Text and Multimedia Mining Project Report
Blog Authorship Attribution

Laurens Kuiper (s4467299)

November 21, 2018

Abstract

This report presents results of using authorship attribution methods on blog data. Blogs are regularly updated web pages, typically run by individuals, in an informational or conversational style. Scalability limitations of conventional methods were tested on selections of blogs from 50-150 authors from a publicly available dataset, and an ablative study on the extracted features was performed. The selections were constructed carefully as to not include authors with a low amount of blogs, or a low median blog length. Finally, the performance of a tree boosting classifier in combination with the conventional methods, inspired by top submissions in various Kaggle competitions, was tested to verify whether this can improve authorship attribution results.

1 Introduction

a[5] b[1] c[3] d[6] e[2] a[4]

2 Resources

3 Methods

4 Results

5 Conclusion

References

[1] Hitschler, Julian and van den Berg, Esther and Rehbein, Ines. Authorship Attribution with Convolutional Neural Networks and POS-Eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58. Association for Computational Linguistics, 2017.

[2] Kapočiūtė-Dzikienė, Jurgita and Šarkutė, Ligita and Utka, Andrius. The Effect of Author Set Size in Authorship Attribution for Lithuanian. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* , pages 87–96. Linköping University Electronic Press, Sweden, 2015.

[3] Sari, Yunita and Vlachos, Andreas and Stevenson, Mark. Continuous N-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273. Association for Computational Linguistics, 2017.

[4] Schwartz, Roy and Tsur, Oren and Rappoport, Ari and Koppel, Moshe. Authorship Attribution of Micro-Messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891. Association for Computational Linguistics, 2013.

[5] Solorio, Thamar and Hasan, Ragib and Mizan, Mainul. A Case Study of Sockpuppet Detection in Wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 59–68. Association for Computational Linguistics, 2013.

[6] Vajjala, Sowmya and Banerjee, Sagnik. A study of N-gram and Embedding Representations for Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–248. Association for Computational Linguistics, 2017.