Finding Social Media Trolls: Dynamic Keyword Selection Methods for Rapidly-Evolving Online Debates

Maya Srikanth¹, Anqi Liu¹, Nicholas Adams-Cohen², Betty Wang¹, R. Michael Alvarez³, and Anima Anandkumar¹

¹Department of Computing and Mathematical Sciences, California Institute of Technology
²Immigration Policy Lab, Stanford University
³Division of Humanities and Social Sciences, California Institute of Technology

{anqiliu@caltech.edu, msrikant@caltech.edu, nadamsco@stanford.edu, pwwang@caltech.edu, rma@caltech.edu, anima@caltech.edu}

Abstract

Online harassment is a significant social problem. Prevention of online harassment requires rapid detection of harassing, offensive, and negative social media posts. In this paper, we propose the use of word embedding models to identify offensive and harassing social media messages in two aspects: detecting fast-changing topics for more effective data collection and representing word semantics in different domains. We demonstrate with preliminary results that using the GloVe (Global Vectors for Word Representation) model facilitates the discovery of new and relevant keywords to use for data collection and trolling detection. Our paper concludes with a discussion of a research agenda to further develop and test word embedding models for identification of social media harassment and trolling.

1 Introduction and Background

According to 2017 survey data from the Pew Research Center, 41% of Americans have themselves experienced online harassment, and 66% say that they have seen online harassment directed at others (Pew Research Center [2017]). Of these Americans who have experienced online harassment themselves, 18% say they have been severely harassed (for example, being stalked, physically threatened, sexually harassed, or harassed over a long period) (Pew Research Center [2017]). Developing effective ways to identify social media negativity, trolling, and online harassment is an important objective that will aid in future research to develop a solution for online abuse. A promising step down this path involves collecting and analyzing social media data to shed light on the nature of dynamic online debates and trolling.

Social media data is often collected by keyword searches, namely, gathering tweets that contain specific keywords or hashtags (e.g., [Barbara et al. (2015)]). There are three main approaches used in past work to compile an expansive set of such keywords. The first involves using a human-created, static set of keywords as a search query, which is not conducive for study of long-term online debates where conversation is constantly evolving. The second is fully-automated keyword extraction, which suffers a lack of transparency and is not optimized for highly-dynamic online debates ([Wang et al. 2016] [Zheng et al. 2017]). Finally there are semi-automated keyword selection methods involving both machine and human input.

We aim to develop a dynamic keyword searching and updating methodology that is fast and efficient (like fully-automated methods), but which provides transparency and is less biased than
non-automated methods. In this work, we explore a new approach for a dynamic keyword search and use it in a preliminary analysis of data from the MeToo movement, a social media movement against sexual harassment.

2 Our Approach and Preliminary Results

We train the GloVe (Global Vectors for Word Representation) embedding model on various corpora consisting of Twitter data, Wikipedia data, and Reddit data to obtain 50-dimension vector representations of words. In this vector space, the cosine similarity between two words indicates linguistic or semantic similarity between two words.

We offer two avenues of keyword extraction: (1) Extract words that co-occur most frequently with the previous set of keywords (2) Use a clustering algorithm to expose "clusters" or "major topics", and for each of these clusters, select a representative keyword using cosine similarity or a ranking algorithm. We demonstrate with preliminary results that this approach allows us to track dynamic topic evolution and uncover differences in conversation across various domains, paving the way for feature engineering in trolling detection.

2.1 GloVE and Cosine Similarity for Keyword Extraction

A round of keyword extraction on #MeToo tweets from May 2017 to August 2017 recovered the following set of keywords which, by the metric of cosine similarity, co-occur most frequently in conversation with "women": ['WomenUnshackled', 'booksovertrump', 'HeforShe', 'womensMarch', 'Join', 'imwithher', 'ChildBrides', 'Impeach', 'metoo', 'UN Women', 'HandmaidsTale', 'love']. Similarly, a round of keyword extraction on #MeToo data from August 2019 identified the following as most similar to "metoo": ['WhoWillYouHelp', 'Wetoo', 'Amplify', 'Endrapeculture', 'falseaccusations', 'metoo', 'preserveinocence', 'timesup', 'SpeakUp', 'BoycottNetflix', 'mansright', 'believevictims', 'WithYou', 'movement']. These examples attest to an evolution of conversation topics over time: the 2017 keyword "imwithher" was used in support of Hillary Clinton commonly during and shortly after the 2016 U.S. presidential campaign, and the 2019 keyword "timesup" entered common conversation in early 2018, persisting as more Weinstein scandals came to light.

2.2 K-means clustering for Topic Modeling and Domain Shift

An alternative method for keyword extraction is the application of the K-means clustering Lloyd (1982) algorithm to the word vector space, which uncovers "sub-topics" under a larger conversation such as #MeToo. Too. For example, a news-oriented cluster obtained from August 2019 #MeToo data is: ['Weinstein', 'DonLemon', 'Burke', 'trial', 'Harvey', 'Manhattan', 'Court', 'Justice', 'SexualAssault', 'James', 'fascinating', 'NewYork', 'Supreme', 'SexualPredator', 'wie' (or Women in Engineering), 'Debate', 'untouchable'], while another cluster consists of MeToo-related hashtags: ['NotSilent', 'BelieveWomen', 'WhyIDidntReport', 'EndAbuse', 'TimesUp', 'WomensRights', 'ImWithHer', 'SupportSurvivors', 'Metooindia', 'EnoughIsEnough', 'AcademiaToo', 'NeverAgain', 'WomensMarch', 'MeTooSTEM].

It is also possible to see differences in conversation across distinct discussion domains. For instance, the closest words to “female” in the top 1000 submissions of Reddit’s Red Pill, an online anti-#MeToo men’s rights community are: ['sexual', 'negative', 'sexuality', 'intercourse', 'self', 'physical', 'dialogue', 'respective'], while the closest words to “female” in August 2019 Twitter #MeToo discussions are: ['companies', 'founded', 'startups', 'desire', 'oppressor', 'employee', 'victims', 'capitalist'], presenting a fairly stark contrast. Wikipedia, a more “neutral” corpus, identifies the following words as closest to “female”: ['adult', 'young', 'woman', 'teenage', 'girl', 'individual', 'age', 'child', 'older']. Understanding such differences sets the stage for cross-platform trolling detection, and sheds light on trending discussions across different platforms.

3 Conclusion

The method we have discussed here, and which we have pilot tested, shows considerable promise. Next steps involve further development of ranking algorithms to identify the relative importance of "most similar" keywords, and better automation of this methodology.
References


