Customer Churn Prediction using Natural Language Processing (NLP)

AFAZ UDDIN AHMED

Abstract

Predicting customer churn is an important consideration for any business, including financial service businesses, because costs of acquiring new customers far outweigh costs of retaining existing ones. Our daily interactions with Siri, Alexa, Hey-Google, and Bixby, which are Natural Language Processing (NLP) based automation systems are currently treated as just another cool feature in our everyday lives. Imagine using this cool feature to solve a fundamental problem for a business – preventing customer churn. Different customers exhibit different behaviours and preferences and cancel their subscriptions for a variety of reasons. Most existing models predict customer churn by using demographic and transactional data of customers, which may not contain a full reflection of customers' intentions. In this paper, a customer churn prediction model is developed using NLP by extracting features and patterns in unstructured data available against customer policies. These unstructured datasets are typically text, calls, and notes, and thanks to the advancement of NLP technology, these datasets can now transform into key information from which we can infer intention for churn. Existing commercial NLP models which predict customer intention based on text, are still in their infancy and researchers are still investigating how to improve such models. With fast emergence of new NLP features, many have become outdated. A case study is presented in this paper to predict customer churn and reason for intention to churn using call data which may not be possible using structured data alone. The model proposed in this paper uses recently available NLP tools and features to develop a customer churn prediction model. The model uses keyword matching to mine expressions of interest and profiles of people corresponding to customer criteria. The proposed model takes advantage of available pre-trained NLP models to perform sentiment analysis. A set of reference sentiments are manually generated and compared with the customer conversation to find the similarity as an index. This index is used as a threshold for a classifier model to identify the reasons for churn for any conversation. The performance shows that NLP has the potential to provide a detailed understanding customers' churn behaviour including why a customer chose to churn.

1 Introduction

In the tech marketplace, it is common to find smart devices that connect with smart assistants like Google, Siri and Alexa. These smart assistants have advanced to a stage where they can efficiently simulate a conversation with users. With advancement of computational power and accessibility of cloud services, these smart assistants can record and extract key features from a voice command and execute tasks in real time. With Natural Language Processing (NLP) tools, it is now possible to extract people's sentiment in conversations through text analysis [1]. Use of NLP tools allow for the extraction of key phrases in a customer's interaction with a service provider and

identify the driving force of customers' behaviour. Customer churn, or loss of customers is a significant issue for any business [2]. When trying to retain customers, it is often best practice to understand as many of the possible reasons which lead a customer to churn. Customers often leave clues of their intention during communication with the business [3]. Nowadays, most of this communication occurs over the phone and online, with these records being stored in an unstructured data format such as voice, text, and emails [4]. NLP holds the potential to extract this information to help predict and reduce customer churn.

Some early work has already found benefits of NLP in addressing different challenges in business. In [5], the authors proposed an automatic labelling model incorporating the Bidirectional Encoder Representations from Transformers (BERT) and word2vec methods to help business automatically label customers' complaint reviews, with the intention of retaining customers. In article [6], an NLP model is used to analyse customer satisfaction by looking at the airline reviews data sourced from TripAdvisor from the year 2016 to 2019. The aim of the study was to guide an airline company in its understanding of its customer feedback. Similar analysis was done in article [7] that explored deep learning and natural language processing technologies to help capture customer perceptions in relation to car reviews. The articles demonstrate that customer sentiment can give insight into actionable items by the business bridging the gap between customers expectation and business performance. Many businesses struggle to apply advanced NLP technologies on their in-house data which can overcome many issues with using structured data such as limited size, noise and imbalance. In article [8], the authors show how a sentiment analysis model uses students' comments to help teachers evaluate online courses. The research used a shallow BERT-CNN model as the students' comments classifier to filter out only the negative comments. It is a pre-trained model which can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as sentence similarity detection. In article [9] an NLP model called Universal Sentence Encoder (USE) is presented for encoding sentences into embedding vectors that specifically target transfer learning to other NLP tasks. USE endeavours to analyse the semantics and syntax of sentences. USE is a strong candidate for the purpose of this article, and it is trained on very large datasets. In theory, USE can then be applied to any downstream NLP task without requiring domain specific knowledge [10-11].

In this article, a customer churn prediction analysis has been performed. The demonstration presented in this article is organised as follows; NLP for understanding Customer Churn in section 2, followed by Methodology in section 3, Results and discussion in section 4 and Concluding remarks in section 5.

2 NLP for understanding Customer Churn

There are many reasons behind a customer churning. If a predictive model is developed using different features of customers from demographic and transactional history, there will be always some features that will show a good correlation with customer churn. These may sometimes lead to a confusing outcome. Churning is an act which requires some effort from a customer as the customer would have likely performed research on alternative products or services before making an active decision to cancel. While doing that, the customer communicates with various business entities. In finance industry, the cancellation typically leaves digital footprints that can be analysed. Digital footprints come in the form of voice conversations,

chats, emails, surveys, feedback, online blog posts, comments, social media statuses and many more. Before deciding to churn, customers tend to show their intention in their conversation with customer service, email queries and social media interactions. Looking at this data, a domain expert can gauge sentiment and assess the intentions of the customer. However, when the number of records exceeds the capacity of humans, an algorithmic approach becomes necessary.

NLP tools can automate processes to perform text analyses that detect pattern and/or sentiment based on desired principles set by a domain expert [12]. NLP converts unstructured text into meaningful data for analysis using various linguistic, statistical, and machine learning techniques. By targeting a particular sentiment, the model can assess millions of call conversations, emails, online comments, and data from various online sources. Many business organisations already store some of these data for quality and training purposes, though, in practice, a very small portion of these data are used. With help of NLP models, customer interactions can be analysed to extract useful information that can help companies determine the best retention strategy to address each customer's concerns. It can also evaluate the performance of different retention strategies to improve customer experience.

Defining and predicting customer churn is not always a straightforward process and involves several complex challenges. A customer may provide different information in their digital footprint. Extracting the right information is quite challenging. Useful information in customers' digital footprint often has a pattern and would expect to be correlated to customer churn. Once such a pattern is identified, it then needs to be validated through testing. Sometimes customers who fit all the valid patterns may not necessarily churn at a given point of time, but later in the future. Therefore, the accuracy of customer churn models is subject to the point of time it is tested for. For such reasons, it is more important to validate and test the extracted pattern using only data from customers who have already churned. Once a set of patterns is validated and established, it is then fed into the NLP model to conduct language processing.

3 Methodology

Customer churn depends on many factors including, but not limited to, the type of product or service and the nature of contract or service accessibility. To understand customer churn, it traditionally requires a domain expert to parse through the customer conversation logs to identify key phrases and comments mentioned. In this article, an NLP model is used to perform the same task. After extracting key phrases and comments manually, the NLP model is used to identify the same phrases and comments mentioned in the remaining conversation log. The requirements of using NLP mandates conversation of any form of unstructured data to be converted into text format. In this demonstration, two NLP models were initially selected for the performance analysis. The models are the Universal Sentence Encoder (USE) and BERT. The BERT model was trained using an online available data-source. However, during the performance analysis, the BERT model performance was not promising. This is not surprising as BERT performs better if it is trained using domain data while USE performs better with heterogeneous data such as call conversations. Therefore, the BERT model has been excluded from the case study.

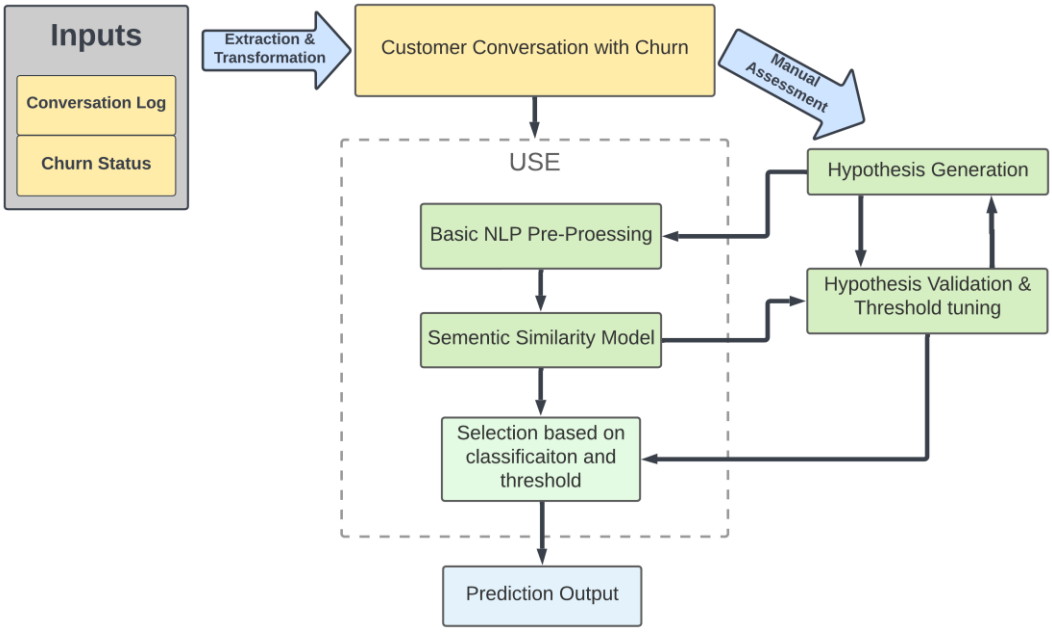

Figure 1. Process flow of the NLP model

Figure 1 shows the structure of the model. From the initial input dataset, the conversation log of all the customers who have churned has been put together. After that, a manual assessment is performed to extract key phrases and clauses which represent the right reason for customer churn. These phrases are considered to be hypotheses. These hypotheses are then tested in the model downstream. After a hypothesis has been selected, it is then passed in pair with the conversation log through each model. The output of the USE model is in the form of a cosine similarity index. In NLP model processing, a conversation log which contains multiple lines or compound sentences with multiple phrases, each sentence or phrase has been considered for separate matching with the target hypothesis and highest index were considered as the representative of that group of phrases/sentences.

The USE model output ranges from 0 to 1 and selecting the right threshold is very important to find the best performance. Depending on this threshold, the recall and precision[1] of the model is calculated. The selection of the right hypothesis is very important for the whole model. If the hypothesis is not working well, it needs to be modified to have a more meaningful match with the conversation log. Once the hypothesis set is calibrated and the threshold is set, the model is ready to go. After the final output of the NLP model, the thresholds are adjusted to obtain the desired confusion matrix. If the objective of the model is to aggressively detect customers with even a low chance of churn, a lower threshold can be selected to ensure a higher true positive rate. This will however also increase the false positive rate. In cases where the model is looking at several sets of hypotheses to validate, each customer conversation is a possible candidate for each hypothesis set. Then, having a higher bar for churn detection will reduce the number of false positives and allow the model to identify the one dominant hypothesis pattern rather than fitting in multiple cohorts.

[1] Recall is the ratio of True-Positives over True-Positive plus False-Negatives
Precision is the ratio of True-Positives over True-Positive plus False-Positives

4 Results

An open-source dataset has been used in this paper which contains customer conversations of a telecommunication company along with their churn status. The conversations between customer and agent discussions are in text format. Customer churn status and conversation log data are collected from an analysis given in [13]. The collected conversation log used in the demonstration are not based on real conversations. The author in [13] has used an unsupervised language model called GPT-2 model to generate customer conversation for the churn data of a telecom company available in [14]. From there, 113 conversation logs have been selected which contained "internet" topic for this case study. Converations generated using the unsupervised language model are not always complete. For completeness of the conversation log, uncompleted sentences were modified by adding finishing clauses. None of these clauses are same as the hypothesis as shown in Table 1, 2 and 3. While the case study has used synthetic conversations, the methods shown to identify reasons for a customer's churn behaviour are applicable to real conversations. The author has successfully applied these methods in commercial scenarios on real data.

The performance of the model is shown briefly in two steps. The first step illustrates the step-by-step process of how the NLP model works, and the second step presents an overall illustration of the case-study.

Table: 1

Canonical (Hypothesis) sentence	**Sentence form conversation**
The internet speed is slow	Well, I am just curious, and I am wondering if you could tell me the truth about what's going on with my internet. The internet speed is too slow for last few days
Internet connection is not working	I've been having some issue with the mobile internet in my property. The internet keeps getting disconnected

In the manual assessment of hypothesis generation, several reasons were found which provided a valid explanation for a customer to churn. One of the reasons identified is poor internet service. To identify the group of customers who have left the company with a possible issue relating to internet service, the two hypotheses mentioned were initially considered. The hypotheses are than selected as a target sentence for all customers that ended in a churn. In NLP, these target sentences are called canonical sentences. After some basic data pre-processing, the conversations are compared with canonical sentences for similarity index using USE. For illustration, Table 1 shows a pair of hypotheses and the corresponding customer conversation.

Figure 2 shows the similarity output of the NLP model. The output in the heat-map shows the performance USE. In the similarity indexing heatmap, the higher indices indicate more similarity. In this case, the conversation that has the best match with the canonical sentences shows a higher index. In this scenario, a similarity threshold with more than 0.6 can provide a desire output.

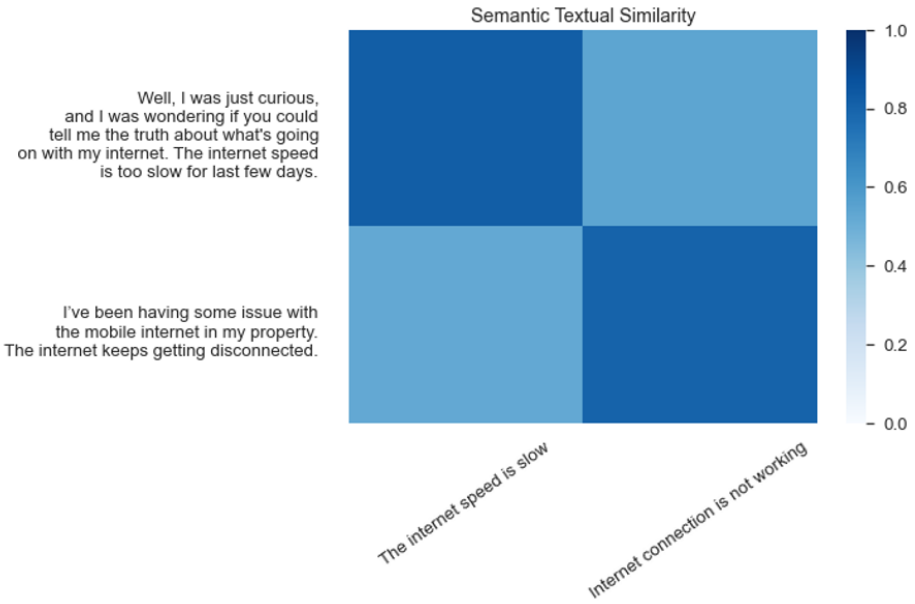

Figure 2. Hypothesis similarity detection outcome

The next demonstration shows how challenging it can be to detect similarity in an oral conversation. Sentences from oral conversations do not always follow the same sentence structure used for USE modelling. This is a challenge common to almost all NLP based applications. One such challenge is to identify contexts containing double negative statements or negative sentiment with opposite meaning. To address such a challenge usually requires various ad hoc steps are considered in the NLP process. To mitigate the confusion around double negative statements an opposite set of hypotheses is considered. If the similarity index for the opposite sentiment has a higher match than the real sentiment, the conversation will be a mismatch for that hypothesis. To illustrate such scenario, a set of sentences are considered in Table 2.

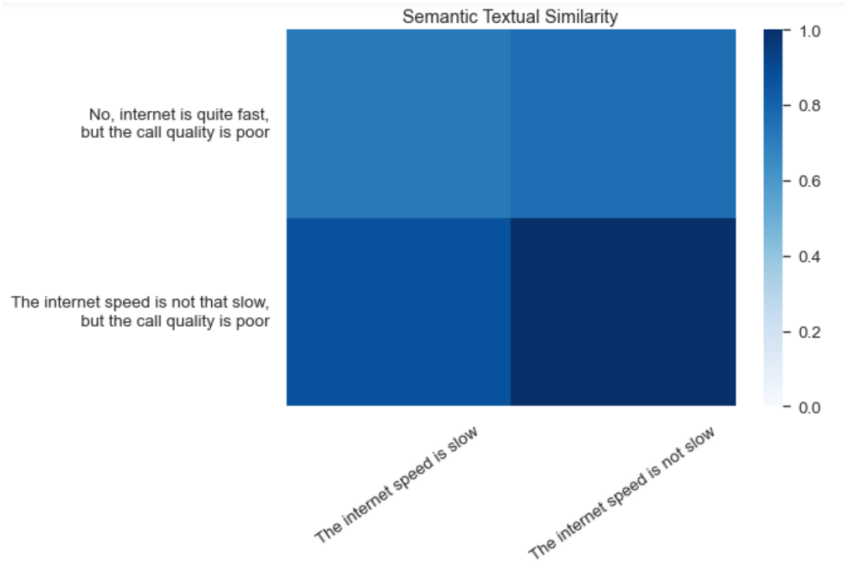

Figure 3. Similarity performance for the hypothesis with opposite meaning

Table: 2	
Canonical sentence	**Sentence form conversation**
The internet speed is slow	No internet is quite fast, but the call quality is poor
The internet speed is not slow (opposite hypothesis)	The internet speed is not that slow, but the call quality is poor

In this scenario, the customers are complaining about the call quality of the service but not the internet speed. Since the NLP models are built to encode word pieces; it directly transforms each chat log embedding vector and assigns weights for each word based on its learning experience. The ability to assign the desired weight for such cases where there are contradictory statements is limited to the availability of such sentences appearing in the training dataset. As these types or oral linguistic techniques are present in the training dataset, the models lead to false indexing of the similarity index and incorrect labelling. However, considering a negative sentiment of each hypothesis helps to reject the false mapping of the hypothesis. Figure 3 shows the performance of the similarity models. For both sentences, the hypothesis with opposite sentiment displays a higher match, which helps avoid the false positive scenarios. In absence of the opposite sentiment, the sentence will incorrectly have a good match with the real hypothesis.

Table 3	
Cohort	**Sentence form conversation**
Internet service issue	The internet speed is slow
	Internet connection is working
	Internet is terrible
	Internet speed is not good
	Not happy with the internet
Opposite sentiment	Internet speed is good
	The internet speed is not slow
	Internet connection is not working
	Internet is not terrible
	Happy with the internet

For the overall analysis, a bigger hypothesis set is generated. Table 3 contains the list of hypotheses that are used to validate the dataset in this demonstration. Figure 4 shows a receiver operator characteristic (ROC) curve. The graph shows the trade-offs between sensitivity and specificity for various thresholds. The ROC AUC curve for the model is shown in the blue line. A model with no predictive capability is represented at the point (0.5, 0.5). The area under the curve (AUC) represents the performance of the model and in this case, the area under curve is around 90%. The closer the score to 100% the better the model distinguishes the classes, if it's closer to 50% then the model performs just as badly as flipping a coin. This indicates that the model does a good job using the hypothesis for "poor internet service". All the chat logs used in this trail include the keyword 'internet'. This means, the model does a very good job by identifying conversations where customers were complaining specifically about the slow internet speed. From Figure 4, the value 0.61 is considered as the best cut-off point or threshold. Figure 5 shows the performance of recall against precision. The graph

also shows that at 0.61 threshold point the model performs an optimum solution. Table 4 and 5 show the performance parameters of the model.

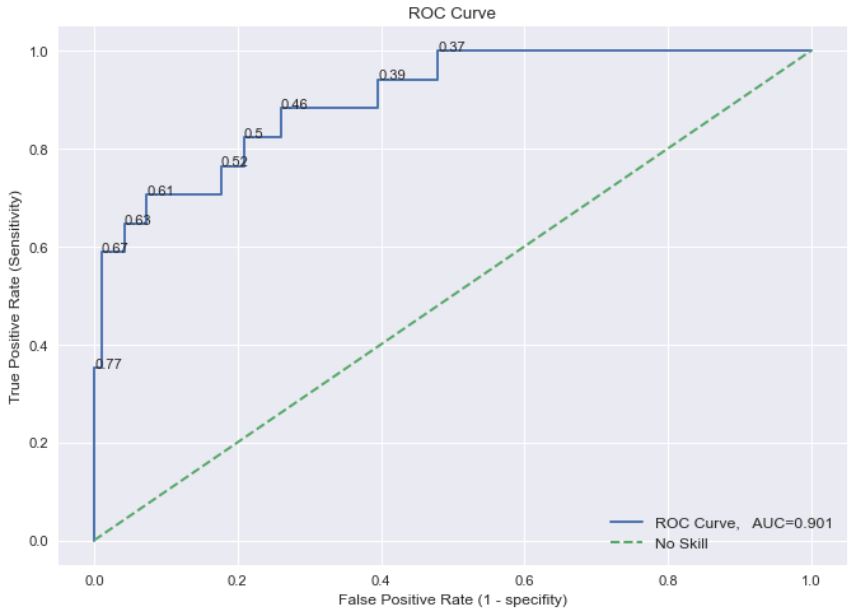

Figure 4. ROC AUC curve

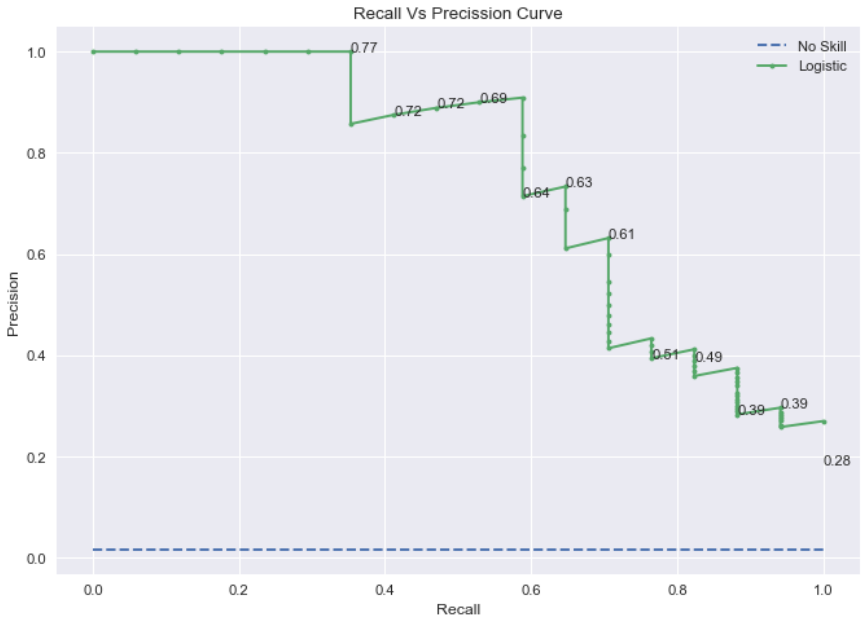

Figure 5. Recall Vs Precision plot for different thresholds

Threshold selection is very important in this model since it is the key factors of the model performance. Selection of threshold can be based on level of important the operation consider for true positive and false position. Let consider the cost function in this case as:

$$R_{th} = C_{FP}\left(\frac{FP}{FP + TN}\right) + C_{FN}\left(\frac{FN}{FN + TP}\right) \quad (1)$$

$$th^* = arg\ min(R_{th}) \quad (2)$$

Equation 2 is a function of threshold th. C_{FP} and C_{FN} are cost co-efficient of false positive and false negative, respectively. Figure 6 shows the cost-function against the threshold for three different combination of cost co-efficient. In this demonstration, the cost of false positive and false negative is considered equal hence selected the 0.61 as threshold.

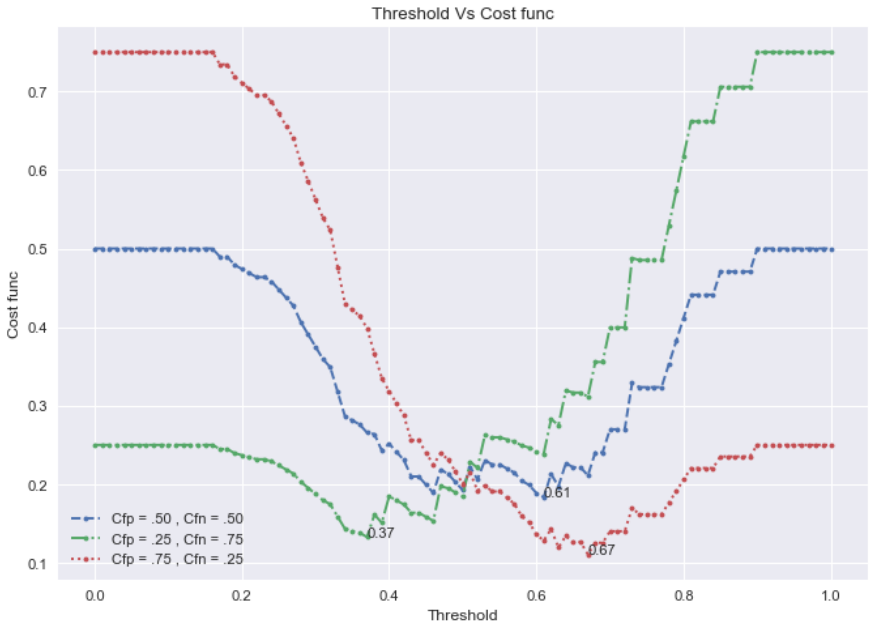

Figure 6. Recall Vs Precision plot for different thresholds

Table 4		ACTUAL VALUES	
		Positive	Negative
PREDICTION VALUES	Positive	TP (12)	FP (7)
	Negative	FN (5)	TN (89)

Table 5	
Performance Outcome	
Precision	63%
Recall	71%
Accuracy	89%
F1 Score	67%

The model shows how the NLP based model can identify customers who will churn expressing a particular type of concern in their conversation with the business. Based on the hypotheses, the model successfully performed a similarity check and identified churn with 89% accuracy. The model is tested on a data set where all the conversations were relevant to the topic in question ('internet'). The dataset was manually verified to make sure that the internet was the only reason of customer churn. With additional pre-processing and post-processing steps, the USE model's performance can be further enhanced.

5 Conclusion

In this paper a case study has been presented to demonstrate a complex language-based task to identify reasons for customers' churn behaviour through sentiment analysis.

The main limitation of the demonstration was the lack of real conversation data. While the case study has used synthetic conversations, the methods shown to identify reasons for a customer's churn behaviour are applicable to real conversations. Some conversations were incomplete which can be misinterpreted by the NLP model. NLP model like USE sometimes struggle to understand verbal conversation as the models are trained using written text. This leads to misinterpretation of words or phrases. With the help of some pre-processing and post-processing steps, some of these can be shortcomings resolved.

The core model presented in this paper can be extended to applications outside predicting customer churn in telecom context. The same method may be applied in financial services and other industries to predict customer churn as well as in other situations such as understanding customer behaviour for marketing, understanding trends in customer sentiment, identifying compliance issues in call-centre, etc.

6 Acknowledgements

The author would like to thank Mudit Gupta, Bahram Nezhad, Rick Shaw, Pok Him Siu, Shobhan Mitra, Welly Han, Joe Lin, Kevin Li, James Yap and Sting Xu for their valuable support and assistance during drafting of the paper.

7 References

1 Mitra, A. (2020). Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies (UCCT), 2(03), 145-152.

2 Chen, K., Hu, Y. H., & Hsieh, Y. C. (2015). Predicting customer churn from valuable B2B customers in the logistics industry: a case study. Information Systems and e-Business Management, 13(3), 475-494.

3 Alibasic, A., & Popovic, T. (2021, February). Applying natural language processing to analyze customer satisfaction. In 2021 25th International Conference on Information Technology (IT) (pp. 1-4). IEEE.

4 Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. Journal of the Academy of Marketing Science, 46(4), 557-590.

5 Tang, X., Mou, H., Liu, J., & Du, X. (2021). Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. Scientific Reports, 11(1), 1-11.

6 Ramaswamy, S., & DeClerck, N. (2018). Customer perception analysis using deep learning and NLP. Procedia Computer Science, 140, 170-178.

7 Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586.

8 Li, X., Zhang, H., Ouyang, Y., Zhang, X., & Rong, W. (2019, December). A shallow BERT-CNN model for sentiment analysis on moocs comments. In 2019 IEEE International Conference on Engineering, Technology and Education (TALE) (pp. 1-6). IEEE.

9 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S.,& Kurzweil, R. (2018, November). Universal sentence encoder for English. In Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations (pp. 169-174).

10 Using NLP to automate customer support, part two. (2020, January 17). FloydHub Blog. https://blog.floydhub.com/automate-customer-support-part-two/

11 Universal Sentence Encoder | TensorFlow Hub. TensorFlow. (2022). Retrieved 18 March 2022, from https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder.

12 Qurashi, A. W., Holmes, V., & Johnson, A. P. (2020, August). Document processing: Methods for semantic text similarity analysis. In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE.

13 Herkert, D. (2021, September 20). Customer churn prediction with text and Interpretability. Medium. https://towardsdatascience.com/customer-churn-prediction-with-text-and-interpretability-bd3d57af34b1

14 Website: https://www.kaggle.com/c/customer-churn-prediction-2020/data