Unsupervised Learning and Online Detection of Internet Trolls in Streaming Data

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. To learn a troll's behavior, we use natural language processing, sentiment, and intent analysis to develop a set of features for our online classifier. We use three classification algorithms: (i) decision tree learning, (ii) naive Bayes, and (iii) Bayesian networks. We have successfully trained all three algorithms on a dataset consisting of 50 million data points from the viral Internet social experiment, Twitch Plays Pokemon. Using MapReduce techniques for preprocessing and feature extraction, we are able to achieve a prediction accuracy of Y% on streaming data as humans communicate with one another live.

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 [3, 4].

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 [1]. Regardless of the circumstances, controversial posts may attract a particularly strong response from those unfamiliar with the dialogue found in some online, rather than physical, communities [1]. 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" [5].

The contributions of this project are as follows. First, we propose a set of features used for identifying trolls in the viral, croudsourced Internet game, Twitch Plays Pokemon [2]. We use features inspired from psychology and sociology to understand human behavior and attempt to identify sarcasm used by Internet trolls. We compare the performance of our features with popular natural language processing techniques.

The second major contribution of this paper is an online classification algorithm. It is initially trained using supervised methods on offline data. When switched to online mode, this algorithm updates as new data points are received from a live stream. We apply this algorithm to the Twitch Plays

[*] This paper was created on April 9, 2014.

Pokemon data stream and visualize the results. The primary purpose of this project, through these two contributions, is to in realtime, distinguish between trolls and humans on the Internet. Future work can be extended to email classifiers, comment moderation, and anomaly detection for online forums, reviews, and other communities.

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

Twitch Plays Pokmon [2] is a "social experiment" and channel on the video streaming website Twitch, consisting of a crowdsourced attempt to play Game Freak and Nintendo's Pokmon video games by parsing commands sent by users through the channel's chat room [1]. Tens of thousands of users are online and active in the chatroom. Next to the chat, a live video displays the current state of the Pokemon game. Users can then input commands into the chat. The game will then execute these commands in a FIFO order. Examples of commands include: up, down, left, right, and start. Up, down, left, and right move the main character in the game and start brings up the main menu. The objective of Twitch Plays Pokemon is to collective receive and process commands from users and attempt to beat the game of Pokemon.

2.2 Dataset

We wrote an Internet Relay Chat robot that collects user messages, commands, time, and user information in realtime. The robot collected data for several months. It is currently 3 GB in size and contains over 50 million data points. A data point is defined as a user's chat message and metadata such as their username and timestamp at millisecond resolution. It is still collecting data at the time we wrote this project proposal.

Users can enter non-commands into the chat as well. When a major milestone is accomplished in the game, it is common to see an increase in positive and ALL CAPS messages. Since a human logs into the system with the same username, we are able to build profiles of each user and learn from their past messages and commands.

3 Feature Extraction

4 Algorithms

4.1 Decision Tree Learning

4.2 Naive Bayes

4.3 Bayesian Network

5 Evaluation

5.1 Cross Validation

5.2 Online Performance

6 Conclusion

References

[1] Wikipedia. `https://www.wikipedia.org/`.

[2] Twitch plays pokemon. http://www.twitch.tv/twitchplayspokemon, April 2014.

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

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

[5] 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.