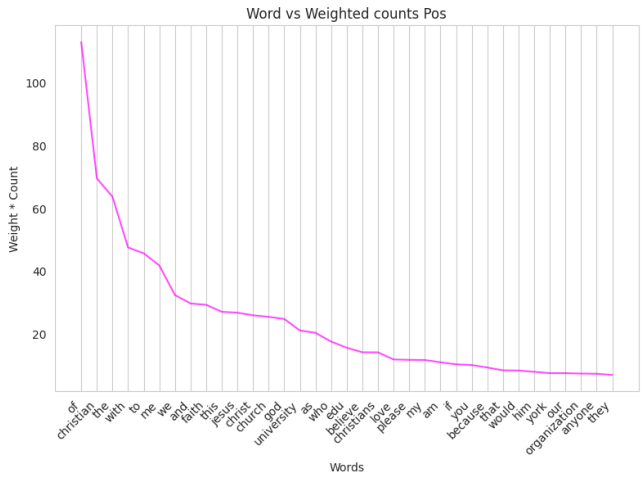

Figure 52: Positive Impact

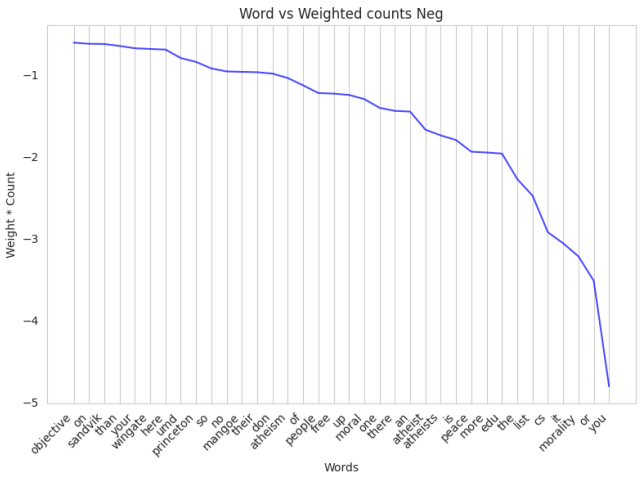

Figure 53: Negative Impact

Feature selection strategy:

While designing a feature selection strategy, we have to choose a hyperparameter 'threshold' which indicates maximum how many words we want to remove (i.e. the stop_words in TFidf Vectorizer) from the training set.
Since (i) we don't want to overfit. (ii) need to balance efficiency along with accuracy - while searching for such words, the number of words to be removed can become very large depending on the training and validation set.
Setting threshold = 6 = 0.5 percent of total number of misclassifying words.

Since, there are important words having high shapley values, and are highly context relevant, we shouldn't remove them as part of feature selection technique.
i.e. *ignore_words= ['christian', 'faith', 'jesus', "christ", "church", "god", "believe", "christians", "atheism", "free", "atheist", "atheists", "beliefs"]*

Trial : First, the focus was on identifying and removing stop words that had both positive and negative Shapley values across different misclassified documents . This trial specifically targeted common words that appeared frequently across the misclassified documents and had notable Shapley values, whose removal increase model accuracy. Result:
Stop Words: ['of', 'the', 'edu', 'say']
Accuracy: 0.9330543933054394
FP: 43, FN: 5
None of the selected words were part of the ignore_words list.

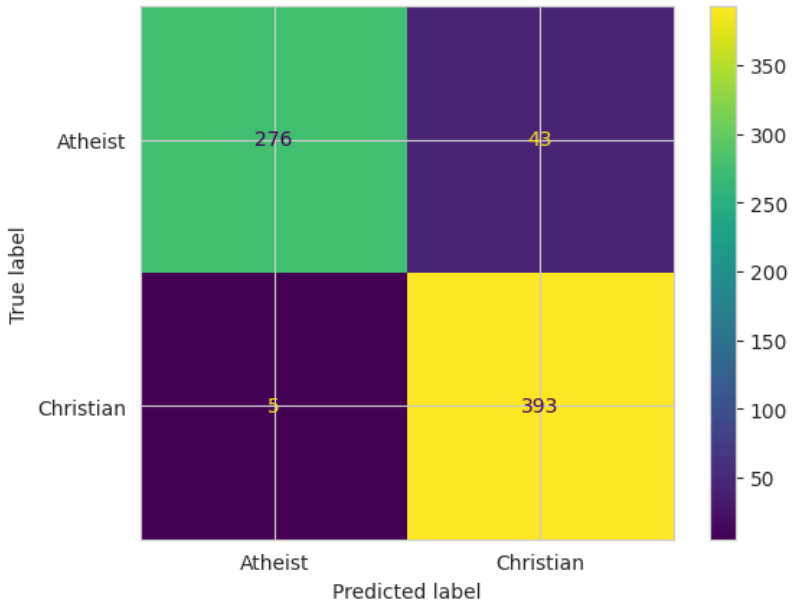

Figure 54: Using Only Common (+ & -) Misclassification Words

Next, attempted to expand the selection of stop words to include all misclassification words. the focus shifted to iterating through all the misclassification words without filtering them based on their direction of Shapley value contribution. It allowed the inclusion of additional words, such as 'york', which might not have been selected in the first trial.

Results:
Stop Words: ['of', 'the', 'edu', 'york']
Accuracy: 0.9330543933054394
FP: 42, FN: 6
None of the selected words were part of the ignore_words list.

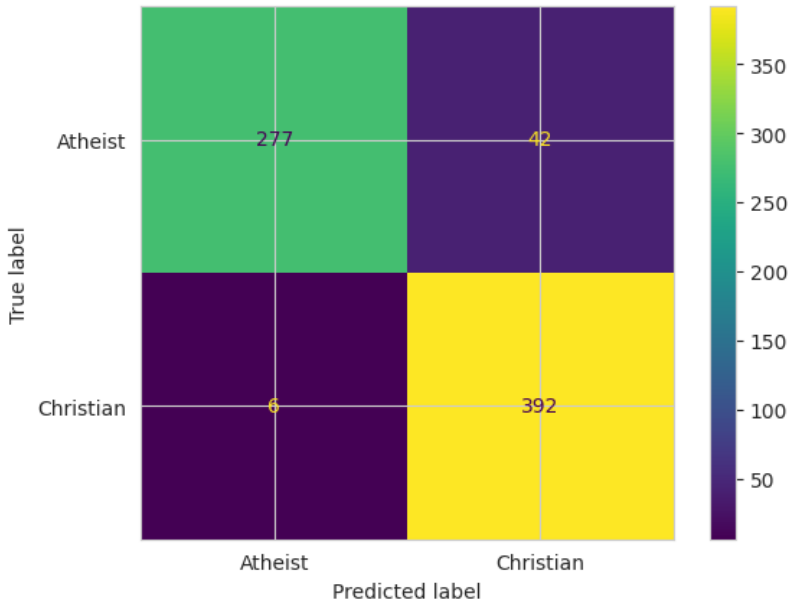

Figure 55: Using All Misclassification Words

Finally, attempted making the feature selection strategy dynamic and conditional. The key difference in this trial was that the stop words were selected based on a conditional logic involving the changes in false positives (FP) and false negatives (FN).

- If FP increases: Remove words with positive Shapley value contributions (words that contribute to false positives).

- If FN increases: Remove words with negative Shapley value contributions (words that contribute to false negatives).

The strategy here aims to balance the model's performance by selectively removing words based on whether the false positives or false negatives are affecting the performance, and whether their removal leads to an improvement in accuracy.

Results:
Stop Words: ['york', 'here', 'princeton', 'of', 'edu', 'the']
Accuracy:0.9358437935843794
FP: 42, FN: 4
None of the selected words were part of the ignore_words list.

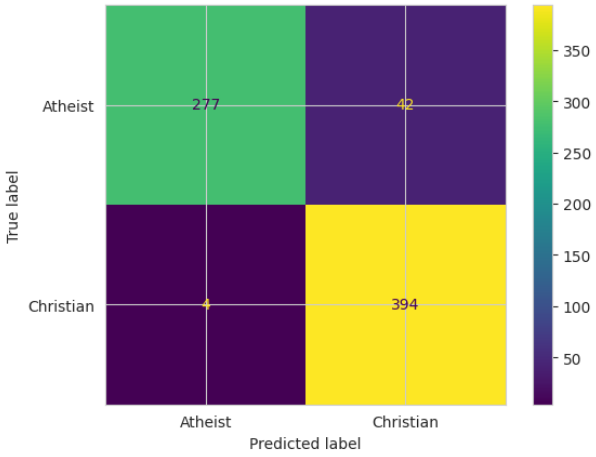

Figure 56: Using Positive & Negative Contribution Words Conditionally

Best Stop Words: ['york', 'here', 'princeton', 'of', 'edu', 'the']
Best Accuracy:0.9358437935843794
Best FP: 42, Best FN: 4

Example that was misclassified before feature selection and that is classified correctly after feature selection:

Figure 57: Correctly Classified -idx62 after Feature selection

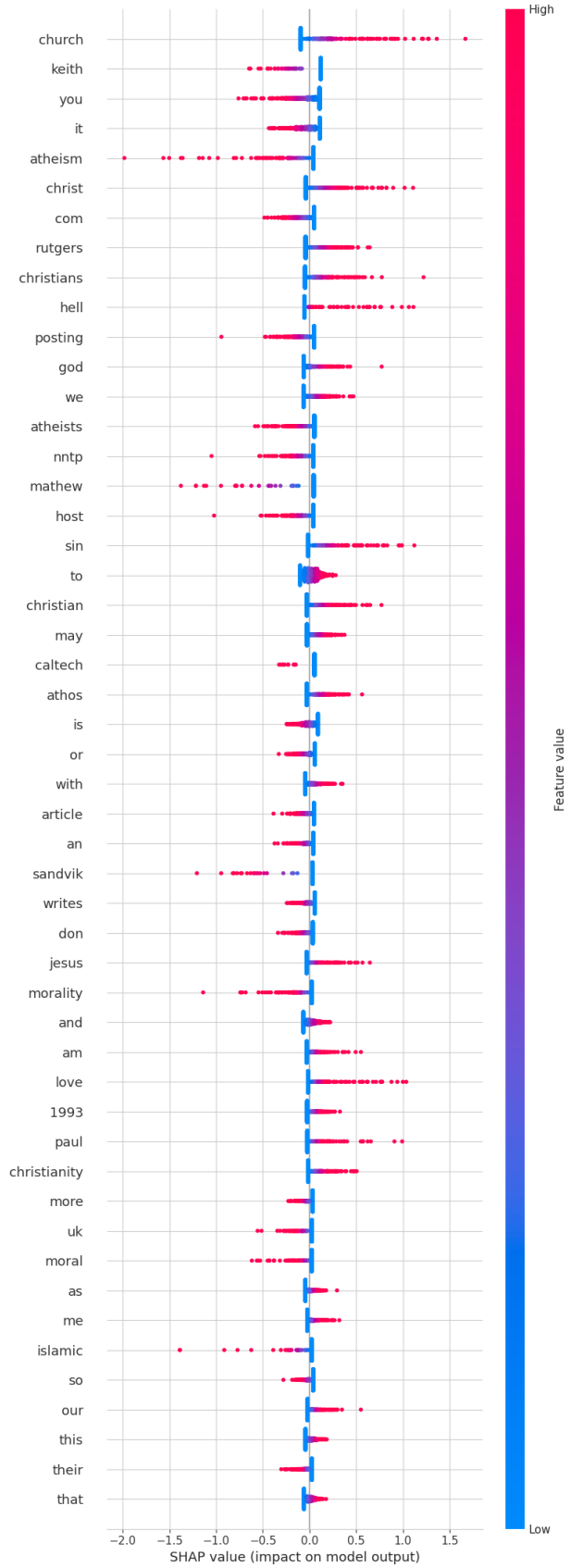

Figure 58: SHAP Values Summary Plot After Feature Selection