CSE556: Natural Language Processing

Assignment 03

Deadline: 6th Nov, 11:59:59 PM

Instructions:

1. The assignment can be submitted in a group of maximum two members. However, it is not mandatory to do assignments in a group.
2. Language allowed: Python
3. You are allowed to use libraries for data processing.
4. For Plagiarism, institute policies will be followed
5. You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Output.pdf.
6. Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
7. Mention your sample outputs in the output.pdf.
8. You are advised to prepare a well-documented code file.
9. Submit code, readme, and output files in ZIP format with the following name: A3_rollno.zip
10. Make sure to use Pickle or any other library to save all your trained models. There will not be enough time during the demo to retrain your model. This is a strict requirement. Use this link to understand more about how to use Pickle.
11. Use classroom discussion for any doubt.

NOTE:

1. **If the assignment is attempted in a group, then, kindly mention the contribution of each member in the report.**
2. **You need to perform both the tasks.**
3. **Evaluation will be based on the number of features incorporated in the feature vector.**
4. **Bonus (max 5 marks) will be awarded for incorporation of additional features (must be informative features considering the task, i.e., sentiment or emotion).**
    a. **Features should be other than the ones mentioned in the two papers.**
    b. **Should have a positive effect on the model. No bonus will be awarded for irrelevant features.**

Assignment 3a: Sentiment Analysis in Twitter

Given a message, classify whether the message is of positive, negative, or neutral sentiment.

Features to implement: You are provided with a paper pdf. You need to construct the feature vector as given in the paper (Section 3.1). [Note: You can skip the 'cluster' feature.]

DATA SET: You can download training and testing data from here.

Dataset contains tweets extracted using the twitter api. It contains the following 6 fields:

1. target: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. ids: The id of the tweet (*2087*)
3. date: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
4. flag: The query (*lyx*). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (*robotickilldozr*)
6. text: the text of the tweet (*Lyx is cool*)

Assignment 3b: Emotion Intensity Prediction

Given a tweet and an emotion X, determine the intensity or degree of emotion X felt by the speaker -- a real-valued score between 0 and 1. The maximum possible score 1 stands for feeling the maximum amount of emotion X (or having a mental state maximally inclined towards feeling emotion X). The minimum possible score 0 stands for feeling the least amount of emotion X (or having a mental state maximally away from feeling emotion X). The tweet along with the emotion X will be referred to as an instance.

Features to implement: You are provided with a paper pdf. You need to construct the feature vector as given in the paper((Section 2.3.1) + N-grams features (N= 1, 2))

DATA SET: https://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html
You are required to use the training sets of the two emotion classes - **Anger and Joy**, for training the model and the test sets of the emotion classes for testing the model.

Machine learning algorithms:

A. **Naive Bayes:** Implement the algorithm from scratch on the provided dataset. Train the model on the train set and report the performance metrics on the test set.
B. **Decision Trees/SVM/MLP**
    a. Use scikit-learn library implementation.

Evaluation:

1. **Sentiment Analysis in Twitter**
    a. Report accuracy, precision, recall, and macro-average F1 score.
    b. Confusion matrix (as shown below). You are not allowed to use any predefined library function.

		Gold Standard		
		POSITIVE	NEUTRAL	NEGATIVE
Predicted	POSITIVE	PP	PU	PN
	NEUTRAL	UP	UU	UN
	NEGATIVE	NP	NU	NN

2. **Emotion Intensity**
    a. Report pearson and spearman correlation. Use the evaluation script available at https://github.com/felipebravom/EmoInt