Dimension of Twitter Trolling: Short Text Classification Using Multiple Correspondence Analysis

Isobelle Clarke (Aston University, UK)

Introduction

Despite the growing body of research on internet trolling (e.g. Donath, 1999; Herring, 2002; Shachaf & Hara, 2010; Hardaker, 2010; 2013; Whelan, 2013; Coles & West, 2016), still relatively little is known about the extent of its repertoires and linguistic properties. This paper reports on the findings of a project using Multiple Correspondence Analysis (MCA) to examine functional linguistic variation in 720 examples of Twitter trolling, with the aim to linguistically distinguish its different types.

Research on multi-dimensional text type analysis (MDA) shows that linguistic features will tend to co-occur in texts which are functionally and/or situationally similar (e.g. Biber, 1989). Typically, MDA takes the relative frequencies of many grammatical features from a number of texts in a language variety and subjects them to a factor analysis. The dimensions revealed from this are interpreted functionally and are then used to cluster the texts into distinct types. Due to the reliance on relative frequencies of grammatical features, most studies employing MDA have dealt with long texts (e.g. Biber, 1988) because it is difficult to accurately estimate the relative frequencies of grammatical features in short texts (Bijhold et al., 2010). Tweets are restricted to 140 characters in length (typically fewer than 30 words), which means that the relative frequencies of features are likely to be inaccurate. One way to deal with short texts is to concatenate them (e.g. Passonneau et al., 2014). However, this is not useful for identifying the functional linguistic variation between texts.

Instead, this study applies a new form of categorical MDA based on an MCA of the simple occurrence of a variety of lexical and grammatical forms in individual Tweets.

Methodology

'Trolling' is used in multiple contexts and describes numerous behaviours. This means that identifying examples of trolling is challenging, and largely depends on how the researcher defines it. Based on the understanding that words gain meaning through their use, I adopt Mihaylov and Nakov's (2016: 403) definition of 'trolling': "those that have been called such by other people", and use this to identify examples. Hence, if something is labelled as trolling, I take it to be such because each use of trolling contributes to its meaning. With this definition, the most inclusive approach for data collection would be to search for 'trolling' in Tweets, detect accusations and proceed to identify instances. However, this approach is labour-intensive and subjective as in multi-message and multi-user discussions, the researcher must decide what post was accused of trolling. To avoid this, I selected the imperative "stop trolling" as one of many possible search strings because, as a directive to stop the current behaviour, it is responsive, suggesting that Tweets prior to this instruction are instances of trolling. Using this search string, 720 Tweets were manually collected by extracting the posts which "stop trolling" was in reply to, or if the "stop trolling" Tweet quoted another users Tweet, then the quoted Tweet was collected. For the latter, the quoted Tweet alongside "stop trolling" suggested that this quoted post was trolling. Following data collection, all of the Tweets were tagged using a Twitter Part-of-Speech (PoS) tagger developed by Gimpel et al. (2011).

Based on the tagged corpus, I then automatically identified occurrences of 86 features in the Tweets. These features are based on basic parts-of-speech, grammatical constructions (Biber, 1988), and additional features specific to trolling (Hardaker, 2013). This resulted in an 86 feature by 720 Tweet binary data matrix. Features occurring in less than 5 percent of the Tweets were removed, resulting in 62 linguistic features. Subsequently, MCA was performed on this data matrix in R using FactoMineR (Husson et al., 2017).

MCA is essentially a dimension reduction method, which aims to represent high dimensional categorical data into a low dimensional space. MCA is predominantly used to analyse questionnaire and survey data, however it has been used for linguistic purposes (e.g. Tummers et al., 2012; Glynn, 2009; 2014). In this study, each linguistic feature has two categories (i.e. presence and absence). The MCA assigns each category a positive or negative coordinate and a value indicating its contribution to the dimension (Le Roux & Rouanet, 2010). The MCA also returns a positive or negative coordinate to each Tweet on each dimension, which can be plotted to visualise the relationship between Tweets. Following Le Roux and Rouanet (2010), each dimension was interpreted by considering the variables whose contributions were above 0.81, the average contribution of a feature (100/124). Subsequently, Tweets with high positive and negative coordinates on each dimension were analysed to check and refine the functional interpretation. Finally, each Tweet's dimension coordinates were plotted to examine if there were distinct types of trolling.

Results

The MCA was used to return three dimensions because they were readily interpretable and subsequent dimensions explained a limited amount of variance. The features most strongly contributing to these dimensions are presented in Table 1.

Dimension 1

Because the relative frequencies of features were not taken into account, the MCA may classify Tweets by length because typically more words means more features. To test this, each Tweet's dimension coordinates were correlated to Tweet length. This revealed that Dimension 1 is strongly positively correlated to Tweet length (*r* =0.74), Dimension 2 is moderately positively correlated (*r* =0.35) and Dimension 3 is weakly positively correlated (*r* =0.19). A closer examination of the linguistic features strongly contributing to Dimension 1 (Table 1) supported this interpretation as positive coordinates were assigned to the presence of features, whilst negative coordinates were assigned to the absence of features. For this reason, Dimension 1 is excluded from further analysis.

Dimension 2

The linguistic features strongly contributing to negative Dimension 2 have an interactive function. *Second person pronouns* suggest that the writer is interacting with a specific person. *Question marks* are associated with interaction as they indicate a question is being asked, and *interjections* are inherently interactive because they are immediate responses to stimuli.

Alternatively, those features strongly contributing to positive Dimension 2 are associated with informationally dense Tweets. *Prepositions*, *numerals*, *quantifiers*, *proper nouns* and *attributive adjectives* are used to provide specific detail. *Nominalisations* are indicative of a high informational load (Biber, 1988). The fact that Dimension 2 has a moderately positive correlation to Tweet length supports this interpretation because informational Tweets tend to be greater in length.

Dimension 2 can therefore be seen to reflect Biber's (1988: 107) "Informational versus Involved Production" dimension. This is supported with examples strongly associated to this dimension.

Table 1: The features strongly contributing to the Dimensions.

Dim		Features
1	+	WH-pronouns, Nominalisations, Prepositions, Past tense, *be* as a main verb, Other adverbs, Public verbs, Determiners, Amplifiers, Auxiliary *be*, Coordinating conjunctions, Quantifiers, Second person pronouns, Analytic negation, Third person pronouns, Other pronouns, Infinitives, First person pronouns, Subject Pronouns, WH-words, Possessive pronouns, Prediction modals, Contrastive conjunctions, Auxiliary *do*, It, Verbs of perception, Object pronouns, Private verbs, Accusative case, Conditionals.
	-	Absence of Nouns, absence of Prepositions, absence of Subject pronouns, absence of Other pronouns, absence of Past tense, absence of First person pronouns, absence of Second person pronouns, absence of Determiners, absence of Other adverbs, absence of Accusative case, absence of Private verbs, absence of Nominalisations.
2	+	Nouns, absence of Accusative case, Prepositions, Attributive adjectives, Determiners, Past tense, absence of Second person pronouns, Articles, Predicative adjectives, *be* as a main verb, Proper nouns, absence of Other pronouns, Quantifiers, absence of Mentioning, Capitalisation, Nominalisation, Comparatives, Hashtags, Numerals, Quoting, Superlatives, Cause subordinators.
	-	Absence of Nouns, Accusative case, Question marks, Second person pronouns, absence of Prepositions, Other pronouns, absence of Attributive Adjectives, Interjections, absence of Proper nouns, absence of Articles, Mentioning, absence of Past tense, Nominative case, absence of *be* as a main verb, absence of Nominalisations.
3	+	WH-pronouns, absence of Predicative adjectives, absence of First person pronouns, Other pronouns, Past tense, absence of *be* as a main verb, absence of Subject pronouns, Second person pronouns, Hashtags, Nominalisations, Public verbs, Auxiliary *be*, Numerals, Time subordinators, Place adverbials, Suasive verbs, Verb-initial, Perfect tense, Question marks.
	-	Predicative adjectives, absence of Nouns, Comparatives, *be* as a main verb, Subject pronouns, First person pronouns, Auxiliary *do*, Contrastive conjunctions, absence of Prepositions, Analytic negation, absence of Past tense, absence of Second person pronouns, Nominative case, absence of Question marks, absence of Other pronouns, absence of Nominalisation.

Dimension 3

The features strongly contributing to negative Dimension 3 have an attitudinal function. *Be* as a main verb, *predicative adjectives* and *comparatives* function to express attitudes towards specific things. The co-occurrence of *first person pronouns* suggests that a personal opinion is conveyed. This interpretation is supported by Examples 1 and 2, which are Tweets strongly associated to negative Dimension 3. Both Tweets express a personal opinion.

(1) @username I don't know letoya but she's better than Beyonce
(2) Rogue One was so bad I don't know if I ever want to watch movies again

By contrast, the features strongly contributing to positive Dimension 3 have an antagonistic function. *Second person pronouns* suggests that the Tweets are targeted. The co-occurrence of *initial verbs*, *question* marks, and *public verbs* indicate that someone's speech is being brought into question. Additionally, *place adverbials*, *time subordinators* and *nominalisations* suggests that there is a high degree of specificity. This interpretation is supported with Examples 3 and 4, which are Tweets strongly associated to positive Dimension 3. These Tweets contain questions antagonistically directed to users.

(3) @username Speaking of which, y'all are sleeping w/ Russia so much you've got bed sores. When's he resigning?
(4) Probably from the hundreds of thousands of businesses and entertainers that produce promo shirts? Are you suggesting someone invented that?

In summary, this dimension represents Tweets functioning to express attitudinal judgement versus antagonistic Tweets. This dimension is in line with Hardaker's (2013) and Merritt's (2012) descriptions of trolling behaviours, specifically that trolls post provocative content and are inherently hostile.

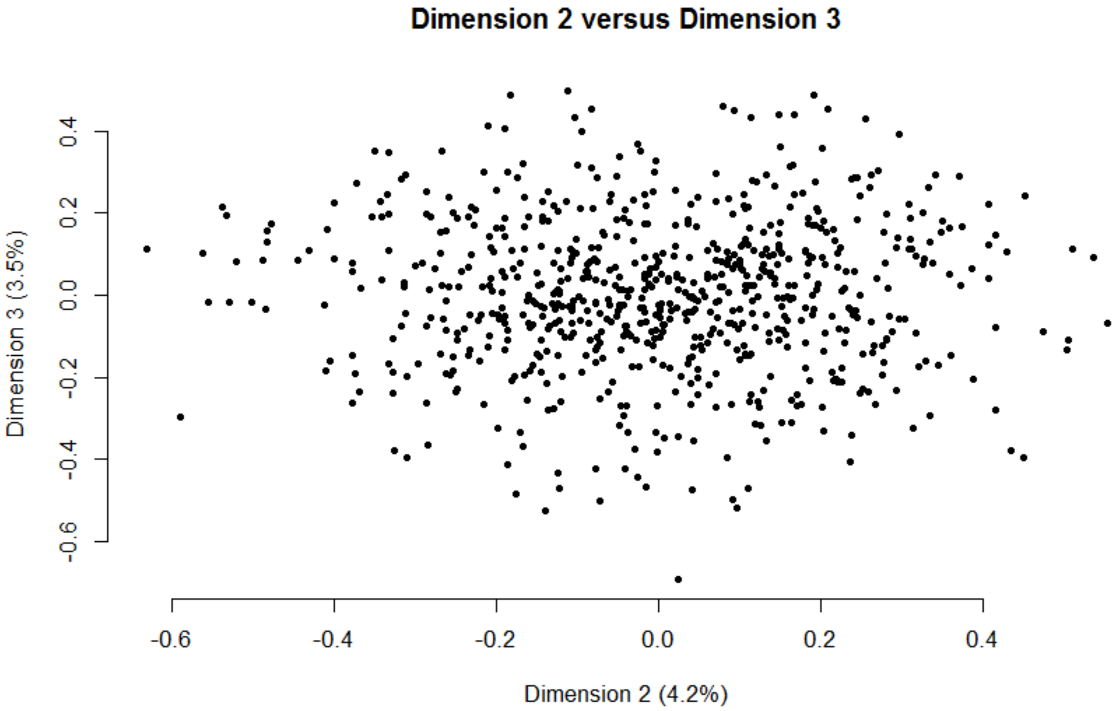
Using R, each text's Dimension 2 and Dimension 3 coordinates were plotted on a map to show where each Tweet lies in relation to these dimensions (see Figure 1). Figure 1 shows that there are no clear clusters, but rather there is a continuous range of linguistic variation.

Conclusion

Based on this analysis, two dimensions of linguistic variation have been identified: namely, *interactive* versus *informational*, and *attitudinal* versus *antagonistic*. The former dimension echoes Biber's (1988) dimension, whilst the latter reflects Hardaker's (2013) and Merritt's (2012) definitions that trolling can be provocative and hostile. By plotting each Tweet's Dimension coordinates, it is possible to see the continuous range of linguistic variation of trolling.

Due to manual extraction, only 720 examples were collected, which is, by today's standards, a comparatively small corpus. Therefore, future research will work on automating this process with other efficient search strings so that more dimensions of linguistic variation can be revealed.

Figure 1: Plot of each Tweet's Dimension 2 and Dimension 3 coordinates.



References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press

Biber, D. (1989). "A typology of English texts". *Linguistics* 27: 3-43

Bijhold, J., Ruifrok, A., Jessen, M., Geradts, Z., Ehrhardt, S. & Alberink, I. (2010). "Forensic Audio and Visual Evidence" In N. N. Daeid & M. Houck (Eds.) *Interpol's Forensic Science Review*, pp. 353-392. Boca Raton, FL: CRC Press Taylor & Francis Group

Coles, B. A. & West, M. (2016). "Trolling the trolls: Online forum users constructions of the nature and properties of trolling" *Computers in Human Behavior* 60: 233-244

Donath, J. S. (1999). "Identity and deception in the virtual community" In M. A. Smith & P. Kollock (Eds.) *Communities in cyberspace*, pp. 29-59. London: Routledge

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J & Smith, N. A. (2011). "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments" In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pp. 19-24. Portland, Oregon: June 19-24, 2011

Glynn, D. (2009). "Polysemy, syntax, and variation: A usage-based method for Cognitive Semantics" In V. Evans & S. Pourcel (Eds.) *New directions in Cognitive Linguistics*, pp.77-106. Amsterdam: John Benjamins

Glynn, D. (2014). "Correspondence analysis: Exploring data and identifying patterns" In D. Glynn & J. A. Robinson (Eds.) *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, pp.443-485. Amsterdam: John Benjamins

Hardaker, C. (2010). "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions". *Journal of Politeness Research* 6(2): 215-242

Hardaker, C. (2013). ""Uh....not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug." An overview of trolling strategies" *Journal of Language Aggression and Conflict* 1(1): 58-86

Herring, S. C., Job-Sluder, K., Scheckler, R. & Barab, S. (2002). "Searching for Safety Online: Managing "Trolling" In a Feminist Forum". *The Information Society* 18: 371-384

Husson, F. & Josse, J. (2014). "Multiple Correspondence Analysis" In J. Blasius & M. Greenacre (Eds.) *Visualization and Verbalization of Data*, pp. 165-184. London: Chapman & Hall/CRC

Le Roux, B. & Rouanet, H. (2010). *Multiple Correspondence Analysis* California: SAGE Publications, Inc.

Merritt, E. R. (2012). "An Analysis of the Discourse of Internet Trolling: A Case Study of Reddit.com". *Unpublished PhD*. South Hadley, MA: Mount Holyoke College

Mihaylov, T. & Nakov, P. (2016). "Hunting for Troll Comments in News Community Forums" In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 399-405. Berlin, Germany: August 7-12, 2016

Passonneau, R. J., Ide, N., Su, S., & Stuart, J. (2014). "Biber Redux: Reconsidering Dimensions of Variation in American English" In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 565-576. Dublin, Ireland: August 23-29, 2014

Shachaf, P. & Hara, N. (2010). "Beyond Vandalism: Wikipedia Trolls" *Journal of Information Science* 36(3): 357-370

Tummers, J., Speelman, D. & Geeraerts, D. (2012). "Multiple Correspondence Analysis as Heuristic Tool to Unveil Confounding Variables in Corpus Linguistics" In *Proceedings of the 11th International Conference on Statistical Analysis of Textual Data*, pp. 923-936. Liege, Belgium: June 13-15, 2016

Whelan, A. M. (2013). "Even with cruise control you still have to steer: defining trolling to get things done" *Fibreculture Journal: Internet Theory Criticism Research* 22:1-36