**IQGateway Assessment**

Name: Rugwesh Kolekar Branch: B. Tech Electronics and Telecommunications

College: Symbiosis Institute of Technology, Pune

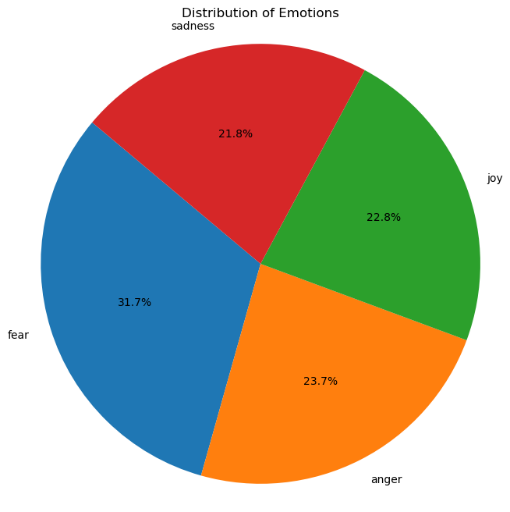
Codalab username: rugways

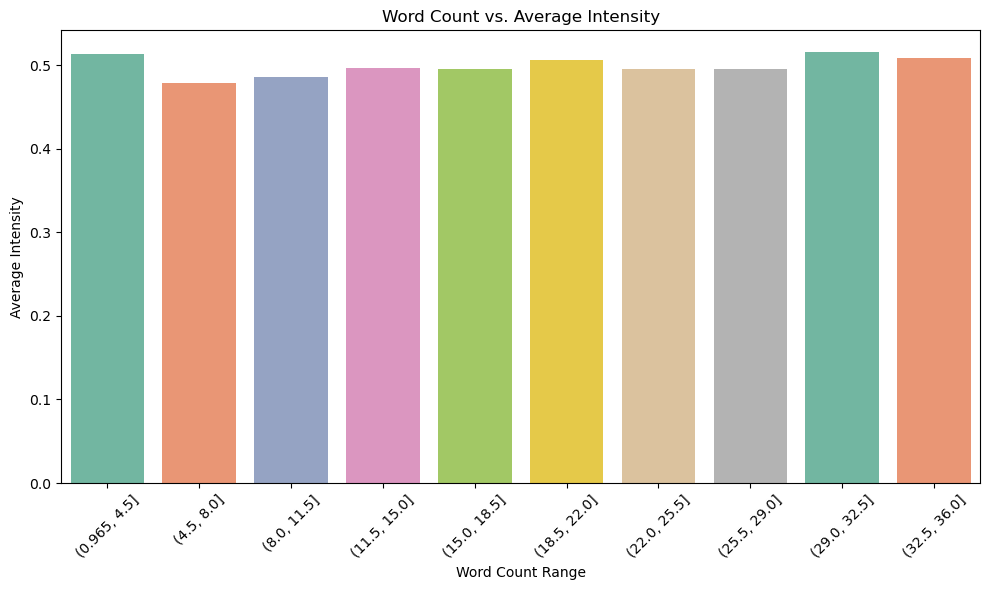
Task 1 EmoInt

**DATA CLEANING**

The Dataset contains short forms of textual data called tweets from Twitter which are not always in systematic format and have grammatical errors. Therefore, it is necessary to clean the raw text to filter noisy data.

For Analysis purpose, each tweet is standardized by converting them to lowercase. Hashtag symbol is removed for each hashtag (#) treating them as single word. Words which act as noise known as Stopwords are removed using NLTK library.





A diagram of a machine learning process

Description automatically generated

**FEATURE EXTRACTION**

Feature Extraction: Since other lexicon features add meaning to the word. Two primary methods for feature extraction from the tweets’ raw text, namely annotated lexicons and pre-trained word embeddings.

It utilizes personalized lexicons for emotion and sentiment intensity. Positive sentiment lexicon features are mapped to some emotions and negative lexicon features to others.

NRC Affect Intensity Lexicon (AI): This lexicon assigns distinct emotion labels to unigrams and provides the intensity at which the emotion is expressed. Each of the emotions evaluated in the EmoInt shared task are represented in this lexicon, and a floating-point intensity score is assigned to each unigram emotion pair.

NRC-Emotion-Lexicon & NRC-Hashtag Emotion Lexicon: These lexicons contain the association of unigrams and Twitter hashtags with eight emotions (inclusive of the four emotions evaluated in this EmoInt task).

NRC -Lexicons:

The first two lexicons associate words with positive/negative sentiment and the other two associate words with similar sentiment labels in affirmative or negated contexts generated automatically from tweets with sentiment emoticons and sentiment-word hashtags. The terms in these lexicons can be unigrams, bigrams or pairs of unigrams and bigrams.

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity. If a word does not have an entry or synonym in SentiWordNet, the positive and negative sentiment scores are assumed to be zero.

DepecheMood: a high coverage and high-precision lexicon of roughly 37 thousand terms annotated with emotion scores.

Word Embeddings: Vector representations of each of the tweets are generated from pre-trained word embeddings. I have used three word embedding sources. Word2Vec models, and GloVe models.

Word2Vec Model- Google News (W2VGN): Word2Vec is a technique for learning low-dimensional word embeddings for words in a corpus, based on the continuous bag-of-words (CBOW) and skip-gram models.

W2V-GN is trained on the Google News corpus containing over 100 billion words. It is a skip-gram model containing 300dimensional embeddings for 3 million distinct words and phrases.

GloVe Model: GloVe builds a word co-occurrence matrix for the entire corpus prior to training.

ML Classifer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotion | Anger | Fear | Joy | Sadness |
| MAE | 0.1211 | 0.1458 | 0.1551 | 0.1340 |
| MSE | 0.0225 | 0.0298 | 0.0357 | 0.0263 |
| RMSE | 0.1502 | 0.1727 | 0.1891 | 0.1623 |

XG Boost:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Emotion | pears-corr | spear-corr | pears-corr-range-05-1 | spear-corr-range-05-1 |
| Anger | 0.5437 | 0.5198 | 0.4556 | 0.4415 |

**MODEL LEARNING**

Linear regression assumes a linear relationship between the input features and the target variable. It aims to find the best-fitting straight line through the data points. code fits a linear regression model to both BoW and TF-IDF representations of the training data and then applies the trained models to make predictions on the development dataset.

Support Vector Regression (SVR):

Support vector regression is a regression algorithm based on support vector machines (SVMs). It aims to find the hyperplane that best fits the data while minimizing the margin violations.

Along with these ML models, Deep learning Model called gradient boosted regression models was implemented. The final system implementation uses the boosted regression implementation provided by the XGBoost library.

**Observation**:

The results of our study show that different features work best for each specific emotion in our task of analysing sentiment. We found that using pre-trained word embeddings, which are representations of words learned from vast amounts of text data, such as Word2Vec and GloVe, tended to perform the best across almost all emotions. Interestingly, including sentiment polarity lexicons (lists of words categorized by their sentiment) helped improve the accuracy of our predictions across all models.