Secure Systems Group, Aalto University

Tommi Gröndahl

Stylometry and information security: A survey of current methods and future prospects

Detecting autorship

- Author verification: providing evidence for or against the claim that two documents are produced by the same author[6]
- Detecting doppelgänger accounts[2]
- Uncovering users behind troll accounts[5]

Detecting trolls, threats and hate-speech

- Classifying a message as trolling/bullying[5][14][16]
- Discovering linguistic commonalities between hate-speech/threat comments

Detecting deception

- Identifying stylistic trends in deceitful language[9][15]
- Detecting fake reviews[13][17]
- Future research: classifying fake news
- Challenge: linguistic cues for deception seem largely domain-specific.

The deanonymization attack

- Identifying authors against their will[12][3][1]
- Potential victims:
  - critics of oppressive governments
  - authors of anonymous complaints[3]

Mitigating the deanonymization attack via style obfuscation

- Changing the text so that meaning is retained but author identification fails
- Methods: manual[1][3], computer-assisted[11], automatic
- Automatic methods:
  - back-and-forth machine translation[1][10][4]
  - lexical and phrasal replacements[8]
  - grammatical changes[7]
- Future research: utilizing automatic paraphrasing and text simplification methods for style obfuscation
- Challenge: how to robustly measure readability and semantic proximity to the original?

References:

[1] M. Almishari, E. Oguz, and G. Tsudik. Fighting Authorship Linkability with Crowdsourcing. In A. Sala, A. Goel, and K. Gummadi, editors, Proceedings of the second ACM conference on Online social networks, pages 69–82, 2014.

[2] S. Afroz, A. Caliskan-Islam, A. Stolerman, R. Greenstadt, and D. McCoy. Doppelgänger finder: Taking stylometry to the underground. In IEEE Symposium on Security and Privacy, 2014

[3] M. Brennan, S. Afroz, and R. Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security, 15(3), 2011.

[4] S. Day, J. Brown, Z. Thomas, I. Gregory, L. Bass, and G. Dozier. Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering. In 25th International Conference on Computer Communication and Networks (ICCCN), 2016.

[5] P. Galn-Garcia, J. de la Puerta, C. den Gmez, I. Santos, and P. Garca Bringas. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In I. Herrero, B. Baruque, F. Klett, A. Abraham, V. Snel, A.C.P.L.F. de Carvalho, P. Garca Bringas, I. Zelinka, H. Quintin, and E. Corchado, editors, International Joint Conference of Advances in Intelligent Systems and Computing, volume 239, 419–428, 2014.

[6]] P. Juola. Stylometry and immigration: A case study. Journal of Law and Policy, 21(2):287–298, 2013.

[7] F. Khosmood and Robert Levinson. Toward automated stylistic transformation of natural language text. In Proceedings of the Digital Humanities, pages 177–181, 2009.

[8] F. Khosmood and R. Levinson. Automatic synonym and phrase replacement show promise for style transformation. In Sorin Draghici, Taghi M. Khoshgoftaar, Vasile Palade, Witold Pedrycz, M. Arif Wani, and Xingquan Zhu, editors, The Ninth International Conference on Machine Learning and Applications, 2010

[9] M. Louwerse, K.I. Lin, A. Drescher, and G. Semin. Linguistic cues predict fraudulent events in a corporate social network. In Proceedings of the 32 Annual Conference of the Cognitive Science Society, 961–966, 2010.

[10] N. Mack, J. Bowers, H. Williams, G. Dozier, and J. Shelton. The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation. International Journal of Machine Learning and Computing,
5(5):409–413, 2015

[11] A.W.E. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt. Use fewer instances of the letter i: Toward writing style anonymization. In S. Fischer-Hübner and M. Wright, editors, Privacy Enhancing Technologies. Volume 7384 of Lecture Notes in Computer Science, 299–318. 2012.

[12] A. Narayanan, H. Paskov, N. Zhenqiang Gong, J. Bethencourt, E. Stefanov, E.C.R. Shin, and D. Song. On the feasibility of internet-scale author identification. In Proc. 2012 IEEE Symposium on Security and Privacy, 300–314, 2012.

[13] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 309–319, 2011.

[14] C.W. Seah, H.L. Chieu, K.M.A. Chai, L. Teow, and L.W. Yeong. Troll Detection by Domain-Adapting Sentiment Analysis. In Proceedings of the 18th International Conference on Information Fusion, 792–799, 2015.

[15] Toma CL and Hancock JT. What lies beneath: The linguistic traces of deception in online dating profiles. Journal of Communication, 62:78–97, 2012.

[16] J.-M. Xu, X. Zhu, and A. Bellmore. Fast learning for sentiment analysis on bullying. In Proceedings of the International Workshop on Issues of Sentiment Discovery and Opinion Mining, 1–6, New York, 2012.

[17] Y. Xu, B. Shi, W. Tian, and W. Lam. A unified model for unsupervised opinion spamming detection incorporating text generality. In Proceedings of the 24th International Conference on Artificial Intelligence, 725–731, 2015.

Aalto University