0.1 Homework 3 DS GA 1017 Joby George (jg6615) Due 5/5/22

1 Problem 1 Online Job Ads

Consider a hypothetical job search website that uses a machine learning system to determine which job openings to show to which users. The system uses historical employment data, and also collects interaction data from its own users: which users were shown which job openings, and which users clicked. The service also receives data from employers, detailing which users were invited to job interviews, and which were hired.

1.1 1A

Give three distinct reasons why gender disparities might arise in the operations of such a system

1. Pre-existing Bias. Several industries, and positions are currently disproportionately dominated by certain genders. In the Table 4 of the appendix in Automated Experiments on Ad Privacy Settings, by Datta, Datta, and Tschantz the five URL+title pairs that the model identifies as the strongest indicators of being from the female or male group are shown. I have added the table below for context.

Title	URL	Coefficient	appears in agents		total appearances	
			female	male	female	male
Top ads for identifying the simulated female group						
Jobs (Hiring Now)	www.jobsinyourarea.co	0.34	6	3	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	0.281	6	2	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	0.247	5	1	29	1
Goodwill - Hiring	goodwill.careerboutique.com	0.22	45	15	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	0.199	19	17	38	30
Top ads for identifying agents in the simulated male group						
$200k+ Jobs - Execs Only	careerchange.com	−0.704	60	402	311	1816
Find Next $200k+ Job	careerchange.com	−0.262	2	11	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	−0.253	0	45	0	310
CDL-A OTR Trucking Jobs	www.tadrivers.com/OTRJobs	−0.149	0	1	0	8
Free Resume Templates	resume-templates.resume-now.com	−0.149	3	1	8	10

Table 4. Top URL+titles for the gender and jobs experiment on the Times of India in May.

1. The fourth most gendered job ad for men, CDL-A OTR Trucking Jobs, is an example of an industry which is dominated by men. According to the Women in Trucking Nonprofit, approximately 7.9% of truck-drivers are women.[1] The highly unbalanced gender ratio in this field would inherently make its way into the model, and the model would not display trucking job advertisements to women as they do not currently reflect a top prospect for this particular job.

2. Different ad-engagement patterns across gender. There are multiple reasons that genders could respond to online ads about employment differently. A 2008 Article from Harvard Business School titled How Female Stars succeed in new Jobs (https://hbswk.hbs.edu/item/how-female-stars-succeed-in-new-jobs) offers insights into one potential reason. Top-performing women, those that would be interested in executive and high paying roles are more cautious and vet potential career shifts more seriously than men. They build a network that is not dominated by people at their own firm, but a group of peers and mentors working at different places. These two reasons could lead to women not engaging with ads for high-paying jobs as their mechanism to get high paying jobs, but rather relying on their network and referrals to land into their next job. On the other hand, the article posits male top-performers tend to build a primarily internal network, and vet potnential future employers less strictly. Thus, when looking for new positions, men would be more likely to respond to an advertisement rather than contact a friend they know who could help. The fact that men would be more likely to engage in an advertisement directly influences the algorithm and would then preferentially treat men for high paying jobs. This is just one hypothesis for why women and men would engage with ads in a gendered manner, but if there is a gendered response in click through behvaior, it would ultimately feed into the algorithm itself in which ads are displayed.

3. Emergent bias. Companies posting for jobs have a culture they are explicitly trying to hire for. Thus as the machine learning system receives feedback from employers on which candidates were invited for interviews and which were hired, it would algorithmically enforce the company and industry biases to potential prospects interested in applying for a job. An example of this would be the video game developer industry. In a Wired article on game-maker Riot Games (https://www.wired.com/story/riot-games-ceo-culture-complaints/), the "boys club" and "fraternity" culture are used to describe top leadership. Women routinely faced a stiffer barrier to full-time employment, promotions, and a toxic work culture. All of these company and industry wide practices will manifest itself in the data, as women convert fewer job interviews that they apply for, making them less valuable prospects for an online job ad market which would pay money for a conversion rather than a click.

   Additionally, as women hear about these stories, they won't engage and will opt to not click on an advertisement even if it's shown to them, lowering the click through rate and further deprioritizing women in the algorithm for a job in this specific industry. This creates a chain effect, where fewer women see ads and click, lowering click through rate even more and as the algorithm gets fed more data, it will reinforce the company or industry's bias that women are not desirable talent for those positions.

1: Footnote for Women in Trucking (http://www.womenintrucking.org/blog/what-have-we-done-to-increase-the-presence-of-women-in-trucking)

1.2 Problem 1B

Suppose that the job search service decides to increase the number of times it presents job

openings in STEM to African Americans. To do so, the service observes that STEM job experience (in years) is positively associated with the likelihood that a user clicks on an advertised STEM job opening: the more years of experience, the more likely a user is to click. Consider the following intervention:

Pre-process the training dataset, replacing the value of the “job experience” feature for African Americans with the best (highest) possible value for the feature in the dataset.

1.2.1 I and II

Under what conditions will this intervention increase the number of times job openings in STEM are shown to African Americans?

Under what conditions will this intervention fail to increase the number of times job openings in STEM are shown to African Americans?

1.2.2 Answer

Some conditions that would increase the ads for STEM job openings would be shown to African Americans are :

1. The ad must still be calibrated for the right job position. If by replacing years of experience to the highest possible value, and all the jobs that are shown to African Americans are now senior and managerial level roles, fresh graduates would not engage with the advertisements, lowering the group's click through rate compared to what the model would expect and thus penalize African Americans from seeing STEM jobs.
2. The system must be optimized for a click-through rather than some conversion event (hiring potential prospect). If the algorithm is trying to optimize for conversion, and companies themselves do not want to hire an African American candidate then the algorithm will not prioritize showing these ads to African Americans. If the algorithm is optimized for click-through rate then any engagement with advertisements will be seen as favorable to both the employer and the algorithm and more ads will be shown to African Americans.
3. For more advertisements for STEM job openings to be shown to African Americans, the advertisement demographics selected in the marketing campaign would have to be relevant. In determining the relevancy to show which ads to which online browsers, there's a component of relevancy and an auction bid itself. If a company believes an online browser has 20 years of experience and has engaged with a lot of STEM related web-browsing they may be willing to bid a high amount to display the ad. However, because of this pre-processing, we may have distorted the bidders understanding of who is behind the screen. Instead of an engineer with 20 years of experience, the web-browser is a middle school science student (who may use their parent's Google account such that their age would align with the experience) there would be little relevance to the actual person seeing the ad. Thus the model could learn that click-through rate for highly experienced African Americans interested in STEM is negatively correlated with Click Through Rate and display fewer ads.

   While the example of a middle school student being behind the screen is somewhat contrived, any person casually interested in science and technology, but pursuing a career in a different field would be misrepresented to the auction bidder and the

career in a different field would be misrepresented to the auction bidder and the algorithm would create a new understanding between years of experience for African Americans and CTR on these ads.

2 Problem 2: AI Ethics: Global Perspectives

I watched [Content Moderation in Social Media and AI (https://aiethicscourse.org/lectures/content-moderation-social-media-ai)](https://aiethicscourse.org/lectures/content-moderation-social-media-ai) video and completed my memo, answering the following questions, below:

1. Identify the stakeholders. In particular, which organizations, populations, or groups could be impacted by the data science issues discussed in the lecture? How could the data science application benefit the population(s) or group(s)? How could the population(s) or group(s) be adversely affected?
2. Identify and describe an issue relating to data protection or data sharing raised in the lecture.Which vendor(s) owns the data and/or determines how the data is shared or used? To what extent is the privacy of users or persons represented in the data being protected? Is the data protection adequate?
3. How does transparency and interpretability, or a lack thereof, affect users or other stakeholders? Are there black boxes?
4. What incentives does the vendor (e.g., the data owner, company, or platform) have to ensure data protection, transparency, and fairness? How do these incentives shape the vendor's behavior?

3 Content Moderation in AI Memo

3.1 Stakeholder Analysis

3.1.1 AI platforms (vendors)

The AI platforms that host social networks (Twitter,Youtube, Meta) are the vendors that have created online spaces for people to engage with each other and the advertisers that fund these platforms. These groups are commercial entities that are seeking to maximize share-holder value, their business relies on scale and utilizing data science to optimize many

elements of their respective products including, ad-delivery, content recommendation, re-targetting strategies (notifications), and **content moderation**. Since the core of these companies products are ADS, they stand to benefit commercially if they can build an engaging product that attracts more users. Ultimately, this creates an arms race between the platforms to create extremely engaging networks. If a competitor can come up with a product or utilize an algorithim that is extremely engaging, a social media network can decay away as the product itself loses utility to the consumers (Google+, emergence of Tik Tok, etc). Content moderation plays a role, a platform with extreme censorship will lead people away from posting content, where a platform with a complete lack of rules and regulations, quickly turns into a space of hate groups and the most extreme content

quickly turns into a space of hate groups and the most extreme content.

3.1.2 Content Creators and Consumers (agents in the platform)

The users that consume and create content (posts, videos, pictures, etc) are the main audience and lifeblood of these networks. Users derive utility from these networks through in many ways, free online messaging, sharing pictures with family, and discovering new content. Creators have a large marketplace that offers them an opportunity to scale globally if their material gets favored by the algorithm as recommended content for a lot of users. Content moderation benefits users when users feel welcome and not disturbed by other content floating on the network. On the other hand, these algorithms can influence users sub-consciously as evidenced in the social contagion article. Furthermore, across all these platforms, content recommendation algorithms tend to recommend content a person finds highly agreeable, leading to an emergent bias where people only consume content they agree with, molding their world-view and creating echo-chambers. This is in effect, content moderation as content people find disagreeable, would not be shown to them through the use of recommendations.

The downside of the data science application for content creators are volatile modelling strategies that are determined by data vendors. For example, YouTube has gone through numerous rounds of criticism from advertisers and consumers when an advertisement for a global company is shown during an offensive or divisive video. In response, YouTube has gone through many revisions of its monetization algorithm (determining which ads are shown when) and content creators that have built a particular niche stand to lose their revenue stream over-night when these changes are made.

3.1.3 Advertisers (revenue generating consumers)

Advertisers are the financial backbone of these platforms. Using a rich dataset that creates marketing profiles, advertisers can create a hyper optimized marketing strategy to help convert customers into buying their product, and partner with content creators to directly advertise to consumers (sponsored content). The main benefit for these companies is the rich data they can use to target consumers, and the revenue this targeted advertising has generated them in conversions. Content moderation generally benefits advertisements. The biggest risk for advertising brands is when an advertiser does not want to be associated with

the content being shown to consumers, but the algorithm believes there is a strong relevance. From class, an example would be InstantCheckmate and criminal record search advertisements being shown disproportionately for African Americans.

3.1.4 Externalities: Governments, other private corporations

The last primary stakeholder are governments, and other private corporations that utilize social media as a way to control society or people individually. With content moderation being a subjective task, and Western companies preferring to react to divisive content rather than pro-actively censoring disputed content there is an opportunity for other companies and governments to use these networks for their own ends as it comes to controlling their

society, or influencing other countries. Russian interference in the 2016 election is a known example. However, foreign governments with large populations and potential user-bases can influence the data vendors to make concessions for certain actions. India is one such example (https://www.npr.org/2021/10/23/1048746697/facebook-misinformation-india), where ruling politicians of a crucial market for Facebook posted videos urging citizens to remove Muslim protestors from the streets of Delhi, resulting in **53** deaths in the following hours after the content was shared thousands of times before being removed.

Lastly, the Chinese government offers a stark contrast to the Western philosophy of decentralized, user-enforced content moderation with a centralized censorship approach. Here the use of algorithms is predictive, and quickly removes content that does not agree with the government's beliefs.

For these large institutions , power and control are what they gain by manipulating these networks billions of users. The downside is that these countries can become victim of the tech companies retaliating for political reasons. Just 7 hours ago, a Business Insider Article (https://www.businessinsider.com/facebook-removed-australia-government-pages-parliament-debating-new-law-wsj-2022-5) details how Facebook blocked access to the Australian government's page as the country was debating legislation that would have regulated Facebook.

3.2 Data Protection & Data Sharing

3.2.1 Who owns the data? How do they determines how the data is shared or used?

The data are owned by the tech platforms, but the data can be argued to belong to the users of hte network. One analogue that could be made are health records, where the patient's data are the information about a given person.

However, in this space, regulation is much less stringent, and therefore the **true data owners** are the people that are collecting and storing the data, which are the tech platforms.

Since the data owners are the AI platform, and user data and user's visual space are the product that these social networks sell to advertisers, they are incentivized to make as much money as possible, without drawing consumer backlash. The networks selling data to data vendors create externalities of use, where downstream consumers can use an extremely rich datasets for unintended consequences. One recent example of how this can happen, is a Denver data vendor recently got highlighted by Vice for selling aggregate data on people who went to Planned Parenthood.

3.2.2 How strong are privacy mechanisms?

The main law in the US that governs content moderation on social networks is Section 230 of the Communications Decency Act of 1996. This absolves the platform from taking responsibility from users who post on Facebook.

In the EU, the idea is the same, but a penchant for regulating content differs due to the

In the EU, the idea is the same, but a penchant for regulating content differs due to the interpretation of the right to speech.

As a part of the challenge in moderating content, data privacy is actually a road-block. In order to truly understand whether a post or piece of content violates terms of use, it is much easier to have the whole graph of a poster. However, data privacy mechanisms in content moderation mean that algorithms have to determine based on a single post, whether or not the content violates terms of use. Just knowing this fact, people can get around censors by posting offensive content bit by bit over many posts.

3.3 Interpretability and Explainability

Interpretability is a huge problem in content moderation. Determining which content is illegal is subjective, and varies amongst different countries. Additionally, with the challenge of dealing with text and content, automating this problem becomes extremely difficult to do accurately.

In order to solve these problems, usually extremely large black box models are used to judge a post to see if it is classified as offensive or not. These models have 100 billion + parameters, so understanding them at an extremely intricate detail is near impossible.

3.4 Incentives for data protection, transparency and fairness

The incentives for data protection, transparency and fairness are low, as determined by the legislating governing these companies. At the end of the video, Dr. Abiterboul notes content moderation completed by the companies is not legitimate as there is little incentive for the company, whereas moderation determined by the government could violate individual liberties. His proposal to fix this non-transparent and potentially unfair is to regulate the platforms, and having society engage with the regulators and tech companies to have all voices heard.

4 Problem 3:

4.1 3A:

Use the provided Colab template notebook to import the 20 newsgroups dataset from sklearn.datasets, importing the same two-class subset as was used in the LIME paper: Atheism and Christianity. Use the provided code to fetch the data, split it into training and test sets, then fit a TF-IDF vectorizer to the data, and train a SGDClassifier classifier.

4.2 3B

Generate a confusion matrix (hint: use sklearn.metrics.confusion_matrix) to evaluate the accuracy of the classifier. The confusion matrix should contain a count of correct Christian, correct Atheist, incorrect Christian, and incorrect Atheist predictions. Use SHAP's explainer to generate visual explanations for any 5 documents in the test set. The documents you

select should include some correctly classified and some misclassified documents.

I provide a screenshot to load the data, fit the model and generate a confusion matrix below.

Problem 3

Variables

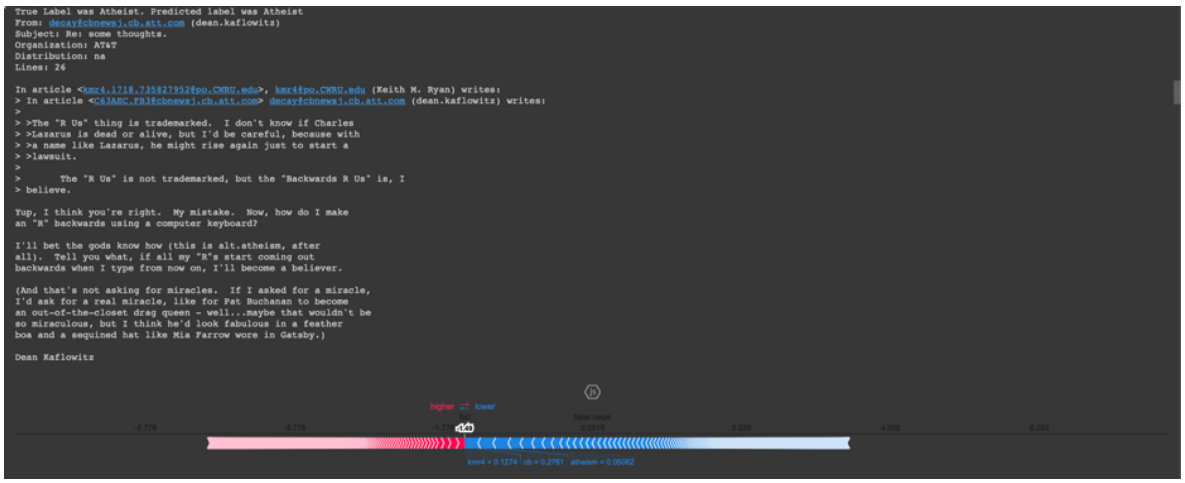
Part (A)

```
[3] # Mark the categories of interest
    categories = ['alt.atheism', 'soc.religion.christian']

    # Fetch the data and labels
    newsgroups_train,y_train = fetch_20newsgroups(subset='train', categories=categories, return_X_y = True)
    newsgroups_test, y_test = fetch_20newsgroups(subset='test', categories=categories, return_X_y = True)

    # Set outcome class names
    class_names = ['atheism', 'christian']
```

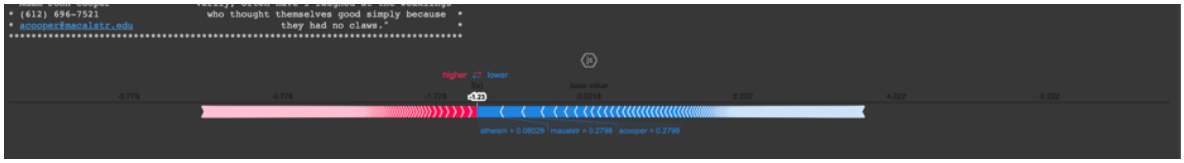
```
[4] # Initialize & fit tf-idf vectorizer
    vectorizer = TfidfVectorizer()
    X_train = vectorizer.fit_transform(newsgroups_train)
    X_test = vectorizer.transform(newsgroups_test)

    classifier = SGDClassifier(loss = 'log')
    classifier.fit(X_train, y_train)
    y_hat = classifier.predict(X_test)

    sklearn.metrics.confusion_matrix(y_test, y_hat)

    array([[276,  43],
           [  4, 394]])
```

For the five explanations on documents, I provide screenshots of SHAP explanations below, with the text that our classifier was trying to predict an Atheist or Christian label.

4.2.1 Review 1

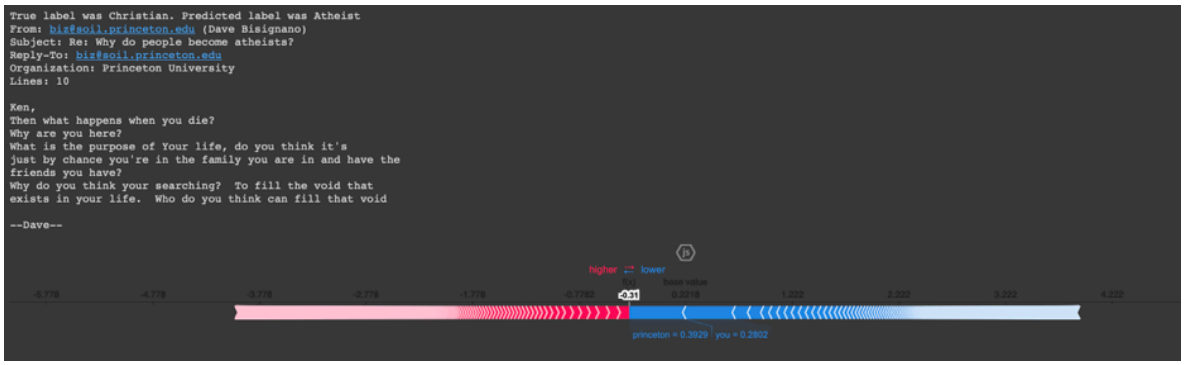


4.2.2 Review 2

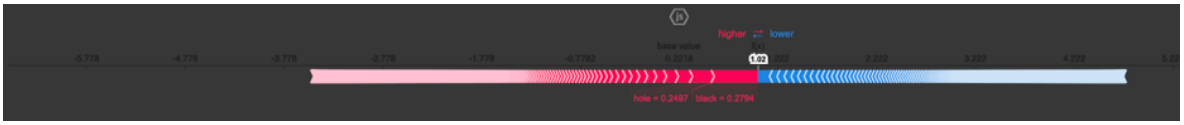


4.2.3 Review 3

4.2.4 Review 4



4.2.5 Review 5



4.2.6 Classification Commentary

With the five selected reviews, four reviews we predicted Atheist, whereas two of the incorrect predictions were truly Christian, but we predicted Atheist. 3 of those were correct, whereas on the only predicted Christian of this sample, we were incorrect as the true label was Atheist.

It's interesting seeing what words are high weights. In our incorrect prediction for the Christian Review, the subject line was "Why do people become atheists." It's simple to see how the phrasing of that subject line skewed the classification to predict Atheist, but interestingly enough the word "Princeton" was the strongest contributor to the incorrect Atheist prediction.

Similarly, a number of digital prefixes, or organizational domains (cwru, kmr4) were strongly associated with the Atheist label. This does not intuitively make sense, but the weighting in our model means as shown by SHAP shows that they are important.

4.3 3C:

Use SHAP’s explainer to study mis-classified documents, and the features (words) that contributed to their misclassification, by taking the following steps:

1. Report the accuracy of the classifier, as well as the number of misclassified documents.
2. For a document doc_i let us denote by conf_i the difference between the probabilities of the two predicted classes for that document. Generate a chart that shows conf_i for all misclassified documents (which, for misclassified documents, represents the magnitude of the error). Use any chart type you find appropriate to give a good sense of the distribution of errors.
3. Identify all words that contributed to the misclassification of documents. (Naturally, some words will be implicated for multiple documents.) For each word (call it word_j), compute (a) the number of documents it helped misclassify (call is count_j) and (b) the total weight of that word in all documents it helped misclassify (weight_j) (sum of absolute values of weight_j for each misclassified document). The reason to use absolute values is that SHAP assigns a positive or a negative sign to weight_j depending on the class to which word_j is contributing. Plot the distribution of count_j and weight_j, and discuss your observations in the report.

4.3.1 3C.1

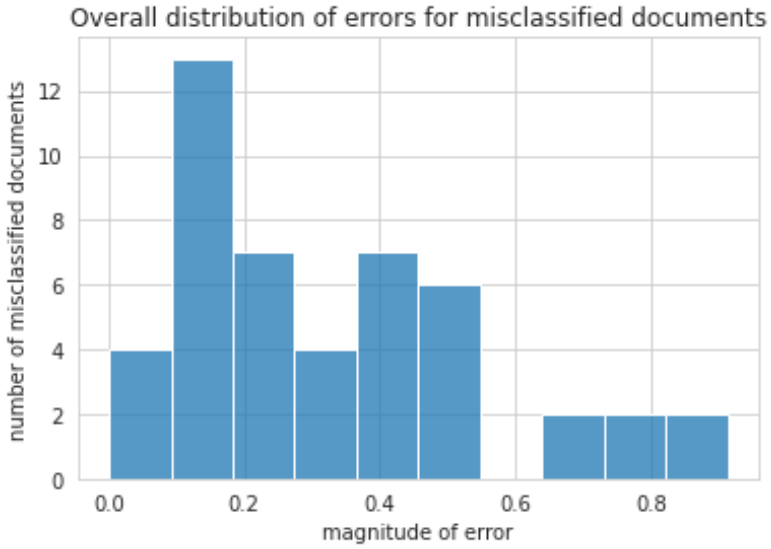
Part (C)

Part (i)

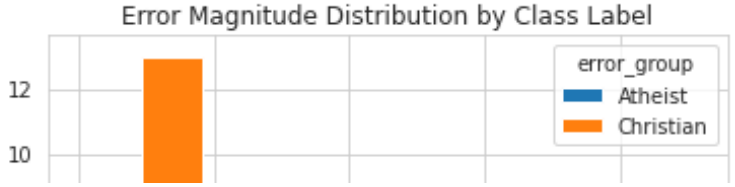
```
[ ] # Compute the accuracy of the classifier and the number of misclassified documents
    #accuracy
    print('Accuracy of our classifier on test set data was ' + str(round(sklearn.metrics.accuracy_score(classifier.predict(X_test), y_test),4)))
    #num misclassified
    print('The number of misclassified documents was ' + str(len(incorrect_guesses)))

    Accuracy of our classifier on test set data was 0.9344
    The number of misclassified documents was 47
```

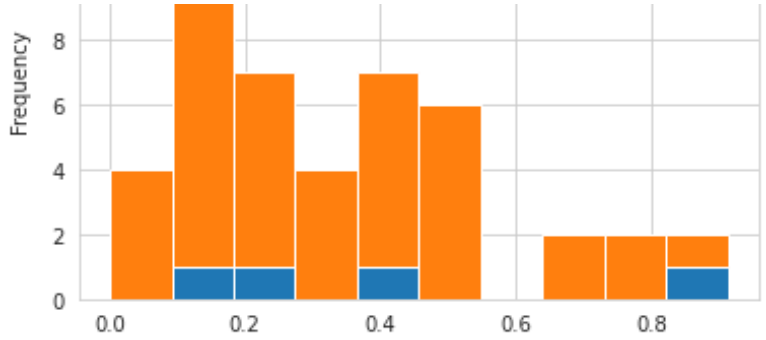
As shown above, the accuracy for an untuned classifier was .9344, with 47 documents being misclassified.

4.3.2 3C.2 Visualization 1: Total distribution of error magnitude



4.3.3 3C.2 Visualization 2: Distribution of error magnitude by Class

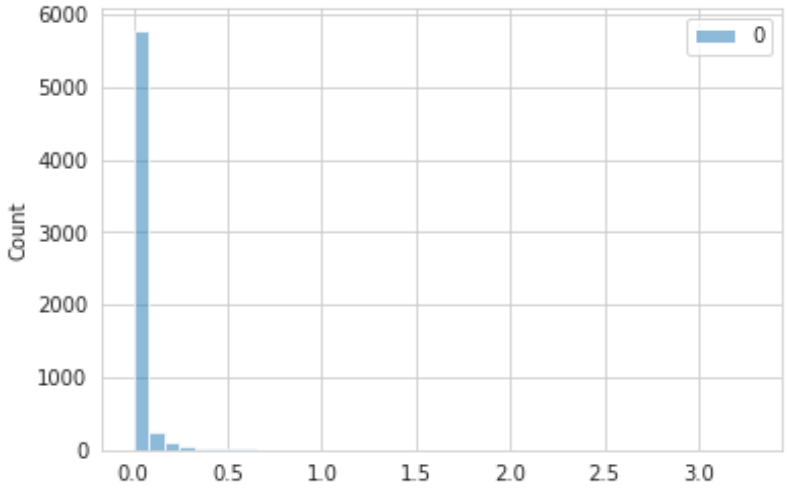




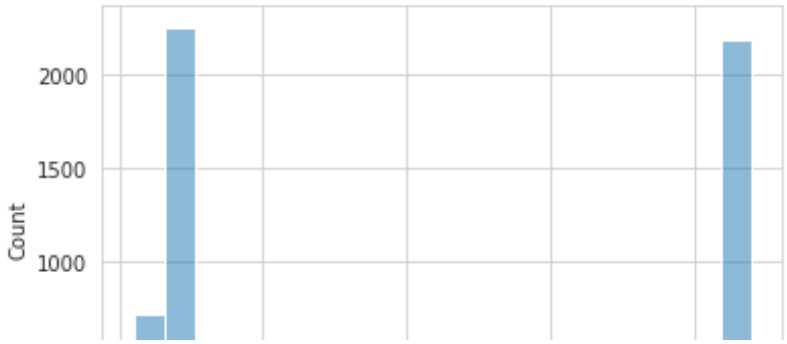
4.3.4 Commentary

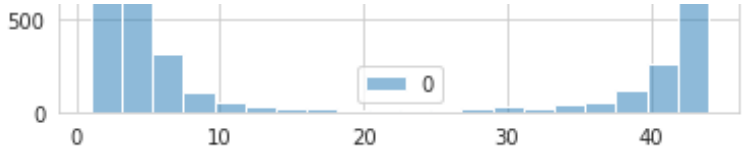
We can see that most of our errors are closely bunched at .1-.2, and that label with the more errors are Predicted Atheist that were truly Christian (**Note: the legend in the second graph refers to the true class, and not the predicted class.**)

4.3.5 3C.3 Visualization of distribution of weights on words that contributed to misclassification



4.3.6 3C.3 Distribution of num words misclassified for words that caused a misclassification





Based on the above graphs, we can see that there are a lot of words that have tiny weights contributing to misclassification, and that the majority of words that have a negative weight on the correct classification either impact a few words, or a large quantity of words.

Given there were only 47 words misclassified, the several thousands of words that were involved in misclassifying 40+ validation data points were likely frequently used words, such as stop words. Whereas the words contributing to a few misclassified examples could have been words that were mis-interpreted.

4.4 3.D

Inspired by this approach of looking at feature weightings for words that were misclassified, i came up with a sequential process to feature selection.

I first, calculated all the words that were used in positively predicting a given datapoint, and calculated the sum of feature weights for each word. I stored this in a dictionary so the key was {word:positive_weight_sum}, **Dictionary A**

I then repeated the process in 3C, finding words and weights for the words that caused in a misclassification. I similarly stored the output of this in a dictionary getting {word:negative_weight_sum}, **Dictionary B**

I then subtracted Dictionary B from Dictionary A, giving me a net weight of a word. I sorted ascending so the words with the strongest net negative impact were first. I then the 3 words in this dictionary, as well as any words that had a weight <= -2.5 which would be on the far right of the graph from 4.3.5 above.

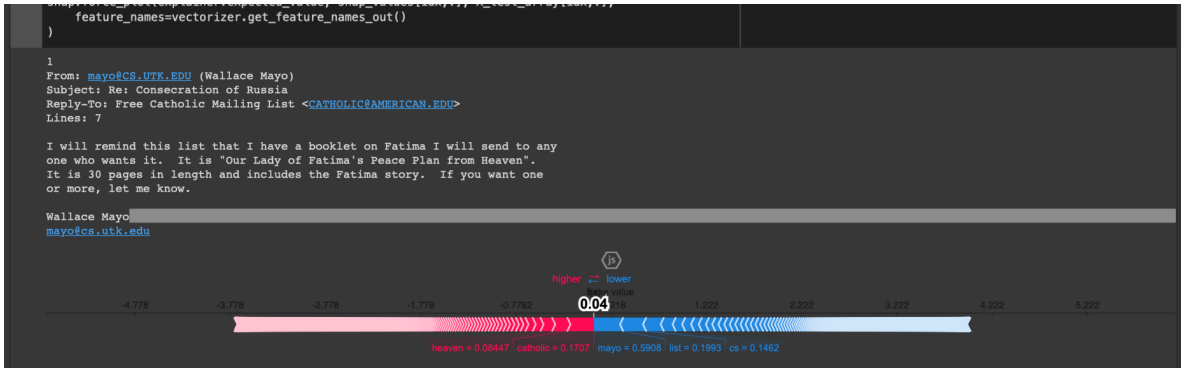
Ultimately, this process caused marginal gains, boosting accuracy from .9344 to a max of .947 after 3 rounds of sequential leading net negative word removal.

{'oser','mayo', 'ncsu', 'oed','thanx','yoking', 'odwyer','thanksgiving', 'ulysses', 'odious', 'yhvh', 'yevgeny', 'ultimate', 'odin', 'tjl', 'odds','thair','oddjob', 'yep','oddities', 'th', 'odder','texts'}

It's interesting to note how many of these words start with o, and contain a d, or have a y.

In looking at an example where this strategy boosted performance, we see the targeted word removal of mayo greatly impacted the prediction of test example 121:

```
idx = 121
print(y_test[idx], newsgroups_test[idx], sep='\n')
X_test_array = X_test.toarray()


explainer = shap.LinearExplainer(classifier, X_train)
shap_values = explainer.shap_values(X_test_array)
shap.force_plot(explainer.expected_value, shap_values[idx,:], X_test_array[idx,:],
```



.

Mayo was likely one of the words that was not very frequently common in causing misclassification, but for this particular example it was the most important feature, by removing it, we correctly predicted the true value of this input, which was Christian.

This hyper-targetted feature removal was also helpful in another example, where the .ncsu domain was the strongest feature in predicting Christian, despite the true label being Atheist.

In [ ]: