Team: Author Finder

LITERATURE SURVEY

AUTHORSHIP ATTRIBUTION FOR NEURAL TEXT GENERATION

Abstract:

There are tremendous advancements in Natural Language Processing, such as the OpenAI chatbot models. These technologies could be used in many ways to help the society but simultaneously could pose certain risks either with or without guilt. Thus this literature survey would analyze some of the research done till date in order to solve the problem. Three problems would be analyzed, that is if two texts are generated by the same NLG, if text is from a human, and classify the texts with respect to the NLG. We consider some of the Neural Language Generation (NLG) models such as GPT-3 and try researching different Linguistic Features such as POS statistics, LICW and so on.

Keywords: *Natural Language Processing, Neural Language Generation, Linguistic Features.*

1. Introduction:

With the developments in the field of chatbots which could mimic the human language, there's a high ground for the models to be used as part of the deep fakes [1]. Thus this survey multiple researches have analyzed and in order to solve the following three questions.

1. Same or not: Whether the generated text is from the same NLG method (human) or not.
2. Human vs bot: Whether the text is from a human or an NLG method.
3. Which NLG method: Whether the text is from the i^{th} NLG method from a set of k methods.

The above questions are typical binary or multi label classification problems. Hence the papers reviewed would deal with the domain of the classification of text using different Machine Learning, Deep Learning models such as the CNN and so on.

2. Survey:

Different papers were reviewed and presented in the following order.

1. **Experiments with Convolutional Neural Networks (CNN) for Multi-Label Authorship Attribution:** by Dainis Boumber, Yifan Zhang, Arjun Mukherjee.
   Datasets: The MLPA - 400 and PAN - 2012.
   Models: A multi layer CNN model that computes probability distribution if a single label task or the average of the probability distributions in case of a multi label task. To control the overfitting the batch normalization technique has been used. The max-pooling method used to keep the maximum value across the sentence. The ELU activation function used in the hidden layers and the softmax / sigmoid activation at the output layer. The sentences are represented using word2vec and glove embeddings.
   Evaluation: The macro F1 score is 0.736, micro F1 score is 0.744 and the accuracy is 65.3%

2. **<u>Style-aware Neural Model with Application in Authorship Attribution:</u>** by Fereshteh Jafariakinabad, Kien A. Hua.
   <u>Datasets:</u> CCAT10 , CCAT50, BLOGS10, BLOGS50.
   <u>Models:</u> Two models. The first one is the Lexical and Syntactic encoding model where the pre-trained Glove embeddings are used for lexical and POS tags for the syntactic purpose. The second one is the Hierarchical model where the inputs from the first model are taken in by two identical CNN models which use temporal max-pooling. Then the output is fed to a sentence level encoder which gives the semantic/syntactic representation of the document. Both representations are fused. Softmax activation function is used for classifying.
   <u>Evaluation:</u> 90.58% on CCAT10, 82.35% on CCAT50, 72.83% on BLOGS10 & 61.19 on BLOGS50.

3. **<u>Exploring syntactic and semantic features for authorship attribution:</u>** by Haiyan Wu, Zhiqiang Zhang, Qingfeng Wu.

   <u>Datasets:</u> CCAT10 , CCAT50, IMDB62.
   <u>Models:</u> A Multi-Channel Self Attention Network (MCSAN) has been used which has four channels i.e., Word, POS, Phrase and Dependency channels. These features are then combined and fed to a multi-channel self attention model which extracts the contextual information and this is fed to a CNN model with max-pooling layer combined with an Long Short Term Memory (LSTM) model with softmax activation function for the classification. Experimentation has been carried out with BiLSTM and TextCNN also.
   <u>Evaluation:</u> 92.89% on CCAT10, 83.42% on CCAT50 and 89.97% on IMDB62.

4. **<u>Leveraging Discourse Information Effectively for Authorship Attribution:</u>** by Su Wang, Elisa Ferracane, Raymond J. Mooney.
   <u>Datasets:</u> novel - 9, novel - 50, IMDB62.
   <u>Models:</u> The CNN2PV model has GR (Grammatical Relations) and RST discourse relations probabilities. The CNN2DE has the discourse embeddings where CNN2 is the baseline model and it uses an embedding layer, a convolution layer, max-pooling and the softmax activation function for the purpose of classification. The Rhetorical STyle(RST) model performs better than GR as it leverages the weight learning in CNN and also provides more fine-grained features.
   <u>Evaluation:</u> 98.8% on novel - 50 and 92.0% on IMDB62.

5. **An evaluation of authorship attribution using random forests:** by Mahmoud Khonji, Youssef Iraqi, Andrew Jones.
   Datasets: PAN12
   Models: Random forests are known to handle noisy data, frequent characters, frequent ngrams, freq wordlen, freq rewriterules, freq wordshapes have been used since decision trees could handle numerical data well compared to the nominal data. The task is divided into three problems namely A with 3 classes, C with 8 classes but bigger size compared to A and I with 14 classes and much bigger size. At first the features for a given document along with its label is derived and used with WEKA to build a classifier.
   Evaluation: 100% on Prob A, 87.5% on C and 92.9% on I.

6. **Authorship Attribution vs. Adversarial Authorship from a LIWC and Sentiment Analysis Perspective:** by Joshua Gaston, Mina Narayanan, Gerry Dozier, D. Lisa Cothran, Clarissa Arms-Chavez, Marcia Rossi, Michael C. King, Jinsheng Xu.
   Datasets: CASIS - 25.
   Models: Preprocessing the data by getting the tf-idf values and standardizing as well as normalizing them to get unit feature vectors. LIWC provides insight from a psychological point of view. All the 93 output variables are used as the features. These provide word frequencies. Three different models Linear Support Vector Machines(SVM), RBF SVM and Multi-Layer Perceptron(MLP) have been used. Stratified 4 fold cross validation was implemented to get good accuracy.
   Evaluation: 78% on MLP, 71% on RBFSVM and 84% on LSVM.

7. **Topic or Style? Exploring the Most Useful Features for Authorship Attribution:** by Yunita Sari, Mark Stevenson and Andreas Vlachos.
   Datasets: Judgment, CCAT10, CCAT50, IMDB62.
   Models: Three kinds of linguistic features have been selected. They are style, content and both. The style-based are punctuation, digits, words and so on. The content-based are bags of n-grams representing the topical words. The style-based have 174 words and 12 punctuation. Both character and word n-grams are limited to bi and tri. Two models were used i.e., the single Feed-Forward Network(FNN) and Logistic Regression(LR). The style features has two sub groups i.e., lexical and syntactic. For FNN softmax activation function has been used.
   Evaluation: Style: on Judgment 91.07%, on CCAT10 76.00%, on CCAT50 72.72%, on IMDB62 95.93%; Content: on Judgment 91.51%, on CCAT10 76.20%, on CCAT50 72.88%, on IMDB62 95.59%; Hybrid: on Judgment 91.21%, on CCAT10 74.80%, on CCAT50 71.76%, on IMDB62 95.26%.

3. Conclusion:

Thus one could conclude that there is a lot of research that has been put into the Author Attribution field. There are various approaches such as the stylometric approaches i.e., lexical or syntactic models or using content based n-gram methods. One could use the psychological aspect of the topic like in LIWC models. One could also take into account the discourse of the topic such as the RST model or combine all of the above features. Also one could even use different Machine learning or Deep learning models such as the CNN or the RNN or the combination of both. Thus with the experimentation and research the accuracy with which the prediction could be done, can be achieved.

4. References:

1. Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.
2. Dainis Boumber, Yifan Zhang, and Arjun Mukherjee. 2018. Experiments with convolutional neural networks for multi-label authorship attribution. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
3. F. Jafariakinabad and K. A. Hua, "Style-Aware Neural Model with Application in Authorship Attribution," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 325-328, doi: 10.1109/ICMLA.2019.00061.
4. Haiyan Wu, Zhiqiang Zhang, Qingfeng Wu, "Exploring syntactic and semantic features for authorship attribution", Applied Soft Computing, Volume 111, 2021, 107815, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107815.
5. Stamatatos E., A survey of modern authorship attribution methods, *J. Am. Soc. Inf. Sci. Technol.* 60 (3) (2009) 538–556.
6. R. Zhang, Z. Hu, H. Guo, Y. Mao, Syntax encoding with application in authorship attribution, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2742–2753
7. Jafariakinabad F., Tarnpradab S., Hua K.A., *Syntactic recurrent neural network for authorship attribution*, 2019, arXiv preprint arXiv:1902.09723.
8. Q. Li, Z. Li, J.-M. Wei, Y. Gu, A. Jatowt, Z. Yang, A multi-attention based neural network with external knowledge for story ending predicting tasks, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1754–1762.
9. Ruder S., Ghaffari P., Breslin J.G., Character-level and multi-channel convolutional neural networks for large-scale authorship attribution, *CoRR* abs/1609.06686 (2016).
10. Elisa Ferracane, Su Wang, and Raymond Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 584–593.

11. R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms", *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161-168, 2006.
12. M. Khonji, Y. Iraqi and A. Jones, "An evaluation of authorship attribution using random forests," 2015 International Conference on Information and Communication Technology Research (ICTRC), Abu Dhabi, United Arab Emirates, 2015, pp. 68-71, doi: 10.1109/ICTRC.2015.7156423.
13. J. Gaston et al., "Authorship Attribution vs. Adversarial Authorship from a LIWC and Sentiment Analysis Perspective," 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 2018, pp. 920-927, doi: 10.1109/SSCI.2018.8628769.
14. Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
15. Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gomez. 2011. Local Histograms of Character ´ N-grams for Authorship Attribution. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT ’11, pages 288–298, Stroudsburg, PA, USA. Association for Computational Linguistics.
16. Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. ACM Transactions on Information Systems (TOIS), 26(2).
17. Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2013. Authorship Attribution with Topic Models. Journal Computational Linguistics, 40(2):269–310.