A SCALABLE MACHINE LEARNING APPROACH FOR SOCIAL NEWS CENSORSHIP PREDICTION

Andrew Quaschnick
Department of Computer Science
University of Wisconsin-La Crosse
1725 State Street
La Crosse, Wisconsin 54601
andrew.quaschnick@gmail.com

Abstract

This project is the development of a machine learning software system that is designed to identify, report, and predict censorship on one of the world's largest social news sites, Reddit. The software system constantly reads in large amounts of real time social news networking data from the Reddit API. It then identifies the articles and posts that were censored by either moderation or by self-censorship. Those articles and posts will then be input as training data for a decision tree machine learning algorithm using Apache Spark. This algorithm will be used to predict whether or not a certain article and post title will be censored given certain variables or keywords in the post's title. This application will also report on all the posts being censored to help curb or increase visibility of such events. The goal of this project is to create a system that can help identify censorship online and to report on trends in censorship in online communities to help enable free and open speech on the internet.

1. Extended Abstract

There are a multitude of social news networks that exist in today's age. A lot of people around the world get their entire news intake from these sites. Many of these sites are moderated by the companies themselves such as Facebook and Google news feeds. Sites like Reddit enable users themselves to moderate the news and in many instances the users are unaware of what or whom is moderating that content. Reddit is currently (according to Alexa [6]) the 5th-most popular website in America and 18th in the world. Gottfried and Shearer [5] estimate that 70% of users on Reddit get news on that platform, the highest among all the social media platforms they surveyed. Many Reddit users are unaware if the content and news they are viewing has been moderated for an ulterior motive. There are many communities on Reddit known as subreddits that cater to specific subjects, like /r/WorldNews, which has around 16 million readers, and /r/Wisconsin, which has around 17 thousand readers. This project is an attempt to provide transparency by identifying censorship on Reddit and reporting those trends in censorship in that online community to help enable free and open speech on the internet.

1.1 - Structure

This project consists of two new and separate applications running on an Ubuntu cluster. Each server within the cluster is running Apache Spark and the cloud monitoring service Datadog. One additional server is running a MySQL database instance.

1.2 - The Controller

The first application, written in Java running as a Linux service on each server, herein will be referred to as 'The Controller'. The Controller, on startup, queries the MySQL instance to ask for a set of subreddits that it should be monitoring. Using a set of stored procedures on the MySQL instance the server selects and returns a set of subreddits to The Controller to monitor. The Controller retrieves the top 50 posts on each specific subreddit in roughly 1/60th of a second depending the size of that subreddit. Due to the limitations of Reddit's API it will never exceed the rate 1/60th of a second. The Controller then watches to see if any of those posts fall from the top 50 to the top 300 or disappear completely. If a post simply falls to the top 300 then the post is considered to be unmoderated and still freely available for users to read and comment on. If a post disappears from the top 50 and top 300 then it is considered to be moderated and is not

considered easily available for users to read and comment on. Users with a link to that post can still visit it but the moderated post will not be available on the front page of that site; thus it can be considered censored because new users cannot view it. The Controller, on shutdown, notifies the MySQL instance that it is shutting down, and other Controller instances on other boxes will pick up the subreddits left behind.

Each post that is considered to be moderated by the application is then stored in the MySQL database. Each time a post is censored there is an additional 5% chance for an unmoderated post from the top 300 to be added to the database as training data. Before the post is to be sent to the database it is sent through the Stanford CoreNLP [1, 2, 3, 4], which parses out keywords and saves them to a separate column in the database.

With work being done every minute by The Controller, large stores of data are being created. Once every day one of The Controllers will generate a 24-hour report of all posts moderated and post it to a subreddit created for this project called /r/CensorshipPrediction/ to increase online transparency.

1.3 - The Predictor

This training data is then fed into a second application herein known as 'The Predictor'. The Predictor is an Apache Spark 2.0.1 application running in Java. The Predictor's objective is to provide a working Gradient Boosted Decision Tree to predict whether a certain keyword or set of keywords would lead to a post being censored. Keywords parsed by the Stanford CoreNLP are fed into a multi-stage pipeline to tokenize, hash, assemble, and finally predict upon. The Predictor receives certain input from the administrative user upon startup that can include a specific subreddit or a specific keyword around which the user wants to cluster the training data. When The Predictor is running its pipeline it shares the work across the entire Ubuntu cluster using Apache Spark. When The Predictor is complete is returns a Gradient Boosted Decision Tree, tests the tree's quality against a set of test data and then returns the accuracy on the test data for that tree.

1.4 - Preliminary Results

As stated earlier the objective of this project is to attempt to provide transparency by identifying censorship on Reddit and reporting those trends in censorship in that online community to help enable free and open speech on the internet.

Some subreddits did not have easily discernible patterns to the data returned and The Predictor did not find a usable decision tree that correctly predicted post moderation more than 30% of the time. It should be noted that merely flipping a coin would give you a 50% chance of correctly predicting the result. Other subreddits, with more consistent themes to post titles, allowed The Predictor to create trees that correctly predicted posts more than 67% of the time.

On subreddits where prediction accuracy is greater than 50% there may be patterns to the censorship of posts. For instance, between February 1st and March 20th 2017, /r/WorldNews had 1,601 moderated posts in the top 50. Of those 1,601 moderated posts, 491 included the keyword 'Trump', accounting for 30.67% of the moderated posts. In this case, we did find that The Predictor created a tree, which was 66.194% accurate, that had a root node of 'Trump'—meaning it was the most descriptive keyword out of the entire dataset to predict whether or not something was moderated.

2. Acknowledgements

This project was created and is maintained by Andrew Quaschnick in his partial fulfillment of the Master of Software Engineering degree at the University of Wisconsin-La Crosse under the advisement of Dr. Martin Allen.

I would also like to thank the Department of Computer Science and the University of Wisconsin-La Crosse for providing a fantastic environment for learning and providing a home for my project.

3. References

1. Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics:* System Demonstrations, pp. 55-60. [pdf] [bib]
2. Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT

Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

3. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
4. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
5. Gottfried, J., & Shearer, E. (2016, May 26). News Use Across Social Media Platforms 2016. Retrieved March 10, 2017, from http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/
6. Reddit.com Traffic Statistics. (n.d.). Retrieved March 10, 2017, from http://www.alexa.com/siteinfo/reddit.com