Fall 2023 IST 557: Data Mining: Techniques and Applications

Instructor: Lu Lin

Individual Project III: Text Classification (20 points + 1 bonus point)
Due Date: 11:59 PM, Monday, Nov 27, 2023

Goal: This project is to expose students to real-world Kaggle Data Mining competition for practicing basic skills on **text mining**. DO NOT CHEAT. You can only learn data mining and machine learning by getting your hands dirty. Enjoy it.

Logistics: To successfully complete the Kaggle competition, follow the rules:

(1) ONLY create ONE account for the competition.

(2) Practice text mining techniques: bag-of-word representation, word2vec embedding, RNN.

(3) You can submit your prediction **10 times each day in maximum**. For this project, we only use a **private leaderboard** for final evaluation, which will be published on Nov 30, 2023. Consider cross validation on training data for model selection.

Kaggle Competition: The kaggle competition for this project can be accessed via this link: https://www.kaggle.com/t/9ab7824d2314466d9a5a208d00a2d41c. This competition is created only for students enrolled in IST 557 to participate. **Please do not spread this link**.

Submission Checklist: Submit the following to **Individual Project III on CANVAS**:

(1) **Code file** (runable Jupyter notebook or python file) used to produce the results and predictions. **(if missing, -5 pts)**

(2) **PDF report** summarizing the following:

- Kaggle account name, your name and PSU email **(if missing, -2 pts)**
- **Screenshot of the following required task** (which is also marked as "YOUR TASK" in the sample code)
  - Implementation of two preprocessing steps choosing from the following **(4 pts)**
    - Turn to lowercase;
    - Remove hashtags start with @;
    - Remove special characters
    - Lemmatization (convert words to base form)

    - Implementation of transforming test texts into BOW representation **(1 pt)**
    - Top-10 most similar words to "bomb" based on word2vec embeddings **(2 pts)**
    - Tune the arguments of Word2Vec and show the top-10 similar words to "bomb" based on updated embeddings **(1 pt)**
    - Implementation of document/tweet embedding **(2 pts)**
    - Implementation of using ML on word2vec embeddings **(2 pts)**
    - **1 Bonus point: implementation and result of RNN training**
  - What did you tried and what were your findings? **(3 pts)**

Grading Rubric: Total 20 + 1 bonus points consists of two parts:

- Performance (Recall) on the **Private** leaderboard **(5 pts)**
  - Accuracy $\in [\mathbf{0.70}, 1.00)$: 5 points
  - Accuracy $\in [\mathbf{0.65}, 0.70)$: 4 points
  - Accuracy $\in [\mathbf{0.60}, 0.65)$: 3 points
  - Accuracy $\in [\mathbf{0.55}, 0.60)$: 2 points
  - Accuracy $\in [\mathbf{0.50}, 0.55)$: 1 points
  - Accuracy $\in [\mathbf{0.00}, 0.50)$: 0 points
- Report (or Jupyter notebook) with the required contents listed above **(15 pts + 1 bonus)**