Twee Emotion Detection

Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. Numerous social media sites have enabled the exchange of opinions among users all over the world that has promoted the use of popular social network site such as twitter, for communication. The users' tweets are highly unstructured, heterogeneous, and vulgar, and they cover a wide range of topics.

Detecting emotions in the sentiment analysis area is one of the most important applications and also serves as an advantage in the digital medium for efficient computing. In the current scenario, sentimental analysis or opinion mining of the twitter emotion detection data-set has derived much attention since the past 10 years. The basic idea is to find the polarity of the text and classify it into positive, or negative and sometimes into neutral. It helps in human decision making.

For more information about emotion detection from tweet please refer to the following:

- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9309291

- https://ceur-ws.org/Vol-3142/PAPER_04.pdf

- https://arxiv.org/pdf/1901.08458.pdf

2. Problem Statement
The goal is to develop an automated solution that predict the emotion expressed in a tweet based on extracted features from the tweet content.

3. Data set:
You will train and test your system using the attached data set. The dataset contains a set of tweets. Every tweet is labeled as positive and negative according to its sentimental polarity using symbol neg and pos. The dataset is balanced and collected using positive and negative emojis lexicon. The dataset consists of two files one for the positive tweets and the other for the negative tweet. The dataset presented in Tab-separated values TSV format.

4. Tasks
A machine learning life cycle describes the steps a team (or person) should use to create a predictive machine learning model. Therefore, you need to work on each stage except data gathering stage. In summery you need to work on the following stages or tasks:

1. Date preprocessing: you must analyze and decide what to do in this stage. In other words, you need to clean and normalize text in the tweet.
2. Features extraction: in this stage you may decide to use automated methods like language models (classical or deep learning) or hand designed features or combined.
3. Training: you need to train and test three classifiers including: Decision trees (J48 or C4 or Random Forests), Naïve Bayes, and any classifier you choose.
4. You need to build model using the classical method (75% training and 25% testing) and using 5-foldes cross validation.

5. Submissions: please submit the following:
1. **Report (in PDF format)**:
- Describe in details your formalization of the problem including the stages of your solution, selected features, classifiers, and results.
- Describe in details how you designed each feature.
- The results you obtained including evaluation method, and measures (confusion matrix, accuracy, precision, recall, and F-measure). You need to include snapshot from the output and show results in graphs. Also, you need to compare between performance of the classifier you trained.

2. **Source Code:** Include all the source code you developed or extended from the program. These need to be submitted only electronically (no hardcopies of the code). The running program needs also to be submitted electronically.

3. **Demo:** You will be asked to demo your work to your instructor. For that you need to be able to work with your program, introduce minor modifications and defend your choices.

Honor Policy: All are required to adhere to the University honor policy and violations will be dealt with according to university regulations.

Good Luck