AI for Business — Lab Assignment 7

Tesla Tweets: NLP, BoW/VADER, and BERT-Based Sentiment Analysis

Instructor: Miao Liu

Overview

This comprehensive lab assignment combines two connected natural language processing (NLP) exercises using a dataset of more than **3 million Tesla tweets** collected in **2018** and **2020**. In the first part, you will clean the raw tweet data, perform exploratory analysis and visualization, and conduct sentiment analysis using a **Bag-of-Words (BoW)** approach with **VADER**. In the second part, you will explore sentiment analysis using **BERT-based models**, including both a default Hugging Face pipeline model and a fine-tuned model trained on **Sentiment140**. Throughout the lab, your goal is to extract actionable business insights and communicate them clearly to a client with no AI background.

Learning objectives

By the end of this lab, you should be able to:

- Clean and prepare large-scale social media text data for analysis.
- Perform exploratory text analysis using tweet counts, engagement metrics, frequent words, bigrams, word clouds, and hashtags.
- Explain and apply sentiment analysis using both a traditional **BoW/VADER** approach and a modern **BERT** approach.
- Interpret sentiment over time and assess how engagement-weighted sentiment can change business conclusions.
- Explain, in plain language, why transformer-based models may behave differently from traditional NLP methods.

How to start (Google Colab + datasets)

1. Go to Google Colab: http://colab.research.google.com/.
2. Upload the provided notebooks for the Tesla tweet analysis.
3. Make sure the Tesla tweet datasets for **2018** and **2020** are accessible in your Google Drive or Colab runtime, and update file paths in the notebooks as needed.

Student Handout

Deliverables

Submit **one PDF report** via Canvas. Your report should include:

- Answers to all required questions below,
- Clean plots and tables (with captions/labels),
- Clear interpretations and discussion of what the results mean for public perception of Tesla.

1 Part A: Tesla Tweets Exploratory Analysis and BoW-Based Sentiment Analysis

This part presents an exploration into NLP using a dataset of more than 3 million Tesla tweets collected in 2018 and 2020. The goal of this analysis is to extract actionable business insights by cleaning the data, performing exploratory analysis and visualization, and conducting a sentiment analysis based on a Bag-of-Words (BoW) approach using VADER.

A1: Data preprocessing

Objective: Clean and prepare the raw tweet data for analysis.

1. Describe the main steps you took to clean the raw data.
2. Discuss potential challenges that might arise if data preprocessing is skipped or done inadequately, and how these challenges might affect business insights.

A2: Exploratory data analysis (EDA) and visualizations

Objective: Understand the overall tweet volume and the distribution of engagement metrics.

1. How many tweets are present in the dataset in total?
2. By plotting the number of tweets per month for 2018 and 2020 separately, what trends or patterns do you observe?
3. Examine the distributions of the following engagement metrics:
   `retweet_count`, `favorite_count`, `reply_count`, and `quote_count`. What do these distributions reveal about the nature of user engagement on Twitter?

A3: Frequent words and bigrams analysis

Objective: Identify the most frequently mentioned words and pairs of words (bigrams) for each year.

1. What are the top 20 most frequently mentioned words in tweets from 2018 and 2020? How do these words reflect the topics of public interest?
2. What are the most common bigrams in each year? Explain what additional context bigrams provide compared to single-word analysis.

A4: Word cloud analysis

Objective: Visualize word frequency using word clouds and draw insights from their patterns.

1. Explain to your clients the intuition behind using a word cloud as a visualization tool. How does a word cloud represent word frequency? You can ask Google or ChatGPT.
2. Compare the word clouds generated for 2018 and 2020. What are the dominant themes in each year?
3. Based on the findings from Steps 3 and 4, what message can you draw about the content of Tesla tweets? What factors appear to drive social perceptions of the company?

A5: Hashtag analysis

Objective: Analyze the role and frequency of hashtags in Tesla tweets.

1. Explain to your clients what a hashtag is and explain its importance in the context of Twitter.
2. Identify the top 10 hashtags for each year (2018 and 2020). How do these hashtags relate to the key topics discussed in the tweets?

A6: BoW-based sentiment analysis using VADER

Objective: Conduct a sentiment analysis on the tweets using a Bag-of-Words (BoW) approach and VADER.

1. Explain to your clients what is a Bag-of-Words approach and why it might be suitable for sentiment analysis in this context.
2. What is VADER, and how does it work? Describe the significance of the composite (compound) sentiment score. You can ask Google or ChatGPT.
3. How do you interpret the overall distribution of sentiment labels in the Tesla tweets? What does this indicate about public opinion?

A7: Sentiment movement over time

Objective: Analyze and interpret the evolution of sentiment in Tesla tweets over time.

1. By calculating the monthly proportion of positive and negative tweets, what trends do you observe over time in 2018 and 2020?
2. Identify any significant shifts or trends in sentiment.

A8: Weighted sentiment analysis

Objective: Refine sentiment analysis by weighting tweets based on engagement metrics.

1. One limitation of the basic sentiment analysis is that it treats every tweet equally. Explain the rationale for incorporating engagement metrics (`retweet_count`, `favorite_count`, `reply_count`, `quote_count`) as weights in sentiment analysis.
2. How would you weight engagement metrics, and why?

3. Compare the results of the weighted sentiment analysis with the unweighted analysis. What additional insights does the weighted approach provide regarding public sentiment?

2 Part B: BERT-Based Sentiment Analysis on Tesla Tweets

Introduction

In this part of the lab, you will explore sentiment analysis on Tesla-related tweets using a **BERT-based model**. The primary objective is to understand how modern transformer-based models, such as BERT, differ from traditional approaches like Bag-of-Words (BoW) in capturing the nuances of language for sentiment classification.

Hugging Face and the Pipeline API

Hugging Face is a leading organization in natural language processing (NLP) that provides an extensive repository of pre-trained models, datasets, and tools for building state-of-the-art NLP applications. Their **Transformers** library is one of the most popular frameworks for working with transformer-based models like BERT, GPT, and others.
In the first part of the BERT analysis, you will use a high-level *pipeline* API provided by the Hugging Face Transformers library. This pipeline API abstracts away the complexities of model loading, tokenization, and inference, allowing you to load a pre-trained model for tasks such as sentiment analysis, question answering, or text generation with a single line of code. For example, the following code initializes the sentiment analysis pipeline:

Listing 1: Initializing the Sentiment Analysis Pipeline

```
sentiment_pipeline = pipeline("sentiment-analysis", device=0)
```

This line does several things:

- It initializes the pipeline using the default pre-trained model for `"sentiment-analysis"`.
- The Hugging Face pipeline for this task defaults to using a model that is fine-tuned for sentiment analysis. Currently, the default model is typically **"distilbert-base-uncased-finetuned-sst-2-english"**, which is a lighter, faster version of BERT that has been fine-tuned on the SST-2 dataset (comprising movie reviews) for binary sentiment classification (positive/negative).
- The parameter `device=0` ensures that, if a GPU is available in the Colab environment, the model will run on the GPU to speed up processing.

Sampling and fine-tuning note

For practice purposes, you will randomly select **500 tweets every month** from the Tesla dataset for the default BERT analysis; this is to save on running time. In real-world practice, a good option to speed up processing is to purchase or use a GPU-enabled environment on Colab.

In the second part of the BERT analysis, you will fine-tune a BERT model using the `bert-base-uncased` model as the base. You will fine-tune this model on the **Sentiment140** dataset, which is labeled with sentiment information for tweets. To save time during training, you will use a subsample of **5,000 tweets** from the Sentiment140 dataset rather than the full dataset. GPUs are required in real-world practices. Still, the fine-tuning can take **1–2 hours to run on CPU**.

B1: Default BERT

1. **Motivation for using BERT:** What is the motivation for performing sentiment analysis using BERT? Specifically, discuss why we expect BERT to perform better than a traditional Bag-of-Words (BoW) approach.
2. **Interpreting the visual outputs:** Examine the figures generated from the default BERT sentiment analysis. How do these visualizations help us understand the distribution of sentiment predictions?
3. **Understanding sentiment scores:** The default BERT sentiment analysis outputs not only a sentiment label (positive or negative) but also a *sentiment score*, representing the probability assigned to the predicted label after applying the softmax function to the logits. Why is this feature absent in traditional BoW models? Additionally, discuss what insights can be drawn from the violin plot that displays the distribution of these sentiment probability scores.

B2: Fine-tuned BERT

1. **Benefits of fine-tuning on domain-specific data:** First, get familiar with the Sentiment140 data: https://www.kaggle.com/datasets/kazanova/sentiment140
   Explain how fine-tuning a BERT model on the Sentiment140 dataset can improve performance for sentiment analysis on Tesla tweets compared to using a default model (e.g., one fine-tuned on SST-2). What are the advantages of using domain-specific data for sentiment analysis, and what challenges might arise when adapting a model from one domain (movie reviews) to another (tweets)?
2. **Interpreting the visual outputs:** Examine the figures generated from the fine-tuned BERT sentiment analysis. Why the results might be different from the default BERT model?

Submission guidelines

- Submit your assignment as a **PDF file** via Canvas.
- Your target audience is a client with no AI background. Your grade depends on clarity, readability, and professional presentation of explanations and visualizations. Do **NOT** include code in the write-up.
- Late submissions will not be accepted unless prior approval has been obtained.