Predicting Gender of Poets with Deep Learning Methods

Samuel J. Mignot
sjmignot@stanford.com

Abstract

Understanding the connections between gender and writing can help explain the way language enforces, impacts, and perpetuates our biases. Poetry is a form that is highly dependent on subtext and the latent features of words. It is also a vastly unexplored textual source in machine learning. The combination of these two features make it an absolutely fascinating medium for deep learning models. This paper tests the efficacy of various neural and machine learning models in predicting the gender of poets. The most effective model tested was a Convolutional Biderectional Long Short Term Memory Model (CBLSTM) that used 50d GloVe word vector embeddings; the model achieved 93.2% accuracy. The trend for gender prediction algorithms for text (and general textual deep learning methods) is towards RNN's and their variations; this is consequent of their ability to capture sequential data. This paper explores the performance of simple RNN's, GRU's, LSTM's, and their more complicated variations. It compares their effectiveness versus simple machine learning methods such as SVM's and Naive-Bayes and earlier computational linguistic models that use n-gram and POS language features.

1 INTRODUCTION

Gender identification through application of machine learning methods is a rapidly developing field. A number of papers have been published in the past few years on gender identification of tweets, blog posts, and longer texts [1][3][5][8][10]. The applications are mainly commercial–the obvious direct application is varied targeting of advertisements based on user profile. However, there are also historic uses for gender classification such as uncovering work by female authors who used male pseudonyms (or attributing anonymous work). In analyzing a poetic corpus this paper seems to remove any commercial application and only provides a slim historic potential. However, this paper has two less obvious main objectives: 1) to bring awareness to the potential and value of poetry as an NLP corpus, 2) to bring awareness and raise questions regarding the impact language has on how we think and interact with the world.

1.1 Poetry as Machine Learning Data

Poetry is a very valuable and vastly unused system for machine learning. Poems are complete small texts that are highly dependent on implicit linguistic features. The essence of poetry is to make words mean more than they do. NLP is about capturing and realizing these subliminal features latent in words and language.

Despite not being commonly discussed poetic features in language are deeply important. Poetry and poetic features appear in our every day lives and impact the way we think and react to language. Two simple but key examples can be seen in pop culture in politics. An example regarding the former involves the frequency with which pop music takes advantage of iambic phrases and meter. An example of the later involves political slogans: the hard 'a' assonance and

meter exploited by Trump's 'Make America Great Again' (or going back further Reagan's 'Let's Make America Great' slogan from which Trump's was plagiarized) . Poetry is neglected data in machine learning but it has a number of very valuable features that could strongly impact the way we think about language.

1.2 Implicit Biases In Language

It is becoming more and more apparent how much of a social construct gender and gender norms are–language greatly impacts this. The goal of this paper is not to demonstrate an implicit differences in a reductive binary gender system. Rather, my intension is to demonstrate the latent differences between male and female poetry that are consequent of societal pressures and the subconscious ways we use language and let it shape us. I had my concerns writing this paper–a fear that the results would be misconstrued.

There is already such a significant bias in how gender in poetry is viewed. The most famous female poets [1] being Sylvia Plath and Emily Dickinson who are both highly emotional, confessional, and contained writers. The most famous male poets being Shakespeare, Whitman [2] who are both unrestrained and exuberant. I hope this paper encourages further research into more interpretable poetic machine learning models that can capture and discuss these biases more explicitly.

2 BACKGROUND

There are nearly no papers on the application of machine learning to poetry (other than a slew of papers on language models based on various poetry corpora [11][12][13]). The only paper considering identification of poetry using neural nets is a 1999 paper by Johan Hoorne titled "Neural Network Identification of Poets Using Letter Sequences" [7]. The paper works to identify the distinct style of two-three different Dutch poets. It compares three different methods: Naive-Bayes classification, k-nearest-neighbor, and a simple Neural Net (with 20-60 hidden units). It also mixes computational linguistic models by employing tri-grams in the learning algorithms. Hoorne's models are dated and are not especially generalizable (especially to the rapid development of modern computational method).

The first main gender classification paper is Mukherjee's which implements POS and a simple SVM systems to achieve 88% accuracy. Since then, there are however a number of papers testing various RNN variations for feature identification in text. Daniel Nguyen implements simple but highly effective linear and logarithmic models for age prediction of twitter users.

A 2014 paper by Zheng and Bartle tests RNN gender classifier; they implement an RCNN and a windowed RCNN and use it to predict gender of blog posts, 19th century novels, and 20th century novels. However, both models do not perform exceedingly well. The RCNN achieves 81%, 67%, and 69% for the three respective genres. The WRCNN performs slightly better at 86%, 73%, and 71%. Their implementations were not able to rival Mukherjee's POS model.

3 APPROACH

3.1 Data Collection

The first 3000 poems were obtained from poetrydb.org–an open source Poetry API. Of these poems 2472 were male and 528 were female. In order to create a balanced dataset the work of a number of female poets was scraped from online poetry databases: the rest of the data was scraped from poemhunter.com. Of the web scraped poems, 1961 were written by women and 50 were written by men. This was done to balance out the data set. The final data set consisted, in its entirety, of a total 2523 poems written by men and 2489 written by women for a total of 5012 samples.

[1] in this author's opinion

[2] once more, in this author's opinion

3.2 Preprocessing

A significant amount of preprocessing was done on the data. Non-ASCII characters were replaced by valid counterparts (when they could be). Punctuation marks were separated from surrounding words. Line-breaks were marked by the token 'lb' and stanza breaks were marked with '@'. Stray punctuation marks and xml tags that found their way into poems were removed. Duplicate poems were also removed. One flaw of the data set was the imbalance between the number of works by different poets. For example, Lord Byron and Emily Dickinson had hundreds of poems in the data set while poets such as Louise Glck had very few.

3.3 Main Model

3.4 CBLSTM

3.4.1 Description and Features

The main Model tested is based on a variation of Chaitanya Joshi's CBLSTM for sentence polarity classification [12]. The model is also an relative of the CLSTM which is described thoroughly by Zhou in his paper "A C-LSTM Neural Network for Text Classification"[18].

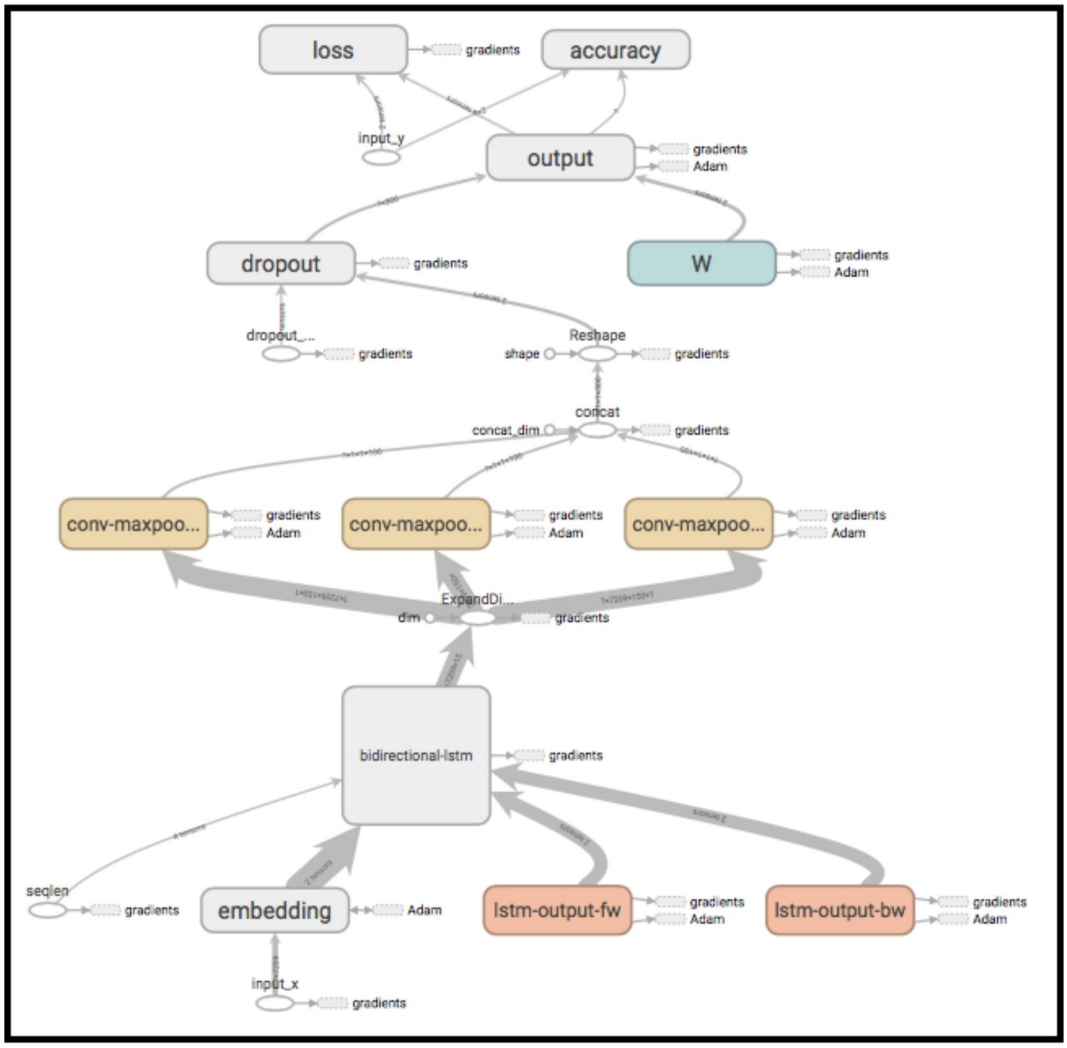

figure. 1 CBLSTM

In textual models that implement both convolutional and RNN layers,t he Convolutional layer serves to extract higher level features from the words and the LSTM layer captures long-term dependencies.

Zhou's CLSTM model runs the words through a Convolutional layer first and then an LSTM. Nearly all documented models that mixed Convolutional and Recurrent layers apply them in that order (this is because the the main use of CRNN seems to involve translation of image to text) [18] [6]. The CBLSTM implemented in this paper operates in the opposite order. It first takes advantage of both forward and backward sequential dependencies by using a Bidirectional-LSTM. The application of a bidirectional LSTM allows words to be impacted both by tokens that come before them and after them.

The result of the BLSTM layer is then splits and passed to three Convolutional maxpool layers.

The model uses Adam optimizers for all back propagation. The model also uses a Softmax output layer and C.E. Loss. Furthermore, a number of regularization techniques are implemented to prevent over-fitting:

3.4.2 Regularization of CBLSTM

L2
The model used L2 regularization. The model performed best with an $\alpha = .3$. L1 regularization was also tried (in conjunction and in replacement of L2). In both cases it was not as effective. Similar tests were done with $L0$ and $L\infty$. The results were the same.

Dropout
Dropout was implemented in the model to prevent over-fitting. A dropout rate of .5 was optimal [17]. Dropout was applied right before the final soft-max and loss calculations. Tests were done applying it in different places but the aforementioned application was the most effective.

Word Embeddings: GloVe
Both random embeddings and GloVe word embeddings were tried for the model. Using GloVe word embeddings did not vastly improve the accuracy but significantly decreased the number of epochs needed to train the model.

4 RESULTS

4.1 Experiment

This paper tested the accuracy of a word embedded CBLTSM versus a Naive-Bayes BOW model, a word-embedded RNN, and a word-embedded GRU (GloVe word embeddings were used for all word-embedded models. The results of each follow:

Accuracy at Poetry Gender Prediction of Machine Learning Models	
MODEL	ACCURACY
Naive-Bayes	83.0 %
SVM	81.2 %
RNN	84.5%
GRU	88.1 %
CBLSTM	93.2 %
POS (on blogs) [1]	88%

4.2 Analysis

4.2.1 Impressive Performance of Simple Models: Naive-Bayes

The simple models were surprisingly effective. This was especially the case for the Bag of Words Naive-Bayes model, which had an accuracy of 83%.This was nearly as effective as the simple RNN.

This is likely a consequence of the reduced importance of word order for poetry, which makes the Bag of Words embeddings significantly more effective than when they are used in prose. There are also definite merits with regard to speed in the use of simpler models (especially since the do not give up too much accuracy). The simple models were trained in around 3 seconds. While some of the more complex models attempted took hours to train. However, of course the time concern only matters once (during the training of the model).

4.2.2 Difficulty in Preventing Overfitting for RNN+GRU model

Attempts to effectively regularize proved incredibly difficult for the simple RNN and GRU models before the implementation of GloVe word embeddings (which were tested in comparision to the main model). This seemed to be a similar issue as the one face by Kim in his Convolutional Neural Net models [15]. Until word embeddings were included the one-hot embeddings combined with a restricted vocabulary (based on the train set) was causing the model to over-fit on features present in the train set.

The word embeddings used were pretrained 50 dimensional GloVe vectors to train embeddings [16]. Unknown words were all assigned to unk embedding. This removed the arbitrariness of creating one-hot encoded vectors from train-data vocabulary.

4.2.3 Effectiveness of RNN Models of Poetry versus Prose

The systems implemented in this paper where consistently more effective then there prose counterparts. A 2014 application of RNN models on text achieved a high of 88% accuracy on blog sites and lower accuracy on both 19th and 20th century prose [18]. This speaks to either increased styalistic variation in the poetry written by different genders or the effectiveness of RNN systems in modeling poetry.

4.2.4 Effectiveness of CBLSTM

Though it took quite a long time to run, the CBLSTM was incredibly effective. It achieved a 93.2% accuracy (and with some extra regularization it could have been even more effective). The effectiveness of the model seems to contradict, to some extent, the impressive performance of the bag of words model. The CBLSTM clearly captures very important latent information regarding the arrangement of words in poems.

5 Possible Sources of Error

5.1 Time-period Differences in data

Collecting a dataset with an equal number of male and female poets was a difficult task due to the unbalanced initial corpus of data, which had 2500 male poems versus 500 female poems. This required a balancing of the data set by web-scraping various poets. However, the majority of female poets with significant amounts of poetry available online (for free) were from the 19th and early 20th century. Consequently in learning between female and male poets the models could have been learning differences in historical language over gendered language.

6 CONCLUSION

Gender classification for poetry seems to be an easier task than gender classification for prose. Simple models work very well and very quickly. More complex RNN models and their variants are highly accurate. The main future explorations that this paper encourages are in questions of visualizing and understanding language bias. In this vein, there are two main directions worthy of further exploration.

6.1 Character Based Language Models

RNN's provide some ability for interpretability–which is the main flaw of Deep Learning methods. A future research direction would involve generating a character based language model on a large poetry corpus and seeing how it picks up unique poetic traits–rhyme, meter, line-breaks. This would be especially interesting if paired with Andrej Karpathy's work on "visualizing predictions and 'neuron firing' in RNN's" [14]. Language models could also be uniquely trained on male and female corpora.

6.2 Poetry GloVe Vectors

GloVe vectors also provide a strong potential for interpretability in Deep Learning methods. Creating GloVe vectors for a very large poetry corpus could expose the unique ways that language is used in poetry versus prose. It could also potentially provide better insight in the biases of language due to the dependence of poetry on subtext. GloVe vectors could also be individually trained on a corpus of female and male poets and the relation between words compared.

Acknowledgments

Thanks to the Richard Socher and the CS224 course staff for what was a very fun and informative class. I close with a poem generated by a character based language-model I trained on the poem corpus I made. It is primed on the characters RNN...

RNN IT But, draigle at soft the orchard likene we follow,
And thus he played her, the gladed breathes, weel of warm:
The step in face or clear rie to sure,
Could Bachhus the Hiven she seen
Through the aught lustre pleguous:
I am ever the weaping, spring there,
That the true and men to dead–anon me!
Be bright had never but love,
But if it then, after frost by the stone,
And former turnhook for deep made
that, his cray to beams –

References

[1] Mukherjee, A., & Liu, B. Improving Gender Classification of Blog Authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[2] Kiperwasser, E., & Goldberg, Y. Simple and accurate dependency parsing using bidirectional lstm feature representations. arXiv preprint arXiv:1603.04351, 2016.

[3] Herring, S. C., & Paolillo, J. C. 2006. Gender and genre variation in weblogs, Journal of Sociolinguis- tics, 10 (4), 439-459.

[4] Sboev, A., et al. 2017 J. Phys.: Conf. Ser. 937 012046

[5] Sboev, A., et al. Deep Learning Neural Nets versus Traditional Machine Learning in Gender Identification of Authors of RusProfiling Texts. Procedia Computer Science, vol. 123, 2018, pp. 424431., doi:10.1016/j.procs.2018.01.065.

[6] Wu, L., Shen, C., & Hengel, A. Personnet: Person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255, 2016.

[7] Hoorn, J.F., Frank, S.L., Kowalczyk, W., & Ham, F.V.D. (1999). Neural net- work identification of poets using letter sequences. Literary and Linguis- tic Computing, 14(3), 311338.

[8] Zheng, R., & et al. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology, vol. 57, no. 3, 2006, pp. 378393., doi:10.1002/asi.20316.

[9] Filho, L., Ahirton,J., Pasti, R. & De Castro, L. (2016). Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. 1025-1034. 10.1007/978-3-319-31232-397.

[10] Nguyen, D., Gravel, R., Trieschnigg D., Meder T.: How Old do You Think I Am? A Study of Language and Age in Twitter. In: Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM), pp. 439–448 (2013)

[11] Colton, S. (2017). Full-FACE Poetry Generation. Available at: http://ccg.doc.gold.ac.uk/papers/colton_iccc12.pdf

[12] Misztal, J. and Indurkhya, B. (2017). Poetry generation system with an emotional personality. 1st ed. Computational Creativity.

[13] Joshi, C. Modelling Context in Word Embeddings. 12 July 2016.

[14] Karpathy, A. The Unreasonable Effectiveness of Recurrent Neural Networks. Hacker's Guide to Neural Networks, 21 May 2015, karpathy.github.io/2015/05/21/rnn-effectiveness/.

[15] Kim, Y. Convolutional neural networks for sentence classification, CoRR, (2014).

[16] Pennington, J., et al. Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, doi:10.3115/v1/d14-1162.