**INTERIM REPORT – TEMPLATE**

Your interim report is also the milestone-1. It should Summarize the 1st 3 deliverables mentioned in the project document. The report structure and brief requirements of each section is listed below.

1. Summary of problem statement, data and findings

Every good abstract describes briefly what was intended at the outset, and summarizes findings and implications.

2. Summary of the Approach to EDA and Pre-processing

Include any insightful visualization you have teased out of the data. If you’ve identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys.

3. Deciding Models and Model Building

Based on the nature of the problem, decide what algorithms will be suitable and why?

Experiment with different algorithms and get the performance of each algorithm.

4. How to improve your model performance?

What are the approaches you can take to improve your model? Can you do some feature selection, data manipulation and model improvements.

Provide your code and as much as visualizations you can share to describe what you have done so far.

1. Summary of problem statement, data and findings

Data summary:

Data and Findings

1. Summary of the Approach to EDA and Pre-processing
   1. EDA
      * Understanding the structure of data
      * Missing points in data
      * Finding inconsistencies in the data
      * Visualising different patterns
      * Visualising different text features
      * Dealing with missing values
      * Visualization for Each groups distribution of data
      * Number of words related visualization
      * Wordcloud for top 5 category
   2. Pre-processing

Discussion points : Are we going to use all language or Only English

* + 1. Remove URL
    2. Remove HTML
    3. Get only Alphabet
    4. Lowercase
    5. Remove punctuation
    6. Remove stop word
    7. Remove space
    8. Lemmetize /Stemming
  1. Vocalbulary and corpus
* Creating word vocabulary from the corpus of report text data
* Creating tokens as required

1. Deciding Models and Model Building

Vectorization methods []

Count Vector

TDIDF

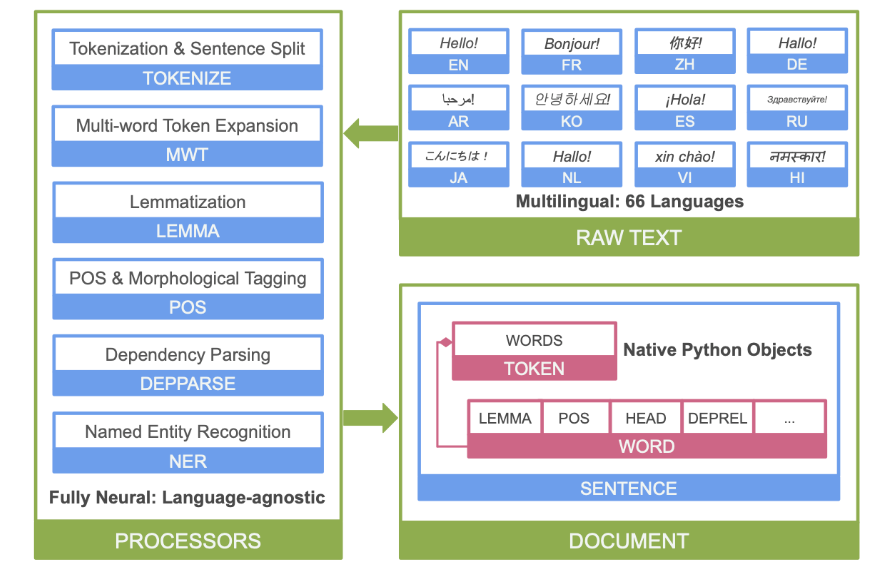
Glove

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Algorithm Name | Accuracy | Precision | Avg Recall | F1-Score |
| 1 | Naïve Bayes |  |  |  |  |
| 2 | Logistic Regression |  |  |  |  |
| 3 | SVM |  |  |  |  |
| 4 | KNN |  |  |  |  |
| 5 | Random Forest |  |  |  |  |
| 6 | XGBoost |  |  |  |  |
| 7 | LSTM |  |  |  |  |
| 8 | Bidirectional LSTM |  |  |  |  |
| 9 | FastText |  |  |  |  |
| 10 | BERT |  |  |  |  |

1. Prediction with one model

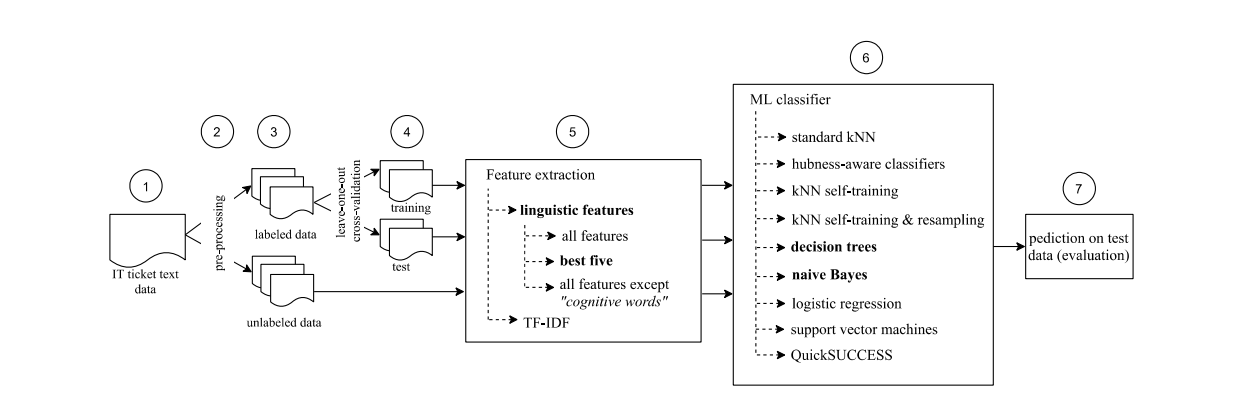
Sequence of steps in Python:

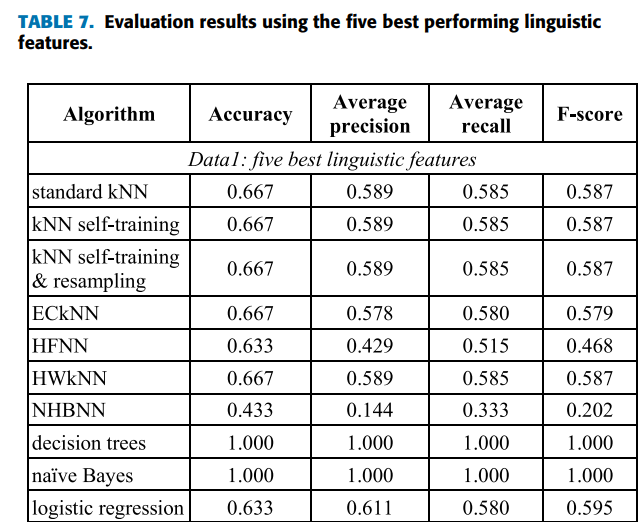
1. Data visualization and Description of all data
2. Data pre- processing on Single method that includes



Sample preprocessing🡪 As per the steps we use, I will update the diagram

Model Evaluation with Various approaches:





Reference:

<https://stanfordnlp.github.io/stanza/>

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9234428