SDA/LING 250 – COMPUTATIONAL TEXT ANALYSIS

Department of Linguistics - Simon Fraser University – Dr. Maite Taboada

Assignment 2

Sentiment analysis

This is a group assignment. You need to work in groups of 2-3 students, to be organized through Canvas. Please work together and submit only one assignment for the group.

Please list the group members' names and student IDs on the first page, together with a list of responsibilities (sample below). Make sure that, as an individual in the group, you understand and agree with the answers submitted.

Name (Last, First)	Student ID	Section contributed	Section edited	Other contributions

For this assignment, you have to perform sentiment analysis on a dataset of your choice. You can go back to the data you collected for Assignment 1, or do your own data collection, scraping from the web or finding interesting datasets. See the appendix for ideas and resources.

You need to submit:

1. A description of your program, the input, the categories, and the output. For example, this is an extremely short description:
    - Supervised machine learning using decision trees and adjectives as features
    - 2,000 positive/negative movie reviews as training; 100 positive/negative movie reviews as test
    - Positive/negative labels for the reviews
    - 84% accuracy
2. The code used to get the output (can be a reimplementation, see Appendix for existing resources)
3. Any resources (existing code, datasets, research). This can be done as a literature review, with a list of references at the end. For sources and bibliographic references, please list all your sources following whatever format you want (APA, MLA, Chicago, or Unified Linguistics). Just make sure you are consistent in following the style.

If you do not want to do much programming, you can use online tools to upload your data and get a result. In that case, you still need to submit the description and the resources (#1 and #3 above). Some suggestions for online services are available in the appendix.

Appendix: existing sentiment analysis resources

You can re-implement an existing system:

- VADER, https://github.com/cjhutto/vaderSentiment
- NLTK (includes a VADER implementation), https://www.nltk.org/api/nltk.sentiment.html

- Stanford CoreNLP (in Java, but you can use an API), https://stanfordnlp.github.io/CoreNLP/
    - https://stanfordnlp.github.io/stanza/
- TextBlob, https://textblob.readthedocs.io/en/dev/
- TextBlob within spaCy, https://spacy.io/universe/project/spacy-textblob

There are many, many blog posts, GitHub repos and how-tos for how to implement various sentiment analysis systems. You are welcome to use them, as long as you list them in your sources. Some examples:

- https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python
- https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk
- https://realpython.com/sentiment-analysis-python/

Appendix: online services

You can use one or two of these for the second option, which involves no implementation.

- IBM Watson, https://www.ibm.com/cloud/watson-tone-analyzer
- Google's Perspective, https://www.perspectiveapi.com/#/home
- Lexalytics, https://www.lexalytics.com/demo
- MeaningCloud, https://www.meaningcloud.com/demos/text-analytics-demo
- Sentiment Analyzer, https://www.danielsoper.com/sentimentanalysis/default.aspx
- OpenText, http://magellan-text-mining.opentext.com/
- SentiStrength, http://sentistrength.wlv.ac.uk/#Test

Appendix: scraping

- Go over the scraping content in Lab 4 for beautifulsoup and other scraping resources.
- Using various APIs:
    - Twitter (https://developer.twitter.com/en/docs/twitter-api/getting-started/guide). Note: you may not need the full Academic account.
    - Twitter through NLTK (http://www.nltk.org/howto/twitter.html) stan
    - Reddit (https://www.reddit.com/dev/api/)
    - Wikipedia (https://en.wikipedia.org/w/api.php)
    - YouTube comments (https://developers.google.com/youtube/v3/docs/comments)
    - New York Times (https://developer.nytimes.com/)
- RSS feeds
    - For instance, this collection of RSS feeds for Canadian news: https://blog.feedspot.com/canadian_news_rss_feeds/

Appendix: datasets

- Existing datasets (warning: some of these may contain offensive material):
    - SFU Review Corpus (https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html)
    - SFU Opinion and Comments Corpus (https://github.com/sfu-discourse-lab/SOCC)
    - Movie Review Data (https://www.cs.cornell.edu/people/pabo/movie-review-data/)
    - Stanford Movie review dataset (https://nlp.stanford.edu/sentiment/index.html)
- Aggregators of datasets
    - Kaggle (https://www.kaggle.com/)
    - FigShare (https://figshare.com/)
    - https://datasetsearch.research.google.com/
    - https://data.gov.bc.ca/

    - https://www.statcan.gc.ca/eng/start
    - https://www.pewresearch.org/internet/datasets/
    - https://www.sfu.ca/~mtaboada/ldc.html
- Parliamentary data. (This is an opportunity to do multilingual analysis)
    - Canadian Hansard
        - https://www.ourcommons.ca/DocumentViewer/en/35-2/house/hansard-index
        - https://openparliament.ca/debates/
        - https://www.isi.edu/division3/natural-language/download/hansard/
    - UN
        - https://conferences.unite.un.org/uncorpus
    - EU
        - https://ec.europa.eu/jrc/en/language-technologies/dcep
        - https://www.statmt.org/europarl/
- Various political debates and speeches
    - https://www.macleans.ca/politics/federal-leaders-debate-full-transcript/
    - https://pm.gc.ca/en/news/speeches
    - https://www.news24.com/news24/columnists/guestcolumn/transcript-boris-johnsons-election-victory-speech-in-full-20191213
    - https://www.gov.uk/search/news-and-communications?people%5B%5D=boris-johnson&order=updated-newest