MIE 1624 Introduction to Data Science and Analytics – Winter 2023

Course Project

Deadline: Sunday, April 2, 11:59pm

Background

Sentiment Analysis is a branch of Natural Language Processing (NLP) that allows us to determine algorithmically whether a statement or document is "positive" or "negative".

Sentiment analysis is a technology of increasing importance in the modern society as it allows individuals, organizations and governments to detect trends in public opinion by analyzing social media content. Keeping abreast of socio-political developments is especially important during periods of war and political instability, such as year 2022, when russia started the unprovoked war in Ukraine. Politicians, media and organizations can benefit from sentiment analysis by making appropriate changes to their campaigning, news coverage, and strategies respectively.

The purpose of this project is to compute the sentiment of text information – in our case, social media posts/tweets and news articles posted recently on the war of Ukraine against russia - and answer the research question: ***"What can public opinion tell us about the russia's war against Ukraine in 2022-2023?"*** The goal is to essentially use sentiment analysis on Twitter, Reddit, Facebook and other social media and news platforms to get insights into the war. For this project, you are encouraged to use data from multiple sources. You are encouraged to develop your sentiment analysis models and compare those to existing models available as Python libraries or on-Cloud.

In addition, based on your sentiment analysis results (especially looking at posts/tweets with negative and positive sentiment, as posts/tweets with neutral sentiment are less informative), you will need to identify factors/reasons/topics that drive sentiment. Consider that you have been hired by a government agency or a news outlet as a consultant to gauge how Ukraine and the war is viewed internationally and suggest actions that could improve Ukraine's image. Your task is to analyze social media posts/tweets and to identify key factors/reasons/topics from those in order to determine how Ukraine is perceived on the international scene and to write a report presenting your findings and your suggestions to the Ukrainian government and international NGOs for changes in their current portfolio of strategies that are projected to have a positive impact on Ukraine's international presence and image. If necessary, you may compliment your analysis of posts/tweets by analyzing news articles from traditional media.

When making your conclusions and recommendations be careful to distinguish sentiment about Ukraine, sentiment about russia, sentiment about the war, sentiment about opinions of Elon Musk, etc.

Learning Objectives

- Develop the ability to work in a team on a consulting project. (You are required to work on the project in the same group as for your in-class presentation. Check the Quercus portal for the list of your group members.)

- Improve on skills and competencies required for performing a full cycle of data science and analytics workflow, i.e., data collection and pre-processing, applying algorithms to analyze data, trend identification, storytelling based on analytics (writing a consulting report and delivering an oral presentation).

Tools Allowed

- You can use Python libraries mentioned in-class as well as any other Python libraries you find during your research. Note that you can only use Python 3.
- For visualizing results in your report and presentation you may use Python or any other outside tool, e.g., Tableau, Power BI, etc.

TO DO:

Finish the following four parts **based on your data analytics:**

Part 1 – Sentiment modeling:

A dataset of classified tweets to train your sentiment analysis models is provided. The *sentiment_analysis.csv* file contains tweets that have had their sentiments already analyzed and recorded as binary values 0 (negative) and 1 (positive). Each line is a single tweet, which may contain multiple sentences despite their brevity. The comma-separated fields of each line are:

0	ID	Tweet ID
1	text	the text of the tweet
2	label	the polarity of each tweet (0 = negative sentiment, 1 = positive sentiment)

The dataset has been collected directly from the web, so it may contain html tags, hashtags, and user tags. You may use this dataset or any other dataset from the web to train and validate your sentiment models.

Prepare the data using TF-IDF or Bag of Words (word frequency) as your feature engineering technique. Train at least four classification algorithms on the training data: logistic regression, k-NN, Naive Bayes, SVM, decision trees, Random Forest, XGBoost, and Deep Learning models, where each tweet is considered a single observation/example, and the target variable is the sentiment value, which is either positive or negative. Evaluate each model on the test data to obtain accuracy measures. Perform hyperparameter tuning and cross-validation.

If you find out that the dataset provided in *sentiment_analysis.csv* file is not suitable for your purposes, feel free to use an alternative dataset available on the web. Among other datasets, you may consider:

- https://www.kaggle.com/code/avinandandutta/twitter-sentiment-analysis-russia-ukraine-conflict
- https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech
- https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

Select the trained model with the best performance to use in Part 2.

Part 2 – Sentiment classification:

Your task is to apply your trained sentiment classification model from Part 1 to datasets related to russia's war in Ukraine. You are also encouraged to use pre-trained sentiment classification models implemented as Python libraries and on-Cloud (AWS, Google Cloud, Microsoft Azure Cloud, IBM Cloud) to compare results of your model from Part 1 with pre-trained models.

You can use datasets and publications suggested here and find your own datasets:

- *Opinion of influencers* – tweets of Elon Musk and responses to his tweets (posted on Quercus);
- *Reddit sentiment analysis* https://medium.com/@suhdong21/sentiment-analysis-of-reddit-comments-on-russia-ukraine-war-with-python-a3632994942b;
- *Tweets analysis* (sentiment, topic, personnel & resource losses) https://omkargawade.medium.com/russia-ukraine-war-tweets-nlp-analysis-bd10b352316c;
- *Conflict detection* (pdf file "NLP Conflict Detection in the Ukraine Crisis" posted on Quercus), data sources: Reddit, social news aggregation, etc.;
- *Kaggle datasets collected from Reddit* https://www.kaggle.com/datasets/gpreda/russian-invasion-of-ukraine and https://www.kaggle.com/datasets/diyacharya/ukraine-russia-war-reddit-data;
- *War in Ukraine in the Perception of the Russian Population* (opinion polls, Levada Center) https://www.discuss-data.net/dataset/947f9970-7a50-493c-bc78-057f0f5eedf7/files/;
- *Ukrainian new source Hromadske* (easily translated to English) https://hromadske.ua/tags/rosijsko-ukrayinska-vijna;
- *News aggregators* (API or web-scraping), e.g. https://www.pressreader.com/search?query=russia%20ukraine&orderBy=Relevance&searchFor=Articles;
- *Google Trends* https://trends.google.com, Kaggle, and similar data/news aggregation portals.

Feel free to experiment with Python libraries and APIs to download news articles, e.g., https://newsapi.org/docs/client-libraries/python

Summarize results of your sentiment classifications for your datasets. For each dataset, make conclusions about your results. Explain those.

Part 3 – Factor and topics identification via Machine Learning:

Based on your sentiment analysis results (especially looking at posts/tweets with negative and positive sentiment, as posts/tweets with neutral sentiment are less informative), you will need to identify factors/reasons/topics that drive sentiment. Those are factors (reasons, topics) that explain sentiment and can be used for decision making and recommendations in Part 4. You can use any Natural Language Processing models for this part of the project (use existing models or develop your own).

Part 4 – Visualizations, storytelling, recommendations:

Your task is to visualize modeling results obtained in Part 1, 2 and 3. Design visualization(s), e.g., wordclouds, that allows decision makers to grasp public sentiment about Ukraine, sentiment about russia, sentiment about the war, sentiment about opinions of influencers, etc.

In addition, based on your analysis of social media posts/tweets and identification of key factors/reasons/topics you need to explain how Ukraine is perceived on the international scene and to develop a narrative (storytelling via presentation and report) presenting your findings and your suggestions to the Ukrainian government and international NGOs for changes in their current portfolio of strategies that are projected to have a positive impact on Ukraine's international presence and image. If necessary, you may compliment your analysis of posts/tweets by analyzing news articles from traditional media.

Note: the scope of the question is quite wide, and it is advised that you narrow it down based on your interests and expertise. Make the work truly yours.

Project Presentations

- Project presentations are scheduled for **Tuesday, April 4, 6:00-9:00pm (room SF 1101)**, those are open to the public.
- **Do not make your presentation overly technical**. Your audience is business-oriented and may know little about data science, people are interested in the insights that you got from your analysis and why your results can and should be used for decision-making.

What to Submit via Quercus:

1. Your Jupyter notebook with appropriate documentation for every step as well as the relevant data files. Comment out any data retrieval processes (e.g., from web scraping, downloading, APIs, etc.) in your code and replace it with code for reading the corresponding data from files. (**Submit all those data files together with your Jupyter notebook**). Consider the Jupyter notebook as what you would report to senior data scientists and machine learning engineers. Documentation and comments in the Jupyter notebook should contain technical details of your data analysis and be understandable by data science professionals if they run it on their own. Make sure that your Jupyter notebook runs on Google Colab https://colab.research.google.com portal and that all needed data files are included in your submission. If the size of the data files exceeds Quercus's

capacity, those should be stored on a cloud drive (e.g., Dropbox, Google Drive), and the link to the directory should be included in the notebook.

2. A 5 to 10-page consulting report in PDF and DOCX formats that summarizes your findings and results (all graphs should have axes appropriately labelled, all visual materials should be understandable and the graphics of sufficient quality to be easily readable.) This report should be business oriented and cover your problem more extensively than your presentation.
3. Your business-oriented presentation slides in PowerPoint and PDF formats. (Each group will present their findings and results during an in-class 7-minute presentation with 1-2 minutes for questions. Presentations will be timed and stopped after 7 minutes.) It is up to you how many group members present (one, two or all).
4. Video recording of your 7-8 minute presentation in mp4 or avi format.

Marking

- The project is worth 20 points (10 points for your analysis and report and 10 points for your business-oriented presentation).
- The presentation will be graded as follows (10 total marks):
  - 3 marks for organization and delivery (e.g. clarity, enthusiasm, poise)
  - 3 marks for content (e.g. proper visuals, high-level ideas, answering questions)
  - 4 marks for the business pitch (e.g. recommendations, solution to the problem)
- The analysis in Jupyter notebook and the report will be graded as follows (10 total marks):
  - 2 marks for narrowing down the problem and searching relevant data
  - 5 marks for the analysis (e.g. cleaning the data, visualizations, applying algorithms)
  - 3 marks for discussion and insight (e.g. how your analysis contributes to the problem, making a decision, storytelling)
- Every group member gets the same mark for the project. It is your responsibility to determine how you split the work inside your group. At least half of your group needs to be present during the in-class project presentations to answer questions.

Notes

- For the deliverables, consider the Jupyter notebook as what you would report to senior data scientists and machine learning engineers, and the consulting report and the presentation as what you would report to CEOs, VPs, PR managers, university officials, government officials and journalists.
- The presentation would be a visual representation of the executive summary of your report.
- The audience for your presentation and report in particular is business-oriented and includes people who are interested in the insights you gathered from your analysis and how your results should be used for decision-making.