Developing a rule-based method for identifying researchers on Twitter: The case of vaccine discussions

Björn Ekström[1]

[1] *bjorn.ekstrom@hb.se*
University of Borås, The Swedish School of Library and Information Science, Allégatan 1, 503 32, Borås, (Sweden)

Introduction

This study seeks to develop a method for identifying the occurrences and proportions of researchers, media and other professionals active in Twitter discussions. As a case example, a dataset from Twitter vaccine discussions is used. The study proposes a method of using keywords as strings within lists to identify classes from user biographies. This provides a way to apply multiple classification principles to a set of Twitter biographies using semantic rules through the Python programming language.

Theory

The theoretical outline is based on rule-based text classification. As described by Glushko (2013, 374), a rule-based system can serve to separate words in terms of tokenization, where textual components are divided using spaces, and stemming, where terms are derived to their word stems. While the rule-based process provides domain-based classification, issues may occur with regards to how punctuation complicates tokenization and how semantic ambivalence can occur from incorrect stemming.

Method

9 647 plain text biographies from Twitter profiles engaged in discussions related to vaccines are studied as a prominent case example. The case dataset is provided through the research project Data for Impact. The method includes a qualitative content rule-based analysis process using the Python programming language and data wrangling software OpenRefine where patterns within the biographies are set to correspond to predefined classes. A set of keywords as strings within lists are represented by variables. Each variable is then matched against the biographies as plain text and returns one of the predefined classes if any of the strings are present.

Strings used to identify biographies are influenced by and partially reused from previous studies (Côté and Darling 2018; Vainio and Holmberg 2017), although amended in order to suit the nature of the biographies used as a dataset in this study. As discussed by Patton (2015), the identification process is performed by working back and forth between the classes and the data in order to verify accuracy. Eleven types of classes are used, as described in Table 1, corresponding with a set of keywords. The class *General public* is used when the biographies does not match any class. Twitter profiles lacking biographies are classed as *Unknown*. Users can also belong to more than one class. Spelling variations are used where needed.

Table 1. Classes, keywords and biography extracts.

Class	**Keyword example**	**Biography extract example**
Science student	student, phd student, phd candidate	[City] University [discipline] Student
Graduated	MS, MA, graduate	[…] Engineering graduate. […]
University faculty	lectur, prof., professor	Professor of [discipline], teaches [subjects].
Other scientist or science-associated group	technician, lab manager, ologist	[…] biologist […]
Education and outreach professionals	curator, teacher, librarian	Language teacher [subject]
Applied science organization	nonprofit, policy officer	[…], nonprofit board member […]
Other professional	recruiter, entrepreneur, manager	Entrepreneur, marketer […]
Media professional	journalis, corresponden, publisher	correspondent for [media outlet]
Policy/decision maker	congressman, senator, parliament	District […] Congressman [year span]
General public		
Unknown		-

Findings

The findings of the classification and their occurrences are presented in Table 2 below.

Table 2. Occurrences and proportions of classes.

Class	No.	% (out of 10255 classes)
Science student	165	1.61 %
Graduated	58	0.57 %
University faculty	191	1.86 %
Other scientist or science-associated group	394	3.84 %
Education and outreach professional	283	2.76 %
Applied science organization	56	0.55 %
Other professional	704	6.86 %
Media professional	1127	10.99 %
Policy/decision maker	23	0.22 %
General public	7188	70.09 %
Unknown	66	0.64 %
Total	**10255**	

As per this case example, academic professionals, organizations and students are engaged to the following extent and order in relation to the total number of classes identified (10 255): *Other scientist or science-associated group* (3.84 %), *Education and outreach* professional (2.76 %), *University faculty* (1.86 %), *Science student* (1.61 %), *Graduated* (0.57 %), *Applied science organization* (0.55 %). The proportion of *media professionals* amounts to approximately a tenth of all classes (10.99 %) while the class of *other professionals* such as recruiters, entrepreneurs and managers amounts slightly lower (6.86 %). A substantial share of the Twitter profiles engaged (70.09 %) does not belong to any of the professionally related classes but rather belongs to the class of *General public.* The classes of *Policy/decision makers* and *Unknown* relates to small proportions (0.22 % and 0.64 % respectively).

The method does provide a certain error margin when examining the outcome through close-reading. For instance, a biography simply mentioning the word "scientist" may be classed as *University faculty.* Although tendencies can be examined on the occurrences and proportions of academic and media voices in the Twitter vaccine discussion, the biographies' free form provides some classification noise.

Conclusion

The rule-based classification process presented provides a method of identifying the occurrences and proportions of researchers and other professionals engaged in discussions related to vaccines based on a set of predefined rules. Keywords as strings within lists are matched to user biographies collected from Twitter. The study has proven to give the project an indication of the relevant share of the collected data. Of these, 7.88 % are academic (class 1 - 4), 3.31 % are academically related (class 5 - 6) and 10.99 % are media related (class 8). 7.08 % consist of other classes (class 7 + 9). 70.09 % are classed as the *General public* (class 10) and 0.64 % are classed as *Unknown* (class 11).

While prior studies have used search-and-replace methods through regular expressions, the method proposed provides a way to apply multiple classification principles to a set of Twitter profile biographies using the Python programming language and data wrangling software OpenRefine. This enables a better understanding of the the occurrences and proportions of researchers as well as other professionals being present in Twitter discussions. Future studies on new classification methods with regards also to natural language processing are needed in order to further develop such methods.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 770531. Special thanks to Ettore Rizza – PhD Student in Information Sciences & Technologies at Université libre de Bruxelles, Belgium – for programming advice.

References

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS.* https://doi.org/10.1139/facets-2018-0002

Glushko, R. J. (2013). *The Discipline of Organizing*. Cambridge, Massachusetts: The MIT Press.

Patton, M. Q. (2015). *Qualitative research & evaluation methods : integrating theory and practice*. Thousand Oaks, California: SAGE Publications, Inc.

Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics*, 112(1), 345–366. https://doi.org/10.1007/s11192-017-2368-0