CS505 Project Choices

March 8, 2021

1 Gun Violence News Classification

This project is related to gun violence news frame classification (see here for background in news frame classification on the gun violence frame dataset). The gun violence frame data can be found in this folder. You can find the description of columns in the annotated data file (GVFC_AnnotatedHeadlines.xlsx) in the two Codebook files. The folder also contains a zip file of lead images for the articles. The names of lead images for the articles can be found in the column **ImageID** of the annotated data file. The annotated data file also contains textual information about the image, including generated captions of the image (generated automatically using Transform-and-Tell image captioning system), as well as Google Visual API tags of the image that contains a string of web entities and pages detected in the image (see here for more details). Project ideas include: (choose 1)

1. Conducting frame classification. You would train models to classify articles into their frames (either column *Q3 Theme1* for multiclass classification, or both *Q3 Theme1* and *Q3 Theme2* for multilabel classification). The current state-of-the-art (SOTA) performance on this dataset (using the articles' headlines) is 84% micro-accuracy for multiclass classification (see here) and 83% micro-F1 accuracy for multilabel classification (see here). You can explore different models for framing classification, and/or use additional data (images, captions, Google Visual API tags, whole text, summary of whole text generated using existing summarization tool, etc.) to compare with these current models. You can also explore text generation models (GPT-2, etc.) for frame prediction e.g., what's the probability that a pre-trained text generation model generates frame titles given headlines? (watch here on why text generation model may be the future of NLP pre-trained models)

2. Conducting relevance classification. Not all lead images of articles are relevant to the frames of the articles' headlines. Column V3Relevance of the annotated data file contains annotations of whether the lead image of an article is relevant (1) or not relevant (0) to the frame of the article's headline. You would train models to predict relevance of lead images to

their headline frames. The highest accuracy obtained so far for relevance prediction is 71% using headlines and image captions. You can explore different models and/or data (images, captions, Google Visual API tags, whole text, summary of whole text, etc.) for predicting relevance. You can also explore text generation models (GPT-2, etc.) for relevance prediction e.g., what's the probability that a pre-trained text generation model generates a particular caption given a headline? Does caption of relevant images score higher than caption of irrelevant images? (watch here on why text generation model may be the future of NLP pre-trained models)

2 COVID-19 News Classification

This project is related to news coverage regarding COVID-19 around the world. Using LDA topic modeling, we have extracted topics in news about COVID-19 over time and from different parts of the world. The visualization of the extracted topics in different parts of the world can be found in this web page. From the web page you can see that topics of news articles about the same issue, which is COVID-19, vary around the world. In this project, you would explore if there is any correlation between topics and public sentiment and/or number of cases/fatalities. Specifically, some project ideas include (1) looking into whether an increase/decrease in number of cases/fatalities is followed by a change in some news topic coverage or vice versa, or (2) looking into whether a change in some news topic coverage is followed by a change in public opinion of the pandemic (as captured by Gallup poll here, for example). You can investigate for correlations and use Granger causality, for example (here is an example of how correlations between a change in the Russian economy and a change in U.S. news coverage in Russian media is explored, and how Granger causality is used to investigate whether these correlations are in fact directed). The list of weekly topics from January 2020 to January 2021 in South Korean news media is available here and in US mainstream media here, with topic classes indicated at the bottom rows of each tab. The topic class description for each country can be found here.

3 Twitter Classification

This project is related to classifying social media texts or users in Twitter. Project ideas include:

1. Multilingual Twitter sentiment classification. Data of English tweets annotated with their sentiment can be found in this link: noisy data automatically annotated with sentiment based on emojis, and in this link: clean data annotated by human (Amazon Mechanical Turk) annotators, from SemEval sentiment analysis in Twitter (task A) – you need to download several files from 2013, 2014, 2015, and 2016. Data of Arabic tweets annotated with their sentiment can be found in this link from SemEval sentiment analysis in Twitter (task A). Project ideas based on these datasets

include: (choose 1)

(a) Comparing the performance on the same test sets (from SemEval 2017), of models trained with noisy data vs. clean data. Then, building models to explore if pre-training models with the noisy data and then training them with the clean data can improve performance (see here for inspiration why this is an interesting question).

(b) Training multilingual models for sentiment analysis: by fine-tuning pre-trained multilingual language models such as multiBERT, XLM-Roberta, etc. on the sentiment prediction task and comparing the performance of such multilingual models with (1) monolingual English model trained on English tweets and test on Arabic tweets translated to English (with pre-trained machine translation or Google Translate) and (2) multilingual model trained on English tweets and test on Arabic tweets (so, zero-shot classification). Do multilingual models benefit from being trained on multilingual data?

(c) Training multilingual models for predicting sentiment in tweets and applying them to predict sentiment of tweets regarding COVID-19. You can download multilingual tweets about COVID-19 from here (you need to hydrate the tweets, so start hydrating early). You don't need to hydrate *all* the tweets, but maybe take tweets from some interesting months like the month that was the peak of the pandemic (April 2020) and comparing them to most recent month tweets (February 2021). Ask some interesting questions from your analysis e.g., are there differences in ratios of positive/negative tweets in different language tweets? Are people becoming more positive now that the vaccine is here? What are some of the most discussed hashtags in the positive tweets vs. in the negative tweets? etc.

(d) Training a model for predicting sentiment in tweets and applying them to predict sentiment of tweets regarding COVID-19 from users of different political leaning. You can search for COVID-19 tweets using these keywords and then classify users based on their political leaning (using methods like here). Ask some interesting questions from your analysis such as what's the overall sentiment of users from different political leaning? What are some of the issues (e.g., from hashtags) users from different political leaning care about? etc.

(e) Training multilingual models for predicting sentiment in tweets and using them to take part in competition with cash-prizes such as this, which requires building models for predicting sentiment of Arabizi tweets (i.e., Arabic tweets written in roman characters). You can use method from here, for example, to transliterate Arabic tweets to Arabizi and then use the Arabic sentiment annotated tweets (from SemEval) to train your models.

2. Multilingual emoji prediction. You would build multilingual models to predict emojis for tweets written in English and Spanish. The overview of

the task and the data can be found here. You can explore different models for text classification, fine-tuning pre-trained multilingual language models such as multiBERT, XLM-Roberta, etc., or by using text generation models such as GPT-2, etc., to *generate* emojis given the tweets (watch here on why text generation model may be the future of NLP pre-trained models).

3. Age, gender, and race prediction of Twitter users. We have a list of 25K unique Twitter handles that will be used for a Network Analysis Paper with colleagues from Boston University Public Health (we may offer co-authorship for the best participating team). These 25K users had tweets about Swisher Sweet cigar from 2018, 2019, 2020. We have annotated some Twitter user handles with their age and gender in the file Twitter_users_labeled_with_age_and_gender.csv (columns: *human.labeled.age*, *human.labeled.gender*) in this folder. You should build models to predict user's gender: Female (F) or Male (M), age age: < 21 or ≥ 21, and race based on say, their 100/500/... most recent tweets. You can use existing models for predicting gender and age such as this or for predicting gender, age, and *race* such as this to compare to the models you build in terms of their accuracies for predicting gender and age on the labeled dataset.

4. Sentiment analysis toward tobacco products. Given the list of 25K unique Twitter handles that had tweets about Swisher Sweet cigar from 2018, 2019, 2020, build models to:

   (a) scrape all their tweets and the date/time stamp of these tweets (using web scraping method such as this)

   (b) determine if they wrote about other tobacco products (by finding keywords in their tweets)

   (c) determine their sentiment of these tobacco products by training sentiment prediction model based on the noisy data or the clean data from SemEval 2017 or using off-the-shelf models such as textblob

4 Data Curation for Tobacco Research

There are several ideas for project on curating data in collaboration with colleagues from Boston University Public Health (best participating teams will be made into co-authors of the resulting research papers): (choose 1)

1. Instagram data curation. Scrape all the Instagram **posts/captions** (and associated comments) from two Instagram sites: Swisher Sweets and Backwoods from 2018, 2019, and 2020. Annotate the posts/captions with: whether/not they contain tobacco use warnings, explicit sponsorship hashtags (#sponsored, #ad, or #paid), and ambiguous sponsorship hashtags (#thanks, #sp, #spon, #ambassador, #collab).

2. Instagram data curation. Scrape all the Instagram **posts of followers** of two Instagram sites: Swisher Sweets and Backwoods from 2018, 2019, and 2020. Build models to:

   (a) predict the followers age, gender, and race based on their posts. You can use age and gender labeled data we have on Twitter users here to train age and gender prediction models. You can use existing models for predicting race from user names from here.

   (b) predict users' sentiment towards the product (Swisher Sweets and/or Backwoods) by training sentiment prediction model based on noisy social media sentiment data or the clean social media sentiment data from SemEval 2017, or by using off-the-shelf models such as textblob

   (c) predict users' sentiment towards tobacco in general (you can use these keywords to filter for users' tobacco-related posts).

3. TikTok data curation for tobacco use. Scrape Tik Tok videos with the keyword Juul (#juul, #juulgang), the keyword Puff Bar(#puff, #puffbar, #puffplus), the keyword Backwoods (#backwoods), or the keyword vaping (#vape, #vaping). Build models to determine sentiment of the comments by training sentiment prediction model based on noisy social media sentiment data or the clean social media sentiment data from SemEval 2017, or by using off-the-shelf models such as textblob.

4. Tiktok data curation for COVID-19 vaccination and pregnancy. Scrape Tik Tok videos with the keywords "(pregnancy or pregnant) AND (covid or coronavirus) AND (Moderna or pfizer or Johnson & Johnson or vaccine or shot or vaccination)" or hashtags "(#pregnant OR #pregnancy OR #covidpregnancy OR #pandemicpregnancy) AND (#vaccine OR #covidvaccine OR #coronavirusvaccine OR #covid19vaccine OR #moderna OR #pfizer). Build models to determine sentiment of the comments by training language models such as BERT to predict sentiment using noisy social media sentiment data or the clean social media sentiment data from SemEval 2017.

5 Financial News Curation and Classification

Inspired by this post, build models to scrape news articles of companies from finviz, predict their sentiment using off-the-shelf method such as Vader or textblob, and correlate the sentiment with the companies' stock performance. The goals of the project is to:

1. curate news articles annotated with their sentiment.

2. annotate news articles with their stock-informed sentiment score (from here). Specifically, if we consider the news of asset of company i at time t, and denote the average of the stock prices over a one minute period prior

to the news being published $p_{b,i}$ and the average of the stock price over the one minute period after the news has been published by $p_{a,i}$, we can define the polarity of the news based on the observed stock log return: $r_i(t) = log\frac{p_{a,i}}{p_{b,i}}$

3. find correlation, if any, between news sentiment and stock performance. For example, using Granger causality to see if a change in the sentiment is followed by a change in stock at different time lags: a minute, an hour, a day, a week, etc. (see a different, but related application of Granger causality here).

6 Tasks from SemEval 2021

The following are suggested projects from SemEval 2021: (choose 1)

1. Word Complexity Prediction. Train models to predict whether a word is simple or complex either independently (e.g., stingy vs. parsimonious) or based on context (e.g., *table* in "put this cup on the *table*" vs. "*table* this motion"). You can find out more about the task of determining complexity of words in context in here.

2. Word Substitution. Train multilingual models to predict whether or not words can be substituted in context. For example, the target word 'souris' can be used in both "la *souris* mange le fromage" and "le chat court après la *souris*" since it is used in the same meaning in both sentences. However, the target word 'mouse' and its corresponding translation into French 'souris' in these two sentences "click the right *mouse* button" and "le chat court après la *souris*" are not substitutable because they have different meaning. You can find out more about the task of determining lexical substitutability in context in here.

3. Toxic Span detection. Train models to extract a list of toxic spans, or an empty list, per text, where toxic span is defined as a sequence of words that attribute to the text's toxicity. Consider, for example, the following text: "This is a stupid example, so thank you for nothing a!@#!@." It comprises two toxic spans, "stupid" and "a!@#!@", which have character offsets from 10 to 15 (counting starts from 0) and from 51 to 56 respectively. Trained models are then expected to return the following list for this text: [10,11,12,13,14,15,51,52,53,54,55,56]. You can fine-tune language models for this task, using example from here that finds text spans of name entities in text. You can find out more about the task of determining toxic spans in here.

7 Low Resource Language Text Classification

Last but not least, you can also train models as part of this text classification challenge. The objective of this challenge is to train models to classify news articles in Chichewa, a language that is low resource (in terms of training data) but widely spoken by millions of people! Chichewa is a Bantu language spoken in much of Southern, Southeast and East Africa, namely the countries of Malawi and Zambia, where it is an official language, and Mozambique and Zimbabwe where it is a recognised minority language. The data contains news articles annotated into categories such as ['SOCIAL ISSUES', 'EDUCATION', 'RELATIONSHIPS', 'ECONOMY', 'RELIGION', 'POLITICS', 'LAW/ORDER', 'SOCIAL', 'HEALTH', 'ARTS AND CRAFTS', 'FARMING', 'CULTURE', 'FLOODING', 'WITCHCRAFT', 'MUSIC', 'TRANSPORT', 'WILDLIFE/ENVIRONMENT', 'LOCALCHIEFS', 'SPORTS', 'OPINION/ESSAY']. Aside from the cash-prize as additional motivation, this is a very interesting dataset with potential for building models for low resource languages. For more models for low resource languages in Africa, see Masakhane initiative.