Supervised Learning and Online Detection of Internet Trolls in Streaming Data Environments

Albert Haque *
Department of Computer Science
University of Texas at Austin
akhaque@stanford.edu

Abstract

With the increasing importance of online communities, discussion forums, and customer reviews, Internet "trolls" have proliferated thereby making it difficult for information seekers to find relevant and correct information. In this paper, we consider the problem of detecting and identifying Internet trolls, almost all of which are human agents. Identifying a human agent among a human population presents significant challenges compared to detecting automated spam or computerized robots. We use two methods for learning troll behavior and features: (i) natural language processing and analysis, (ii) Markov random fields, (iii) probabilistic graphical models. Using this algorithm, we have successfully trained on a dataset consisting of 50 million data points from the viral Internet social experiment, Twitch Plays Pokemon. Using MapReduce techniques to process the data, we are able to achieve a prediction accuracy of 00% on live streaming data as humans communicate with one another. Future work can be extended to email classifiers, comment moderation, and anomaly detection.

1 Introduction

In Internet slang, a "troll" is a person who sows discord on the Internet by starting arguments or upsetting people, by posting inflammatory, extraneous, or off-topic messages in an online community (such as a forum, chat room, or blog), either accidentally or with the deliberate intent of provoking readers into an emotional response or of otherwise disrupting normal on-topic discussion [1, 2].

Application of the term troll is subjective. Some readers may characterize a post as trolling, while others may regard the same post as a legitimate contribution to the discussion, even if controversial. Like any pejorative term, it can be used as an ad hominem attack, suggesting a negative motivation. Regardless of the circumstances, controversial posts may attract a particularly strong response from those unfamiliar with the robust dialogue found in some online, rather than physical, communities. Experienced participants in online forums know that the most effective way to discourage a troll is usually to ignore it, because responding tends to encourage trolls to continue disruptive posts hence the often-seen warning: "Please don't feed the trolls" [3].

The major contributions of this paper are as follows. After an initial survey of feature extraction techniques for human text mining, we (i) identify several dataset characteristics that will act as features for (ii) our several candidate supervised learning algorithms. We then (iii) develop a MapReduce technique for offline dataset processing which is then translated into (iv) an online algorithm for classification of live human chat behavior.

The remainder of this paper is organized as follows: In Section 2 we conduct a review of the current literature and summarize related efforts. Section 3 outlines how we extract features from our dataset.

*This paper was created on March 31, 2014.

These features are then used by algorithms outlined in Section 4 and are evaluated in Section 5. We then conclude in Section 6.

2 Background

2.1 Related Work

3 Feature Extraction

4 Algorithms

5 Evaluation

6 Conclusion

References

[1] T. Campbell. Internet trolls. http://web.archive.org/web/20011026130853/http://members.aol.com/intwg/trolls.htm, July 2001.

[2] H. Fosdick. Why people troll and how to stop them. http://www.osnews.com/story/25540, January 2012.

[3] C. Heilmann. De-trolling the web: dont post in anger. http://christianheilmann.com/2012/06/04/de-trolling-the-web-dont-post-in-anger/, June 2012.