#HeOrShe? - Online Data & Gender Privacy Issues

Shubham Agarwal
Saarland University, Saarbrücken

Aleena Thomas
Saarland University, Saarbrücken

ABSTRACT

In today's Internet, there are certain non-technical user-centric factors such as age-group, sex, nationality, etc. that drives most of the web-based monetization models, such as targeted advertisements, stored preferences, etc. However, the methods associated with collection & utility of these data related to specific individuals has been questioned upon by privacy-concerned researcher, organizations and individuals for a very long time.

This project work takes into account the previous works done on author profiling and gender-based data classification by Flekova et al. in PAN at CLEF'13 as a base study. We compare and contrast the studies done so far and discuss the classifications performed on different types of data, i.e., blogs, tweets, and chats. We further employ our ML experimental-models based on word embeddings in LSTM and CNN to perform classification on a dataset that includes a conversation between two individuals used for base study, obtained from PAN'13 and observe the accuracy with which it can classify the gender.

We report the significant difference observed among the performance and accuracy reported in our experiments, the base work and other associated studies and also discuss the consequences and challenges faced when performing classification on the conversational dataset.

1. INTRODUCTION

Ever since the introduction of Web 2.0 and humongous rise among users on Social Media Platforms, there have been continuous discussions among security researchers about the privacy of user-identifying information held by these Social Web Applications. We also witnessed the infamous "Facebook - Cambridge Analytica Scandal" where user-identifying information of millions of people on Facebook was used to promote political motives through biased advertisements. It is still not always clear to the users how these applications utilize their personal information for targeted content and more importantly, what if this data is accessed by unwanted individuals.

The technological architecture of applications on the web, today, are such that they carry unique footprints of the user and may be further be used to categorize them in gender, race, class, caste, income and other such classes as also stated by the proposed Social Web Gendered Privacy Model in [9]. However, a user might not always want this data to be shared with anyone else. We look at two specific cases which necessitate gender-anonymity - Online Gender-based Violence & Anonymous Posts.

One of the many social evils that flourish digitally on the Internet is Gender-based Violence (GBV). Though there have been certain regulations put forth for filtering sensitive contents, the social media continue to be the biggest carrier of posts related to GBV. ElSherief et al. analyzed specific nuances of GBV-related posts in his work in [3] and reported that there is increased engagement with GBV-related tweets found.

The concept of anonymity on the Internet was synonymous to private web browsing until the concept of Anonymous Social Media was recently introduced, e.g., *Whisper* & *Secret*. On these platforms, people usually discuss sensitive issues anonymously with regards to their relationship, health, and sexual activity, irrespective of their sociocultural background. Though the identifying information of users is hidden and not included in the post, merely the content of the post can also sometimes be used to find some statistical patterns and thus, conclude user's personal information such as gender or nationality.

We take motivation from the work done by Flekova, L & Gurevych, I in [5] and build our machine learning experimental models which takes input as a conversation dataset between two individuals fed in the form of word embeddings. We further classify the gender of the speakers in the conversation based on language metrics, lexical and contextual patterns listed in the base study. We further identify a few related follow-up studies across the different dataset and compare our findings.

To summarize, the key contributions in this project are as follows:

- We discuss the previously related research studies and different types of data considered for classification, their findings as well as the challenges faced in their approach in Section 2.

- Based on the lead study, we build our machine learning experimental models to classify the gender of the individuals in the conversational dataset by considering textual and semantic patterns that are unique to a specific gender.

- We report the significant differences in accuracy and performance observed by the result of classification by our deployed model and highlight the consequences and challenges faced.

2. RELATED WORKS

Kucukyilmaz et al, in [6], investigate the feasibility of predicting the gender of a text document's author using linguistic evidence. They processed chat texts obtained from chatbots and evaluated based on term and style based classification techniques. They could achieve accuracy of up to 84.2% with the simple evaluation framework when working on small chat-styled texts. The authors asserted general characteristics of text patterns across genders such as e.g., female tend to use longer and more content-bearing words than male. This was one of the earliest work done in this domain.

In [7], Marquardt et al investigated author profile attributes and examined the predictive quality of texts in terms of age and gender of several sets of features extracted from various genres of online social media. They worked on reducing multi-label classification to single-label classification by power transformation and chained classifier to predict age as per the output of gender label. The dataset used in this work are tweets extracted from Twitter API. They report better results of a classification in Spanish when compared to English language and highlights the limitations concerning several features used for classification for a particular dataset.

In [1], Bsir et al employ a variant of the Gated Recurrent Units (GRUs) architecture to classify gender based on texts obtained from Twitter and Facebook of Arabic Users. They implement a combination of unsupervised and supervised techniques to learn word vectors capturing the syntactic and semantic meaning of words and their relationship to respective classes. The text in the dataset was a short text of around 15-20 words. The model results in 62.1% accuracy for texts from Facebook and the authors mention that the gated recurrent neural networks can efficiently adapt to natural language processing tasks.

Chen at al, in [2], work on data obtained from mobile devices, specifically Android and further investigates gender-specific usage of emojis among Android users. The authors argue that male and female individuals not only use emoji in different frequencies but also have different preferences for selecting which emojis to use. They extract their dataset from an Android data input application, Kika Keyboard. Their model consistently results in over 80% accuracy for a different set of languages. Their work emphasizes that with the language-independent characteristic, the use of emojis can be a reliable indicator for users in different languages, and the competitive performance of the emoji-based model is generalizable to non-English users.

We observe a significant difference between the architecture of ML models, their target features and accuracy results over time as discussed above. Based on previous works, we further compare the neural and non-neural approach of gender-based classification on a dataset based on conversations.

3. OUR EXPERIMENT

We carry out two individual experiments to carry out gender-classification on the chosen dataset. The main intuition behind conducting two experiments is to set up the baseline of accuracy that we obtain from traditional supervised learning model, i.e., SVM in our case, and then further compare against the result obtained from the more advanced approach of deploying Neural Networks to classify data. The two experiments are described as follows along with the dataset and preprocessing techniques.

We borrowed certain features from our base study - surface features such as length of words and sentences, term frequency, etc, semantic features such as the relationship between n-grams and other entities and other linguistic features into consideration to determine patterns across the text that could be used to identify gender-specific traits and classify on its basis.

3.1 Experiment 1

The first experiment aims to determine how good a bag of words approach is in identifying genders of authors and decide if it serves as a weak baseline. The binary classifier used is Support Vector Machine. The input to the SVM is feature vectors that are derived from the input text. Here, only the top 10,000 features are considered. The reason is because of the nature of the training corpus which contains a large number of words occurring just once or twice. The assumption is that these words would not have enough discriminative power to detect the gender since they are occurring in very few training samples.

The hyperparameters considered during training are the kernel and C (penalty parameter). We have built two variant of this model using kernel types; linear and radial basis function. In total, we run over six values of C in the range of 0.001 - 1000 as well the number of iterations ranging from 5 -2000 to obtain best accuracy value by relaxing decision boundaries as well as to consider overfitting or underfitting in any sense.

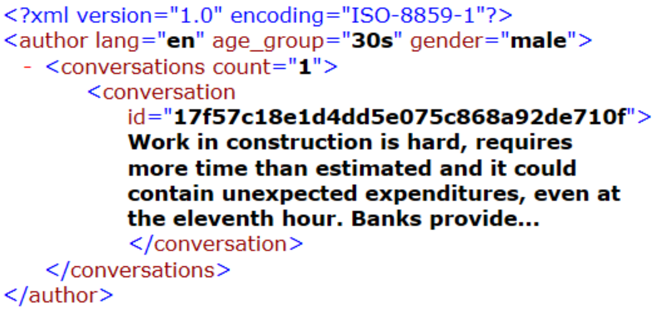
3.2 Experiment 2

The second experiment aims to determine the effectiveness of incorporating patterns in the input text for gender classification. For this, a neural approach is taken to build the classifier. The input layer is the embedding layer. The word embeddings used are GloVe (Global Vectors). The pre-trained embeddings are used and the embedding weights are not trained as part of the network. The top 50,000 features in the dataset are considered. The reason for increasing the number of features is because the neural net is capable of benefiting from huge amounts of data. As part of hyperparameter tuning, word embeddings with dimensions of 100 and 300 are considered. These are fed into a series of convolution layers with max-pooling and ReLU activation function.

The size of the CNN filter used is 5 and the total number of filters is 128. The learning rate for Adam optimizer is set to 0.01. The pattern extracted by these layers is then fed into a fully connected layer and a final softmax layer. Additionally, we have used batch normalization to speed up training and dropout for regularization. The dropout keep probabilities considered are 0.7 and 0.8.

3.3 Dataset

The dataset is extracted from the PAN'13 online portal and it contains training corpus labeled with gender and age of the author. The dataset contains documents having conversations about views/opinions/suggestions on different topics posted online. There are 2,36,000 documents, in total, in the dataset. We used 1,50,000 documents among them to train our model while 10,000 documents for testing. From the selected training dataset, we could obtain 77K unique tokens spread over the corpus. An example of a document in our dataset is displayed in Figure 1.



```
<?xml version="1.0" encoding="ISO-8859-1"?>
<author lang="en" age_group="30s" gender="male">
  - <conversations count="1">
        <conversation
            id="17f57c18e1d4dd5e075c868a92de710f">
            Work in construction is hard, requires
            more time than estimated and it could
            contain unexpected expenditures, even at
            the eleventh hour. Banks provide...
            </conversation>
    </conversations>
</author>
```

Figure 1: Example: Text Document Structure in our Dataset

3.4 Preprocessing

As shown in Figure 1, we filter all the XML and HTML tags and metadata and extract the conversational text relevant for our work. We further employ widely used natural language processing techniques such as tokenization, n-gram generation, stemming and lemmatization. After the basic preprocessing, we further filter out words that are not present in the WordNet English dictionary available online to rule out usage of specific native words in the text. Further, we also filter out the words whose frequencies are comparatively insignificant and appear just once in the whole dataset to ignore specific proper nouns such as names, places, etc.

4. RESULT

Considering the baseline accuracy of 50% as per random classifier, we further trained our ML-models to obtain better results. For the first experiment, we tried with two variants of SVM kernels - radial basis function as well as linear kernel, to understand the nature of dataset and classification patterns spread over the corpus. We obtained an improved yet insignificantly accurate results from both the results. The results obtained from both the kernels are listed in Table 1 and Table 2, respectively.

We observe that linear kernel performs achieving an accuracy of 52.9% when compared to 52.0% as reported by RBF kernel. However, we see that the number of iterations, as well as the values of C, suggests that the model allows some errors and uncertainty while classification.

Table 1: Model 1 - SVM - RBF Kernel

Iterations	Hyperparameter, C	Accuracy
2000	1000	51.4%
1500	1000	52.3%
1000	100	52.0%
500	**1**	**52.0%**

Table 2: Model 1 - SVM - Linear Kernel

Iterations	Hyperparameter, C	Accuracy
200	100	52.4%
250	100	52.7%
300	100	52.8%
100	**1**	**52.9%**

We further proceeded with the second experiment and trained our CNN model. The results obtained from the second model is listed in Table 3.

We observed that the accuracy reported by the second model is approximately 59.3%, which is also better than the best results submitted in PAN'13 Workshop - 58%, as can be verified in [8]. We infer that the reason behind low accuracy is dataset inconsistency and fair distribution of considered features across classes. Also, the implemented models fail to identify long-range dependencies that possibly exists in the dataset.

Table 3: Model 2 - CNN

Dropout	Epochs	Accuracy
0.2	11 - 20	58.7%
0.2	40 - 50	57.0%
0.3	21 - 30	59.0%
0.3	**41 - 50**	**59.3%**

5. CONCLUSION

In this project work, we tried to understand previous research works done on author profiling and gender-based classification on text articles of different types, sizes, and properties over time. We compared the implementation techniques of each of the discussed work as well as their accuracy to determine the best-suited model for our project. We further implemented two of our experimental models to compare and contrast the performance of traditional machine learning models and CNN. We borrowed the training and test dataset from the Internet openly available on PAN'13 website.

From our traditional ML-model, i.e., SVM, we infer that the Bag-of-Words Approach to determine gender from text data is insignificantly better than the results of the random classifier with a baseline accuracy of 0.5. We further modeled the patterns spread in the text data in our second model, i.e., CNN, to determine whether we could obtain comparatively better results when leveraging patterns. We obtained an accuracy of 59.3% accuracy, which is above the maximum accuracy achieved in our base study as stated in [5] as well as the first model. We conclude that the patterns in text data are an essential ingredient to classify gender from text input.

Also, we observe that our model is not able to capture long-range dependencies in the dataset, which are vital for the pattern-based classification. We then find that, in [4], the authors implement the RNN architecture to classify gender and thus, achieves 90% accuracy in doing so by appropriately modeling the long-range dependencies not considered in our work.

6. REFERENCES

[1] B. Bsir and M. Zrigui. Enhancing deep learning gender identification with gated recurrent units architecture in social text. *Computación y Sistemas*, 22(3), 2018.

[2] Z. Chen, X. Lu, W. Ai, H. Li, Q. Mei, and X. Liu. Through a gender lens: learning usage patterns of emojis from large-scale android users. In *Proceedings of the 2018 World Wide Web Conference*, pages 763–772. International World Wide Web Conferences Steering Committee, 2018.

[3] M. ElSherief, E. Belding, and D. Nguyen. # notokay: Understanding gender-based violence in social media. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[4] R. Felipe, S. Dias, and I. Paraboni. Combined cnn+rnn bot and gender profiling. In *Technical report*. PAN CLEF Workshop 2019, 2019.

[5] L. Flekova and I. Gurevych. Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*. Citeseer, 2013.

[6] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining for gender prediction. In *International Conference on Advances in Information Systems*, pages 274–283. Springer, 2006.

[7] J. Marquardt, G. Farnadi, G. Vasudevan, M.-F. Moens, S. Davalos, A. Teredesai, and M. De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 1180:1129–1136, 2014.

[8] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.

[9] M. Thelwall. Privacy and gender in the social web. In *Privacy Online*, pages 251–265. Springer, 2011.