uHack Sentiments 2.0: Decode Code Words - MachineHack

Problem Statement

The last two decades have witnessed a significant change in how consumers purchase products and express their experience/opinions in reviews, posts, and content across platforms. These online reviews are not only useful to reflect customers' sentiment towards a product but also help businesses fix gaps and find potential opportunities which could further influence future purchases.

The challenge here is to analyze and deep dive into the natural language text (reviews) and bucket them based on their topics of discussion. Furthermore, analyzing the overall sentiment will also help the business to make tangible decisions.

The data set provided to you has a mix of customer reviews for products across categories and retailers. We would like you to model on the data to bucket the future reviews in their respective topics (Note: A review can talk about multiple topics)

Overall polarity (positive/negative sentiment)

`Train: 6136 rows x 14 columns`

`Test: 2631 rows x 14 columns`

`Topics (Components, Delivery and Customer Support, Design and Aesthetics, Dimensions, Features, Functionality, Installation, Material, Price, Quality and Usability)`

`Polarity (Positive/Negative)`

Skills

```
 * Text Pre-processing – Lemmatization , Tokenization, N-Grams and other relevant methods
* Multi-Class Classification, Multi-label Classification
* Optimizing Log Loss
```

Solution Approach

```
 * Text Preprocessing using Different Pipelines
* Text Features based approach using Feature Engineering
* Transfer Learning Using Simple Transformers using different models. Eg BERT, Roberta
* Multi label Straified Cross Validation
* Optimizing Log Loss using Ensemble approach to reduce Variance
```

Steps to reproduce

- Run Below Notebooks in below Sequence
- **mh-uhack_text_feats_v5**

```
 * Approach
    1. Cleaning Text using regex ,removal of stopwords using NLTK package
    2. feature Engineering For text Based Features like, word length, density, upper case count, pos tagging, sentiment extrac
    3. Understanding different Topics & creating features based on those
    4. Bag of Words Approach using TFID
    5. OneVsRestClassifier with LGBM Modelling using Multi Label Stratification 10 folds
* Local CV & Score
    * Mean Local CV Score - 3.48 Log Loss
* Solution File - MH_uhack_s5.csv
```

- **mh-uhack-transformers_v10** ```text
    - Approach
        1. Cleaning Text using regex ,removal of stopwords using NLTK package
        2. Understanding Rarewords in train & test data
        3. Simple transformers using MultiLabelClassificationModel using Multi Label Stratification 5 Folds
        4. Hyperparameters - Model - Roberta (Roberta-large) | Epochs-7 | Max Seq Length - 300 | Learning Rate - 3e-5
    - Local CV & Score
        - Mean Local CV Score - 3.00 Log Loss
    - Solution File - mh-transformers-v10.csv ```
- **mh-uhack-transformers_v13** ```text
    - Approach
        1. Cleaning Text using regex ,removal of stopwords using NLTK package, Keeping Train & Test Vocab in sync
        2. Simple transformers using MultiLabelClassificationModel using Multi Label Stratification 5 Folds
        3. Hyperparameters - Model - BERT (Bert-large-Uncased) | Epochs-5 | Max Seq Length - 256 | Learning Rate - 3e-5
    - Local CV & Score
        - Mean Local CV Score - 3.06 Log Loss

    - Solution File - mh-transformers-v13.csv ```
- **mh-uhack-transformers_v14** ```text
    - Approach
        1. Cleaning Text using regex ,removal of stopwords using NLTK package
        2. Simple transformers using MultiLabelClassificationModel using Multi Label Stratification 5 Folds
        3. Hyperparameters - Model - Roberta (Roberta-large) | Epochs-7 | Max Seq Length - 300 | Learning Rate - 3e-5
    - Local CV & Score
        - Mean Local CV Score - 3.14 Log Loss
    - Solution File - mh-transformers-v14.csv ```
- **mh-uhack-transformers_v16** ```text
    - Approach
        1. Cleaning Text using regex ,removal of stopwords using NLTK package
        2. Simple transformers using MultiLabelClassificationModel using Multi Label Stratification 5 Folds
        3. Hyperparameters - Model - Roberta (Roberta-base) | Epochs-10 | Max Seq Length - 300 | Learning Rate - 5e-5
    - Local CV & Score
        - Mean Local CV Score - 3.19 Log Loss
    - Solution File - mh-transformers-v16.csv ```
- **mh-uhack-transformers_v18** ```text
    - Approach
        1. Cleaning Text using regex ,removal of stopwords using NLTK package, Keprt Casing intact
        2. Simple transformers using MultiLabelClassificationModel using Multi Label Stratification 5 Folds
        3. Hyperparameters - Model - BERT (Bert-large-Cased) | Epochs-5 | Max Seq Length - 256 | Learning Rate - 4e-5
    - Local CV & Score
        - Mean Local CV Score - 3.11 Log Loss
    - Solution File - mh-transformers-v18.csv ```
- **Ensemble Approach** ```text
    - Approach
        1. Weighted Average of mh-transformers-v10.csv, mh-transformers-v13.csv, mh-transformers-v14.csv based on CV score (ensemble_1.ipynb)
        2. Weighted Average of mh-transformers-v10.csv, mh-transformers-v14.csv, mh-transformers-v16.csv, MH_uhack_s5.csv based on CV score (ensemble_2.ipynb)
    - Final Solution File - ensemble_2.ipynb (ens_1.csv) ```

Leaderboard

```
 * Public Leaderboard 2nd Rank
* Private Leaderboard 2nd Rank
```

Portfolio

Site	Links
Linked In	https://www.linkedin.com/in/rajat-ranjan24/
GitHub	https://github.com/rajat5ranjan/
Website	https://rajat5ranjan.github.io