Project 2

Sentiment Analysis

Due Date: 09:00 PM EST., March 22, 2024

MIE 1626, Data Science Methods and Statistical Learning

University of Toronto, Winter 2024

Background

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

Sentiment analysis is a technology of increasing importance in modern society, as it allows individuals and organizations to detect trends in public opinion by analyzing social media content. Keeping abreast of sociopolitical developments is especially important during periods of policy shifts such as election years when both electoral candidates and companies can benefit from sentiment analysis by making appropriate changes to their campaigning and business strategies respectively.

The purpose of this project is to compute the sentiment of text information - in our case, tweets posted recently on Canadian Elections - and answer the research question: **"What can public opinion on Twitter tell us about the Canadian political landscape in** 2021**?"** The goal is to essentially use sentiment analysis on Twitter data to get insight into the Canadian Elections. For this project, we have pulled tweets regarding the Canadian elections from the announcement of the 2021 election to the day before the election for your analysis.

Central to sentiment analysis are techniques first developed in text mining. Some of those techniques require a large collection of classified text data, often divided into two types of data, a training data set and a testing data set. The training data set is further divided into data used solely for the purpose of building the model and data used for validating the model. The process of building a model is iterative, with the model being successively refined until an acceptable performance is achieved. The model is then used on the testing data in order to calculate its performance characteristics.

Produce a report in the form of a **Jupyter notebook** detailing the analysis you performed to answer the research question. Your analysis must include the following steps: **data cleaning, exploratory analysis, model preparation, model implementation, and discussion**. This is an open-ended problem: there are countless different ways to approach each part of the analysis, and therefore the motivation for each step is just as important as its implementation. When writing the report, make sure to explain (for each step) what it is doing, why it is important, and the pros and cons of that approach. Include in your Jupyter file the key findings from the exploratory analysis, model feature importance, or model results.

Two sets of data are used for this project. 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 second data set, *Canadian_elections_2021.csv* contains a list of tweets regarding the 2021 Canadian federal elections. The fields of each line are:

0	text	the text of the tweet
1	sentiment	can be “positive” or “negative”
2	negative_reason	reason for negative tweets. NaN for positive tweets

Both datasets have been collected directly from the web, so they may contain HTML tags, hashtags, and user tags.

Learning objectives

1. Implement functionality to parse and clean data according to given requirements.

2. Understand how exploring the data by creating visualizations leads to a deeper understanding of the data.

3. Learn about training and testing machine learning algorithms (logistic regression, k-NN, decision trees, random forest, SVM, XGBoost, etc.).

4. Understand how to apply machine learning algorithms to the task of text classification.

5. Improve on skills and competencies required to collate and present domain-specific, evidence-based insights.

To do:

1. Data cleaning (10 marks):

The tweets, as given, are not in a form amenable to analysis – there is too much ‘noise’. Therefore, the first step is to “clean” the data. Design a procedure that prepares the Twitter data for analysis by satisfying the requirements below. Remember to use the same pipeline for both datasets.

- All HTML tags and attributes (i.e., $/ < [>]+ > /$) are removed.
- HTML character codes (i.e., &...) are replaced with an ASCII equivalent.
- All URLs are removed.
- All characters in the text are in lowercase.
- All stop words are removed. Be clear in what you consider a stop word.
- If a tweet is empty after pre-processing, it should be preserved as such.

2. Exploratory analysis (15 marks):

- Design a simple procedure that determines the political party (Liberal, Conservative, New Democratic Party (NDP), The People’s Party of Canada (PPC)) of a given tweet and apply this procedure to all the tweets in the Canadian Elections dataset. A suggestion would be to look at relevant words and hashtags in the tweets that identify certain political parties or candidates. What can you say about the distribution of the political affiliations of the tweets?
- Present a graphical figure (e.g. chart, graph, histogram, boxplot, word cloud, etc.) that visualizes some aspect of the generic tweets in sentiment_analysis.csv and another figure for the 2021 Canadian Elections tweets. All graphs and plots should be readable and have all axes that are appropriately labeled. Discuss your findings.

3. Model preparation (15 marks):

Split the generic tweets randomly into training data (80%) and test data (20%). Prepare the data to try seven classification algorithms – logistic regression, k-NN, Naive Bayes, SVM, decision trees, Random Forest, and XGBoost, where each tweet is considered a single observation/example. In these models, the target variable is the sentiment value, which is either positive or negative. Try four different types of features, Bag of Words (word frequency), TF-IDF, word embedding, and N-grams on all 7 models. (Hint: Be careful about when to split the dataset into training and testing sets.)

Resources for Bag of Words, TF-IDF, Word Embedding, and N-grams:

https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22
https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/
https://en.wikipedia.org/wiki/Word_embedding
https://web.stanford.edu/~jurafsky/slp3/3.pdf

4. Model implementation and tuning (60 marks):

4.1. Using four types of features (Bag of Words and TF-IDF, word embeddings, and N-grams), train models (7 models) on the training data from generic tweets. Perform hyperparameter tuning and cross-validation. Apply the model to the test data to obtain an accuracy value. **(40 marks)**

(a) Evaluate the trained model with the best performance on the Canadian Elections data. How well do your predictions match the sentiment labelled in the Canadian elections' data?

(b) Propose three other metrics you could use to evaluate the models. Then evaluate the models accordingly. In one to two sentences, provide reasoning for each metric.

(c) Choose the model that has the best performance and visualize the sentiment prediction results and the true sentiment for each of the 4 parties. From this model, discuss your findings and whether NLP analytics based on tweets is useful for political parties during election campaigns. Explain how each party is viewed in the public eye based on the sentiment value. Suggest one way you can improve the accuracy of this model.

4.2. Split the **negative** Canadian elections tweets into training data (80%) and test data (20%). **Use the true sentiment labels in the Canadian elections' data instead of your predictions from the previous part**. Choose one algorithm from classification algorithms (choose any model from logistic regression, k-NN, Naive Bayes, SVM, decision trees, RF, XGBoost), to train a multi-class classification model to predict the reason for the negative tweets. Tune the hyperparameters and choose the model with the best score to test your prediction reason for negative sentiment tweets. **(20 marks)**

(a) Provide a few reasons why your model may fail to predict the correct negative reasons. Back up your reasoning with examples from the test sets.

(b) Suggest one way you can improve the accuracy of your selected model. Implement it and show the improvement.

(c) Combine similar reasons into fewer categories, as long as you justify your reasoning. You are free to define input features of your model using word frequency analysis or other techniques.

Please clearly label each section of your work. Significant marks for each section are allocated to discussion. Use markdown cells as needed to explain your reasoning for the steps that you take.

Bonus (up to 5 marks): For tree-based approaches, use sentence embeddings from a pre-trained BERT model that is accessible and loadable via $tensorflow_hub-$.

Tools:

- **Software**
  - **Python Version 3.X** is required for this project. Python Version 2.7 is not allowed.
  - Your code should run on the Google Colab cloud.
  - All libraries and built-ins are allowed but here is a list of the major libraries you might consider: Numpy, Scipy, Scikit, Matplotlib, Pandas, NLTK.
  - No other tool or software besides **Python and its component libraries** can be used to touch the data files. For instance, using Microsoft Excel to clean the data is not allowed.
- **Required data files**
  - **sentiment_analysis.csv:** classified Twitter data containing a set of tweets which have been analyzed and scored for their sentiment.
  - **Canadian_elections_2021.csv:** Twitter data containing a set of tweets from 2021 on the Canadian elections, which needs to be analyzed for this project.
  - The data files cannot be altered by any means. The Jupyter Notebooks will be run using local versions of these data files.

What to submit:

1. Submit via Quercus portal a Jupyter notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention:

   lastname_studentnumber_project2.ipynb

   Make sure that you comment your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase and underscores are used as shown. **A program that cannot be evaluated because it varies from specifications will receive zero marks.** Include in your Jupyter file the key findings from exploratory analysis, model feature importance, or model results.

2. Submit PDF of the same Jupyter file with all code, output, texts, and comments.

   Use the following naming conventions:

   lastname_studentnumber_project2.pdf

Late submissions will receive penalty as per syllabus.

Tips:

1. You have some freedom with how you approach some of the steps and what function you want to use. As open-ended as the problem seems, the emphasis of the project is for you to be able to explain the reasoning behind every step.
2. While some suggestions have been made in certain steps to give you some direction, you may decide to use a different, but justifiable approach. Following instructions, however, guarantees full marks if implemented and explained correctly.

TA:

Ahmad Sajedi

Please post your questions to **Piazza** (as a public or a private post). The instructor team will not answer project questions via email.