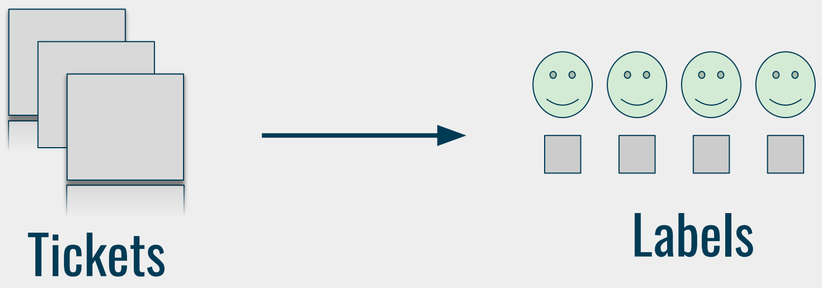
MACHINE LEARNING BASED TICKET CLASSIFICATION - PRATT & WHITNEY

Team: Aishwarya Sudhakar, Ariel Reches, Chris Watson, Kruti Chauhan | Mentors: Abe Handler, Nicholas Monath | Prof Andrew Mccallum

ABSTRACT

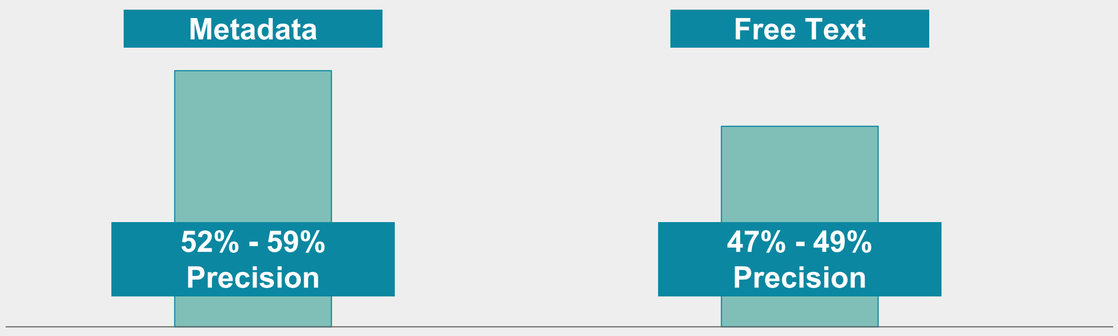
The goal of the project is to build a machine learning based ticket classification model. This model would take the incoming tickets in a system and classify them to assignees. The model uses descriptive and categorical domain specific text for learning. Classification can be supervised learning by assigning a ticket to a specific analyst or unsupervised learning where the tickets are put in clusters, the better performing model is to be chosen.



CHARACTER N-GRAM BASED CLASSIFIER

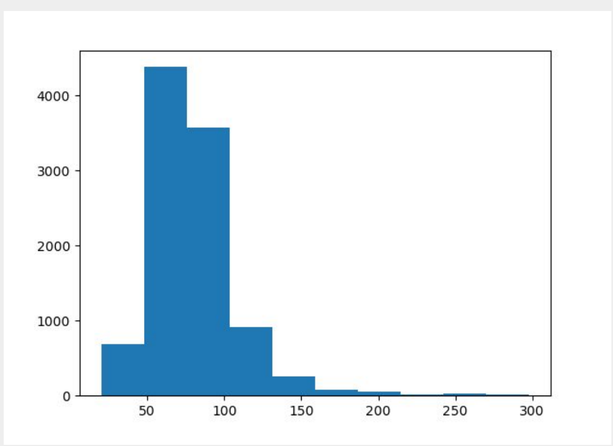
Used FastText - Facebook's Hierarchical Classifier.

- Unbalanced classes
- E.g:Getcontentdimensions | Browserelementchildpreload
- Sub-word embeddings / character n-grams
- Obtain better accuracy with categorical data



DATASET

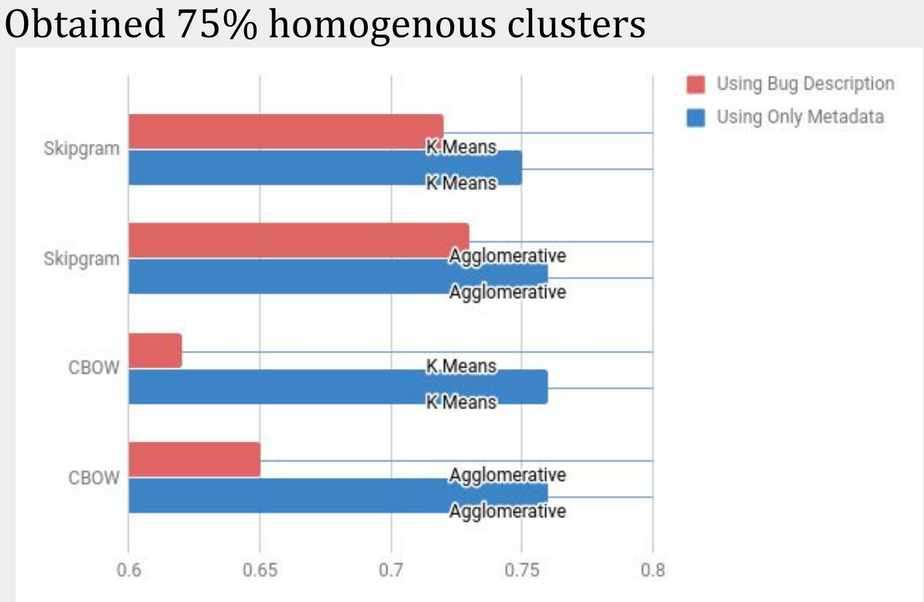
Mozilla Bug Dataset (9937 Open Bugs)
Cateogrical, Free Text and Unique Id Columns



Free Text Column - Length - Histogram

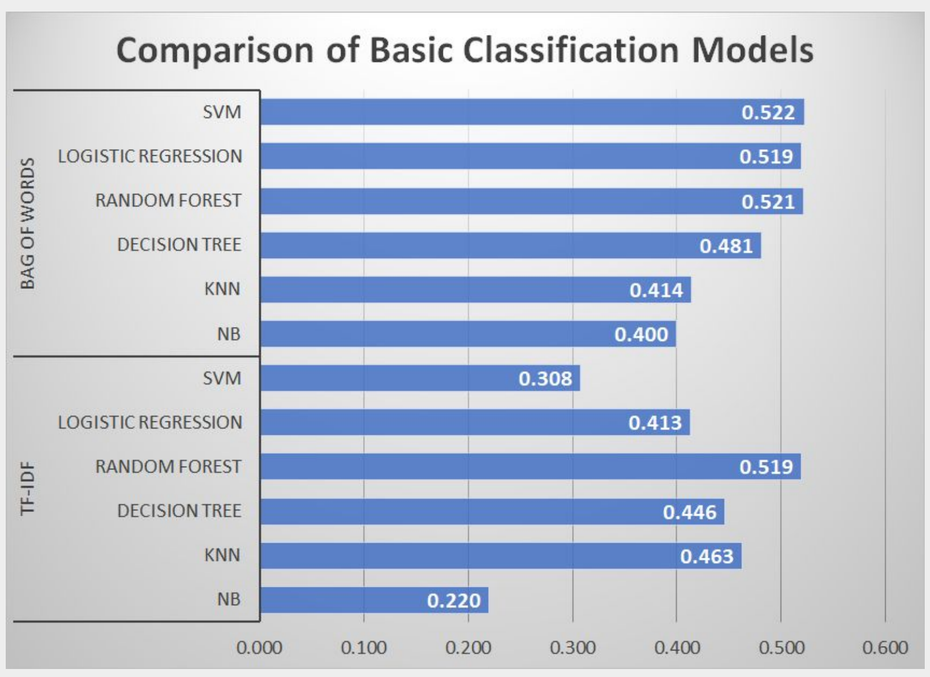
CLUSTERING

- Cluster input data making each analyst their own cluster
- Obtained 75% homogenous clusters



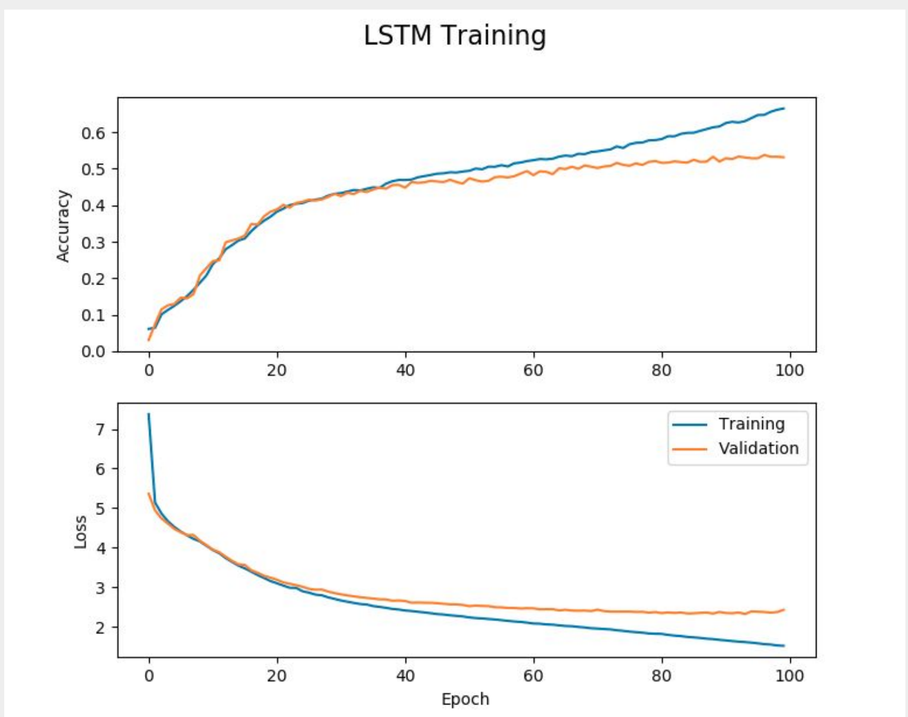
BASELINE MODELS

We used basic sklearn vectorizers and classifiers and got baseline results. Performance was good with SVM and Bag Of Words.



DEEP LEARNING

54% Accuracy achieved by LSTM model



BAD FREE TEXT REPRESENTATION

On querying the word vectors from the basic vectorizers, the free text representation of bugs were bad. Using only metadata from categorical columns seem to provide more meaningful vector representations.

- Query: security
- **Free Text Actual: javascriptvalidatelogin (1 bug), presenting (1 bug), awarded (1 bug)**
- **Free Text Expected: process, sandboxing, enterprise, information**
- **Metadata Actual: process, sandboxing, enterprise, information, caps, risk**
- **Metadata Expected: process, sandboxing, enterprise, information**
- Reason: Enterprise Process Sandboxing (35 bugs), Enterprise: Information Security(20 bugs)

DELIVERABLE

- Pratt and Whitney obtained 80% accuracy on their dataset using the models we built vs 60% on our public dataset!

WHAT DID NOT WORK

- No more than 60% accuracy on mozilla dataset
- Topic Modelling using LDA
- Clusters as a feature for classification

FUTURE WORK

- Clustering multiple analysts into a single cluster
- Using FastText as vectorizer for our supervised models
- Deep Learning
- Matrix Factorization