30th January 2024

As a group we decided to each select our top three topics for the term paper.

I told the group I have worked on the complexity topic before and that I have some data and web scraping scripts I made for an project in previous year.

My top three choices were: 1. Complexity 2. Frequency 3. Questions

One disadvantage about the complexity topic is that I remember some tools that can be used for structural analysis being fairly compute intensive (such as the stanford parser and STANZA), which can make it harder to write and run code for analyses in a timely manner.

The complexity topic seems to allow an approach with minimal information coming from outside the text, i.e. the analysis can focus on comparing the textual properties/quantities (measures of complexity) across different texts/contexts, i.e. comparing a set of formal texts to a set of social media texts.

A challenge I remember about measuring linguistic complexity is that testing the validity of a metric is difficult.

weekly reading

The Golbeck et. al. (2011) paper entitled "Predicting Personality from Twitter" showed that openly accessible data from twitter was highly revealing about personality types. The authors did not address the sampling bias in the paper, namely that participation was solicited on twitter, so the types of people to take part may have been biased towards particular personality types such as openness to experience or agreeableness. Perhaps they would have seen different results had they selected the subjects at random from a larger population, e.g. randomly selecting from the set of citizens of a country. Generally I thought the methodology was not well motivated, and that the authors neglected to think about/advocate for the individual's right to privacy.

term project preference ranking

Our team has settled on the Affect topic as our first choice for the term project.

1st February

lecture thoughts

We conceptualise a natural language as a countably infinite set of acceptable sentences. This implies a function f which ascribes membership of the set, and therefore we also get the complement of the set for free. That we model the set as countable means that a mechanical process for ascribing membership is decidable, and empirically we know that people can decide deterministically

about the membership of a sentence in the language, but this argument may be circular, especially if we define the natural language relative to its speakers.

term project allocation: Advice

The essay topic is around responses to medical advice as available in public discourse. Perhaps we will have to broaden our focus to advice in general.

Potential sources of responses to medical advice:

- Reddit
  - r/shittymedicaladvice could be a source but from a scan of about 10 posts I have seen no responses to advice, only solicitations and the advice itself
  - r/AskDocs
  - r/Advice looks useful. It seems the initial post usually solicits advice, and ofter posters will reply to the advice in the comments.

In order to collect a dataset from Reddit we will need to review and comply with the Reddit API Developer Terms as well as the Data API Terms.

Additionally API access requires OAuth authentication, so I am beginning the process of attaining API access. I have set up a 'reddit app' and retrieved an app client_id and a client secret, and tested, it can be used to get posts from r/Advice. Next steps will be to decide on the type of dataset we want to create (which attributes to keep/discard).

An aspect of analysis we could pursue is features of advice solicitation. We could for instance assume all first posts on r/Advice are advice soliciting, yielding a supervised data set, and then try to develop means to detect advice solicitation in other contexts.

behavioural considerations

A patient's response to authoritative medical advice may be correlated with that patient's personality. While the delivery of the advice could influence the response, they may be other stronger predictors of the response, such as the severity/impact of the illness/measures advised, the patient's agreeableness and skepticism more generally, the patient's feelings towards authoritative figures.

6th February 2024

Summary: Recognition of Affect, Judgement, and Appreciation in Text

The authors develop a rule-based algorithm for recognizing attitude in terms of affect (negative and positive), judgement (negative and positive), and appreciation. The algorithm systematically composes the classifications of constituent sentence

parts, such that entire sentences can be classified. Their theory and algorithm applies at three levels. At the top level are just three classifications, positive, negative and neutral. At the mid level, positive and negative classifications are further subdivided into, affect, judgement, and appreciation. At the most fine-grained level affect is divided into interest, joy, and surprise (positive affect) and anger, disgust, fear, guilt, sadness, shame (negative affect). The algorithm accounts for negation by several means, amplification of sentiment, neutralisation of sentiment etc. A corpus of human-annotated sentences is collected and compared to the results of the automated system. The automated system was found to have greater agreement with the human annotations than a baseline system that simply selected the most intense token annotation. The authors discuss some of the failures and failure modes of their system. The accuracy of the automated systems (at predicting human annotations) is highest at the top level of the classification hierarchy (positive, negative, neutral), and the accuracy decreases as we move to finer grained classifications.

- If the algorithm proposed in the paper is available as code it would be useful for our textual analyses, but it would too complex to implement from scratch given the time available.
- If an implementation exists it may be possible to make adjustments leveraging other techniques, for instance using word embeddings from a large language model like BERT to make atomic phrasal classifications, and then continuing the authors' algorithm.
- It may be generally true that finer grained analyses/classifications are more difficult than coarser analyses, so the hierarchical approach is useful for validating the methodology at different levels.
- If the corpus described in the paper is available it could be useful to our project.

Summary: Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup

The authors collect a dataset from various subreddits in which Reddit users have described breakup and divorces. For each Reddit user who described a breakup they also collect a set of related and unrelated posts from Reddit. Features from the texts are extracted using LIWC and the authors find that before and after the breakup pos, there are observable changes users' language use. The authors focus on measuring the influence of the breakup on analytic thinking, cognitive processes, self-focus, and collective-focus. Thus the study amounts to an analysis of emotional and cognitive responses to and indicators of an impending or recent breakup. The data collected were posted spontaneously, which has advantages when compared to data collected by solicitation; the authors of the text are perhaps less conscious of being studied and therefore less hesitant/biased. However, the collection of spontaneous data is biased towards people who post about intimate topics on public social media, and this tendency may be correlated with the factors under analysis.

The authors opportunistically engage in a further analysis of the effect of writing about a breakup on recovery to normal behaviour. Their analysis seems to line up with other literature suggesting that a small amount of reflecting on the events is associated with a faster return to normal, whereas excessive reflection (rumination) can prolong the process.