Measuring meaning, capturing personality? Auditing language-assessing algorithms

Cornelius Puschmann, Zeppelin University/Humboldt Institute for Internet & Society
Marco Bastos, University of California, Davis
Wiebke Loosen, Hans Bredow Institute for Media Research, Hamburg
Jan-Hinrik Schmidt, Hans Bredow Institute for Media Research, Hamburg

Does my punctuation give me away as neurotic, or is my overuse of the first person pronoun a sign of depression? How do the algorithms behind common social media platforms interpret experimental poetry, popular song lyrics, or switching between languages? What happens to artistic literary style or computer-generated newswriting when they are thrown into the clinical bag-of-words blender? We propose an algorithm audit consisting of testing a wide range of pre-assembled corpus materials on common social media platforms. Compiling samples from different genres should not be difficult, neither should be chopping them up and 'feeding' the material to selected algorithms, though automating this process could be more challenging. Both computationally generated text and unusual natural material in non-standard linguistic varieties could be used for this purpose.

Unprecedented volumes of textual data are produced in social media services each day. From status updates and tweets to product reviews and Wikipedia entries, this material is no longer read just by humans, but summarized, categorized, mined, and scored by a wide range of language-assessing algorithms (LAAs). Computational techniques range from traditional approaches such as latent semantic indexing (Deerwester et al, 1990) to (conceptually) more novel procedures such as sentiment analysis (Pang & Lee, 2008). Increasingly, procedures such as Linguistic Inquiry and Word Count (LIWC) are used not just to classify documents, but also to evaluate users based on correlating word usage with psychological well-being (Pennebaker, Booth, & Francis, 2007). LIWC's origins lie in clinical psychology, and originally the approach was tested using diaries and other traditional written genres, rather than the short texts that today dominate social media (Argamon et al, 2007). Such issues complicate the increasingly popular integration of LIWC into products such as the Facebook News Feed (e.g. Kramer, Guillory, & Hancock, 2014).

The application of LIWC and other psychological measures to textual data is a novelty both in terms of the mechanisms such methods employ (e.g. the combination of bag-of-words approaches with machine learning techniques), and the degree to which they operationalize abstract functional or psychological categories (the relevance of a review, the well-being of a user) through particular signals, such as words, punctuation, or use of emoticons (Campbell & Pennebaker, 2003; Gill, 2011). In contrast to areas such as credit scoring, interpreting texts and making inferences about the psyche of authors has historically been the domain of the humanities, rather than computer science. Language analysis, whether manual or computational, is subject to a number of potential problems, ranging from the polysemy of words and phrases and the usage of irony or sarcasm, to issues such as ellipsis (leaving out information considered given), or the complex contextual knowledge needed for interpretation. The channel and participant structure of the discourse also play pivotal roles.

Algorithmically associating language use with behavior, ethnicity, or socioeconomic status is problematic in a number of ways though, admittedly, humans have been doing this for thousands of years. Such procedures are also dangerous because they were never developed for the linguistic diversity they now encounter in social media spaces. Dialectal features, code-switching between multiple languages, and countless other idiosyncrasies not anticipated by computational linguistics throw up a long list of questions, not least whether it is legally discriminatory to classify on the basis of language. Another tension is the operationalization of communication as behavior. The original texts that LIWC was developed for had a very different communicative purpose than the genres to which it is now increasingly applied. The idea that LAAs can reveal hidden issues of which speakers are themselves not aware turns language into a diagnostic, something that in Goffman's terms is *given off,* rather than deliberately *given*. The view of algorithms as overhearers deserves more scholarly attention, in addition to the more practical issue of how to conduct LAA audits.

Workshop goals

1. Practical aims in relation to algorithm auditing:
    a. Share useful tools for conducting algorithm audits;
    b. Establish a resource to document algorithm audit recipes in the form of lab notes;
    c. Create a database for describing and sharing audit results;
2. Discuss ways of comparing audit results transnationally and across languages;
3. Trace the underlying social assumptions made by algorithms and examine how they operationalize these assumptions in practice.

We are very interested in applying the above-mentioned tools and recipes to the algorithmic scoring of news. Together, we are currently preparing a project to be situated at the Hans-Bredow-Institute for Media Research in Hamburg that will explore the role of algorithms in creating new types of preferential publics. Marco has noticed considerable differences across countries and languages when examining what information is presented to dummy Twitter users in of different nationalities that he created for his research on local differences in social news consumption. The "default news feed" offered by Twitter clearly varies from country to country, not just in terms of news items (which obviously are bound to vary strongly), but also in terms of editorial emphasis. It would be great to conduct a larger transatlantic collaboration aimed at assessing how Facebook and Twitter score news via accounts in North America and Europe. In Germany, the Volkswagen Foundation[1] currently hosts a program that seeks to support collaborations of journalists and scholars, which could be useful for such an effort, though perhaps the focus on journalism is too narrow for the workshop.

References

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2007). Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday*, *12*(9).

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. doi:10.1111/1467-9280.01419

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Gill, A. J. (2011). Personality and language in computer-mediated communication. *Information Design Journal*, *19*(3), 250–257. doi:10.1075/idj.19.3.06gil

Goffman, E. (1959 [2007]). The presentation of self in everyday life. In J. M. Henslin (Ed.), *Down to earth sociology: Introductory readings* (pp. 135–146). New York: Simon & Schuster.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. doi:10.1073/pnas.1320040111

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. doi:10.1561/1500000011

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC2007*. Austin, TX. Retrieved from http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual.pdf

[1] See http://www.volkswagenstiftung.de/en/funding/communicating-science-and-research/science-and-data-driven-journalism.html