

---

# PREDICTING TECHNICAL APTITUDE USING CUSTOMER CONVERSATIONS

---

Sponsor Client – Dell Technologies



MAY 14, 2022

GROUP 11

Namit Agrawal | Neha Dipali | Bret Jaco | Meha Mehta | Vinay Pahwa | Aditya Soni

## Contents

Executive Overview.....	2
Business Context, Problem and Expected Outcomes .....	3
Context.....	3
Problem.....	3
Expected Outcomes .....	3
Exploratory Data Analysis .....	4
Analysis of Data Shared by Sponsor Client Team .....	4
Data Pre-Processing .....	6
Manual Labelling of Data for Evaluating Model Performance.....	7
Customer Technical Aptitude Prediction .....	8
Term Frequency – Inverse Document Frequency (TF-IDF) Based Approaches .....	9
Model 1: Dividing Transcripts between Agents and Customers.....	10
Model 2: Scoring Entire Transcripts .....	12
Neural Network Based Approaches .....	13
Generation of Word Embeddings .....	13
Neural Networks for Classification and Regression .....	13
Clustering .....	15
Customer Persona Identification .....	16
Insights, Recommendations and Future Scope .....	17
Acknowledgments.....	18

## Executive Overview

Customer centricity lies at the heart of Dell Technologies. The objective of this project is to gain a deeper understanding of the customers seeking post-sale support. The project aims to specifically answer the following business questions –

- Is it possible to use the internal unstructured data sources to assess the technical aptitude of the customers?
- Is it possible to identify dominant customer groups to enable enhanced personalization of support services?

This project will help in developing targeted solutions for the customers, which will allow the customer service agents to provide effective solutions while providing quicker solutions to the problems.

Approximately 5GB of customer conversation data in the form of chat transcripts and emails between the customers and customer service agents was shared by the client sponsor team. The data belonged to the customers based in the United States and had been captured between Q1 and Q3 of FY2021. The data comprised of over one million unique chat transcripts, accompanied by additional data pertaining to the location, date and time of the conversation.

The student team manually labeled 300 chat transcripts due of the unavailability of prior labelled data. Some notable challenges faced by the team included dealing with anonymized and truncated transcripts, along with limitations on computing resources.

This project focuses on two different approaches for evaluating the technical aptitude of the customers. The first one capitalizes on TF-IDF methodology, which quantifies the importance of words in a document amongst a collection of documents. The second approach is based on word embeddings and neural networks.

TF-IDF approach resulted in an accuracy of approximately 61% while classifying customers as technically adept or otherwise, with a 41% correlation in the numeric scores. The neural network approaches, although more promising, were limited in terms of performance because of limited labelled data. For the scope of this project, the neural network-based approaches have been used only as a proof of concept.

The recommendations of the student team include the following –

- Increasing the character limit of the chat transcripts would help capture the information more accurately and yield more fruitful results.
- Augmentation of Labelled Dataset to improve the performance of neural network-based techniques, using services such as Amazon Mechanical Turks to manually label the data.

The code used for analysis can be found [here](#).

## Business Context, Problem and Expected Outcomes

### Context

With a vast product and service line comprising of computers, other electronics, cloud services and business solutions, customer-centricity lies at the heart of Dell Technologies. While Dell's vision is "Delivering a better tomorrow," the code of conduct emphasizes most on the customers.

### Problem

The objective of this project is to gain a deeper understanding of the customers seeking post-sale support. This will help in developing targeted solutions for the customers, which will allow the customer service agents to provide effective solutions while providing quicker solutions to the problems.

The project aims to specifically answer the following business questions –

- Is it possible to use the internal unstructured data sources to assess the technical aptitude of the customers?
- Is it possible to identify dominant customer groups to enable enhanced personalization of support services?

Understanding the customers better will help Dell customize their services to enhance the customer experience through personalization, minimize chat durations, improve customer satisfaction, and boost customer loyalty.

### Expected Outcomes

The scope of the project includes the following –

- Initial analysis of the data sources shared by the client sponsor
- Exploration of different analytical techniques for unstructured data, and development of technical proficiency score
- Identification of dominant cluster groups of customers

## Exploratory Data Analysis

The exploratory data analysis can be divided into three parts which are stated as follows –

- Analysis of data shared by sponsor client team
- Data pre-processing
- Manual labelling of data for evaluating model performance

Each of the aspects of exploratory data analysis are elaborated in the following pages.

### Analysis of Data Shared by Sponsor Client Team

Approximately 5GB of customer conversation data in the form of chat transcripts and emails between the customers and customer service agents was shared by the client sponsor team. The data belonged to the customers based in the United States and had been captured between Q1 and Q3 of FY2021. The data shared comprised of over one million unique chat transcripts, accompanied by additional data pertaining to the location, date and time of the conversation. Please find below a sample of the same –

Fiscal Quarter	Fiscal Week	Group Name	Region Name	Sub Region Name	Case Number	Chat Create Date	Chat Transcript Body
2021-Q1	2021-W01	Client Basic Support Consumer	AMERICAS	United States	49683308	2/2/2020 18:36	<p align=center>Chat NAME: {{NAME}}, {{NAME}} 03, 2020, 00:06:38 (+0530)</p><p align=center>Chat NAME: NA.DB.INTP.CLI.CO. EN.PRESP.BLR</p><p align=center>Agent {{NAME}} {{NAME}}</p>( 16s ) {{NAME}} {{NAME}}: Hi, Thank you for contacting {{NAME}} technical expert, how may i help you.

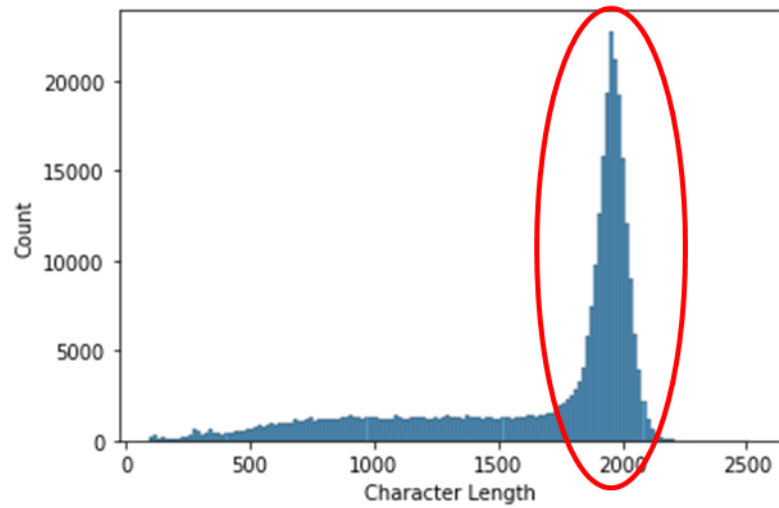
The data shared also included the agents' comments regarding the customer service request, along with additional customer information such as product type, customer type (commercial v/s consumer), warranty information, reason for contacting customer service, and satisfaction scores (if available). The data had been stored in SQL servers and had been shared in the form of 7 csv files.

Owing to the limited availability of the computational resources, the student team was unable to utilize all the data shared by the sponsor team and had to work with 100,000 randomly sampled transcripts.

Some of the challenges faced by the student team from a data perspective are as follows –

- **Data Anonymization:** Owing to safety concerns, the names of the customers and the customer service agents in the transcripts were anonymized with the identifier “{{NAME}}.” While this was essential for protecting the privacy of the customers, it became challenging to separate agent conversations from that of the customers. This would have a negative impact on the performance of the models, as the technical proficiency of the agents would intervene with that of the customers.
- **Data Truncation:** As mentioned earlier, the data shared for analysis was stored in SQL servers. The chat transcript body column had a 2000-character limit, because of which many conversations were truncated mid-sentence. While there were a considerable number of transcripts with length

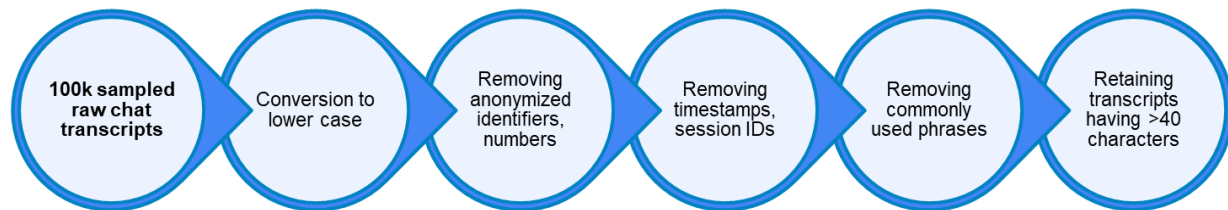
exceeding two thousand characters, it was likely because of the anonymization process done later. Please find below the distribution of the character lengths of the chat transcripts.



Data truncation adversely affects the performance of the models as the transcripts capture the initial bits of conversation that include pleasantries, while omitting the meat of the conversation which would provide insight about the customer's technical aptitude.

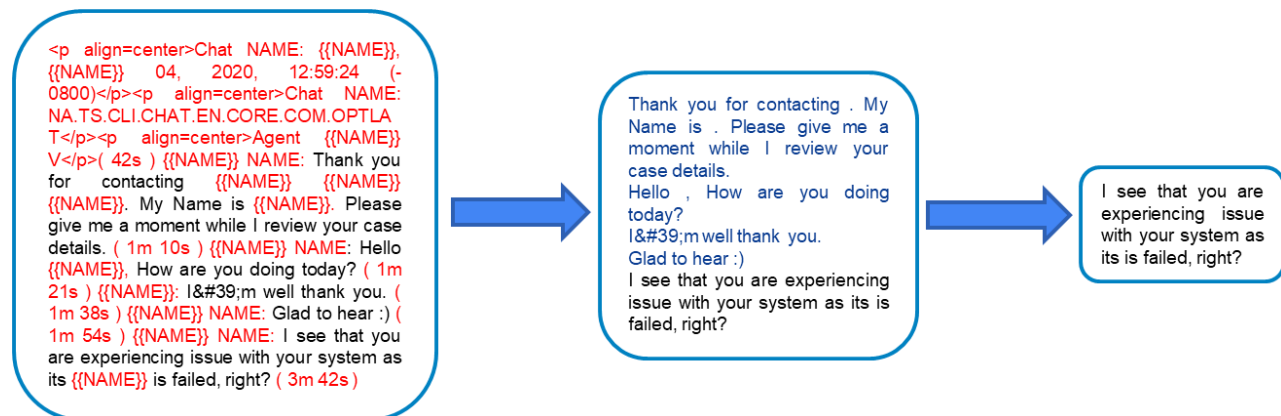
## Data Pre-Processing

The data cleaning and analysis was done using Python. The data cleaning process is as follows –



- **Random Sampling of Transcripts:** Considering the limited availability of computational resources, the student team had to work only with a subset of 100,000 randomly sampled chat transcripts from Q1 FY2021.
- **Conversion to lower case:** Considering that Python is a case sensitive language, the sampled transcripts were converted into lower case. This would ensure that multiple instances of a single word would be treated similarly if a difference in case existed.
- **Removal of anonymized identifiers and numbers:** Considering that anonymized identifiers such as “anonymous@domain.com” would not add value in terms of determining customer aptitude, they were removed.
- **Removal of timestamps and session IDs:** Session IDs and timestamps were excluded as well because they would not add contribute to any information regarding the customer’s technical proficiency.
- **Removal of commonly used phrases:** Pleasantries and greetings were excluded as well, as they would not help in accurately determining the level of customer’s technical knowhow.
- **Retention of transcripts with at least 40 characters:** Over 93% of the conversations were found to have a length greater than 40 characters after the removal of commonly used phrases. Conversations shorter than 40 characters were excluded as they were unlikely to paint a picture about the customer’s technical knowledge.

The data cleaning process can be diagrammatically summarized as follows –

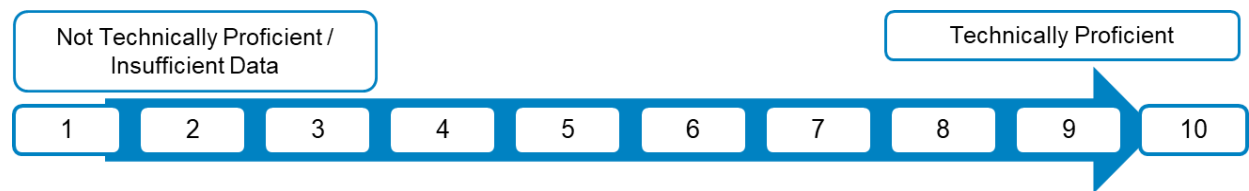


## Manual Labelling of Data for Evaluating Model Performance

Considering that the customer technical proficiency score was a novel project, one of the major challenges faced was the unavailability of the labelled data for evaluating the performance of the logics developed. The student team hence had to read several chat transcripts and score them manually. Considering the time constraints, a total of three hundred conversations were scored.

For assessing the technical proficiency of the customers, the student team considered several factors including but not limited to their ability to reinstall or update drivers and perform troubleshooting steps, and their knowledge of BIOS and error codes.

The transcripts were manually scored on a 10-point scale. The conversations which lacked sufficient information about the customers or had customers with low technical proficiency were scored between 1 and 5, while the conversations with technically adept customers received a score between 6 and 10. Please find below the representation of the same.



For future scope of this project, one recommendation would be augmentation of the labelled dataset with the help of services such as Amazon Mechanical Turks.

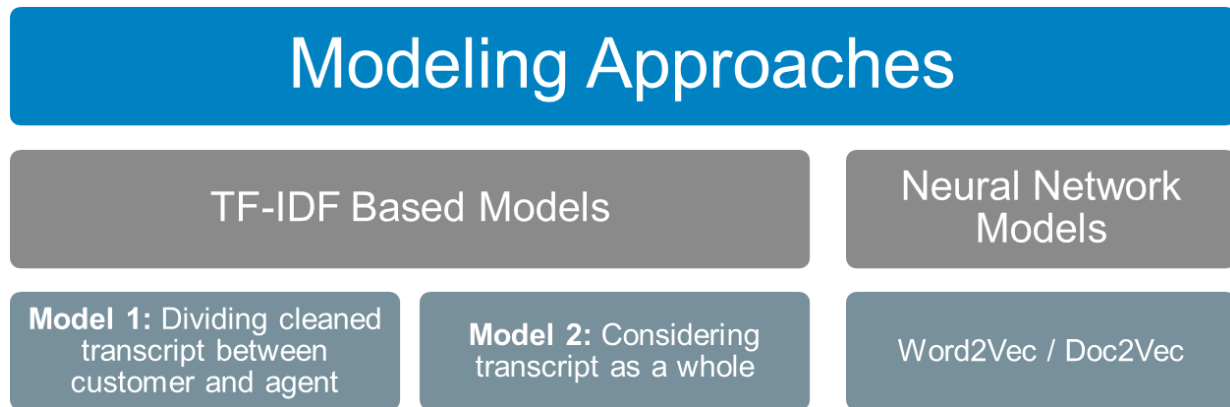


## Customer Technical Aptitude Prediction

The following modeling approaches were considered for evaluating the technical aptitude of the customers –

- Term Frequency – Inverse Document Frequency (TF-IDF) based approaches
- Neural Network based approaches

The diagrammatic representation of the same is as follows.



The following pages will cover these aspects of the analysis –

- Brief introduction to the approach and appropriateness of the analytical method used
- Assumptions used in the analysis and modeling process (if any)
- Detailed explanation of the modeling approach
- Model performance evaluation
- Insights from the analysis

## Term Frequency – Inverse Document Frequency (TF-IDF) Based Approaches

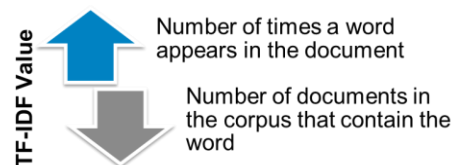
TF-IDF is a popular natural language processing technique that quantifies the importance of words in a document amongst a collection of documents. Based on TF-IDF, unique and important words should have high TF-IDF values in a certain document. It should be possible to leverage the text weight to extract the most important words of a document.

TF-IDF, as its name suggests, is made up of two complementary components:

- **Term Frequency (TF):** This component awards a word higher score for appearing more times in a document
- **Inverse Document Frequency (IDF):** This component penalizes words that occur in many documents.

While the TF component helps identify words that could be keywords in a certain document, the IDF component helps to remove the focus from words that are common across all documents. This results in a removal of words such as 'a,' 'the' and other commonly used English words that may be unimportant to the analysis and do not provide any insight into the technical aptitude of the customers.

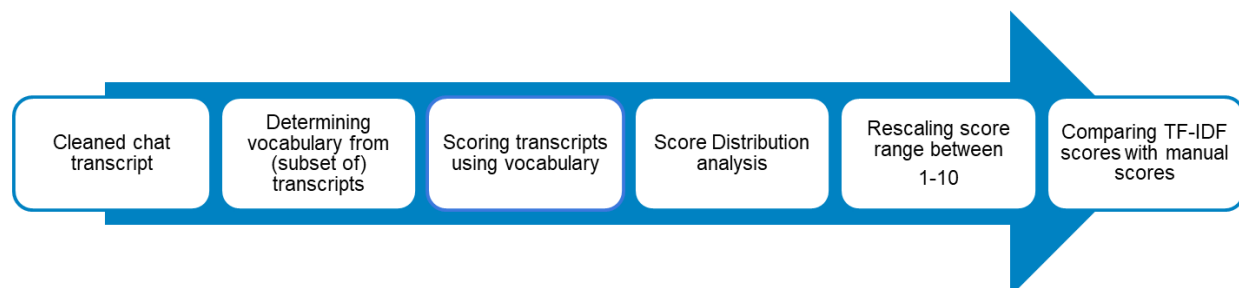
Essentially



The TF-IDF package is available in the Python NLTK library. Within the TF-IDF approach, the following two models have been considered.

- Dividing the transcripts between the agents and customers
- Considering entire transcripts for evaluating customers' technical proficiency.

Beginning with the cleaned chat transcripts after the pre-processing step, the methodology for both the models is identical, except for the part of the transcripts used to build the vocabulary. The process is represented as follows –

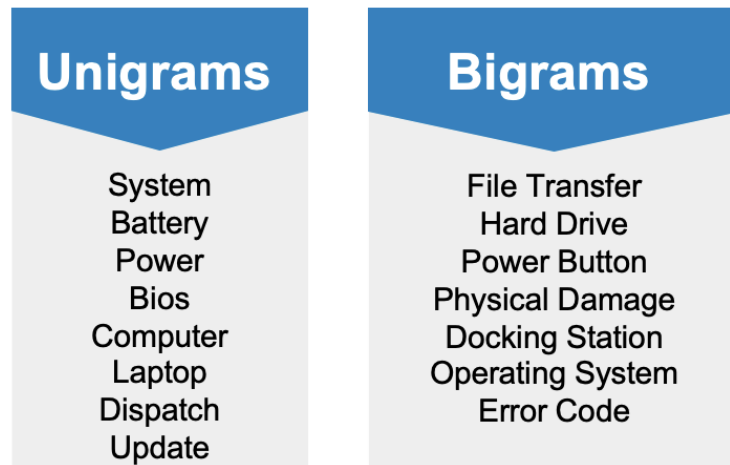


- **Extracting keywords from (a subset of) the transcript:** TF-IDF is traditionally used for single-word phrases (unigrams) but may also be extended to two-word phrases (bigrams). In the case of Bigrams, the two-word phrases are counted as one term and their frequency is only incremented

when those two words appear together in the same order. Some examples of the keywords extracted are as follows:

- a. **Unigrams:** power, system, battery
- b. **Bigrams:** docking station, power button, hard drive

A few more examples are shown in the following figure.



- **Using keywords to score (the other part of) the document:** Once a list of unigrams and bigrams was extracted, the documents were parsed again to identify customers having a higher frequency of these words. This was based on the idea that customers having a higher occurrence of the technical keywords identified in the vocabulary would have a higher technical proficiency score. Similarly, customers having a lower frequency of these keywords would have a correspondingly low score.
- **Rescaling the scores to lie between 1 and 10:** TF-IDF scores for transcripts ranged between 0 and 13. On consultation with the client sponsor team, a scale between 1-10 was chosen to enhance interpretability. The scores obtained from TF-IDF logic were thus rescaled to lie between 1 and 10 and then rounded so that they all took integer values.
- **Measuring model performance:** The conversations manually tagged in the previous step were scored using the TF-IDF model. The model performance was assessed in two ways – correlation of the numeric scores, and the accuracy using the confusion matrix (considering customers scoring above 5 to be technically proficient).

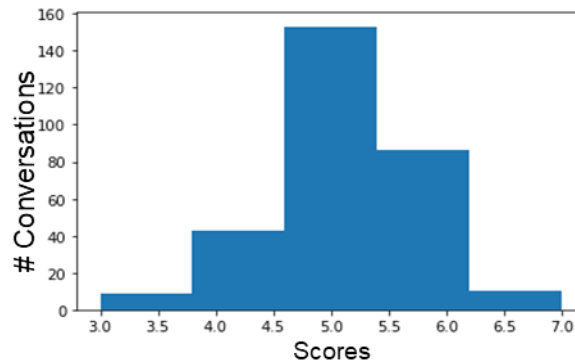
#### Model 1: Dividing Transcripts between Agents and Customers

This modeling approach relied on three key assumptions, which are stated as follows –

- **Technical Proficiency of the Agents:** By default, it was assumed that the agents were technically proficient. The vocabulary of technical keywords (unigrams and bigrams) was determined from the transcripts of the agents.
- **Technically Proficient Customers use more Technical Terms:** The vocabulary of key words identified from the agent transcripts was then used to identify technically adept customers.

- **Agents initiate conversations:** To analyze customer data, the cleaned chat transcripts were first divided into separate transcripts for Agents and Customers. Considering the data anonymization, it was impossible to determine if the conversation was initiated by the customer or by the agent. It was hence assumed that the first sentence was spoken by the agent and then the speakers alternated thereafter. This assumption is poor because agents and customers both were often found to send multiple messages in a row, and thus the model would suffer from limitations.

The scores obtained from TF-IDF model followed a normal distribution with a mean and median of 5. The distribution of the scores of the validation set is shown in the following histogram –



The performance of this logic was measured in two ways:

- **Correlation:** The correlation between manually tagged scores and predicted scores was found to be 0.41 implying a strong positive correlation. This showed that the model performed well
- **Accuracy using Confusion Matrix:** The TF-IDF method turned out to be quite effective with an average error of approximately 2 points and accuracy of 61%. The confusion matrix is as follows:

		TF-IDF Logic Output	
		FALSE	TRUE
Manual Scoring Output	FALSE	True Negative 36 11.9%	False Positive 102 33.8%
	TRUE	False Negative 16 5.3%	True Positive 147 48.8%

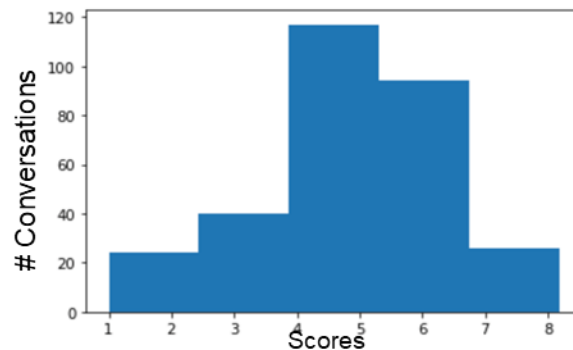
The shortfalls of this TF-IDF approach are as follows –

- **Assumptions:** While the TF-IDF model performs well for this use case, it is crippled by the assumptions made about the speakers of each sentence. The model performance is expected to improve significantly with the availability of the deanonymized data.
- **Limited Variety of Customer Issues:** Another limitation of this model could be the corpus of documents used to build it. TF-IDF identifies keywords by how often they appear within documents and penalizes them if they appear in many documents. If the corpus of documents used to generate these keywords is diverse, the model will identify topical keywords within each document accurately. However, if the corpus used to train the model is not diverse, there is a possibility that keywords would be treated at par with stop words. For example, the model may

penalize frequently occurring terms such as “boot” using the IDF score, even though it should ideally be recognized as a keyword.

#### Model 2: Scoring Entire Transcripts

In this model, the entire conversations were scored without dividing them between the agents and customers. The distribution of the TF-IDF scores of the manually labelled dataset are as follows –



This model led to a slightly better performance on the manually tagged dataset, with a correlation of 43% and accuracy of 65%. The confusion matrix is as follows –

		TF-IDF Logic Output	
		FALSE	TRUE
Manual Scoring Output	FALSE	True Negative 97 32.3%	False Positive 41 13.7%
	TRUE	False Negative 62 20.7%	True Positive 101 33.7%

The shortfall of this TF-IDF approach is the premise. While the performance on the manually labelled dataset has shown marginal improvement as compared to the first TF-IDF model, this approach would not be recommended as the premise itself is flawed. The agent conversations are used to score the customers. This hence may not paint an accurate picture about the technical capability of the customers.

## Neural Network Based Approaches

### Generation of Word Embeddings

The second analytical method implemented was by creating Word Embeddings to encode the semantic meaning of the transcripts analyzed. The underlying idea was to capture the semantic meaning of what was said by the customers so that a proficiency score could be assigned more accurately.

Each raw word was represented by a vector with  $n$  dimensions ( $n$  being a hyperparameter determined by cross-validation). Word embeddings were used for this specific problem because it was important to understand what words separated technical customers from non-technical and how they communicated as a whole.

The Word2Vec function from the GenSim library helped in extract word embeddings and to train over all chat transcripts. Once extracted, the word embeddings were used in a variety of applications ranging from neural networks to clustering.

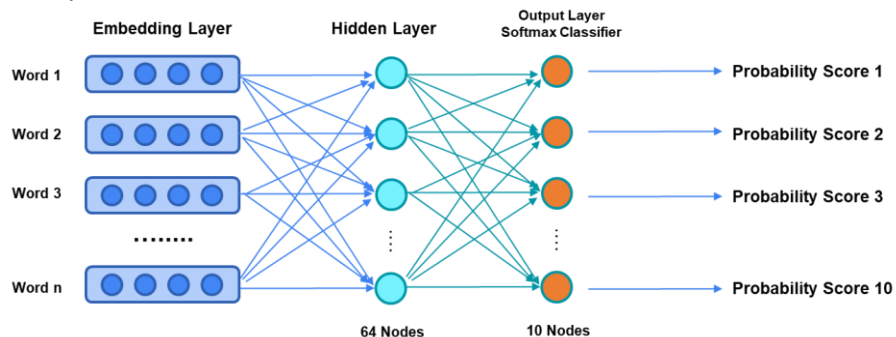
Embedding Dimension: 300

bios	.787	.981	.076	-.872	.231	-.112	.543	...	-.912
display	.235	-.674	.863	-.984	.467	-.653	.299	...	.101
disconnect	-.653	.451	.347	.276	.873	.277	-.786	...	.982
...									
warranty	.488	.286	.423	-.652	.568	.111	.375	...	-.854

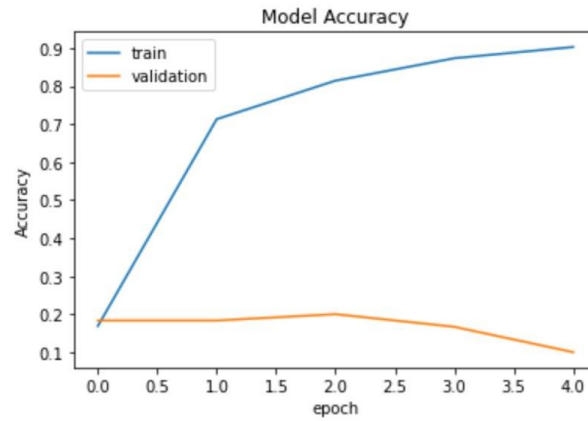
### Neural Networks for Classification and Regression

Word embeddings were generated for customer sentences, agent sentences, and complete transcripts. The central idea was to analyze just the customers since the task was to generate scores for the customers themselves. As a result, a neural network was built using TensorFlow with an embedding layer, where this layer was not trainable (i.e., no gradient updates) along with a couple of dense layers.

This problem was posed as a classification and a regression problem to evaluate performance. The architecture of the neural network model used for classification was as follows. In case of regression problem, the output layer had a single node instead of ten nodes. The model used was simplistic because of limited availability of labelled data.



The loss functions used hence were MSE and sparse cross-entropy. The classification neural network using only customer data performed poorly with an accuracy of 15% and MSE of 6.5.



The results do not have much merit simply because only used 300 labelled transcripts to train the models. A neural network requires thousands of data points to learn the underlying patterns. As a result, this model was simply used a proof-of-concept. The performance of neural network-based algorithms would be significantly better than the TF-IDF model with the availability of labelled data.











## Clustering

The final approach in the neural network-based approaches was clustering the word embeddings, with the assumption that the customers using technical words would be separated from those using non-technical words.

Each transcript and its associated embeddings were averaged across each of the  $n$  dimensions. This exercise resulted in the representation of each transcript in the  $n$ -dimensional space. K-means algorithm was used to cluster the transcripts.

The silhouette score metric was used to determine the distance of the generated clusters. The underlying concept was that the clusters that were further apart were more distinguishable from the rest. The silhouette scores range from -1 to 1 where scores closer to 1 are better. The model resulted in a score of approximately 0.07, indicating that a transcript is remarkably close to the decision boundary that differentiates whether a point should be in one cluster or another.

The ten clusters were obtained from this process are as follows –

<b>0</b> Service Tag		<b>1</b> Warranties		<b>2</b> Replacement Parts		<b>3</b> Software		<b>4</b> Abbreviations	
aservicetagg plz dsp aservicetaggb sir aservicetagw aservicetagi aservicetagm aservicetagh drt		warranty purchase purchasing warrenty reseller therefore discount order buy pay		part ost cx mobo insists declined denies refused insisted replace		bios driver window installs o rebooted utility reboot try generic		ty nah sweet kk partselection pn coolio improve yw man	
<b>5</b> Hardware		<b>6</b> Customer Support		<b>7</b> Power Issues		<b>8</b> Display Issues		<b>9</b> Conversation	
system zone valid active est caused ensure confirming near physical		sent addresscontact dispatch tech part technician quote perfect mailed absolutely		power plugged unplugged plug adapter light plugging turn powered charger		computer system laptop monitor device display detects recognizes pc dock		response reply chatthank received minute disconnect conversation reponse connected kindly	

Instead of clustering similar customers, this exercise resulted in clustering of similar issues. While cluster 1 is concerned with warranty issues, cluster 7 pertains to power issues.

Word embeddings in general gave inconclusive results as no insights could be gathered from it. The major reason for this is the poor assumption made about the data is split between the customers and the agents. Moreover, the text itself is extremely disorganized as misspelled words were found. This model can perform significantly better with the availability of quality labelled data.



## Customer Persona Identification

Once a technical proficiency score was of the customers was determined, it was possible to assign a persona to them. These personas would allow customer service representatives to serve the customers in a quicker and better fashion, as they would be able to tailor their approach.

Based on the technical aptitude scores and the customer type (commercial or consumer), the customers were grouped in one of the following four categories –

- Proficient Commercial Customers
- Non-Proficient Commercial Customers
- Proficient Consumers
- Non-Proficient Consumers

Dividing customers based on their technical proficiency would allow representatives to communicate more effectively with the customers. In case of technically proficient customers, the agents would be able to resolve solutions quickly. On the other hand, agents may need to dive into greater depth regarding the details in case of a less technically proficient customer.

Dividing customers based on their consumer status is also important, as the issues faced by many commercial customers are managed by an internal IT department. With a higher probability of technical proficiency, they would have already tried some troubleshooting techniques before reaching out to for help.

## Insights, Recommendations and Future Scope

The insights of the analysis are summarized as follows –

- The TF-IDF model was able to accurately classify customers on 61% of the occasions, given the constraints
- More labelled data would result in improved performance of the Neural network-based techniques. Considering the scope of this analysis, the Neural Network and Word2Vec models have used only for POC purposes.
- Customer Personas can help agents serve customers more effectively.

A few recommendations for effective implementation of this project in the client team's ecosystem are as follows –

- **Data Storage:** Increasing the character limit of the chat transcripts would help capture the information more accurately and yield more fruitful results.
- **Augmentation of Labelled Dataset:** This would tremendously help improve the performance of neural network-based techniques. This could be done using services such as Amazon Mechanical Turks to manually label the data.

The future scope of this project is as follows –

- **Retraining models using deanonymized data:** This will allow a more accurate evaluation of the technical aptitude of the customers.
- **Training the models using a larger sample:** The current analysis was done using only a sample of 100,000 transcripts owing to the limitation on computing power. Training with a larger dataset can help overcome this issue to create more robust models.
- **Enrichment of the vocabulary of keywords:** Additional unigrams and bigrams can be incorporated into the vocabulary based on the sponsor team's definition of technical aptitude.
- **Aggregation at customer level:** For the scope of this project, the scores have been generated at a case level. In case of multiple interactions with the same customer, the scores can be aggregated at a customer level to provide a more holistic picture about technical aptitude.
- **Enrichment of personas with other variables:** Depending on the requirement, additional variables can be used to enrich the customer personas and paint a more holistic picture of the customers

## Acknowledgments

We would like to take this opportunity to extend our gratitude to the MSBA Program at McCombs School of Business for giving us the opportunity to apply the concepts learnt in the courses to impactful industry applications.

We are extremely grateful to Prof. Daniel Mitchell and Prof. Robert Hammond for their mentorship and encouragement throughout the course of the project. We offer our sincere appreciation for the learning opportunity provided by them.

We would like to express our thanks to Michelle Slick, Tom Jennings, Edi Edwards, Rachel Meade, and Connor O'Brien from the client sponsor team at Dell for their proactiveness, guidance and constant support through the course of the project.