CS 590 ETHICAL FALL 2024 HW5 PRIVACY

DRASHTI SNEHALKUMAR PATEL

**Homework Understanding :**

This task exemplifies how text classification models may leak privacy despite the lack of obvious identifiers like usernames. It shows the importance of indirect pointers like language patterns towards revealing secret information. Under practical exercises, we aim to detect such leaks and perform anonymization such that privacy remains protected without jeopardizing model effectiveness.

---------------------------------------------------------------------------------------------------------------------------For the dataset , we use the politics dataset (train and test ) which contains social media posts labeled with political affiliations (Democrat or Republican) from various users, with user IDs removed, and the text content reflects topics and issues commonly associated with either political party, such as healthcare, taxes, and national defense.

**​​Part 1: Privacy Leak in Classification Model**

**Part a: Exposing the Leak**

**My approach:**

Initially training the lr model was an easy and simple task in which I trained a Logistic Regression model, using TF-IDF to turn raw text into meaningful features, and evaluated its performance on both the training and test sets

Uptil here it was pretty clear nd the result is as discussed below and then we move on to the part b which focussed on demonstrating the privacy leak and showing that the howing that the model could predict a user’s political party based purely on the content of their tweets.

**We had a few hints provided before hand:**

Hint 1: As Donald is party neutral, he may provide a starting baseline to view the model through.

Hint 2: Use the knowledge about what Republicans and Democrats tweet more about to help explore the model outputs. Do not rely on the usernames alone as Donald is neutral and it will still only assign one label. (Use only the provided information, do not scan the training data or testing data).

These hints helped me to concentrate on content-based analysis over user identifiers, employing neutral content (Donald's tweets) as a starting point to investigate how the model reacts to uncertain data. They also encouraged me to utilize keywords associated with each political party (such as "america" for Republicans and "change" for Democrats) to evaluate the model's dependence on text content in making political predictions and give me an idea for further processing

Once we had trained the model, we proceeded to demonstrate the privacy leak by examining the model's political leaning predictions based on the tweet content. From what we know Republicans and Democrats typically tweet, i.e., "america" for Republicans and "change" for Democrats, we filtered the test data using these keywords.

We then applied the trained model to make predictions for political alignment of the tweets with the predict() and predict\_proba() functions.

Upon examination of the predicted labels and probabilities, we observed that the model had been able to make predictions for political alignment based solely on the content text, even without explicit user IDs.

This suggested that the model leaked political information by inferring topics being discussed in the tweets. In order to probe further the model's behavior, we used Donald's neutral tweets as our baseline and showed how the model still labeled neutral content as having political affiliation. This provided us with clear numerical proof that the model was leaking political party affiliation on the basis of tweet content.

**Results produced for Logistic Regression Model:**

| Metric | Value |
| --- | --- |
| Accuracy | 0.8768 |
| Precision | 0.8851 |
| Recall | 0.8791 |
| Confusion Matrix | [[8295, 1192]  [1263, 9185]] |

The Logistic Regression model also performed very well, with a test data accuracy of 0.8768, which means it was able to differentiate quite easily between Republican and Democrat tweets.

That the precision was 0.8851 tells us that most of the model's predictions of being a Democrat were accurate, and the recall figure of 0.8791 tells us the model did reasonably well at distinguishing actual Democrats from the data.

The confusion matrix also shows that there were fewer false positives (Republicans predicted as Democrats) and false negatives (Democrats predicted as Republicans), further demonstrating the model's reliability.

Robust results like these, however, also point to the privacy leak issue. Even if user IDs have been stripped, the model can still make political affiliation predictions based on the text of the tweets, e.g., using keywords like "america" (for Republicans) and "change" (for Democrats). This means that text data alone is enough for the model to make accurate predictions, even in the absence of explicit identifiers. While the model performs well, it is questionable how personal or sensitive information could inadvertently leak through text patterns, even when user data is obscured.

**Few samples of our data**

| Text | Predicted Label | Predicted Probability(Democrat) | Predicted Probability(Republican) |
| --- | --- | --- | --- |
| American families need help affording things like health care. | Democrat | 0.9989 | 0.0011 |
| As the housing affordability crisis continues, we need action. | Republican | 0.0588 | 0.9412 |
| American democracy is at a crossroads. @SenBlue needs support. | Democrat | 0.9133 | 0.0867 |
| No one should have to choose between a paycheck and their health. | Democrat | 0.9954 | 0.0046 |
| Methane traps heat in the atmosphere up to x times more than CO2. | Democrat | 0.9964 | 0.0036 |
| Tune now to the @GuyBensonShow. We’ll be discussing the latest in conservative politics. | Republican | 0.0095 | 0.9905 |
| Everyone who pays taxes in America will see their tax rates lowered. | Republican | 0.1313 | 0.8687 |
| My heart is with all who lost loved ones in Afghanistan. | Democrat | 0.9889 | 0.0111 |
| People working full-time in New Hampshire should be able to afford basic needs. | Democrat | 0.8806 | 0.1194 |
| We are facing the largest surge of migrants at the border. | Republican | 0.0600 | 0.9400 |
| @WilliamD We know that if detected early enough, this virus can be cured. | Democrat | 0.9556 | 0.0444 |
| We were honored to join Jack Holder Pearl Harbor Survivor for a moment of remembrance. | Republican | 0.3474 | 0.6526 |
| This legislation will complement the broader defense strategy to secure America. | Republican | 0.0972 | 0.9028 |
| @TinaK Inflation is a pernicious tax on every American. | Republican | 0.0002 | 0.9998 |
| For years, the @USNationalGuard has stood ready to protect our country and our citizens. | Republican | 0.2558 | 0.7442 |

**My take on describing the output :**

The Logistic Regression model indicates political orientation can be deduced from tweet text without needing access to user IDs, exposing how classifiers in text can leak sensitive information.

Opinions on Republicans

A mention of "lower tax rates" was predicted in a tweet with a probability of 0.9412 for Republican. The model connects keywords like tax reduction, economic liberty, and national defense to Republican views. The word "America" was also highly responsible, as the tweets that contained it were correctly predicted as Republican, showing how the model is reliant on words of patriotism.

Opinions on Democrat

One of the "health care" tweets had a 0.9989 probability of being Democrat. Healthcare reform is central to Democratic policy, and the model has correctly sensed this. The word "change" was also a very heavy indicator, and those tweets with it having a high likelihood of predicting as Democrat (for example, a 0.9964 probability for climate change tweets).

The model repeatedly produces correct political affiliation predictions based on text content, flagging keywords like "tax cuts" as Republican and "health care" or "change" as Democrat. The predicted probabilities (0.9989 for Democrats on health care, for instance) show that the model isn't making wild guesses, but rather making good predictions based on clear patterns of content.

The political orientation prediction in the model from text yields a privacy leak, as it taps into political cues in the text which is , "America" for Republicans and "change" for Democrats. This means individual political opinions can be inferred from publicly revealed data, which is a privacy threat.

In Part 1, there were no significant issues that we faced in the ,model building of the Logistic Regression model. It was a relatively straightforward process, and the TF-IDF vectorization worked correctly in converting the text data into numerical feature.

### **Part b: Fixing the Leak**

**My approach:**

In Part 1b, the goal was to reduce the privacy leak seen in Part 1a, where political affiliation was predictable from usernames. To obviate this, I anonymized usernames by replacing them with generic names like "USER1", "USER2", and "USER3". This removed sensitive user-specific information but left intact the core content of the tweets, which remained effective for predicting political affiliation.

After I anonymized the data, I trained a Logistic Regression model on anonymized training data and validated it using anonymized test data. In order to check whether the privacy leak was still present, I filtered the test data with keywords like "america" (for Republicans) and "change" (for Democrats). This allowed me to check if the model was still able to predict political party using text content alone.

I implemented this solution because removing usernames while keeping the text content intact would de-risk privacy by removing individual identifiers. By replacing usernames with placeholders instead of removing names in totality, I ensured that the model could still make use of political signals like "America" and "change" without losing performance. The solution was effective in preventing the risk of privacy leakage while allowing the model to still be precise at making predictions.

This approach has the advantage that sensitive user information is not being used in predictions by the model, but without preventing the model from predicting political affiliation solely from textual content.

**Results:**

| Metric | Before Anonymization(Part1) | After Anonymization(Part2) |
| --- | --- | --- |
| Accuracy | 0.8768 | 0.8589 |
| Precision | 0.8851 | 0.8680 |
| Recall | 0.8791 | 0.8619 |
| F1 Score | 0.8821 | 0.8650 |
| Confusion Matrix | [[8295 1192] [1263 9185]] | [[8118 1369] [1443 9005]] |

**Discussion on results**

Upon anonymizing the usernames of the tweets, we had been able to demonstrate that the privacy leak had been successfully tackled. In the original data, political orientation had been predicted according to the content of texts, and usernames like @WilliamD and @TinaK were adding extra context that could potentially influence the predictions.

For example, a tweet mentioning "lower tax rates" was predicted with a probability of 0.9412 as Republican, and a tweet mentioning "health care" was predicted with a probability of 0.9989 as Democrat. After anonymization, user names like @WilliamD and @TinaK were replaced with the symbolic USER, and the model continued to make the same types of predictions, such as predicting the Republican label for a tweet about "America" with a 0.9400 probability and the Democrat label for a tweet about "health care" with a 0.9954 probability.

This means that the model continues to rely on content-based patterns—keywords like "America" and "change"—in making political affiliation predictions. Even with anonymized politically neutral Donald tweets who, the model still labeled as Democrat based on the tweet content. This confirms that the privacy leak which previously allowed the model to infer political party from user-specific information no longer occurs after anonymization.

The forecast in the model now depends on the topic of the tweets talked about, which confirms that anonymization had eliminated the risk of revealing sensitive political information associated with individual users successfully.

**Comparative Study**

Before anonymization, the model could predict political affiliation based on content in the text as well as user-specific information (e.g., usernames). For example, health care tweets were strongly predicted as Democrat, and tweets that included "America" were predicted as Republican with high confidence.

Once anonymized, we removed the usernames, but the model still accurately predicted political orientations of tweets depending on what was spoken about, such as "health care" for Democrats and "America" for Republicans. The predictions remained the same, showing that the model relies on what was being discussed rather than the users themselves. This is an indication that the privacy leakage no longer persists following anonymization, as now the model can only predict text content.

The results show that anonymizing the usernames had a moderate effect in lowering the performance of the model, with a marginal reduction in accuracy from 0.8768 to 0.8589. The performance of the model, however, is still good with good precision, recall, and F1-scores. This means that the model relies heavily on content-based signals like "America" for Republicans and "change" for Democrats. While anonymization reduced the privacy leak, it did not eliminate the model's ability to infer political affiliation from text content. This indicates the challenge in striking a balance between privacy preservation and model performance.

**A description of why the privacy leak is no longer present (with numerical values to back up your claim).**

The privacy violation is no longer present after anonymization because, although the model still produces accurate predictions, the key factor of user-specific identifiers (like @WilliamD, @TinaK, @DonaldW) has been removed.

The model still accurately predicted political orientation in the anonymized data set but now based solely on content-based signals like "America" for Republicans and "change" for Democrats.

Its accuracy on the anonymized test data is 0.8589, a fall from 0.8768 for the baseline model, a sign that prediction confidence has dropped slightly. Precision (0.8680), recall (0.8619), and F1-score (0.8650), however, remain strong and show that the model can still classify tweets based on political content instead of personal identities.

This means that, even though the anonymization process impacted the confidence of the model, it no longer relies on user information, effectively eliminating the privacy breach.

**PART 2 :Better Data Anonymization**

**My approach**

For Part 2, the goal was to improve the anonymization of the training and test data to reduce the privacy leak while still maintaining the model’s accuracy as much as possible.

I thought that while doing this it is important to consider that too much anonymising can also cause to reduction in data so it needed to be considered as well.

Checking if the Accuracy Loss by Deleting All Names

We wanted to check if deleting all names (i.e., words that are in capital letters) from the training and test data made any difference. We applied this anonymization technique to both anonymized training data and test data in an attempt to test how much the performance suffered when we removed all the words in capital letters, which we supposed were names.

After we deleted the names, we retrained a new Logistic Regression (LR) model on this modified data and verified the model performance. We were curious to know how much the accuracy would drop because of this anonymization process. We believed that the removal of names would cause some kind of loss in performance because capital words are likely to carry significant context information that the model could be using for determining political allegiance.

Creating a Better Anonymization Strategy

The second objective was to design and implement a better anonymization strategy. Instead of eliminating all capitalized words (names) altogether, we decided to substitute such names with placeholders (e.g., USER1, USER2, etc.).

This method tried to preserve the syntax and intent of the text without anonymizing the names. With placeholders, we safeguarded the text content against replacement and guaranteed the existence of politically important keywords (e.g., "change" for Democrats, "America" for Republicans) for the model to make predictions upon.

With this approach implemented, we saved the freshly anonymized datasets and retrained the model. The goal was to check if this smarter anonymization method would reduce the privacy leakage while maintaining model accuracy.

This method was supposed to strike a correct balance between anonymization and model performance in a way that the model would still be able to predict political alignment accurately without exposing sensitive user-specific data.

**Results**

**1.Table for “ removed data “ by LR model**

| Metric | Before Anonymization(Part1) | After Anonymization(Part2) | After removing all names |
| --- | --- | --- | --- |
| Accuracy | 0.8768 | 0.8589 | 0.8094 |
| Precision | 0.8851 | 0.8680 | 0.8230 |
| Recall | 0.8791 | 0.8619 | 0.8106 |
| F1 Score | 0.8821 | 0.8650 | 0.8168 |
| Confusion Matrix | [[8295 1192] [1263 9185]] | [[8118 1369] [1443 9005]] | [[7665 1821] [1979 8469]] |

**2.Table for “better data” by LR model**

| Metric | Before Anonymization(Part1) | After Anonymization(Part2) | After removing all names | After better anonymisation |
| --- | --- | --- | --- | --- |
| Accuracy | 0.8768 | 0.8589 | 0.8094 | 0.8094 |
| Precision | 0.8851 | 0.8680 | 0.8230 | 0.8230 |
| Recall | 0.8791 | 0.8619 | 0.8106 | 0.8106 |
| F1 Score | 0.8821 | 0.8650 | 0.8168 | 0.8168 |
| Confusion Matrix | [[8295 1192] [1263 9185]] | [[8118 1369] [1443 9005]] | [7665 1821] [1979 8469]] | [7665 1821] [1979 8469]] |

The results show that while removing all names from the text data leads to a decrease in accuracy (0.8094), the better anonymization technique, which maintains the structural integrity of the data, achieves the same accuracy.

First, we experimented with the technique of removing all the names from the text. This resulted in a noticeable degradation of the performance of the model on key metrics: accuracy dropped from 87.68% to 80.94%, and precision, recall, and F1 score all experienced comparable drops. This revealed that the deletion of names can impede the model's ability for accurate prediction, as it loses keywords and context helpful for classification.

Next, we employed a better anonymization technique, in which names were replaced with placeholders, while maintaining the structure and meaning of the original text without revealing sensitive information. Interestingly, this method preserved the model's performance, with accuracy, precision, recall, and F1 score all remaining at 80.94%, just as they had after names were removed. This result demonstrates that the enhanced anonymization method successfully blocked the privacy leak without undermining the model's ability to predict political affiliation from text content to any appreciable extent.

Overall, the contrast between these methods illustrates the trade-off that exists between privacy and model performance. While removal of all the names leads to loss of accuracy, the enhanced anonymization approach is a midway solution, managing to overcome the privacy concerns effectively without compromising on the model's predictive power.

Difficulties:

1.The challenge was ensuring the text content remained useful for the model’s predictions after anonymizing names. Stripping too much context could hurt the model’s ability to make accurate predictions

2.After removing names, some tweets turned into empty strings, causing issues when applying the model. We had to ensure that tweets with no text left after anonymization were properly filtered out before training and testing  
3Overall, it was a good task to do with minor coding understandings needed.

**Conclusion:**

Lastly, this case shows that text models are also able to infer sensitive information like political beliefs even after user data has been anonymized on the basis of text content only. Even trivial techniques like removing names may curb privacy risks but at the expense of affecting the model's accuracy. Another better approach like replacing names with blanks helps maintain the model's performance while enhancing privacy. This highlights the need to strike a balance between protecting privacy and upholding the precision of the model.

Note: Use of Ai and chat gpt leveraged for some coding understanding and report structuring.