Determining Topic Sentiment Across Social Network Platforms

Craig Dick

2378675

Proposal

Motivation

In recent times, but more notably throughout COVID-19, there has been a lot of disinformation campaigns around certain topics within the pandemic. This includes misinformation around important topics that can put the public at risk, such as vaccine efficacy, face masks not working and slander on pharmaceutical companies. This is extremely important as spreading false information on such sensitive topics can influence someone's decisions or feelings, which could potentially be the difference between life and death.

Aims

The aim is to investigate whether topic sentiment around terms such as vaccines, lockdowns, and other queries, can identify disinformation being spread online. It involves gathering a collection of data from Reddit and Twitter respectively, and comparing the sentiment amongst certain topics between the two social network platforms. It aims to compare sentiment on the search queries across the social networks, and see the different public reaction to the same terms on the sites. Investigating the topic sentiment in Reddit also involves looking at subreddits, and seeing whether the topic sentiment within a particular subreddit can identify whether this subreddit is, for example, anti-vax. The project involves first gathering the data for each term we are interested in looking at from Reddit and Twitter, saving this to a database, and then performing filtering and analysis on this by determining the topic sentiment and comparing it to other topics and social networks.

Progress

- Language chosen along with the APIs to collect the data from Twitter and Reddit: Python is the language the project is coded in, with the Tweepy API and Reddit Pushshift API to collect data on topics. VADER is being used to analyse the sentiment of the social network posts.
- MongoDB being used to store the results of the data collections.
- Researched previous topic sentiment papers on social networks.
- Project was changed to investigate topics surrounding the misinformation of COVID-19, to try and determine if sentiment could suggest where misinformation was being spread.

- Code to collect data from Reddit and Twitter is implemented, successfully pulls back data on a search term and stores this into a cloud database on MongoDB.
- Used this code above to search for seven important terms around covid, and subreddits are stored as well as the posts. There is potential to do this for more topics if time permits
- Basic sentiment analysis done on terms collected so far for Reddit and Twitter separately.

Problems and risks

Problems

- The original Reddit API used to gather data would not work, as it could only gather the most recent posts up to a number of 1000. This held up production for a week or two until I was able to find the Pushshift API.
- Finding a library in Python to analyse sentiment accurately was difficult. I have settled on VADER for now but this has the potential to change if there is a more accurate way of doing so.
- When gathering the data on search terms, the code was breaking after a certain amount of time. Try except statements were added as there were problems in the Reddit code that when searching for a post, it had been deleted since so was unretrievable. It was also breaking in the twitter code, but have changed this to respect the rate limit.

Risks

- A lot of reddit data has terms like [removed], [deleted], or text that says the post was removed by admin of the subreddit. This could skew the data as it may perceive this to be positive, negative or neutral. **Mitigation:** See what the most common terms are from reddit, and remove these from the collection if they are from deleted or removed posts and comments.
- VADER may not be the most accurate way of sentiment analysis on this data. **Mitigation:** VADER says it is reliable with social network posts, could maybe compare to tool such as BERT and see whether the results are completely different if time permits.
- People may not be against vaccines, but their posts may have terms that are deemed as negative. For example, “Feeling ill and rubbish after my vaccine yesterday!”. **There is no clear way of mitigating this data.**

Plan

Semester 2

- Week 1-2: Run the topic on any more terms for comparison, also test VADER against BERT to see how they compare.
    - **Deliverable:** Two pieces of code, one using VADER and the other BERT to see the sentiment difference.
- Week 3-4: Investigate any subreddits that were found to be the most negative and positive, see if there are any patterns that could suggest misinformation.
    - **Deliverable:** Subreddit analysis on the most negative and most positive around certain covid search queries, could this influence public opinion? Are they coming from subreddit names you would expect?
- Week 5-6: Have all graphics done on comparisons between the social networks, as well as analysis on the subreddits.
    - **Deliverable:** Clear visualisations produced for the social networks and the features interested in.
- Week 6-10: Collate previous papers together based on the same topics for background, and begin the write up. If time permits, could begin the write up at start of semester 2, and add to this as I go along at same time as finishing implementation.
    - **Deliverable:** First draft to supervisor by this time at very latest, hopefully by week 8/9