Predicting Author Gender from Political Writing

The Real Wuglet Research Institute's Pet Shark Happiness Research Group

Adriana Knight, Bayliss Wagner, Jake Chanenson, Naomi Park

Intro:

We want to see if we can train a language model to be able to identify whether an author is male or female[1]. We also want to use our NLP tools such as sentiment analysis to compare writing by women with writing by men; this will include creating a list of most common unique words used by men and women; compare usage of certain words by men vs. by women (removing most popular words to take Zipf's law into account); we will investigate the stereotype that women use more emotional language than men while also looking for other prescriptive assumptions based on gender; and differences in topics that each group focuses on.

Central Hypothesis:

We intend to build and train a classification system that will predict whether an author is male or female based on analysis of a text's syntax, sentiment, word choice, among other features.

Background:

In "Gender as a Variable in Natural-Language Processing: Ethical Considerations," Brian Larson writes on the theoretical and ethical considerations for using gender as a variable in NLP studies. The paper brings up the concern that the gender binary is an exclusive way to frame gender identity that needs to be explicitly defined by researchers. While the framework of gender is not inherently bad, a lack of transparency or acknowledgement of which theoretical framework is in use is unethical. Larson lays out three views of gender and gives examples of applications that could use those frameworks. The first is a folk view of gender which "conflates sex… with gender." This is an alienating view of gender, particularly for transgender individuals. Larson suggests that this might be appropriate for researchers exploring people's use of language related to their own conceptions of their own gender. The second is a performative framing which derives from Judith Butler's theory in which **gender is a repetitive practice that adheres to normative conventions which works to further construct that gender identity.** This may be useful in studying "cases where persons of one gender attempt to appropriate conventional communicative practices of another gender without adopting a transgender identity." The third is to think about the gender binary as given from social psychologists' framework to examine — in NLP — consumers and marketing. The paper emphasizes the importance of transparently stating the use of gender and which theory of gender. In part because it destabilizes the validity of the research — since the central terms are not defined properly. **This is useful to us and our project**

[1] In an effort to simplify our computational task we will be reducing our judgement to a binary classifier, based on Butler's basic principles of masculine and feminine performance and presentation. Once we finalize our datasets we will be able to more clearly and explicitly outline how Butler's framework applies to our data and analysis.

in that we are making gender a central question and we need to make clear how we are using gender and why we are using gender.

The 2011 conference paper published by John D. Burger et al., titled "Discriminating Gender on Twitter," lays a good groundwork for the sort of task that we aim to perform. Concentrated on Twitter, due in part to the website's accessible API as well as its widespread popularity across the world, this team scraped a corpus of approximately 213 million tweets from 18.5 million users in a wide variety of languages. The paper describes tweets as a form of "micro blogging," small snippets of free form text. The team also scraped individual profiles, gathering fields such as name, username, description, and any links hosted on a user's profile that could lead to another facet of their online presence. Using links to other blogging platforms with more robust fields for profile information, the team generated a corpus of 184,000 users labeled with gender. The researchers treated gender disambiguation as a binary classification task, using character and word n-grams for n=1 to n=5 as their features. To reduce the size of this processing task, the researchers preprocessed their data to convert each feature pattern into a number, and represented each tweet as a single vector of these vector numbers. To further reduce the complexity of this task, users were condensed into a single vector representing the union of all tweet vectors. Classifiers worked only with these numerical vectors, and at no point were made to decode the numbers back into their alphanumeric origins. The researchers applied multiple classification methods to optimize their methodology: on the development set, Naive Bayes performed at 67.0% accuracy, Balanced Winnow2 at 74.0% accuracy, LIBSVM achieved 71.8% accuracy, but took considerably longer than either other implementation, and hand annotation from Amazon Mechanical Turk (AMT) workers, depending on how it was aggregated, performed with between 60.4-68.7% accuracy. All of these did better than the baseline of 55%, which is naively assuming that all tweets were authored by female users (who constitute just over half of the website's user base at time of this experiment, and for whom 995 of the top 1000 features are more strongly associated). Ultimately, the researchers were able to produce a machine based classifier that performs better than about 95% of the 130 humans presented with the same task. They close with a proposal of future work in similar classification tasks, this time based on other features like age or location.

Mukherjee and Liu's "Improving Gender Classification of Blog Authors" proposes a novel scheme of text classification using a mix of selected features as well as an examination of part of speech (POS) patterns of variable length, which they argue is a more robust approach than n-grams. The researchers applied their research to blogs, which contain more casual speech than formal publication, but contain more robust text than some other forms of social media. They scraped blog posts from blog hosting sites and blog search engines such as blogger.com, technorati.com, and others, producing a data set of 3100 blogs. The researchers manually labelled each blog with the gender of its author based on information posted in the blog profile. The resulting data set consists of about 3100 blogs. The researchers used a wide breadth of features, such F-measure, stylistic features, gender preferential features, factor analysis and word classes, as well as a new class of features based on POS sequence patterns. To avoid making

their experiment intractable, the researchers applied a method of ensemble feature selection (EFS) to produce a subset of features which are the most discriminatory in the classification task. Features were applied values based on either a boolean scheme (for stylistic features like words) or a term-frequency (TF) scheme based on the appearances of the pattern in the POS tagged document. With the features and corpus then identified, the researchers applied three different learning algorithms for their classifier to see which combination worked the best. These three algorithms are: SVM classification, SVM regression, and Naive Bayes. All three performed at or better than comparable classifiers of the time, with SVM classification (with an EFS-selected subset of features) performing the best at an accuracy of 88.56%, a full 10% higher than classification systems proposed by other researchers in this period.

Sobev et al's "Deep learning network models to categorize texts according to author's gender and to identify text sentiment" reports on the efficacy of a deep learning network for both text classification based off gender classification and text sentiment for russian documents. Sobev et al notes that previous research has employed parameters such as frequency of words to help determine gender but thinks that this approach "might not be appropriate to use for corpora of texts of other genres." Indeed, Sobev et. al is wary of previous results being overoptimistic because of "non-stylistic factors such as topic bias in gender that can make the gender detection task easier." As such, they use the RusPersonality Corpus. This corpus consists of "Russian-language texts of different genres, which are samples of a natural written speech (e.g. description of a picture, essays on different topics, etc.) labelled with information on their authors" such as gender. With the data selected, they selected a set of syntactic and morphological features that were "more or less topic-independent." The syntax features were syntactic relations identified in the Russian National Corpus and the morphological features were counting POS tags. Additionally, they used some emotion-based features specific to Russian--the paper they cited for emotion based features was in russian so we are unsure what exactly Sobev et al did. These 141 parameters were then used to train a neural net. This neural net had a score value of 0.86 +/- 0.03 for gender identification and a F1-score of 0.86.

Argamon et al present "a new type of lexical feature for use in stylistic text classification" in their paper "**Stylistic Text Classification Using Functional Lexical Features.**" This stylistic text analysis can be used to identify and profile an author of a given text. To accomplish this task, Argamon et al uses ATMan to tokenize a text, assign tokens a POS, extract lexical units from a text, compute relative frequencies of "semantic attribute values for [a] text," and then use machine learning to compute and generate results of the stylistic text classification. With this new approach to stylistic text classification, Argamon et al also provide a taxonomy that underpins the functional lexical features that they use. This taxonomy and its accompanying method was then run on a corpus of 20 19th century texts with 8 unique authors. This experimental system was able to successfully perform authorship identification of a given chapter of text and gender characterization on the corpus.

In "**How to Use Less Features and Reach Better Performance in Author Gender Identification," Juan Soler Company and Leo Wanner** aim to build an author gender

classification tool that is generalizable enough to use on various types and genres of text, citing as a major issue that many such systems overfit to the style of the training set and as a result cannot retain accuracy on texts of different style. They chose to use roughly 800 *New York Times* Opinions articles as a training set because of the wide range of topics covered, from science to politics to food. They found that the most predictive features identified text as positive or negative (as in sentiment analysis) and identified patriotic language, as well as syntactic analysis and dictionary features. Compared to their baseline bag-of-words, content-based approach, the accuracy of their many-feature system fared 20% better, and the system with fewer features was even more accurate, at 29% better than the baseline. This suggests that content-based approaches cannot suffice for such an analysis.

Montero et al's "**Investigating the Role of Emotion-based Features in Author Gender Classification of Text**" the authors extract emotions as features instead of bag of word features. These authors found that adding this small set of emotion based-features to bag of words positively influences the performance. This was accomplished by mining diary texts for emotional information with the help of a support vector machine. Results show that this combination results in an 80% cross validation accuracy. Moreover, the authors were able to achieve 75% cross validation accuracy when classifying the gender of an author of a blog post. While these types of data aren't within the scope of what we hope to accomplish with our project, the method that Montero et al employs will be useful when conducting our project.

Data & Proposed Analysis:

We are reaching out to Juan Soler Company and Leo Wanner, the authors of "How to Use Less Features and Reach Better Performance in Author Gender Identification" to see if it would be possible to use their corpus of *New York Times* articles annotated with authors' gender. If that is not available, we will instead seek out a corpus through the Library of Congress dataset and combine it with the VIAF directory, which compiles authors from the Library of Congress and other authorities and lists their names, birth sex, and other attributes. We can then make sure our training set has an equal number of texts written by male as by female authors so that we don't bias our system. Additionally, drawing on concerns presented in [6] we will attempt to avoid gendered topics such as romance novels in our corpus.

To approach this problem we will first implement a Naive-Bayes binary classifier to use as our baseline for accuracy. We will then implement a logistic regression or neural classifier with features measuring syntactic dependency, words used, degree of positive/negative connotations of words (sentiment analysis)[2], and other features to see if it is possible to make a significant improvement in our classifier. Following the cues of previous work, we anticipate that we will likely be prioritizing a small and specific set of features to avoid overfitting our model.

Our experiment would then be testing to see if our classifier deviates from our baseline in any significant way, as well as how accurate individual assessments are. Potential stretch goals

[2] Juan Soler Company and Leo Wanner cite these features as most effective for predicting gender in their paper.

include attempting to optimize our classifier as well as analyzing what features would cause gender classification to deviate in this way.

Timeline:

Nov 9-11: Identify, download dataset (Bayliss & Jake); explore prebuilt packages such as nltk (Adriana & Naomi). Write up a notes document/cheat sheet of what was found in the prebuilt packages
Nov 12-18: Figure out the best way to import the data set (Jake & Bayliss); run baseline model and draft code/attempt preliminary analysis (Adriana & Naomi).
Nov 22-29: Fall/Thanksgiving Break
Nov 29-Dec 2: Fine tune analysis (Adriana & Naomi) and prep midpoint presentation (Jake & Bayliss)
Dec 3-Dec 8: Wrap up, start presentation and paper (Adriana & Everyone)
Dec 9: Finish and practise presentation (Jake & Everyone)
Dec 10: Presentation & Have draft of paper (Bayliss & Everyone)
Dec 15: Paper submission (Naomi & Everyone)

Works Cited:

[1] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan, "Stylistic text classification using functional lexical features," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 802–822, 2007, doi: 10.1002/asi.20553.

[2] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating Gender on Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., Jul. 2011, pp. 1301–1309, Accessed: Nov. 08, 2020. [Online]. Available: https://www.aclweb.org/anthology/D11-1120.

[3] J. S. Company and L. Wanner, "How to Use Less Features and Reach Better Performance in Author Gender Identification," http://www.lrec-conf.org/proceedings/lrec2014/pdf/104_Paper.pdf.

[4] B. Larson, "Gender as a Variable in Natural-Language Processing: Ethical Considerations," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain, 2017, pp. 1–11, doi: 10.18653/v1/W17-1601.

[5] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, Oct. 2010, pp. 207–217, Accessed: Nov. 08, 2020. [Online]. Available: https://www.aclweb.org/anthology/D10-1021.

[6] A. Sboev, T. Litvinova, I. Voronina, D. Gudovskikh, and R. Rybka, "Deep Learning Network Models to Categorize Texts According to Author's Gender and to Identify Text Sentiment," in

2016 International Conference on Computational Science and Computational Intelligence (CSCI), Dec. 2016, pp. 1101–1106, doi: 10.1109/CSCI.2016.0210.

[7] C. Suero Montero, M. Munezero, and T. Kakkonen, "Investigating the Role of Emotion-Based Features in Author Gender Classification of Text," in *Computational Linguistics and Intelligent Text Processing*, Berlin, Heidelberg, 2014, pp. 98–114, doi: 10.1007/978-3-642-54903-8_9.