Radical Reddits - Into the mind of online Radicalised Communities

Paul Verhaar[a] Maya Sappelli[b]

[a] *Utrecht University. Domplein 29, 2512 JE Utrecht*
[b] *TNO The Hague. Anna van Buerenplein 1, 2595 DA Den Haag*

Abstract

We have investigated language use in radical communities using automated approaches. A Naïve Bayes classifier was capable of predicting whether a Reddit post was radical or not with 75% accuracy. Moreover, the language use showed characteristics of a virtual community, based on its salient language features. This work is a step towards automatic detection of radicalism online.

1 Introduction

Nowadays there is an increasing amount of radical communication that is taking place on online platforms such as Reddit, Facebook or Twitter. Recently, Twitter closed more than 200.000 accounts for violating policies related to violent threats and promotion as terrorism. In this thesis I investigate the use of online media for radical communications from the perspective of new media and linguistics.

2 Method

The conveyance of radical beliefs has been widely researched offline [2]. However, the amount of research for online sources remains scarce [8]. Recent studies looking into the effect of the internet on radicalism have shown that non-US based extremists were more likely to learn through virtual tools [4]. The web enhances possibilities for radicalised groups to communicate, organise and plan activities [3]; they use the power and the freedom of social media platforms to exercise pressure, power and influence across the globe [5]. Using social media in such a way mostly leads to a virtual community, which shows that all information that is spread (online) contributes to the shaping of communities and a "we" sense [7]. It shows that several studies have sought to understand how radicalism online comes to exist and grow. Typically, these studies are focused on keyword analysis. In this thesis we also explore machine learning algorithms for analysing and predicting salient language features within online radicalised media. Radicalisation not only poses a threat, but also creates the necessity for further research within social media platforms.

In this thesis, I demonstrate a proof of concept for a classifier that predicts whether a post is radical or not. The characteristics of the classifier give insight in salient language features within online radical discourse. Moreover, we analysed the language use in radical discourse from a linguistic perspective. Data was gathered from the social media platform Reddit as part of the VOX-Pol project and TNO The Hague. The data were crawled via the Reddit API from October 2007 to May 2015 and was divided in two classes: radical and non-radical in nature. It contains 105,930,239 topics collected from 239,773 sub-reddits. The radical class contained clear distinguishable radicalised sub-reddits such as FeministHate, nationalism and WhiteRights, which clearly promotes radicalism in a variety of beliefs such as anti-feminism or white supremacy. For this thesis, a million posts per class were used for analysis, making the two classes equal in size. Using Python software libraries as NLTK and SciKit, a Naïve Bayes algorithm combined with TF-IDF was trained.

3 Results and Conclusion

Results showed that the classifier achieved an accuracy of 75% on correctly labelling documents as belonging to the radical or non-radical discourse. The radical class contains salient language features, such as “white male”, “rape victim” or “black person”. These words are frequently used in combination throughout all radical posts. The data display a stark difference in the distribution of collocations such as “sexual assault” and “hate women”. The collocations within the radical dataset contain important features typical of their discourse within their virtual community [5]. The linguistic analysis shows that the language within the radicalised communities are similar but different from the language in non-radicalised communities. The virtual communities act in a similar way; the notion of community is less apparent in the non-radical discourse, which is shown by the difference in language patterns compared to the radical discourse [1]. This not only shows a ‘we’ sense as part of the sub-reddit, but also displays this in terms of language use within the radical virtual community [1].

This thesis shows that multidisciplinary methods yield interesting results and should be further explored. This work successfully builds on previous theories from the field of new media and linguistics. This work sheds light on research methods applicable for detecting online radicalism within the Reddit domain. The techniques used in this paper allow for use on other (radicalised) communities as well as building towards automatically detecting radicalism online.

References

[1] B. Anderson. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London, UK: Verso, 1991.

[2] D. Barton and C. Lee. 2013. Language Online: Investigating Digital Texts and Practices. *Studies in Second Language Acquisition, 224*, 2013.

[3] C. Easttom and J. Taylor. Computer crime, investigation, and the law. *Course Technology,* Boston, MA, 15-28, 2011.

[4] P. Gill, C. Emily and A. Thornton. 2015. What Are The Roles Of The Internet In Terrorism?

[5] C. Graham. Terrorism.com: Classifying Online Islamic Radicalism as a Cybercrime, 2013.

[5] H. Rheingold. *The Virtual Community*. MIT Press, 2000.

[6] R. Rogers. *Digital Methods*. Cambridge: MIT Press, 2013.

[7] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal and L.H. Ungar. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS One, 8(9), e73791*, 2013.