Technical Guide

Project Title: Detecting brigading on social media with machine learning

Student Name: Adam Pegman

Student ID: 14351741

Stream: CASE4

Project Supervisor Name: Ray Walshe

motivation

Reddit is a social media and news aggregator site with over 250 million users. Users on reddit can create, moderate and subscribe to ‘subreddits,’ which are forms on which links and text can be posted, each with a thread of comments. Each subreddit has it’s own topic, culture and rules. Use of language also varies considerably, with some subreddits using their own unique terminology.

The moderators of each subreddit remove dozens or hundreds of comments from their subreddits every day, and must keep track of frequent offenders so they can be banned from the subreddit. This task is made more difficult by the fact that creating a new account takes less than a minute, so bans can be easily avoided.

There is an existing bot called automoderator that moderators can use to remove posts automatically, but it uses regular expressions to detect unsuitable comments. This makes the system difficult to use for many moderators, and limits the scope of what can be detected. The system should partially automate the moderation of a community in an easier to use and more versatile way, saving time and resources.

The purpose of the project is to monitor a subreddit and take actions such as creating posts, removing posts and banning users automatically. The system would assist volunteer moderators keep unsuitable posts off the subreddit. The system will achieve this by using the reddit api to retrieve posts and comments, and to automatically remove posts and users.

research

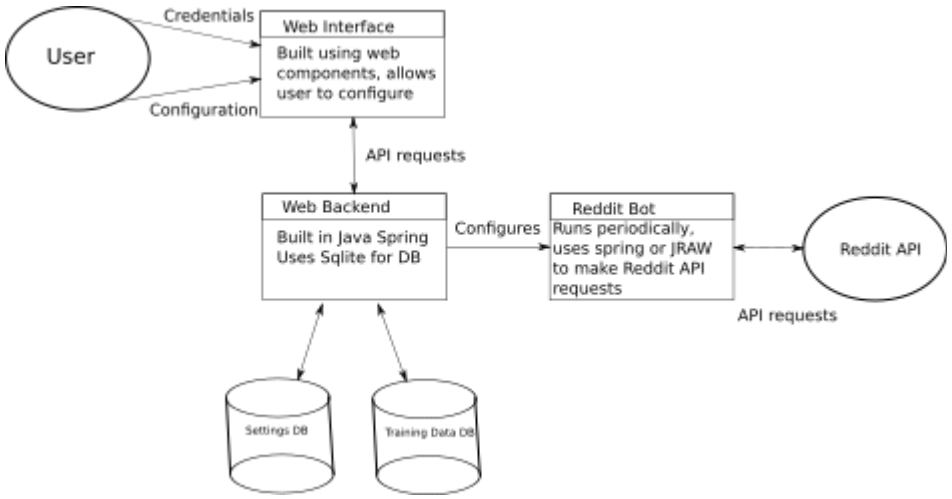
I researched the problem of classifying comments online, as well as in the notes for the CA4010:Data Warehousing & Data Mining module, in order to find algorithms that I could implement and compare.

I found a paper on baysian spam filtering, and decided to try implement filtering with greyham's method to combine the probabilities, as it seemed the most straightforward way to implement a baysian filter and it avoids divide by 0 errors.

I decided to implement a k-nearest algorithm using the difference in the number of occurrences of each word that exists in either comment as the distance metric. This was straightforward to implement with the notes I had. I looked at word usage statistics from a few Subreddits beforehand and they seemed different enough that this should work.

I finally I decided to try anomaly detection, but many of the techniques I found weren't suitable for the task. Idecuided to go with multivariate gaussian as it seemed like the most general.

design



results

From ad-hoc testing on https://www.reddit.com/r/sample_cat_subreddit/, the baysian classifier works best, followed by the k-nearest. The Anomaly detection algorithm seems to label almost everything it hasn't seen an anomaly, probably because the length of comments vary too much, there is not enough data, or I need to calculate the threshold per-dataset.