Final Project Proposal: Controversial Topic Extraction

Author: Ryan March (UUID: ryanm14)

Many topics are not "too" controversial in the grand scheme of things; few would have a strong argument against the idea that "the sky is blue" or that "tomatoes are red." Those are just statements of fact. However, this idea can extend to more ambiguous statements – even moral and ethical proposals such as "murder is bad" and "it's wrong to steal" are widely agreed upon. Of course, there are subsets of topics within these categories that do introduce some gray areas (ie what about self defense? Is theft *always* wrong?) I think one statement that is not too controversial would be that many subjects are much more controversial than others. Political and religious beliefs are one obvious area where controversy is common, but even subjects like music preferences, food tastes, and sports team loyalties can serve as grounds for debate.

My goal is to come up with a general framework using techniques covered in class to take online discussions as input and derive "controversial topics" from them. By limiting myself to online discussions, I think I will retain many benefits. Let's consider an example controversial statement: "chocolate chip cookies are the superior cookie choice." An article about cookies will dilute many of the factors involved in determining if the subject is controversial. However, if that article has a comment section in it, users will more likely have debates in the comments about the cookie quality thesis. My hypothesis is that the more controversial topics will have longer debates, with sentiment that is either overwhelmingly negative (users angrily arguing over the subject) or a fairly balanced sentiment (some positive statements mixed in with negative statements). A mostly positive sentiment in a comment thread seems less likely to denote controversial underpinnings. However, this hypothesis demands research, and if it is to be applied in a more general framework will need to be considered more formally. I want to come up with some kind of model that aggregates a number of quantities, such as rewards for comment thread length (maybe with some diminishing returns, like a Term Frequency measure might do), to determine if subject matter being discussed is controversial.

A number of datasets seems to have the qualities I am looking for. News sites like CNN, NBC, or Fox, can sometimes have comments sections in them. Forum sites like Reddit or Quora deal with a wide swath of topics while also having high community/user engagement. Twitter also seems like a perfect forum for this, where statements are generated by users to be directly discussed and commented on. A tweet about chocolate chip cookies could have hundreds or even thousands of tweets in response to the initial statement. This would allow easily generating a simple topic model for the first tweet, which can then be categorized either as "controversial" or "not controversial" based on the responses.

My hope would be to use as little labeling as possible; can I go with a fully unsupervised approach to detect some soft quantity of a "controversial" metric? Unfortunately, probably not. Thus I can also leverage other subjects that are interesting to me: weak supervision, distant supervision, or other methods to reduce the labeling overhead.

As far as tools and technologies, I plan to use Python with various NLP/machine learning libraries. I will need to hook into various APIs (like Twitter's) and maybe write some simple web scrapers to divulge comment threads on other sites. My expected outcome is some sort of model that can take in a forum page, a single post, etc, as input, and output a set of topics (might just be single topic words) that are "controversial" from that input. I would then have to evaluate this as a human judge empirically, determining whether the returned results make sense.

There are a lot of complexities and unknowns here, including the fact that the controversial element of a post may not be its core subject – it may be a secondary part of the statement that incited discussion and debate. I want to see how far I can get even given these unknowns.

Since I will be working by myself (Ryan March, UUID: ryanm14), I need to have achievable deliverables (1 week will equate to roughly 3-5 hours):

Task	Level of Effort Estimate
Research Prior Art on Controversial Topic Classification	1 week (and ongoing throughout project)
Collect data from a number of sources (Twitter, Reddit, News Sites)	1 week
Create hypothesized model (theoretically and in code)	1 week
Test hypothesized model on data	1 week
Verify results, add friendly user interface (frontend or friendly CLI)	1 week
Final Project Delivery/Video	1 week