

Automatic Customer Ticket Classification Model Design and Implementation

MOTIVATION

An average telecom network operator receives nearly 300,000 customer complaint tickets in a year. The latency time between complaint reception and resolution is very high, thereby incurring huge operational costs on the network operator side and dissatisfaction on the customer side. This project in collaboration with Tupl Inc. aims to maximize the benefits for the network operators and their customers. Customer complaints can be designated to 9 different issues. Machine learning models process these complaint tickets and classify them into appropriate categories. This model has the potential to automatically process nearly 80% of the tickets without increasing Engineers' workload.

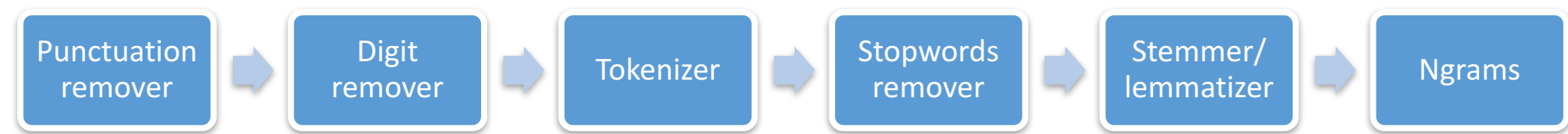
APPROACH

Dataset: 11562 customer complaint tickets.

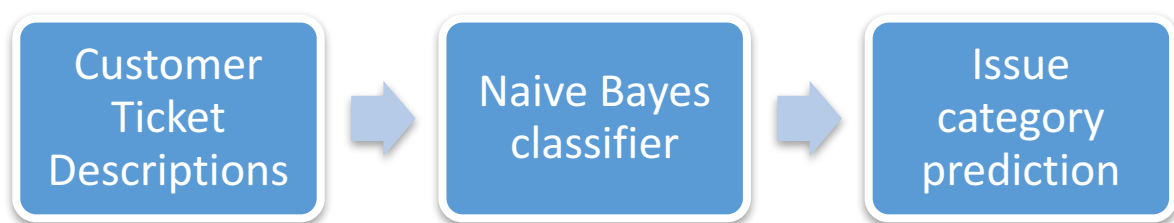
Train/ Test split: 80% training, 10% validation, 10% testing.

Multinomial Naïve Bayes: Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of text/ document classification.

- Pre-processing:** The raw data is pre-processed by applying several feature transformers. These include punctuation removal, digit removal, tokenization, stop words removal and stemming/ lemmatization.



- Classification:** Multinomial Naïve Bayes classifier model was used with smoothing.

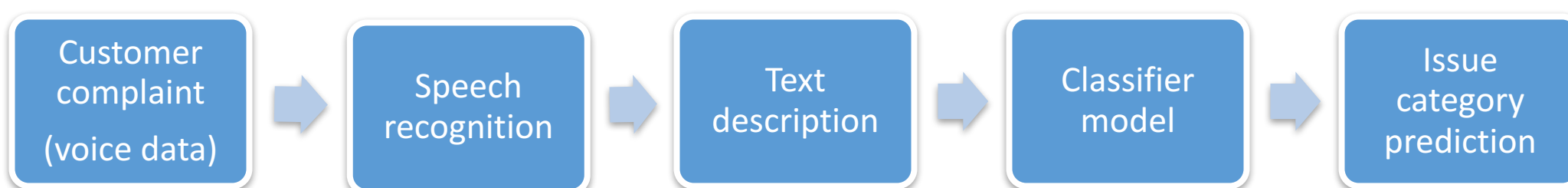


Long Short Term Memory Networks (LSTMs): A special kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies.

- Pre-processing:** Punctuations and stop words are removed from the raw data.
- Word-embeddings:** Word-embeddings are obtained using word2vec algorithm.
- Visualization:** The word vectors are visualized using t-distributed stochastic neighbor embedding (t-SNE) for dimensionality reduction.
- LSTM Model:** Word vectors are fed into LSTM and trained to obtain a classifier that performs optimally.

System Integration:

- Speech Recognition:** Open source voice recognition software Pypl is used to convert speech to text.
- Front-end/ Back-end:** JavaScript and React was used to design the front-end. The back-end framework was developed using Flask.
- Website Development:** REACT was used to build the website.
- System Design:** The API is developed using REST API and integrated with Naïve Bayes Classification model.



SOFTWARE/ PACKAGES USED

Apache Spark, Apache Zeppelin Notebook, Python, Pandas, NLTK, Numpy, Tensorflow, Pypl, React, REST API, JavaScript.

PERFORMANCE METRIC

Accuracy is the evaluation metric used for the project. The goal is to maximize this value.

$$\text{Accuracy} = \frac{\text{\#no of cases (predicted labels)} }{\text{\# actual labels}}$$

RESULTS AND ANALYSIS I

Pre-processed data: Several combinations of feature transformers were tested before deciding the final pre-processing feature transformers. Lemmatization instead of stemming improved the performance of the model slightly.

```
| Punctuation_removed | Digit_removed | Tokenized | Stop_wordsremoved | wrongspellings_removed |
|-----|-----|-----|-----|-----|
| cx is having prob... | cx is having prob... | [cx, is, having, ...] | [cx, problems, sl... | [problems, slow, ...]
| cust keeps losing... | cust keeps losing... | [cust, keeps, los... | [cust, keeps, los... | [keeps, losing, s...
| (customer opened u... | (customer, opened... | (customer, opened... | (customer, opened... | (customer, opened...
| weak signal KEN... | weak signal KEN... | [weak, signal, ke... | [weak, signal, ke... | [weak, signal, we...
| (incoming calls o... | (incoming calls o... | (incoming, calls... | (incoming, calls... | (incoming, calls...
| (created 123 Stree... | (created, streetn... | (created, streetn... | (created, mm, che... | [created, mm, che...
| (has extreme slow... | (has extreme slow... | [has, extreme, sl... | [extreme, slowes... | [extreme, slowes...
| (not able to load... | (not, able, to, l... | [able, load, web... | [able, load, web... | [able, load, web...
| (keeps on loosing... | (keeps on loosing... | [keeps, on, loosi... | [keeps, loosing, ... | [keeps, loosing, ...
| (customer is not a... | (customer, is, no... | (customer, able, ... | (customer, able, ... | (customer, able, ...
| (cannot call numb... | (cannot call numb... | [cannot, call, nu... | [cannot, call, nu... | [cannot, call, nu...
| (cx can not send a... | (cx can not send a... | [cx, can, not, se... | [cx, send, receiv... | [send, receive, p...
| (voicemail giving... | (voicemail, givin... | [voicemail, givin... | [giving, error, n... | [giving, error, n...
| (cannot receive a... | (cannot receive a... | [cannot, receive... | [cannot, receive... | [cannot, receive...]
```

Keywords: Term frequencies and mutual information was used to obtain keywords from the data.

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

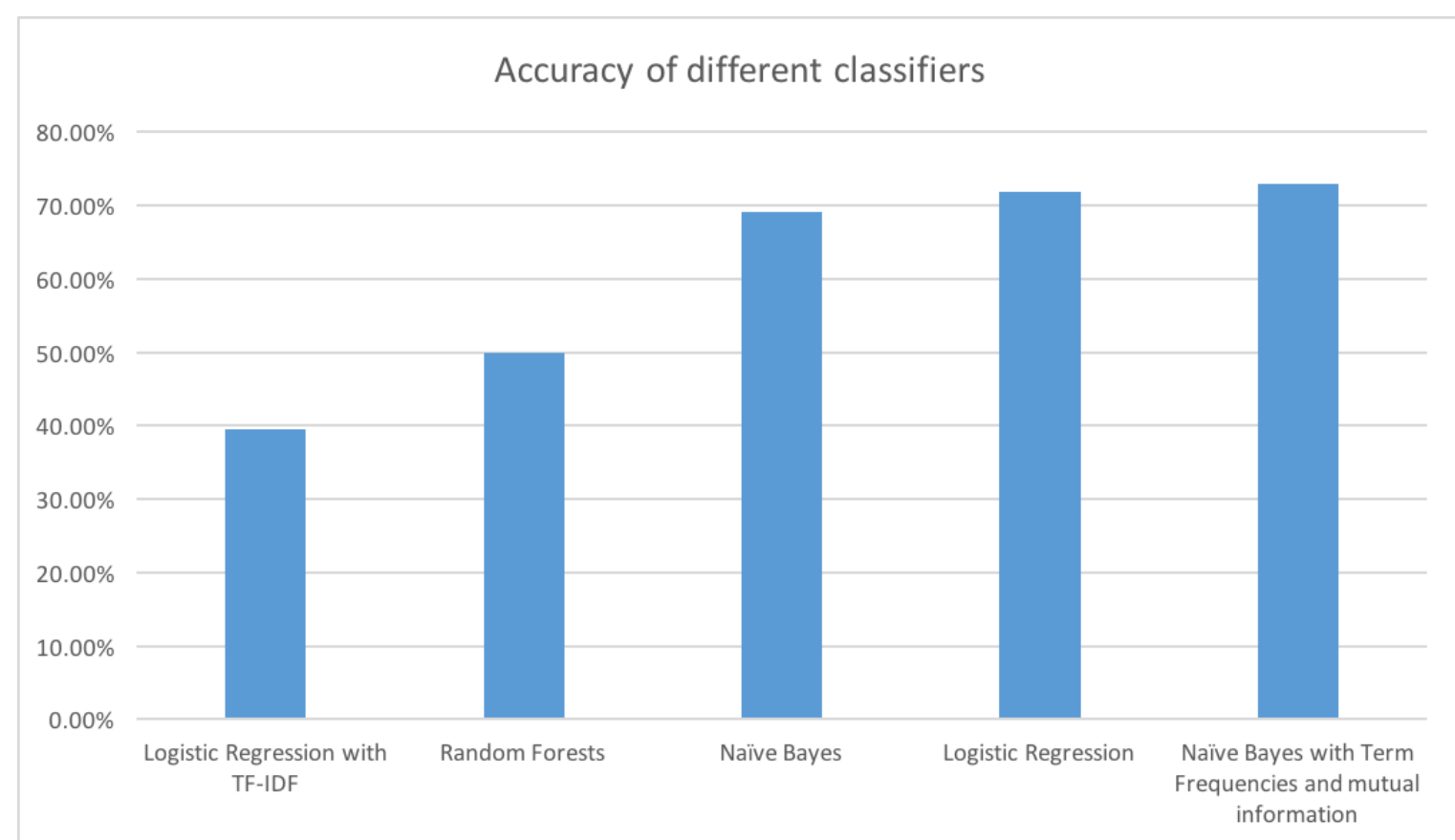
Unigrams using mutual Information:

KeyWords_mi
signal account mile sunday immediately protective lines intermittent line message continuous home around service report antenna accessibility of area visit house otis different fire lines devices redundancy startedhead intermittentlycustomer bird orwell priem long roguard experientiel matchsubarber webonly phys useareception experiencing dayg phonebefore registration receivedchecked before probably wania warenten violence million servicesesdomily balancecancelled houseasus madul group address kiewly locationw reflection homeasny pignout neeag intermittentlyhas outagsewing recent customergarbed coverage issueasues degradation message wellsometimes atb catifollowed people retainability connection usable coworker abou inasurablum kello node mram seest mckinley vm nut voicemail call route incoming straight go receive directly disturb drop ring signal forward data audio mail unconditional block number use numb home speed poor miss center ar failurebeing action locationreset instantly handsize mainin customer accepted sherton grande callcenteru received ric rebroadcast immediately apnd dispatch working interuaring sto dic ereds yesterdayer last followed subter hander number right auto vntagen cancelo vigh yearly verified total callforwarding signing batch behavior ferences update canteo rametella herca one routing me address get service outside pho unlocked month
data connection connect call stay maintain agree hr need otis doc signal audio dpo prio sleepy multiply thye that interrupt calculate reference antelope correctly losplease gasline dta sierra choiced robe antisepticy intermediate witties started dta usable intermittent period music drop 3r mobile unit educate receive limited loaner and sus poked meter catch song location save locationpoweredon go stream one unavailable ticket message exp cc cannot apps master connectivity finally troubleshot do cycle phone apps power intermittent pandora specific sometimes indicates stash turn location awf data internet call connection drop access connect signal audio web apps reset number receive setting load g cancel app limit hear byond tie able make mobile congept webpage vm site browser way hap

Bigrams using Mutual Information:

KeyWords_mi_bigrams
drop call data connection access internet get data connect data connect internet able access able connect access data data data internet connection cant access call drop get internet site registerte data mobile data radius please sudden loss unknown tower customer internet customer data unable connect check local internet reset loss data please t domestic data customer check access web can turn internet phone data limit data service wifi call show g utill cant data network internet everywhere data roam issue internet active data auto issue setting cancel reset apps home mta check setting provision reset setting data reset way audio cannot connect connection fail web browser connection specific internet test data plan lower seusa signal booster make call bar signal reset network roam der data call call tenus rm network type apps send related apps internet disconnect internet pas present cancel limit data data data web stream issue could fx indicator present non device online trinary where pmr connection state anywhere any online unknown internet connect services cannot establish
drop call issue drop call drop customer drop experience drop call everywhere get drop let drop call home customer experience call every call constantly customer unable device touchscreen call last get selection case cover customer able call day call within still drop rm update touchscreen wake manual network get service call matter multiple drop experience lot remove case call area issue remove slow data call time based test data version software data able get drop minute call line phone drop customer signal drop time device software signal issue able connect phone verify call location c location data speed call arrival call start call happens wake call month drop keep drop update device anywhere where go per day call rm data roam device drop call billing call month connect network drop m receive call customer cover device rm pm able make billing address every call signal home line drop register network make device receive exchange device good bar rm software many drop drop drop drop today call per customer cannot say happen

Comparison of Classifiers: Naïve Bayes with Term frequencies and mutual information had the best accuracy compared to all other models.



RESULTS AND ANALYSIS II

Integrated System: The interactive website takes in an audio recording/ description and classifies the ticket into the right issue category.

Trouble Ticket Classifier Demo

Recording

Ticket Description:

Classify

Category:

Trouble Ticket Classifier Demo

Recording

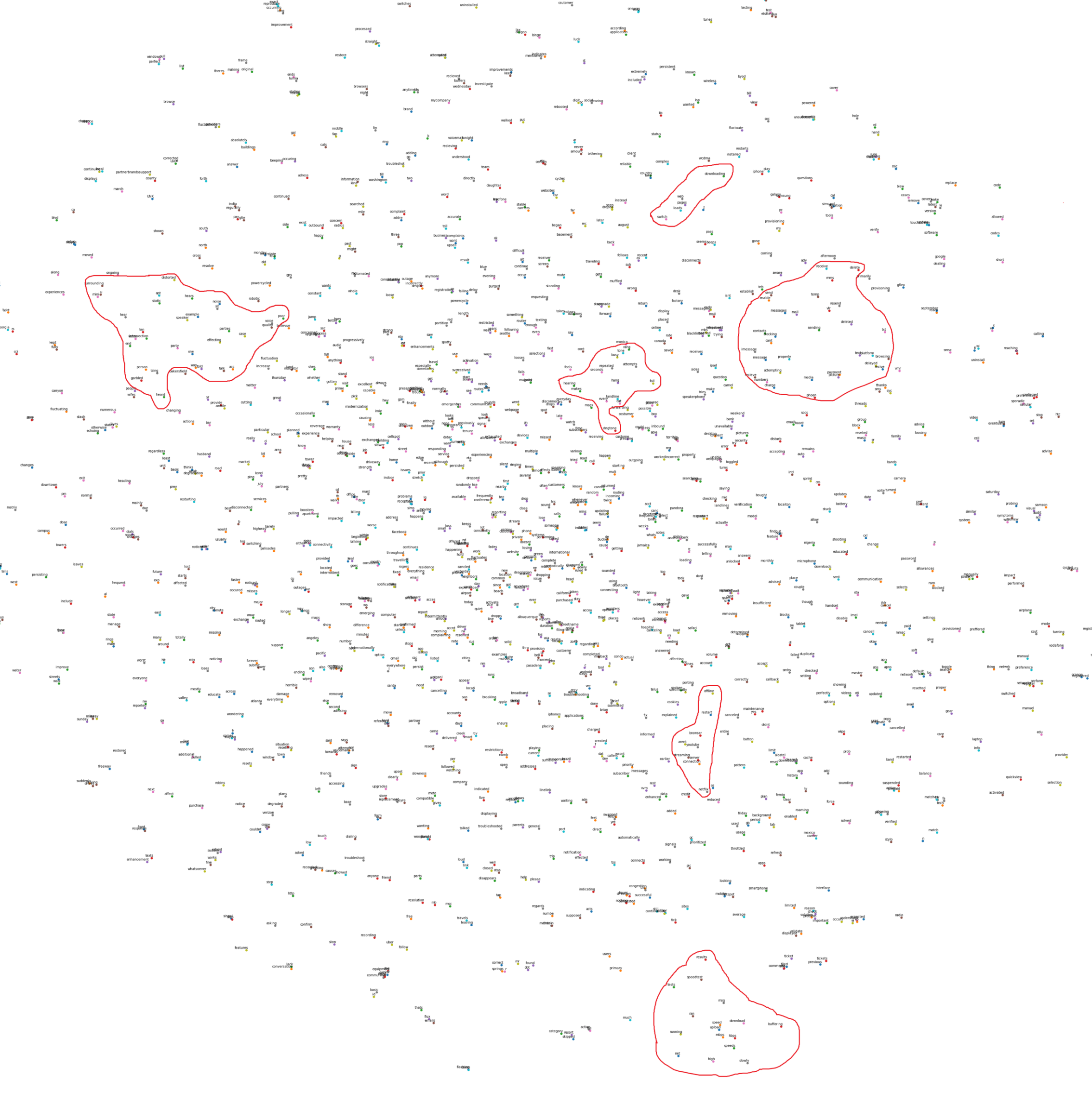
Ticket Description:

my phone is not working well I cannot access to Internet since last Friday

Classify

Category: Connection Fail

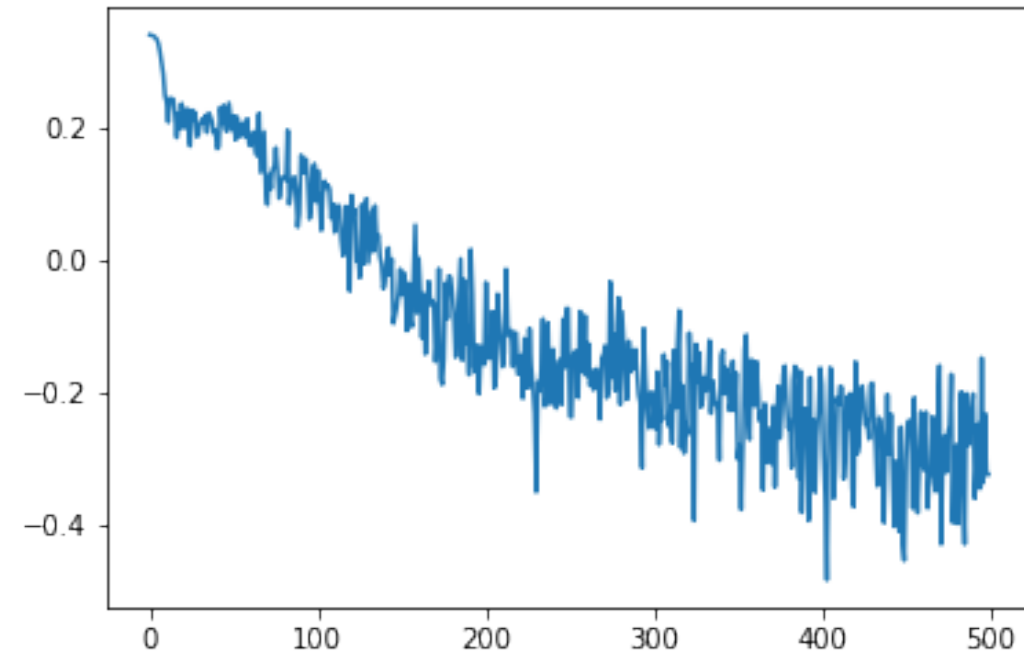
Embedding Visualization: The word embeddings were visualized using t-SNE dimensionality reduction technique. It was found that several words related to each other clustered together in the embedding space.



RNN Pre-processing: All the tickets were padded with starting and ending tokens until all tickets were of equal length. The original word lengths were also computed and fed to the RNN model.

splitWords	wordLengths	wordsPadded	wordsStd
[customer, issue, able, get, one, bar, ph, home]	8	[<s>, customer, issue, able, get, one, bar, ph...	[872, 109, 790, 176, 69, 761, 553, 1768, 3609,...
[states, past, month, 03, 5am, experienced, fl...]	15	[<s>, states, past, month, 03, 5am, experience...	[872, 3110, 1138, 330, 738, 2103, 1547, 888, 2,...
[make, receive, calls, signal, bars, verify, n...]	8	[<s>, make, receive, calls, signal, bars, veri...	[872, 2156, 2718, 1265, 2554, 2993, 716, 2153,...
[customer, keeps, dropping, calls, alot, even,...]	13	[<s>, customer, keeps, dropping, calls, alot, ...	[872, 109, 3259, 3414, 1265, 3543, 1137, 2839,...
[cx, able, make, rcv, calls, going, calls, dia...]	59	[<s>, cx, able, make, rcv, calls, going, calls...	[872, 589, 176, 2156, 3165, 1265, 2535, 1265, ...

RNN Training costs: Cross entropy was the loss measure used for the RNN.



$$H_{y'}(y) = - \sum_i y'_i \log(y_i)$$

DISCUSSION

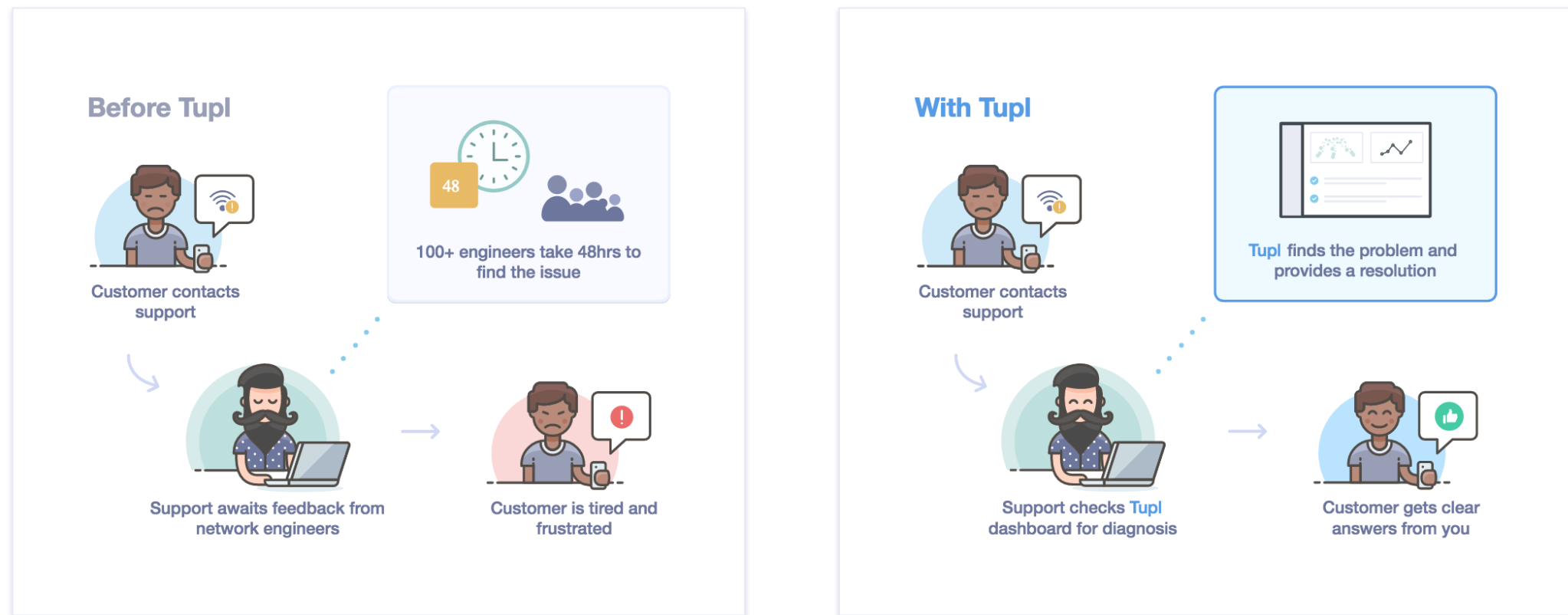
Classifier Accuracy:

Naïve Bayes Classifier – 75% on test data

LSTM Model – 65% on test data

Currently, the Naïve Bayes classifier outperforms the LSTM Model. This is partly because deep learning requires a lot of tuning and careful selection of hyper-parameters. These include embedding layer size, number of embedding layers, learning rate, drop out, number of epochs etc.

POTENTIAL IMPACT OF THE PROJECT



- On a scaled level, automation of customer care complaints can save up to 25% of Engineers' workload.
- 80% of complaint ticket resolutions can be automated.
- 10x improvement in responding to customer queries.
- Reduced operational costs for telecom network operators.
- Improved work efficiency for customer support team.
- Improved customer care experience.

CONCLUSION/ FUTURE WORK

Intelligent Process Automation is a very promising idea that leverages machine learning and big data technologies to offer solutions to various problems. It has a vast potential in the telecom network operations side. The long term goal of this project is to deliver a machine learning classifier with optimal accuracy. The accuracy of the RNN model has to be improved since deep learning is proven to be an interesting direction for several machine learning projects.

ACKNOWLEDGEMENT

We gratefully acknowledge the Department of Electrical Engineering, Industry capstone course staff and Tupl Inc. for providing the opportunity to work on a real-world problem and provide solutions. We would also specially like to thank our mentors Xiang Chen, Professors Payman Arabshahi and Sreeram Kannan for their guidance.

REFERENCES

- <https://www.tensorflow.org/tutorials/word2vec>
- <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>