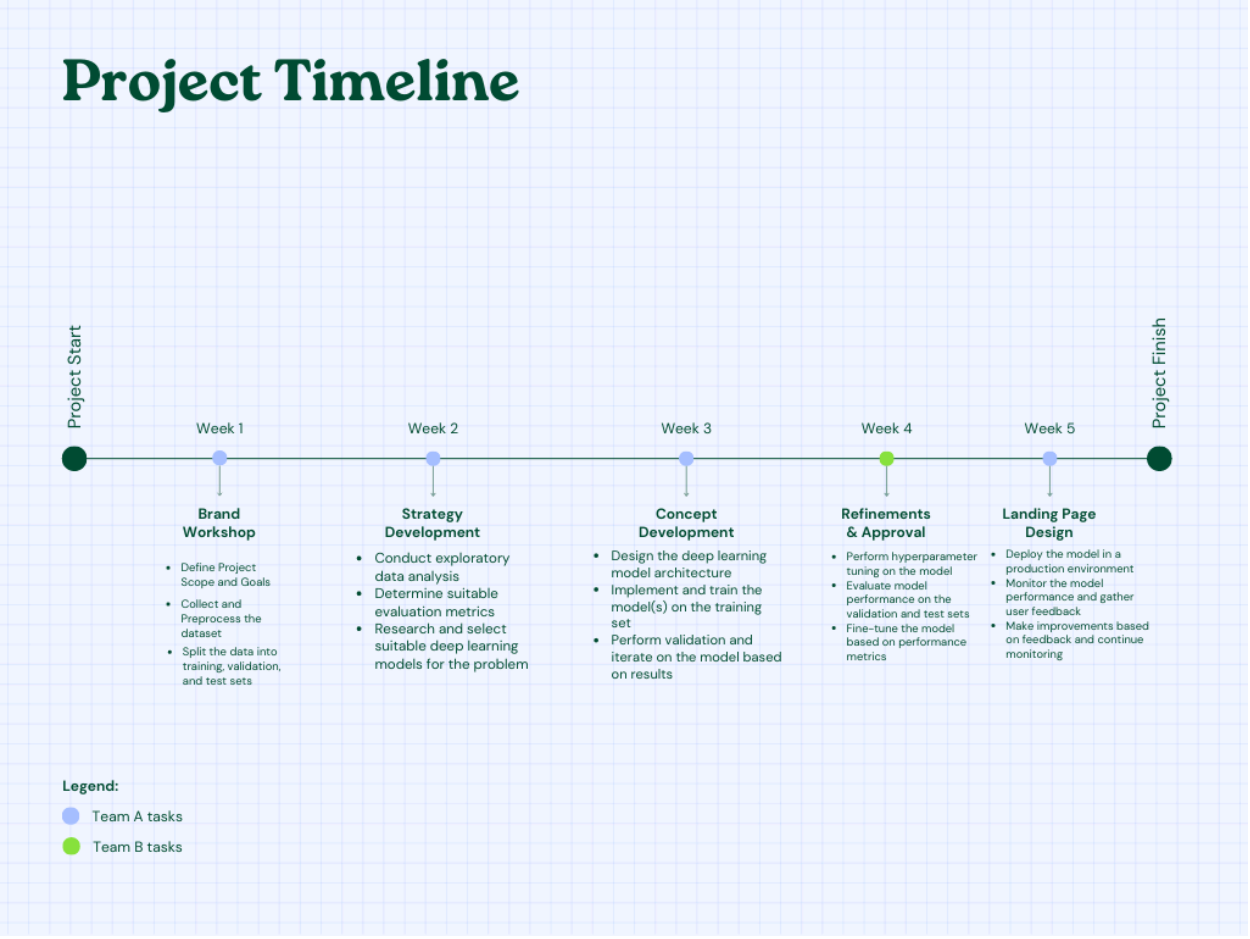
DL Project Summary

1. Group Number: 1
2. Group Members:
    a. **Jash Damani B013 70021019015**
    b. **Saksham Seth B045 70021019052**
    c. **Vedant Misra B091 70011019052**
    d. **Riya Adsul B100 70021018157**

3. BTech Computer, SEM 8
4. Subject Code:
5. Problem Statement: To use Deep Learning to respond quickly and accurately to citizen inquiries by offering real-time recommendations based on rapid classification of customer complaints and feedback, and to understand public sentiment by analyzing feedback to assist in decision-making based on citizen perspectives.T
6. Title:**Improve Citizen Engagement and Handle Complaints with Deep Learning**
7. Gantt Chart:



8. Dataset Link:
   https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data

   520K complaints in the form of text data, classify the complaints into one of 11 categories.

9. List of software used:
   Google Colab
10. Preprocessing methods: Normalization

    Remove X: Using Regular Expressions, removes all instances of 'X' and any dates as well.

    Clean Text: Removes all symbols and punctuation from the text

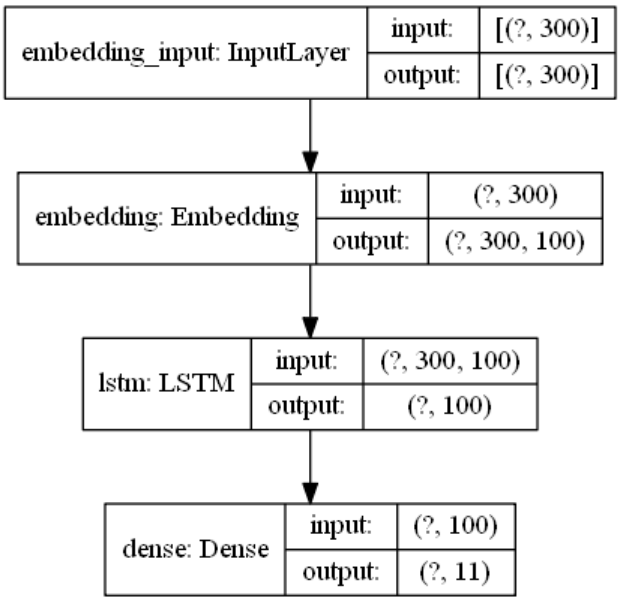
    Tokenize Text: Tokenizes the text into words that can be used for processing

    Lowercase Text: Changes the case of all words to lowercase to ensure consistent handling (Bank --> bank)

    Lemmatize Text: Lemmatizes words by reducing them to their base form (Ex. walking --> walk, better --> good)

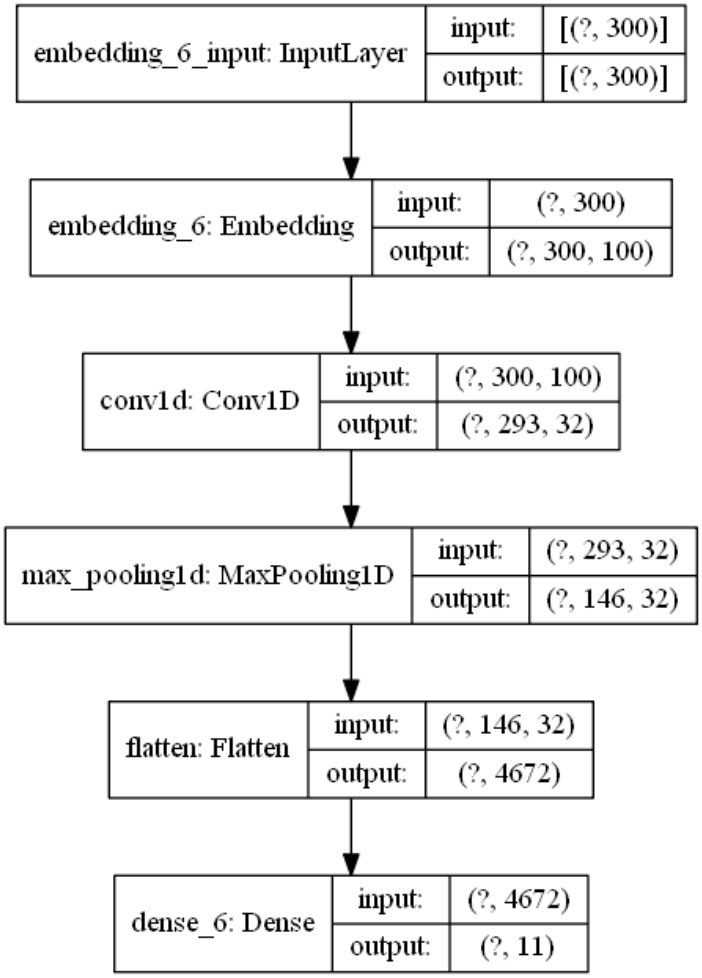
    Remove Stopwords: Removes stopwords such as "the", or "I" which are commonly found in the English language.

11. Architecture:
    In the project, we are showing a comparison between 2 different methods that can be used to analyze customer feedback and subsequently use that analysis to understand the general sentiment of the consumers. In doing so, 5 distinct methods have been implemented. These are: DistilBERT, Bi-Directional LSTM and LSTM, CNN, LDA with Logistic Regression. Out of these LSTM achieved the highest accuracy with 97%.
    LDA+LR
    LSTM:



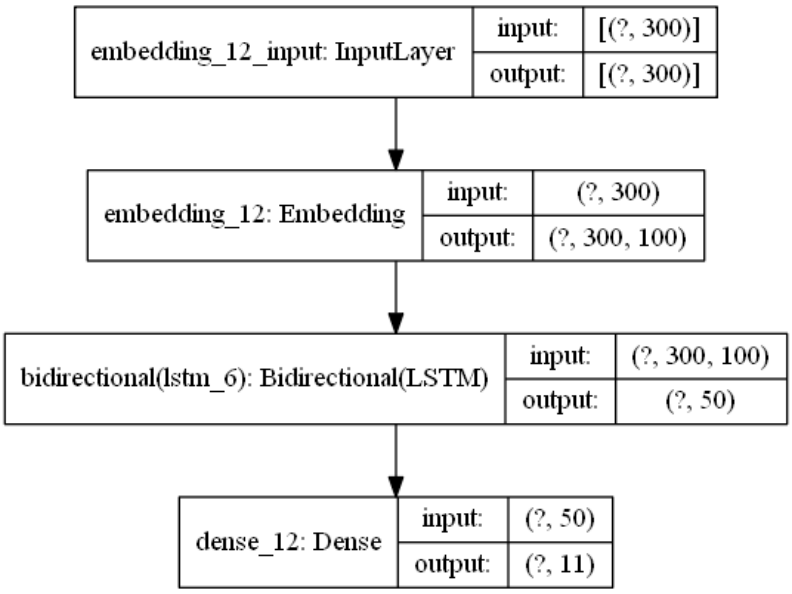

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 300, 100)          10000000
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_1 (Dense)              (None, 11)                1111
=================================================================
Total params: 10,081,511
Trainable params: 10,081,511
Non-trainable params: 0
```

CNN:




```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (None, 300, 100)          10000000
_________________________________________________________________
conv1d (Conv1D)              (None, 293, 32)           25632
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 146, 32)           0
_________________________________________________________________
flatten (Flatten)            (None, 4672)              0
_________________________________________________________________
dense_6 (Dense)              (None, 11)                51403
=================================================================
Total params: 10,077,035
Trainable params: 10,077,035
Non-trainable params: 0
```

BiLSTM

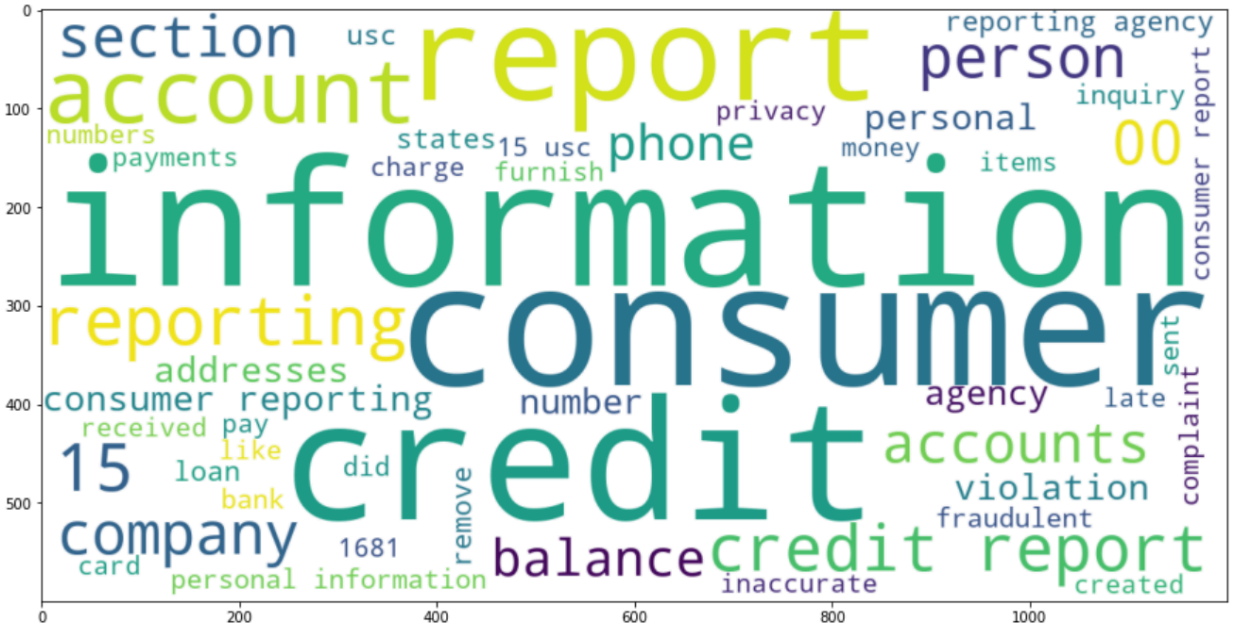



```
Model: "sequential_13"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_13 (Embedding)     (None, 300, 100)          10000000
_________________________________________________________________
bidirectional_1 (Bidirection (None, 50)                25200
_________________________________________________________________
dense_13 (Dense)             (None, 11)                561
=================================================================
Total params: 10,025,761
Trainable params: 10,025,761
Non-trainable params: 0
```

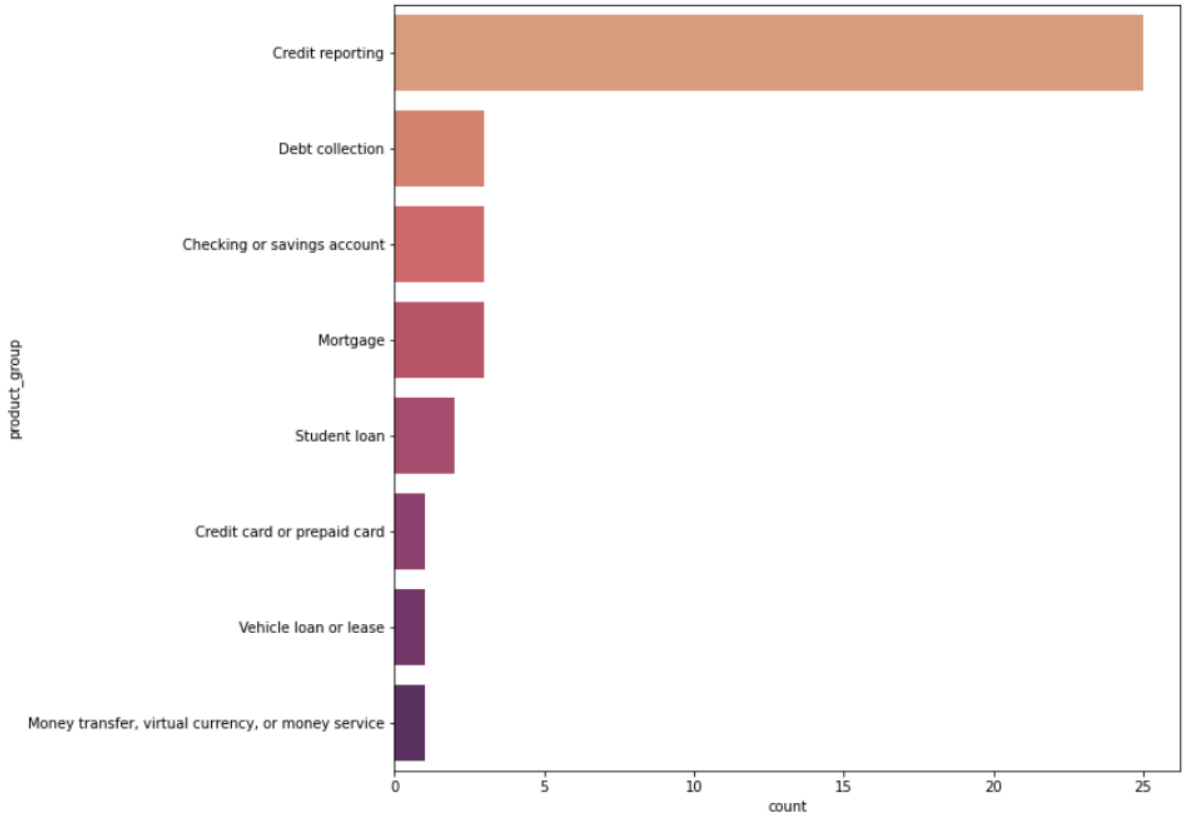
12. Pre-Training Models (if any):
13. Algorithms(Names):
    - Latent Dirichlet Allocation (Topic Models) + Logistic Regression (Supervised Classification)
    - Long Short-Term Memory Network (LSTM) Model
    - Convolutional Neural Network (CNN) Model
    - Bi-Directional LSTM (BiLSTM) Model
    - DistilBERT

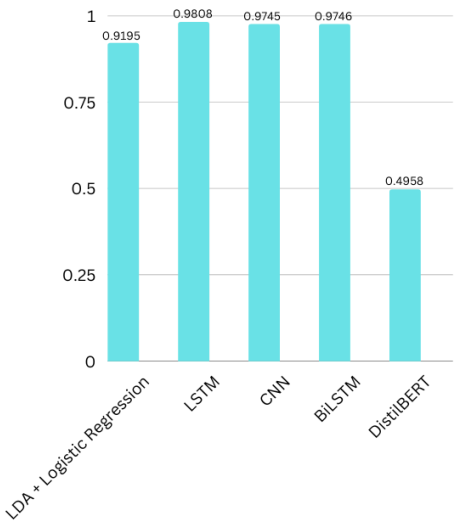
14. All metrics(snapshot of all graphs with proper X, Y axis)

Word Cloud



Bar Graph for Categories of Products present





15. Limitations:
    While other pretrained models are giving a decent auc accuracy
16. Future Scope:
    Attention Layers - Experiment with adding attention layers to improve model performance
    Transformer Neural Network - Given time and computational power, one possible approach would be to approach this problem with a transformer such as BERT, ELMO, or GPT-2
    Semantic Role Labeling - It would be interesting to analyze the more complex and lengthy complaints using semantic role labeling, and potentially researching how this can improve text classification. This would help in understanding how multiple customer service agents followed up with a customer's requests and the steps they took to resolve a complaint. This way, we would be able to see who was able to impact whom in a given complaint.
    Scaling Up - Given that this data was ~ 520K records, large-scale data management was not required. In the future, for a dataset consisting of ~ 1M or more records, tools such as PySpark and SparkNLP can certainly improve performance and will be necessary to some degree. Additionally, GPU computation on an AWS, Azure or Google Cloud instance could significantly improve training times for models.

17. Conclusion:
    Through this project we have compared various models for customer complaint segmentation and concluded that LSTM works well for the given datatset to categorize the complaints into 9 product categories, we also observe that transformer based approach for this dataset is not suitable.