AN APPLICABLE APPROACH TO AUTOMATE CUSTOMER COMPLAINTS

A Thesis

by

Yücel Doğan

Submitted to the
Graduate School of Sciences and Engineering
In Partial Fulllment of the Requirements for
the Degree of

Master of Science

in the
Department of Industrial Engineering

Özyeğin University
August 2022

AN APPLICABLE APPROACH TO AUTOMATE CUSTOMER COMPLAINTS

Approved by:

Assistant Professor Erinç Albey, Advisor,
Department of Industrial Engineering
Özyeğin University

Professor Örsan Özener,
Department of Industrial Engineering
Özyeğin University

Professor Mehmet Güray Güler,
Department of Industry Engineering
Yıldız Technical University

Date Approved: 11 August 2022

To my family, the greatest happiness of my life…

ABSTRACT

Customer complaint management is critical and time-consuming process for institutions. For an effective management and increased customer satisfaction, developing an instant and automated reply mechanism is essential. This phenomenon leads the motivation which helps data scientists to work for developing chatbots. For complex chatbots state of the art techniques of Natural Language Processing is essential to catch intend whereas its high costs. On the other hand, basic machine learning algorithms which is enhanced with NLP techniques can be more applicable for limited data resources. Also even with the help of Regular Expression techniques, quick and effective solutions can be achieved. This thesis covers the development process of a primitive chatbot which is developed by using the customer complaints which are in Turkish language.

ÖZETÇE

Müşteri şikayeti yönetimi firmalar için kritik ve zaman tüketen bir süreçtir. Etkin bir süreç yönetimi ve müşteri memnuniyetini arttırmak için anlık ve otomatik bir cevap yapısı kurulması kritiktir. Bu fenomen, veri bilimcilerin chatbot'ları geliştirmek için motivasyonlarını sağlar. Gelişmiş chatbot'lar için Doğal Dil İşleme alanındaki en modern teknikler niyeti belirlemede başarı için gerekli olduğu kadar aynı zamanda geliştirmesi de maliyetlidir. Diğer taraftan doğal dil işleme teknikleri üzerine kurulu temel makine öğrenmesi algoritmaları limitli veri kaynakları üzerinde daha uygulanabilir olabilir. Hatta Düzenli İfadeler (Regular Expressions) ile bile hızlı ve efektif çözümler üretilebilmektedir. Bu tez Türkçe dilindeki müşteri şikayetlerinin kullanılarak başlangıç seviyesinde bir chatbot oluşturulmasını kapsamaktadır.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor Assistant Professor Dr. Erinc Albey for guiding and supporting me during my thesis period.

Besides I also wish to thank the committee members: Professor Dr. Örsan Özener and Professor Dr. Mehmet Güray Güler for the time allocated to this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Customer complaint management is a crucial process for institutions in various aspects. Handling complaints with minimum time and managing human resources more efficiently are both essential for increasing customer satisfaction and enhanced cost control. The first step of the manual solution pipeline is categorizing the complaints into predetermined categories. According to the selected complaint category, complaints are directed to related departments to solve complaints by taking the proper actions. If the manually chosen category is wrong, this may lead to inappropriate actions or prolonging in the resolution period. Therefore, in order to increase customer satisfaction and more importantly avoid inappropriate actions, customer complaints needs to be correctly classified according to their topics in the first touch point.

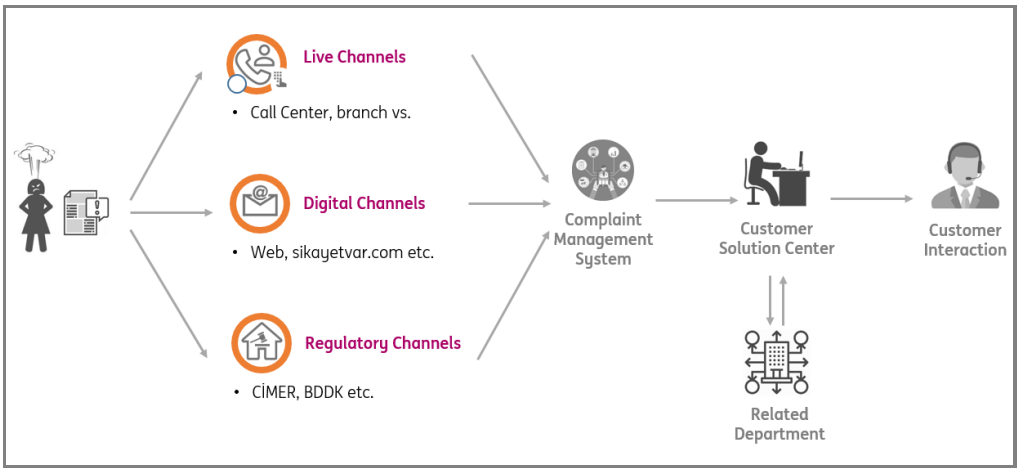


Figure 1: Complaint handling process before automation

The objective of this thesis is to detect the topics (classes) of customer complaints which are originated from digital channels (website, mobile application) by building a predictive classification model pipeline. Currently, classification is being done manually by back office agents. After implementing the pipeline, categorization will be automated.

Following the automated classification, solution of the complaint is also going to be automated with Robotic Process Automations (RPA) solutions. This automation will not be applied to all of the complaints that are classified. Thus the confidence of the prediction is going to be compared with a defined threshold. If the probability of the prediction is high enough and the specific complaint process can be automated in risk-wise, then necessary RPA solutions are also going to be implemented.

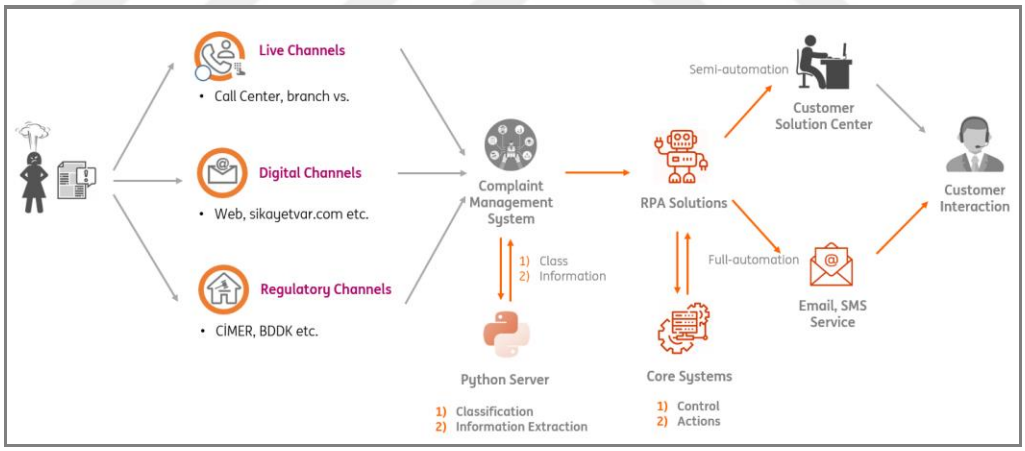


Figure 2: Target complaint handling process after automation

CHAPTER II

LITERATURE REVIEW

In this thesis, NLP based text classification is performed over Turkish complaints. During this work, several supervised classification algorithms have been applied. In this section, I have listed several relevant studies that I have reviewed before executing the technical steps. My literature review is mainly focusing on algorithms and difficulties of NLP studies with Turkish text.

2.1 Algorithm Reviews

Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets [1]

The most straightforward predictive approach for classification is Naïve Bayes algorithm. In this paper, text classification of English tweets via Naïve Bayes algorithm has been covered. Algorithm has been applied after regular preprocessing steps. Authors are focusing to the easiness of implementation besides interestingly the high accuracy rate of their trials. On the other side, there are just 2 classes, positive and negative. Thus I am really skeptical about getting the same evaluation metrics with increased number of classes.

Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification [2]

In this study, more 2.6 million comments having been sent to an e-commerce application have been classified. On the label side there are five classes. After the

preprocessing steps, authors are applying following algorithms; Naïve Bayes, Random Forest, Decision Tree, SVM and Logistic Regression. In the paper, it is stated that these algorithms have selected because of the popularity of the algorithms in text classification. In the evaluation section Pranckevicius and Marcinkevicius select accuracy as the main metric. According to the results of their experiments Logistic Regression is the best algorithm with 58.5% accuracy whereas Decision Tree is the least successful one.

Text Classification and Classifiers:A Survey [3]

Korde and Mahender are firstly mentioning the importance text handling in data generated via online channels. According to their beliefs more than 80% of the online data is in text format. Thus to create the commercial value, they are looking for analytic alternatives to increase the data handling ratio. In their study, they are applying following algorithms; Rocchio's Algorithm, K-Nearest Neighbors, Naïve Bayes, Decision tree, Decision Rule, SVM, Neural Network, LLSF, Voting, Associative classifier, Centroid based classifier, Additional classifier. In this study, it is stated that Naïve Bayes and SVM algorithms work well with textual and numerical data. They are also comparing the algorithms in the context of implementation and they are in favor of using these 2 in implementation perspective.

A comparative analysis of text classification for Turkish language [4]

In this study, authors are dealing with news data. Data consists of 7 categories (economy, policy, sports etc.). In each category there are 700 instances. In the study, preprocessing steps like stop word cleaning and morphological analysis have been applied. As most of the researchers in the same field, they are trying Naïve Bayes classification.

Authors state that some preprocessing steps like stop word cleaning and morphological analysis are not effecting the results. Differently they are focusing on the importance of feature selection. For this feature selection process they used Knowledge Gain and ChiSquare approaches.

Character-level Convolutional Networks for Text Classification [5]

In this research, authors deal with eight different datasets. These datasets are AG's News, Sogou News, Amazon Review etc. These datasets contain between 120000 to 360000 instances. Number of classes are also changing between 2 to 14. The main aim of the paper is observing the performance of character level classification in addition to the traditional methods. Researchers executed approaches like bag-of-words, TF-IDF, stop word removals, n-grams. As a predictive algorithm, they first tried RNN. Then they have applied Convolutional Networks and interestingly concluded that CNN is more useful for extracting information from raw signals, ranging from computer vision applications to speech recognition and others.

TLS-ART: Thai Language Segmentation by Automatic Ranking Trie [6]

Authors are stating that NLP is developed for computers to understand the language humans speak to each other. Besides usage areas of the field is also steadily increasing. In the research, NLP is categorized under four groups:

- Lexical analysis: Sentences are split into smaller parts and these parts are called as tokens.

- Syntactic analysis: Sentence structure or syntax of all the tokens are checked to examine whether the syntax structure of the sentence is proper or not.

- Semantic analysis: In the 3rd group, meanings of the sentences are analyzed.

- Output transformation: In the final stage, data is transformed into an output to predict the context of the sentence.

Automatic Complaint Classification System Using Classifier Ensembles [7]

In his study Fauzi is dealing with Indonesian complaints which have been sent to city management of Malang. Briefly he applies different ML algorithms like Naïve Bayes, Maximum Entropy, K-Nearest Neighbors, Random Forest and Support Vector Machines. Afterwards he also checks whether ensembling methods increase the performance or not. He applies both hard and soft voting approaches. Results show that performance increases when it is compared with outputs of the standalone algorithm tries. Accuracy percentage is used as main evaluation metric.

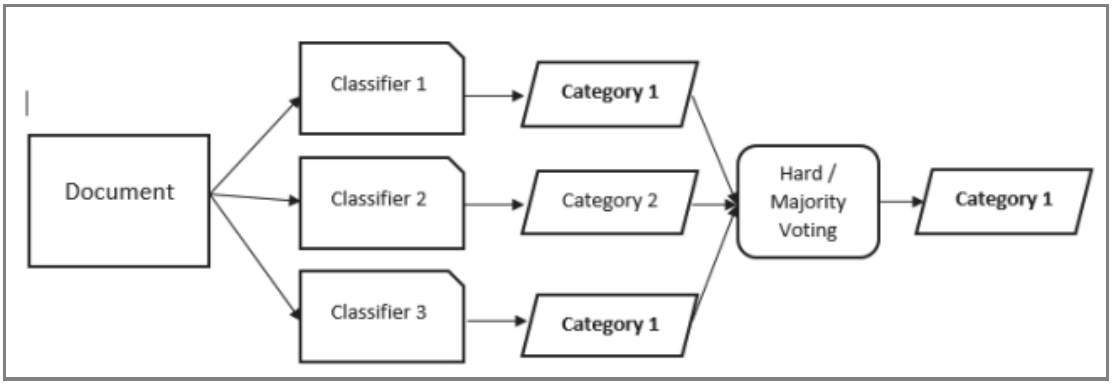


Figure 3: Example of an ensemble via hard voting

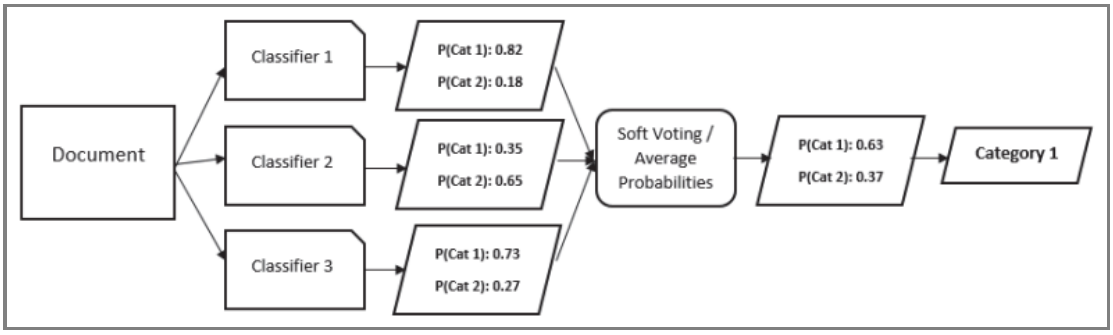


Figure 4: Example of an ensemble via soft voting

Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches [8]

In this research paper, authors are working with customer complaint data that are directed to a global food brand. Complaints are incoming from various digital channels. Afterwards these are manually classified and processed. Besides the analytic side of the projects, researchers are also emphasizing the importance of the project from business perspective. The key points that they are mentioning are:

- Manually labelling the data is so time consuming. When the number of complaint increases process efficiency is also decreasing.

- Customer satisfaction is critical for firms. To ensure the satisfaction of the customers, response time should be decreased.

2.2 NLP with Turkish Text

Zemberek, an open source NLP framework for Turkic Languages [9]

In this research paper A. Akın and M. Akın are comprehensively covering how to deal with Turkic Languages, not just Turkish, in NLP studies. According to their resources Turkic Languages are being spoken by more than 140 million people. They are introducing Zemberek framework which provides basic NLP operations such as spell checking, morphological parsing, stemming, word construction, word suggestion, converting words written only using ASCII characters and extracting syllables.

In this research, Zemberek library has been used for Turkish morphological analysis. This analysis is summarized as follows:

- Zemberek parses the tokens with root dictionary-based parsing. At this stage parsing is based on the probabilities of the root. They have named this process as Direct Acyclic Word Graph (DAWG) tree as may observed below.

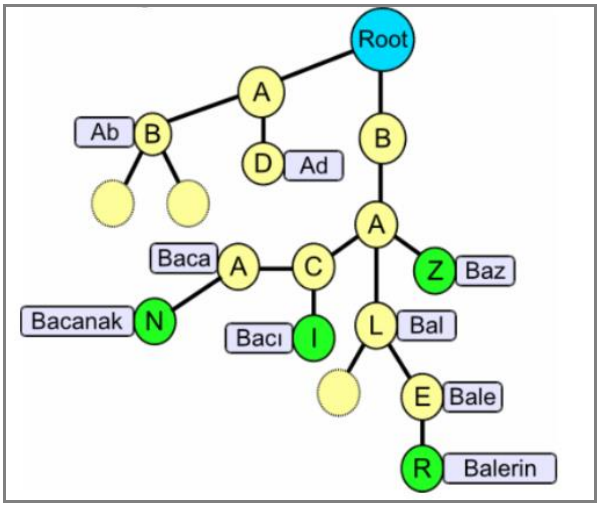


Figure 5: Direct acyclic word graph

- Another problem in NLP studies with Turkish Language is the changes of the root of Turkish words in case of suffixes being add. This phenomena has been shown with the below example:

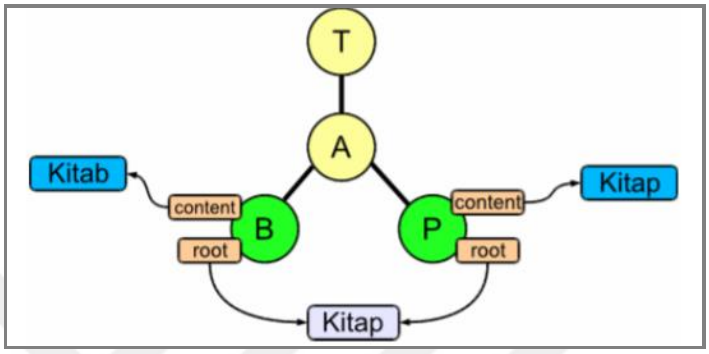


Figure 6: Example for a change in the root of a word in case a suffix is added

The impact of preprocessing on text classification [10]

In this research paper, the importance of preprocessing has been detailed for different languages, English and Turkish. For these analysis they have used email and news data. They have applied following preprocessing steps for the classification problem; stop-word removal, lowercase conversion, tokenization and stemming. Afterwards feature selection and classification have been executed. SVM was used as predictive algorithm and micro-F1 as evaluation metric of the results.

They have tried to measure the impact of each preprocessing step for each language. The results show that:

- NLP pipeline also includes preprocessing steps which is an important component of the studies. Researchers should not just focus on main analytic

approaches like feature extraction, feature selection or application of classification algorithm.

- Even simple approaches like stop-word cleaning should also be handled sensitively. Stop words might be reviewed with use case specific process.

- Lowercase transformation is not a must for all languages. However while handling Turkish data, it shouldn’t be skipped.

CHAPTER III

METHODOLOGY

3.1 Data Collection and Details

The main data in this thesis is the complaint content (text) and date-time of the given complaint. According to these features, complaints are classified into categories.

Table 1: Sample data showing the content of the project data

Feature	**Type**	**Sample Data**
Customer Unique ID	Numeric (ID)	1234567
Complaint Unique ID	Numeric (ID)	C004324D20210512T1047
Complaint Date-Time	Date-time	2021-05-03 18:47:03
Complaint Text (In Turkish)	String	I want to increase my credit card limit. Could you please set the new limit as 30000 TL

There are more than 200 different classes in the process. Since building a sole multi-class classification model will increase misclassification rates, I grouped the classes in a higher level. At the end, solution approach of the project consists of 2-step sequential modelling design.

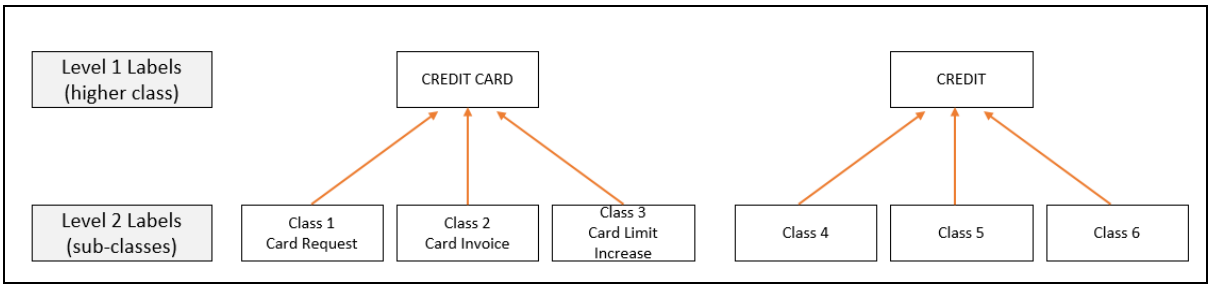


Figure 7: Illustration of Leveled Class Design

Level 1 Data:

Data used in model development (train + validation) is between October 2018 and October 2020. Test data is between January 2021 and May 2021. The following table summarizes the time window for level 1 input data:

Table 2: Time window for Level 1 Classification

Data set	Start Date	End Date	# of complaints
Train	11/10/2018	11/10/2020	6404
Validation	11/10/2018	11/10/2020	2135
Test	02/01/2021	05/05/2021	2623

Level 2 Data:

Since the data size for each Level 2 model for 2 year time-window is not sufficient in data preparation, level 2 data set time-window is extended to 4 years for Level 2 classification. For whole level-2 models data sets with following time periods are used:

Table 3: Time window for Level 2 Classification

Data set	Start Date	End Date	# of complaints
Train & Validation	01/01/2017	01/01/2021	50911
Test	02/01/2021	05/05/2021	2623

As may observed in the upper table, 50911 instance (train + validation) is directly used in modelling process with no exclusions.

3.2 Feature Engineering

3.2.1 Text Based Features

Datasets below are gathered to generate the features for the thesis:

Table 4: Datasets and features

Dataset	**Feature**
Complaint text-based features	TF-IDF vectors from complaint text
Complaint date features	Target encoding feature of date buckets

Since the thesis aims to categorize complaints considering their context, words are tokenized as features. After several preprocessing steps, features are generated by TF-IDF vectorizer. Due to the linguistics of Turkish language, several preprocessing steps have been applied on Turkish complaints. Preprocessing steps are detailed in section 3.3.

3.2.2 Date Based Features

Text based features are composing the main set of input features. In addition to this, thanks to the inputs of business unit, I suspected that the incoming date-time of the complaint might also be statistically significant in classification process. For example certain types of complaints might be generated in certain days or hours. Vice versa only certain types of complaints might be led in certain times of the day. To give a concrete example, the probability of a fraud inquiry is more likely in off-day hours (1-3 am) whereas the frequency of credit requests are decreasing in the same time interval.

Thus in addition to text based features, date-based features have been generated to reveal the importance of time. There are 2 types of date features that have been engineered.

- First feature that is engineered is related to information of the day itself when the complaint is being initiated. Flag features like national holiday, religious holiday, salary day or working day are generated.

- The second set of features that is related with date is named as date bucket features. According to the incoming hour of the complaint, I assigned the complaint into 1 of the 4 predefined date buckets.

    - Date bucket 1 for between 22:00 and 04:59
    - Date bucket 2 for between 05:00-07:59
    - Date bucket 3 for between 08:00-18:59
    - Date bucket 4 for between 19:00-21:59

Determination of hours for buckets was made based on the analysis in below.

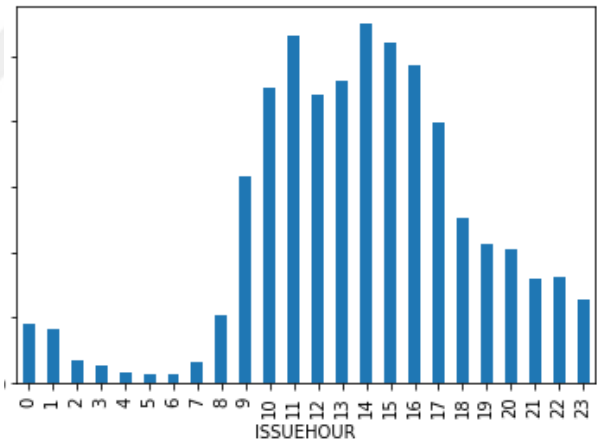


Figure 8: Number of complaints in hours

The figure in above shows the number of complaints received in each hour. It can be seen that the workhours are the most intense hours. Hence, the working hours are defined as one bucket. In the hours after work hours have slightly high number of complaints, so determined as another bucket. Lastly, late hours and early hours of a day were decided as the remaining two buckets based on the business unit advisory.

There are 17 level 1 classes in the thesis. Encoding 18 x 4 = 72 binary date bucket feature would increase the dimension. To overcome this, target encoding [11] has been applied. I took the counts of the number of total complaints for each main category and each bucket within these 3 months. Those counts are divided by the total number of complaints for each bucket within the last 3 months. Therefore, we get a ratio of each main category within the last 3 months for each bucket.

Bucket features have been illustrated in the following figure:

Table 5: Examples of date bucket features

FIRST_DAY	BUCKET	CLASS 1	CLASS 2	CLASS 3	CLASS 4	CLASS 5	CLASS 6	CLASS 7	CLASS 8	CLASS 9	CLASS 10	CLASS 11	CLASS 12	CLASS 13	CLASS 14	CLASS 15	CLASS 16	CLASS 17
01/04/2021 00:00:00	BUCKET_1	0.241	0.053	0.107	0.107	0.011	0.064	0.016	0.016	0.155	0	0.027	0.075	0.043	0.005	0.005	0	0.075
01/04/2021 00:00:00	BUCKET_2	0.133	0	0.133	0.2	0.067	0.067	0	0.133	0.067	0	0.067	0	0	0.067	0.067	0	0
01/04/2021 00:00:00	BUCKET_3	0.191	0.028	0.094	0.089	0.028	0.039	0.029	0.027	0.079	0.004	0.288	0.029	0.016	0.008	0.017	0.008	0.025
01/04/2021 00:00:00	BUCKET_4	0.312	0.029	0.098	0.092	0	0.04	0.035	0.035	0.156	0.012	0.017	0.052	0.017	0.012	0.017	0	0.075

In this figure, we have 4 rows for bucket 1, bucket 2, bucket 3, and bucket 4 of April 2021. Each row shows the bucket information and the ratio of each main category within total complaints from January, February, and March. The sum of the rows of main categories is 1 for each bucket. If there is no complaints from a bucket and a main category, it is filled with 0.

3.3 Data Preprocessing

Before starting the modelling, customer complaint texts are processed with several steps. Those steps are summarized in below:

1- Stops words are removed. In this process, common Turkish stopword list is enriched with the business unit to eliminate company specific words as well.

2- Name Entity Recognition (NER) [12] models are applied to detect name, location and organizations stated in the complaint text. Text that is being handled is containing several personal names, organizations and locations. These types of names area anonymized with named entity recognition (NER) model call "Zemberek NER model". Zemberek is an open source NLP framework for Turkish Language. It contains spell checker, tokenizer, normalization, and NER tools that were built based on Turkish language properties. Zemberek NER uses Averaged Perceptron models for each class of named entities (PERSON, ORG, LOC). Thanks to the framework, personal names are replaced as 'PERSON', locations as 'LOCATION', and an organization as 'ORGANIZATION'.

Table 6: Sample NER Replacements

Original	**Replaced**
…Besiktaş Branch	[Location] Branch
…Levent Branch	[Location] Branch
…person called Ece in your branch	…person called [Name] in your branch

3- All texts are converted to lower case.

4- Turkish characters such as 'ş', 'ğ', 'ü' etc. are replaced with English ones.

5- Similar words are replaced with their corresponding main word. At this step, similar words are extracted with Word2Vec [13] trained on complaint data set. They are checked out by business owners to be sure the similar meaning of words. With this way it is aimed to detect meaningfully close words. Then, meaningfully close words are replaced with a main word. To illustrate any word

meaningfully similar to ‘temsilci’ -means agent in English- such as ‘çalışan’ (‘employee’) are replaced with ‘temsilci’. Thus, variety of words with same meaning is decreased. With this manual process, business wise dimension reduction is applied

6- Manual corrections are made for typos in the texts. Typo errors are also commonly occur in the complaint texts. To overcome this, similar words are replaced with the more common ones. To find close words, the output of Word2Vec model is used. A list of similar words considering embedding suggestions is prepared and shared with the business owners. With their expert judgements, word replacements are executed. Example set of replaced words are documented in table:

Table 7: Sample synonym, typo replacement

Main word	**Replaced words**						
dilekce	dilekcesini	dilekceyi	dilekcemi	dilekcenin	dilekcede		
ertelemek	erteletmek	ertlemek	otelemek	ertelemk	ertelenmesini	ertemek	
gise	gisede	gisedeki	banko	vezne	bankoda		
harca	harcamaya	harcamaa	haracama	harcam	harcayin	harcamasina	harcamaniza

7- Stemming is applied. [14] One of the main problems with preprocessing is dealing with prefixes and suffixes. Turkish is a rich language that has lots of prefixes and suffixes. Suffixes in words are omitted with the help of Zemberek package [9]. Examples of this removal process is demonstrated in below:

Table 8: Example suffix removal

Original Word	**After Removal of Suffix**
Beylikdüzü'nden	Beylikdüzü
Şişli'nin	Şişli
İstanbul'un	İstanbul
Ali Bey'den	Ali Bey

8- Complaints with subcategories that were exist in past but not used anymore are removed from the dataset.

After applying all preprocessing steps, a TF-IDF vectorizer [15] is built to calculate the term and document frequency with complaint texts. Text-based features have been generated from TF-IDF vectorizer considering the importance of tokens/words in texts. Finally Date features are generated and included to the feature set.

3.4 Feature Selection

Before feature selection, TF-IDF vectorizer has been used from sklearn with the limitation of maximum 10000 features.

Level 1:

Chi-Squared test [16] is performed by aiming to select the features which are highly dependent on target. 0.05 is selected as as p-value threshold for Level 1 categorization. Features with p-value less than 0.05 is kept for the model training. At the end, we have more than 3000 features from TF-IDF vectorizer.

Level 2:

The same feature selection method in Level 1 categorization is being applied for level 2 as well. However considering the number of features, different p-values for each category has been selected. p-values of each Level 2 model are recorded in the following table:

Table 9: p-values for Level 1 and Level 2 models

	Level 1	**Level 2 models**							
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
p_value	0.05	0.04	0.04	0.04	0.04	0.04	0.02	0.02	0.04

Selection of P Values

Since a p-value is greater than the conventionally accepted significance level of 0.05 (i.e. $p > 0.05$), we fail to reject the null hypothesis. Null and alternative hypotheses are defined in below:

- Null Hypothesis (H0): There is no relationship between the variables
- Alternative Hypothesis (H1): There is a relationship between variables

P-values given in upper table is used to have reasonable number of features and features with strong relation between target variable.

Although Level 1 p-value is selected as 0.05 as common sense, uncommon p-values like 0.02 or 0.03 are also used in level 2. The reason is the high number of features, in order to limit the dimension to overcome sparse matrix problem. So, I avoided to confuse model with unrelated words/features.

Selected Features

In Level -1 categorization, we have 3509 features from TD-IDF method. Top 10 features (important keywords) of them are published for each main category:

Table 10: Final Feature List of Complaint-based Features

Main category	# of selected features	Top 10 features
Credit Card	244	ekstre,harc itiraz,itiraz,kart,kart limit,kredi kart,limit,kart iptal,adres,kart temazsiz
Credit	214	ertelemek,ertelemek iste,ipotek,kkb,kredi ertelemek,rapor,kredi,ipotek kaldirma,tapu,yapilandirma
##Other classes have censored due to company privacy		

In Level 2 categorization, we have different features from TD-IDF process. The number of total text-based features for Level 2 models are in the following table:

Table 11: Number of text-based features for each Level 2 model

Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
2697	1133	396	566	227	137	378	523

Day and bucket features are common in both level 1 and level 2 models.

3.5 Algorithmic Approach

Classification models are built to learn important factors that determine the predefined topics of complaints. Initially first level prediction model is being run. After getting the predictions from the first level model, the second level prediction is being made.

Second level modelling aims to predict level 2 classes of the complaints whose main category is assigned by the first level model. To illustrate, if a complaint is categorized as Card in the first level model, the complaint is conveyed to second level model that predicts

which subcategory of card that complaint belongs to. For example, assume that a customer made a complaint about card fee objection. This complaint is expected to be categorized as Card in the first level model. Next, it is conveyed to the second level models which are changing according to the main category. Therefore, this particular complaint is proceeded to the 'Second Level Card Model' to predict which subcategory of card it belongs to.

The following figure shows sequential modeling pipeline. After preprocessing steps, Level 1 category is predicted. Then, the complaint is conveyed to the proper Level 2 module.

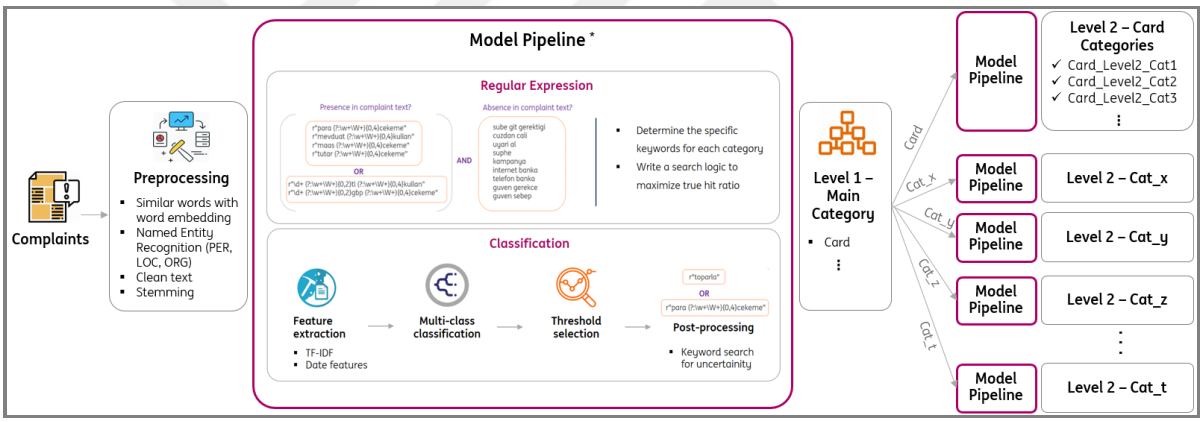


Figure 9: Sequential modeling pipeline

To be more precise:

- After preprocessing steps, model development of first level is started.
- Feature selection is performed.
- Several classification algorithms are applied and the best performing one is selected according to micro-averaged F1-score. [17]
- Next, development of second level model has begun.

- Same preprocessed complaint texts and date features are used in this phase to train models for selected main categories.

- To summarize, a complaint will be first preprocessed then will be convey to first level model and next, according to its main category prediction it will be transmitted to second level models to predict its subcategory.

3.6 Model Selection and Hyperparameter Tuning

3.6.1 Algorithms Used

Following classification algorithms have been tried in the thesis:

- XGBoost [18]
- ANN – feedforward [19]
- RNN [20]

XGBoost is one of the best boosting algorithms that provides fast and high performance. Thus, it is widely used in ML cases. Moreover, among the applied models, the best results for the main category model were obtained by XGBoost Algorithm with One-vs-Rest considering micro F1, train duration and deployment easiness.

3.6.2 Cross Validation and Testing Setup

Level 1

During the development phase of the level 1 model, the complaints were divided into two groups as training and validation sets randomly. Train and validation sets include

0.75 and 0.25 of complaints, respectively. The hyperparameters were decided by 5 fold Grid Search cross validation on the training set.

Level 2

While developing the level 2 models, the train and validation sets are constructed in the same way as done in level 1 model. 5 fold Grid Search cross validation has been used in hyperparameter search.

3.6.3 Hyperparameter Search:

The hyperparameter search (Grid Search) space and the selected parameters of the final model of Level 1 are shown in following tables respectively:

Table 12: Parameter search of Main Category Model

Parameter	Search Space
max_depth	10, 11, 12, 13
learning_rate	0.03, 0.04, 0.05, 0.06
n_estimators	20, 25, 30, 35, 40
reg_lambda	100, 200, 300, 400
reg_alpha	50, 100, 200
scale_pos_weight	0.2, 0.3, 0.4, 0.6, 0.8

Table 13: Hyperparameters of Final Main Category Model

Parameter	Value
max_depth	10
n_estimators	35
reg_lambda	100
reg_alpha	50
learning_rate	0.05
scale_pos_weight	0.8
random_state	0

Same process has also been applied for level 2 as well. XGBoost model with One-vs-Rest model was performing better according to selected criteria. After feature selection, we apply hyperparameter search and find the best parameters for each Level 2 model.

In total there are 4800 possible combinations. By considering both data sizes and computation power, the search space should be considered as sufficient.

CHAPTER IV

COMPUTATIONAL RESULTS

4.1 Performance of Predictive Models

Performance metrics:

Precision of each category, AUC score, and micro-averaged F1-score of the model were selected as evaluation metrics during the model development. It is aimed to have reliable predictions as much as possible. Reason for the selection of micro precision is that to include volume of categories to evaluation since not every category has the same number of complaints. Since the final aim of the pipeline is to decrease the workload, with this method I increases the focus on performances in categories having high complaint size.

Results of Level 1:

The model results on the validation set are evaluated by Recall and Precision metrics for each category in the following table:

Table 14: Model performances on validation set in main categories

Main Category	Precision	Recall
Class 1	0.66	0.75
Class 2	0.75	0.65
Class 3	0.77	0.89
Class 4	0.47	0.27
Class 5	0.62	0.32
Class 6	0.61	0.58
Class 7	0.63	0.75
Class 8	0.8	0.87
Class 9	0.83	0.85
Class 10	0.6	0.61
Class 11	0.62	0.61
Class 12	0.56	0.56
Class 13	0.6	0.64
Class 14	0.46	0.26
Class 15	0.84	0.72
Class 16	0.33	0.12

Micro-AUC based on ROC curve and micro-averaged F1-score of the model are reported on the validation and test sets to demonstrate the model performance in the following table:

Table 15: Performance metrics of Final Main Category Mode

Data set	Model	Micro-AUC	Micro-averaged F1-score
Train Set	XGBoost	0.84	0.71
Validation Set	XGBoost	0.82	0.66
Test Set	XGBoost	0.81	0.63
Train Set	ANN	0.87	0.73
Validation Set	ANN	0.83	0.66
Test Set	ANN	0.80	0.62
Train Set	RNN	0.88	0.73
Validation Set	RNN	0.83	0.67
Test Set	RNN	0.81	0.63

As a conclusion, I did not take the significantly different evaluation metric scores on validation and test sets for XGBoost algorithm. Therefore, it can be stated that the model results are consistent over time.

Results of Level 2:

For each level 2 model, model performances of subcategories on the validation set is documented in below:

Table 16: Performances of level 2 models

Level 1 Label	Level 2 Label	Precision	Recall
Class 1	0	0.73	0.33
	1	0.83	0.74
	2	0.78	0.64
	3	0.76	0.79
	4	0.5	0.18
Class 2	0	0.17	0.08
	1	0.82	0.82
	2	0.8	0.71
	3	0.83	0.75
	4	0.5	0.03
	5	1	0.2
Class 3	0	0	0
	1	0.81	0.74
	2	0.5	0.06
	3	0.81	0.85
	4	0.75	0.2
	5	0.87	0.46
	6	0.76	0.57
	7	0.88	0.76
	8	0.79	0.73
	9	0	0
	10	1	0.3
Class 4	0	0	0
	1	0.84	0.64

Level 1 Label	Level 2 Label	Precision	Recall
Class 5	0	0.39	0.52
	1	0.82	0.77
	2	0.58	0.25
	3	0.7	0.33
	4	1	0.07
	5	0.8	0.86
Class 6	0	0	0
	1	0.83	0.85
	2	0.83	0.51
	3	0.81	0.23
	4	0.78	0.35
	5	0.75	0.1
Class 7	1	0.79	0.93
Class 8	0	0.43	0.14
	1	0.83	0.99
	2	0.66	0.88
	3	0.51	0.42
	4	0.66	0.68
	5	0.64	0.55
	6	0.5	0.28
	7	0.63	0.35
	8	0.67	0.4
	9	0.67	0.44
	10	0.65	0.76
	11	0.29	0.07
	12	0.63	0.65
	13	0.17	0.08
	14	0.63	0.53

Micro-AUC based on ROC curve and micro-averaged F1-score of the model on the validation and test sets are listed below table. This table can be observed as the summary of the project performance:

Table 17: Performance metrics of Final Level 2 models

	Train		Validation		Test	
Level	**Micro Average F1 Score**	**Micro Auc**	**Micro Average F1 Score**	**Micro Auc**	**Micro Average F1 Score**	**Micro Auc**
Class1	0.61	0.79	0.763	0.872	0.598	0.686
Class2	0.74	0.84	0.686	0.811	0.523	0.727
Class3	0.77	0.85	0.721	0.825	0.687	0.782
Class4	0.73	0.85	0.668	0.817	0.467	0.707
Class5	0.69	0.81	0.583	0.749	0.578	0.735
Class6	0.79	0.87	0.70	0.82	0.611	0.674
Class7	0.86	0.51	0.855	0.496	0.763	0.382
Class8	0.72	0.79	0.762	0.821	0.709	0.75

4.2 Technical Bottlenecks

Following topics were created a confusion during the thesis period:

1- Stemming errors: Due to linguistics of Turkish language, stemming is a necessity in the preprocessing period. On the other hand it has a main drawbacks. The problem is losing the information of suffixes. A sample case might be observed in the below table. For both of the cases, stems are the same whereas they don't imply exactly the same in original forms.

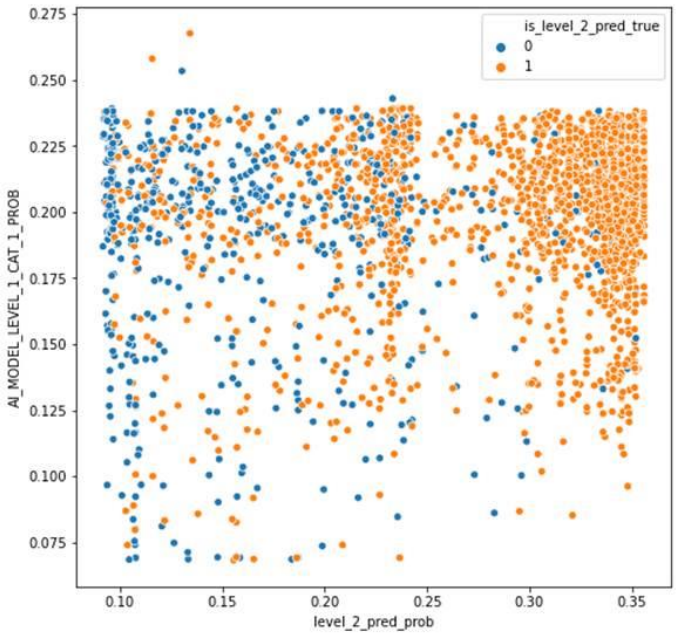
Table 18: Stemming examples

Original word	After Stemming
Kampanyalarım – my campaigns	Kampanya - campaign
Kampanyalarınız – your campaigns	Kampanya - campaign

2- Integration of the results of level 1 and level 2 models: Level 1 and level 2 models are running in a sequential pipeline. First running level 1 model. According to the result of level 1 model, I am just running the related level 2 model,

whose prediction probability was the highest in level 1. This approach is helping to reduce the runtime of the pipeline since I am not running all level 2 models for each instance. On the other side, I am losing the information of level 1 models like 2nd highest probability class. In addition to this, I am not calling other level 2 models which I might get a more significant prediction.

Figure 10: Sample prediction probabilities for level 1 model and level 2 model for the same instances



3- <u>New categories in the complaint process:</u> Model training has been done with historic data belongs to 2018-2020 data as stated in section 3.1. Due to the dynamic environment of the business process, there are always new categories emerging. Since those classes are defined after the train period, it is impossible to create a prediction for newly emerged classes with a static model.

CHAPTER V

CONCLUSION

The main objective of this thesis is to create an automated reply mechanism for customer complaint process. The initial point of this ultimate target is developing a predictive classification process. I am not stating this developing a single model, since there are multiple different models in this classification process.

At the end, I have developed a theoretical pipeline. However to evaluate the success rate of the pipeline, we need to analyze each category separately. Success of the each category depends on 3 main factors.

First factor is the significance of the prediction. This significance can be evaluated with the related model precision. As a rule of thumb, as the prediction probability increases, precision is also increases. Thus threshold selection is also an important task that needs to be handled sensitively.

Second factor is related with the business process of the complaint. Automation of the solution process might be done in a certain extent for most of the classes. On the other hand automation level might differ from class to class. Thus success of the automation cannot be evaluated without considering the business process.

The last success factor is the business-wise risk of the automated complaint solution for each class. Unless data scientist ensures 100% precision level, there might always be a risk cost of the automated solution. For different classes, this risk cost is also

differentiating. So success evaluation should also be done by considering the risk of misclassification.

CHAPTER VI

FUTURE EXPANSION AREAS

There are several points that will help to improve the process. 2 of them are quite important, however these are beyond the scope of the thesis. These are:

1- Continuous retrain for new categories: As stated in section 4.2 Technical Bottlenecks, one of the problem is the emerging of new categories. To overcome this, an automated retrain pipeline needs to be implemented. Thus, I will overcome the problem related with new categories.

2- Enhanced NLP libraries for Turkish Language: Success of all of the work in this thesis is heavily based on the correctness of preprocessing phase which is done thanks to Zemberek library. Morphology and linguistics of Turkish Language lead data scientists to use similar libraries. This type researches are also going to be improved, parallel to enhancement of these specific languages.

APPENDIX A

Top 25 frequent complaint classes are composing more than 83% of the all complaints.

Figure 11: Distribution of complaints

APPENDIX B

Percentage of the share of the complaints are varying within the day.

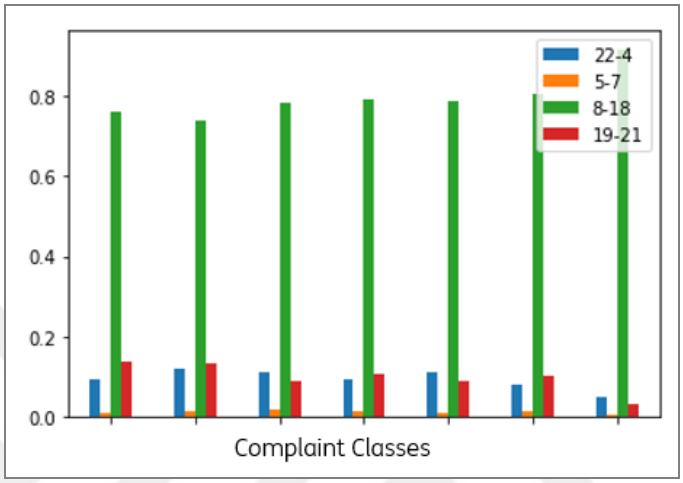


Figure 12: Distribution of the complaints over time by sample classes

APPENDIX C

Length of the complaints are significantly differentiating. Besides as the length of the complaint increase, overall accuracy of the pipeline decreases. There several reasons of this decrease. The main one is, in several complaints customers are trying to tell more than 1 issue in the same complaint.

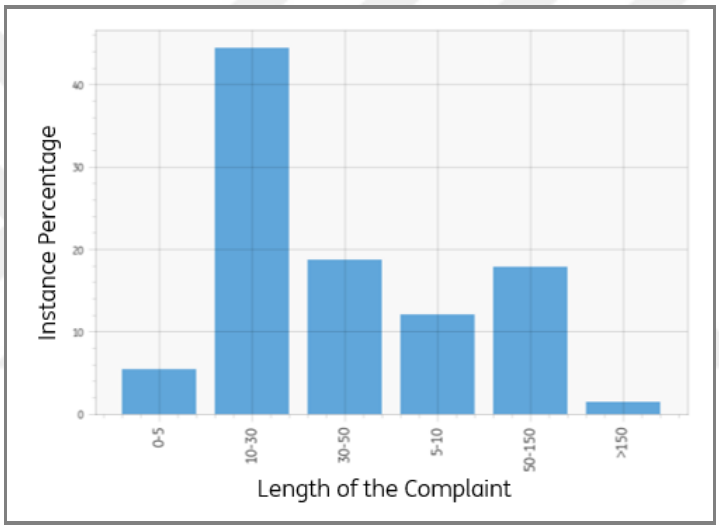


Figure 13: Length distribution of the complaints

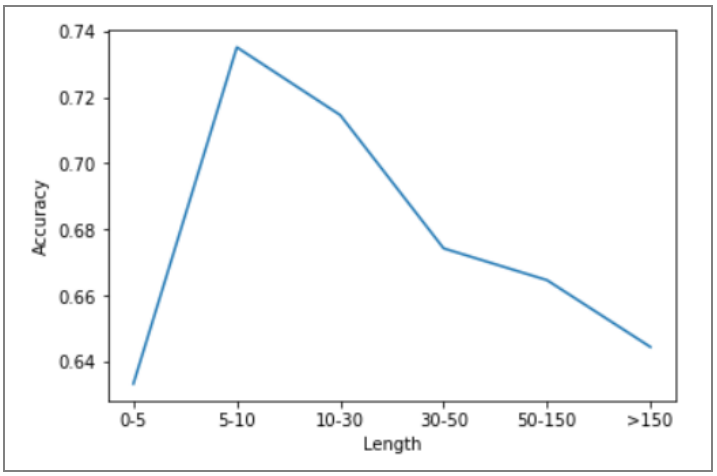


Figure 14: Accuracy by length

APPENDIX D

As expected, predicted probability of the top class has a significant effect over prediction being correct. Thus instead of selecting the class with top probability, several thresholds might be applied in decision process.

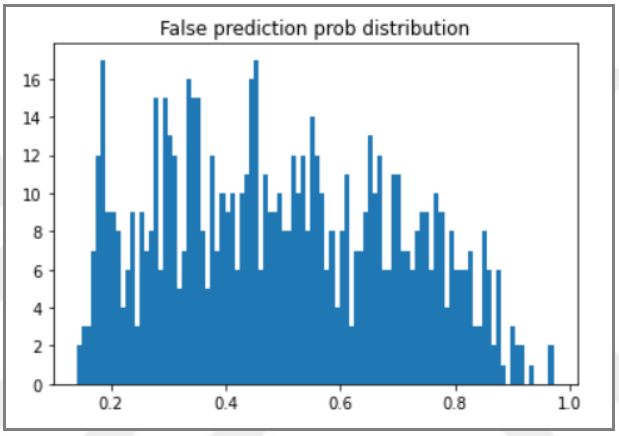


Figure 15: Probability distribution of false predictions

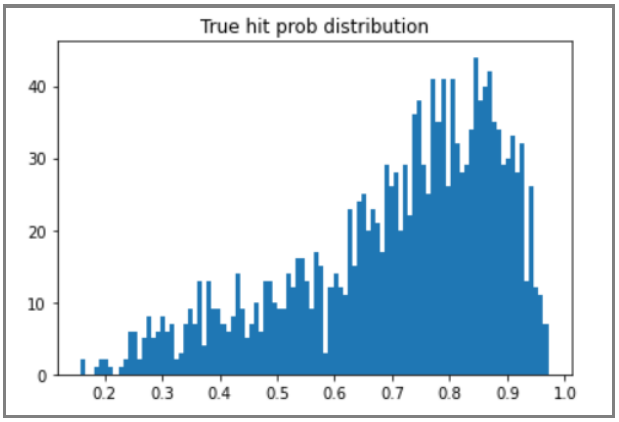


Figure 16: Probability distribution of true predictions

APPENDIX E

Even in the confusion matrix which is observed at the earlier stages of the project, it can be clearly observed that some classes have close context relations. Unfortunately this creates a confusion in predictions.

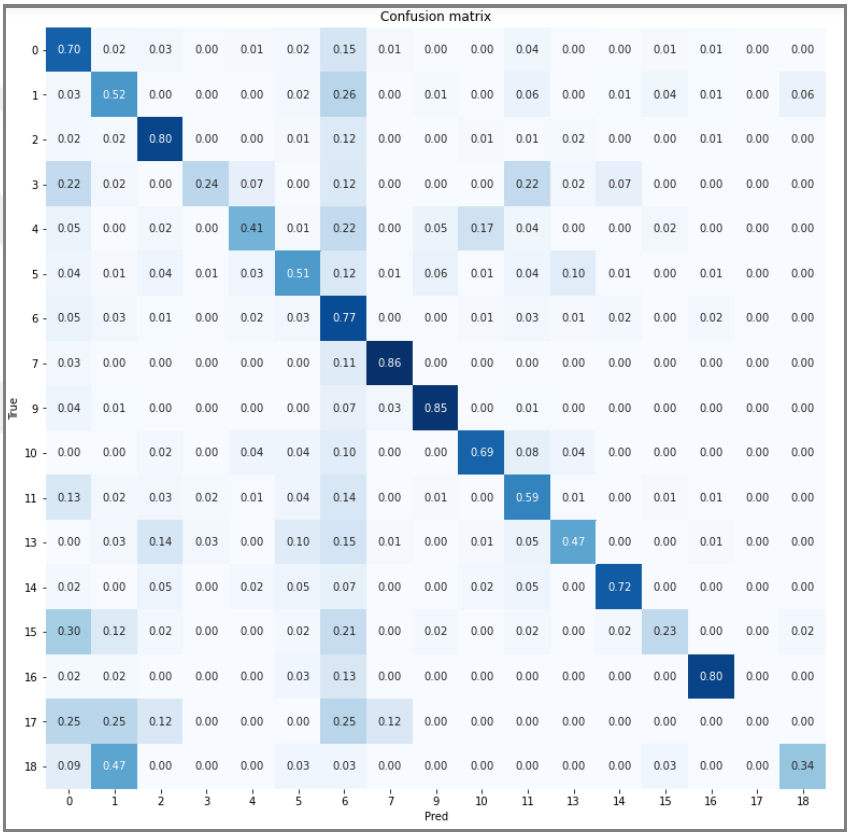


Figure 17: Illustrative confusion matrix of level 1 predictions

BIBLIOGRAPHY

[1] P. Gamollo and M. Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets," in SemEval, Dublin, Ireland, 2014.

[2] T. Pranckevicius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," Baltic Journal of Modern Computing, pp. 221-230, 2017.

[3] V. Korde and N. Mahender, "Text Classification and Classifiers:A Survey," International Journal of Artificial Intelligence & Applications, pp. 85-99, 2012.

[4] S. Yıldırım and T. Yildiz, "A comparative analysis of text classification for Turkish language," Pamukkale University Journal of Engineering Sciences, pp. 879-886, 2018.

[5] X. Zhang, J. Zhao and Y. LeCun, "Character-level Convolutional Networks for Text Classification," NIPS, 2015.

[6] C. Tapsai, P. Meesad and C. Haruechaiyasak, TLS-ART: Thai Language Segmentation by Automatic Ranking Trie, Cala Millor, Spain: AutoSys 2016, 2016.

[7] M. A. Fauzi, Telfor Journal, pp. 123-128, 2018.

[8] F. Bozyiğit, O. Doğan and D. Kılınç, "Categorization of Customer Complaints in Food Industry," Journal of Intelligent Systems: Theory and Applications, pp. 85-91, 2022.

[9] A. A. Akın and M. D. Akın, "Zemberek, an open source NLP framework for Turkic Languages," Structure, pp. 1-5, 2007.

[10] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," Information Processing and Management, pp. 104-112, 2014.

[11] F. Pargent, F. Pfisterer, J. Thomas and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," Computational Statistics, pp. 1-22, 1 4 2021.

[12] D. Küçük and A. Yazıcı, "Named Entity Recognition Experiments on Turkish Texts," in International Conference on Flexible Query Answering Systems, Berlin, 2009.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases," Advances in neural information processing systems, no. 26, 2013.

[14] A. G. Jivani, "A Comparative Study of Stemming Algorithms," Int. J. Comp. Tech. Appl, no. 2, pp. 1930-1938, 2014.

[15] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Proceedings of the first instructional conference on machine learning, vol. 242, pp. 29-48, 2003.

[16] M. L. McHugh, "The Chi-square test of independence," Biochemia medica, vol. 242, no. 1, pp. 143-149, 6 5 2013.

[17] Z.-H. Deng, S.-W. Tang, D.-Q. Yang, M. Z. Li-Yu Li and K.-Q. Xie, "A Comparative Study on Feature Weight in Text Categorization," in Asia-Pacific Web Conference, Berlin, 2004.

[18] T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," R package version, no. 4, pp. 1-4, 11 6 2015.

[19] P. L. Prasanna and D. D. Rao, "Text classification using artificial neural networks," International Journal of Engineering & Technology, no. 7, pp. 603-606, 1 2018.

[20] Online Source: J. Y. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks" https://arxiv.org/pdf/1603.03827.pdf. (accessed March 12, 2016).

VITA

Yücel DOĞAN received his B.Sc. degree in Industrial Engineering from Bogazici University in June 2009. In his professional career, he has worked as management consultant in Peppers&Rodgers, risk consultant in Marsh&McLennan Companies, limit risk manager in Turkcell. Currently he has been working in ING as Head of Data Engineering and Artificial Intelligence. By 2018, he joined Master of Science program in Data Science at Ozyegin University and has been working under supervision of Asst. Prof. Dr. Erinç Albey.