MEF UNIVERSITY

ANALYZING THE DRIVERS OF CUSTOMER SATISFACTION VIA SOCIAL MEDIA

Capstone Project

Kadir Kutlu Yücel

İSTANBUL, 2019

MEF UNIVERSITY

ANALYZING THE DRIVERS OF CUSTOMER SATISFACTION VIA SOCIAL MEDIA

Capstone Project

Kadir Kutlu Yücel

Advisor: Asst. Prof. Dr. Utku Koç

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Analysing the Drivers of Customer Satisfaction via Social Media

Name/Last Name of the Student: Kadir Kutlu Yücel

Date of Thesis Défense:

I hereby state that the graduation project prepared by Kadir Kutlu Yücel has been completed under my supervision. I accept this work as a "Graduation Project".

Asst. Prof. Utku Koç

I hereby state that I have examined this graduation project by Kadir Kutlu Yücel which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

Director
Of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Utku Koç	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name Date Signature

EXECUTIVE SUMMARY

ANALYZING THE DRIVERS OF CUSTOMER SATISFACTION VIA SOCIAL MEDIA

Kadir Kutlu Yücel

Advisor: Asst. Prof. Utku Koç

SEPTEMBER 2019, 29 pages

Social media became a great influence force during the last decade. Active social media user population increased with the new generations. Thus, data started to accumulate in tremendous amounts. Data accumulated through social media offers an opportunity to reach valuable insights and support business decisions.

The aim of this project is to understand the drivers of customer satisfaction by public sentiments on Twitter towards a financial institution. Data was extracted from the most popular microblogging platform Twitter and sentiment analysis was performed. The unstructured data was classified by their sentiments with a lexicon-based model and a machine learning based model. The outcome of this study showed machine learning based model successfully overcame the language specific problems and was able to make better predictions where lexicon-based model struggled.

Further analysis was performed on the extreme daily average sentiment scores to match these days with prominent events. The results showed that the public sentiment on Twitter is driven by three main themes; complaints related to services, advertisement campaigns, and influencers' impact.

Key Words: Sentiment Analysis, Text Classification, Turkish Twitter Analysis, Machine Learning, Prediction

ÖZET

ANALYZING THE DRIVERS OF CUSTOMER SATISFACTION VIA SOCIAL MEDIA

Kadir Kutlu Yücel

Tez Danışmanı: Asst. Prof. Utku Koç

EYLÜL 2019, 29 sayfa

Sosyal medyanın etki alanı geçtiğimiz yıllarla birlikte giderek artmıştır. Yeni jenerasyonlarla birlikte aktif olarak sosyal medya kullanan nüfus artış göstermiştir. Bu sebeple büyük veri birikimi artmıştır. Sosyal medya üzerinden oluşan büyük veri şirketlerin iş yapış şekillerine yönelik değerli kavrayış ve karar alma mekanizmalarına destek fırsatları sunmaktadır.

Bu çalışmanın amacı bir finansal kurumun müşterilerinin memnuniyet seviyelerini sosyal medyada oluşan algıyı kullanarak anlamaya çalışmaktır. Çalışma kapsamında kullanılan veri popüler mikro-blog sitesi Twitter üzerinden derlenmiştir. Yapılandırılmamış bu veri sözlük tabanlı ve makine öğrenmesi tabanlı iki model kullanılarak analiz edilmiştir. Çalışma sonucu makine öğrenmesi tabanlı modelin sözlük tabanlı modelin karşılaştığı Türkçe kaynaklı sorunlardan daha az etkilendiği ve daha başarılı tahminler üretebildiğini göstermiştir.

Analizin sonraki aşamasında ortalama sonucu aşırı uçlarda çıkan günler aynı günlerde ortaya çıkan olaylar ile eşleştirilmiştir. Ortaya çıkan sonuçlara göre müşteri memnuniyeti sosyal medyada ortaya çıkan üç temel faktörden etkilenmektedir. Bunlar, şikâyet yönetimi, kampanya yönetimi ve sosyal medya fenomenlerinin etkisi olarak tanımlanmaktadır.

Anahtar Kelimeler: Duygu analizi, Metin Sınıflaması, Türkçe Twitter Analizi, Makine Öğrenmesi, Tahminleme

TABLE OF CONTENTS

.

1. INTRODUCTION

One of the most popular development that has been occurred in the area of technology in last two decades is the growth of social media. After worldwide usage of internet established, people figured the ways of communication and sharing their emotions or ideas with social media. The sheer size of the people using social media and the information that has been accumulated, creates a unique opportunity for the companies to acquire insights with new angles that suits the needs of new customers.

There are different definitions of social media available, according to most established and agreed definition, social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content [1].

The applications mentioned on this definition reached to large numbers. There are more than 65 active social media platforms as of 2019 and this does not include special networks such as connecting people with certain school graduation or work alumni. Some of the most dominant social media platforms and their trademark specifications are as follows:

- Facebook: Facebook is one of the most active platforms started the boom of social media expansion. It allows members to connect and keep in touch with friends, family and other members by sharing status, photos and videos.
- Instagram: Instagram is a photo and video sharing network that created a total new visual advertisement environment while enabling people to share their memories or experiences in a visual way.
- YouTube: YouTube is the largest video-sharing social networking site in the world. It enables users to upload, share, view videos and add comments about them.
- Stack Overflow: Stack Overflow is a network that connects people with programming experience or people try to improve their skills on coding.
- WhatsApp: WhatsApp is the most popular instant messaging application that disrupted the telecom companies' SMS services by making it free to send and receive messages through mobile phones.

- Twitter: Twitter is a microblogging site that allows people to share short messages called tweets. Users can also read tweets posted by other users.

The main innovation implemented by the social media is the multi way communication that enables people or entities to interact with each other whereas traditional media such as TV or Newspaper deliver a message but unable to collect feedback directly.

This project paper focused on microblogging platform Twitter. The tweets are classified into specific emotions that indicates positive or negative perception. Each tweet contains messages up to the limit of 140 (recently updated 280) characters. Twitter is one of the most popular social media tools that enables people interact with each other or companies by writing their feelings on any topic. According to website's own statistics Twitter had 330 million active users as of first quarter of 2019. These users type 500 million tweets on average each day to express their feelings. Since most of the content is publicly viewable by others, it is also important for the companies to handle this media by interacting back with the customers. This paper focused on Yapı Kredi, one of the four largest private banks of Turkey, and the relationship between customer satisfaction and social media perception by utilizing a sentiment analysis.

1.1. Sentiment Analysis

Sentiment analysis is the process of matching and categorizing of words with their sentimental probabilities of being positive or negative in order to discover any patterns within the texts. These patterns can provide information regarding the composers' feelings or attitudes towards the subject of the text.

Although various approaches are available, in this study a basic polarity exploration through word by word matching and a supervised machine learning based classification technique are performed.

1.2. Theoretical Background

Sentiment analysis on text messages can be described as a classification problem. This classification problem is addressed either with lexicon-based method (unsupervised approach) and machine learning based method (supervised approach) [2]. In this study, both methods were performed on the same dataset in order to have comparable results.

Rapid growth of the social media created different platforms for people to express their feelings on any topic. The abundance of information naturally attracted attention of both academic and commercial researchers. One of the main reasons of this attraction is the predictive power of sentiment analysis. Predictive models have been utilized in various areas such as hotels rating predictions based on customer reviews [3] and depression levels of users [4]. Social media feeds can also be used for predictive purposes because people tend to decide on their actions based on existing social media perceptions of the topic and this creates a collective wisdom that anyone can contribute and benefit at the same time [5]. Many users, upon reading an article or buying a product, feel the need to share their opinion online about this [6]. This collective wisdom created by people who have first-hand experience about the subject has an increasing influence on the sales is important for the companies. Thus, understanding and managing this area is a profit generating event. Companies invest significant amounts of money and time to sustain customer satisfaction on social media platforms as well as traditional activities. Predictive modelling is a significant tool to analyse and follow which type of events drives positive emotions and which areas needs to be improved for a company's success.

Sentiment analysis can be categorized into two: First one is based on a polarity lexicon and the second is machine learning based techniques. Lexicon-based techniques depends on pre-compiled sources containing words and word groups with sentiment probability scores. There are studies that calculate the polarity of product reviews by identifying the polarity of the adjectives within text messages [7]. Machine learning based techniques do not depend on any pre-defined lexicon. Instead, they try to solve the problem by deploying classification algorithms as an attempt to construct computational models of the separation boundary between the positive and negative sentiment. Pak and Paroubek [8] performed a classification study on random tweets by deploying a binary classifier with n-grams and POS features structure, which will be defined in section 4.2. Their model was trained on instances that had been annotated according to the existence of positive and negative emotions.

2. ABOUT THE DATA

In this project tweets that mentioned @YapiKredi between 2017-2018 is studied. A word count frequency matched with two basic emotional states which are positive and negative. Additionally, a machine learning based classification model is deployed for the same dataset.

2.1. Twitter Data

There are different ways of collecting tweets from Twitter database including free and paid services. The main difference between the public APIs and paid services is that only paid services can provide historical data feed whereas public APIs generally accumulate the data with desired criteria during on active period. In order to observe evaluation through time and impacts of different events, a historical approach that includes two years period have been chosen.

Twitter dataset for this project contains tweets with @YapiKredi mentions composed between 01.01.2017 and 10.04.2019. The dataset included 49,790 tweets and was kindly provided by the company itself via the commercial API.

2.2. Pre-Processing

Tweets included up to 140 characters without any limitations of format or alphabets. Twitter also enables users to use different facilities such as link/image sharing or retweet which means users can re-share an already existing message to show support or solidarity with the original message.

In order to unify each tweet format and clear content that holds no emotions, a pre-process pipe is utilized. The aim of this cleaning process is to eliminate the content with no useful information for either of approaches. Table 1 summarizes the content eliminated during this process. The relevant R code for this step can be found in appendix.

Table 1 Unwanted Content

Content	Action
URL: Web links within text message	Removed
Mentions: User references starts with "@"	Removed
Hashtags: Any word that starts with "#"	Removed
Digits: Numerical information	Removed
Punctuations	Removed
Uppercase Characters	Converted
White space	Removed
Non-Latin alphabet words	Removed
Words with 3 repetitive characters	Removed
Words with 2 or less letters	Removed

The second part of the pre-process include the final steps before vectorizing words in each tweet in order to create features of the dataset. Another important step is to remove the stop words. Stop words are defined as words that help building ideas but do not carry any significance themselves [9] such as “ve” (and) or “acaba” (I wonder if…) in Turkish language. These stop words are eliminated by running an R code using tidytext and SnowballC libraries. SnowballC library also provides a function to stem the remaining words. In order to match the words with their emotional states, stemming is important since the sentiment library that has been used for this project does not include words with their attachments. Table 2 includes a sample of original and processed versions of the same tweets.

Table 2 Original and Processed Tweet Sample

Original Tweet	Processed Tweet
@YapiKredi hayatımda olumlu sonuçlar almak da varmış. Yıllık üyelik ücreti iade edildi.	hayat olumlu sonuç al var yıl üye ücret iade edildi
Dikkat @YapiKredi @TwitterSafety https://t.co/rAAyVLDvON	dikkat
Bu sahtekarlıktan haberiniz oldu mu? @YapiKrediHizmet @YapiKredi https://t.co/T0S9obugYV	sahte haber oldu

3. PROJECT DEFINITION

3.1. Problem Statement

Yapı Kredi, as one of the largest lending banks in Turkey, is using social media to interact with its customers. This project focuses on extracting predictive information to improve performance of the company. A predictive approach can provide valuable information to shape company's strategy to improve its financial success.

Being a financial institution attracts more negative comments compare to other companies. After recent financial crisis, public sentiment towards financial institutions polarized further to negative side of the scale. In response financial institutions tried restoring this sentiment through socially positive events [10]. However, as of today, emotion pattern is still on far negative side of the emotion scale.

This paper aimed to provide a better way to improve social media sentiment of financial institutions by asking what needs to be improved or which events create more negative / positive sentiment towards the company image.

3.2. Project Objectives

There are two main objectives for this project. The first objective is to create a classification model that separates positive and negative emotions towards the subject company. The second objective is to understand if any significant events occurred during this time of period effecting the brand image of the subject company. By achieving these objectives, it is targeted to answer the question of what type of inputs required to create positive sentiments towards the companies.

3.2. Project Scope

This project includes 49,790 tweets that mention @YapiKredi composed between 01.01.2017 and 10.04.2019.

Due to nature of the dataset, this project completely excludes other factors besides social media sentiment such as economical concerns or other advertisement efforts performed by companies. However, it is still possible to objectively measure brands image within social media which companies are heavily investing in recent years.

4. METHODOLGY

As mentioned in the previous chapters, in this project each of the two main classification approaches, lexicon-based (unsupervised) and machine learning based (supervised) techniques have been used.

4.1. Lexicon Based Approach

The main advantage of lexicon-based approach is that there is no requirement for labels in a dataset. This means, collected tweets can be used directly without any training after cleaning steps in order to map the emotional equivalents of the words [11]. Emotional equivalents are available in terms sentimental probability. By replacing each word with their sentimental probability, it is possible to estimate each tweet overall possibility of being positive or negative.

4.1.1. Bag of Words

The bag of words is a reduced and simplified representation of a text, based on specific criteria such as word frequency. Commonly used units for text learning projects are called n-grams. n-gram can include one to n-numbers of words depending on the analysis.

Turkish is a complex language such that chain of words may indicate a very different emotion compare to emotions associated to individual words. However, analysis on chain of words require both linguistic and psychological know-how. Thus, in this project unit of analysis defined as 1-gram which means every single word have individual emotional state of their own.

The bag of words approach creates a vocabulary specific to dataset. Each text in this vocabulary represented by their frequencies. The below examples show how texts are represented in bag of words:

Text 1: Computer in the classroom finished running the model.

Text 2: The girl in the classroom should be a model.

After eliminating the stop words, bag of words for a dataset that includes just above texts is {Computer, classroom, finished, running, model, girl}. Feature vector would be as follows:

Text 1: {1,1,1,1,1,0}

Text 2: {0,1,0,0,1,1}

Tidytext and SnowballC libraries of R used for both vectorizing and eliminating the stop words.

4.1.2. Word to Emotion Mapping

Although there are several emotion mapping libraries available, resources for Turkish language are limited. SentiTurkNet, an open source library which includes 15,000 words or word groups in Turkish each of which has three polarity score, used to determine polarity scores of each tweet [12].

Table 3 includes the examples that are extracted from SentiTurkNet library with their polarity scores. In order to create a scale between -1 (absolute negative) and 1 (absolute positive) all probabilities net-off as "Sentiment Score". Since objective probability cannot be linked to any specific emotional state, objective probability is disregarded for this project.

Table 3 Sentiment Score Sample from SentiTurkNet

Synonyms	Negative Probability	Objective Probability	Positive Probability	Sentiment Score
Iştah	0.06	0.872	0.068	0.008
Iştahlı	0.06	0.462	0.478	0.418
Iştahsız	0.48	0.452	0.068	-0.412

Another mapping activity was performed on "emoji" characters separately. Emojis are small digital images or icons used to express an idea or emotion. These special characters are represented with their Unicode expression in the text dataset. Another R code utilized to extract and create another bag of words from emojis. Since emojis are visual icons their sentiments are free from any language barrier. Table 4 represents examples of emojis and their sentiment equivalents [13].

Table 4 Sentiment Score Sample from Emojis

Icon	Unicode Codepoint	Negative Probability	Objective Probability	Positive Probability	Sentiment Score
	0x1f602	0.247	0.285	0.468	0.221
	0x2764	0.044	0.166	0.79	0.746
	0x1f620	0.564	0.172	0.265	-0.299

4.2. Machine Learning Based Approach

Although machine learning based techniques require labels, there is an important advantage of this approach as compared to lexicon-based approach. The lexicon-based approach usually struggles in detecting sarcasm especially on Turkish language. Generally, tweets tend to have high level (even complicated levels) of sarcasm that is almost impossible to detect with 1-gram lexicon-based approach. As an example, here is a tweet from actual dataset which presents sarcasm:

"RT @HakanYilmaz: @YapiKredi kızıl toprak şubesi yavaşlıkta zirvede. Tebrikler." (Yapı Kredi Kızıl Toprak branch is on top at being slow. Congratulations.)

This message clearly complains about how slow the services provided by a certain branch of the bank, but words like "zirve" (top) and "tebrikler" (congratulations) have high positive probability. Thus, lexicon-based approaches would tag this tweet as a positive message. Different machine learning algorithms explored for this paper in order to detect sarcasm as well.

4.2.1. Labels

Supervised learning algorithms requires labels for classification. 1,000 randomly selected tweets labelled by 10 individuals in two subsets. Each subset includes 5 people and 500 tweets. Labels are defined as -1 being negative, 0 being neutral and 1 being positive in three likert scale.

Google Forms grid question structure is used with Survey Monkey's online survey tools in order to collect information. All individuals are between the ages of 25 and 35 and all are active social media users.

Since the aim of the project is to distinguish negative sentiment from the positive, neutral labels are combined with positive ones in order to have labels in binary form.

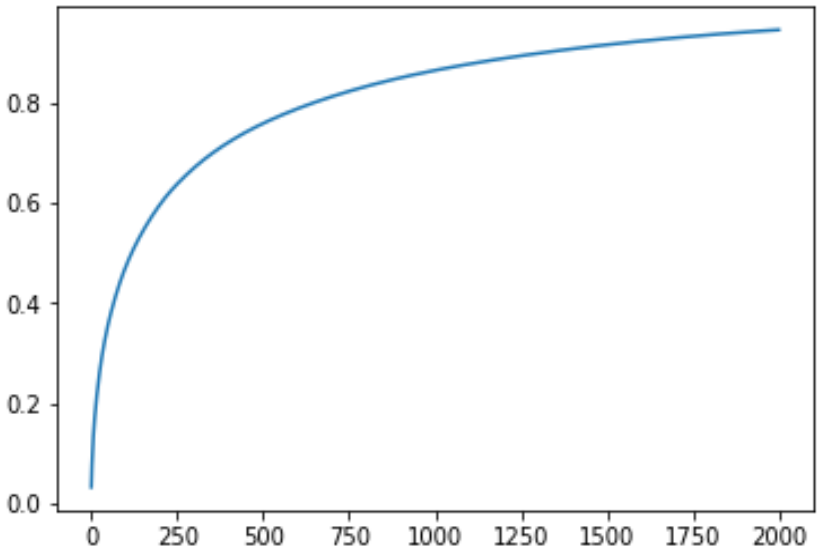
4.2.3. Dimensionality Reduction

Although less frequent words were eliminated from the bag of words in pre-process step there were still 4,657 features in our dataset. Reducing the number of features would improve the run time for model training. Another benefit of dimensionality reduction would be decreasing the chance of over-fitting.

Principal Component Analysis (PCA) is a statistical way that converts correlated variables into a set of linearly uncorrelated variables. PCA works by finding linear combinations, a1′x, a2′x, …, aq′x, called principal components, that successively have maximum variance for the data, subject to being uncorrelated with previous {a}k′{x}s. [14]

Exploratory analysis showed 2,000 features out of 4,657 features could explain 95% of the variance. Thus, feature set transformed with PCA to 2,000 features. Figure 1 shows the explained variance accumulation with each additional feature.

Figure 1 Explained Variance Ratio per Feature



4.2.4. Dealing with Imbalanced Data

One of the biggest challenges of this dataset is the imbalanced structure. As mentioned earlier, due to general public perception towards the financial institutions, most

of the observations are in the negative side of the scale. In our labelled dataset, 65% of total observations are negative. Thus, both under and over sampling techniques are explored in order to create a balanced dataset

Imblearn library of Python provides different options to balance the dataset. Nearmiss and Smote algorithms were chosen for their simple implementation process and providing a straightforward solution to balance the data classes. In order to be prudent, 20% of the labelled data saved as test subset and both methods were applied to only 80% of the labelled data.

Nearmiss adds some heuristic rules to select samples and implemented in three different options [15]. Option 2 had been applied since it selects the positive samples for which the average distance to the N farthest samples of the negative class is the smallest. After application of the algorithm the train subset has 50:50 distribution of each class having 293 observations.

Smote (Synthetic Minority Over-sampling Technique) increases the under sampled class not by replacement but instead creating synthetic examples by joining any/all the k-minority class nearest neighbours [16]. Smote algorithm also provided a 50:50 distribution on each class but this time each class have 493 observations.

4.2.5. Classification Algorithms

All algorithms were imported from Python Scikit Learn library version 0.21.3 [17]. All project was carried out on a personal computer with single Intel Core i5-6200U 2.3GHZ processor with 2 cores (4 threads), 16 GB RAM, running Windows 10 64-bit operating system.

The most successful and popular classification algorithms on text classification were explored as Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Gradient Boosting Machines (GBM) [18], [8].

- **SVM** creates several hyperplanes depending on the number of classes to be classified in a high dimensional space [17]. Successful classification on SVM can be defined with a functional margin. Functional margin is the distance between hyperplane and the nearest point to it. Classification would be clearer as functional margin increases. SVM operates with different kernels such as linear, gaussian

radial or polynomial. Although all kernel types have been explored for this study, the best results achieved with a gaussian radial kernel (RBF). Mathematical function behind RBF kernel is as follows where x represented as feature vectors and $||x- x'||^2$ represents squared Euclidian distance between two feature vectors.

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = \exp(-\gamma \|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2) \text{ for } \gamma > 0$$

- **DT** is a classification algorithm resembles a tree with its branches. Each internal node represents tests on features, i.e. if "thank you" feature is 1 decision should be 1 as positive. Each leaf nodes represents a class label depending on the result of the test. Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items [19]. Decision tree algorithm studied in this project defines the best with Gini impurity. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset. It can be formulated as below where p_i is the probability of an item with label i misclassified within J classes.

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

- **RF** is an ensemble machine learning method. It fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. RF model in this study also uses a bootstrapping technique where the sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement.
- **GBM** is also an ensemble machine learning method. GBM combines week learners usually in the form of decision trees to a single strong learner by iterating over each learner. It builds an additive model in a forward stage-wise fashion to allow for the optimization of arbitrary differentiable loss functions.

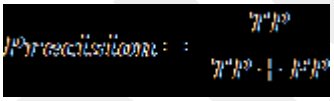
4.2.6. Training Methods & Performance Evaluation

The target of this study is to predict whether the tweet has positive or negative sentiment. That requires a clear classification through the accuracy of the prediction. Accuracy as a performance metrics, would perform well in a balanced dataset. Since our

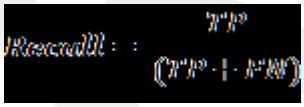
training subsets were balanced by application of Nearmiss and Smote, this metric is chosen to evaluate the performance of the models.

However, test subset was separated before under/over sampling techniques were applied. Thus, it had an imbalanced structure with 128 negative cases and 69 positive cases indicating a 65:35 distribution. This imbalanced distribution means, without any modelling if all cases were labelled as negative in this subset, accuracy would be 65%. As solution three additional metrics were introduced to test subset.

- **Precision** is defined as the ratio of true positives to the sum of true positives and false positives. As precision approximates to its maximum value of 1, model would less likely to misclassify a negative case as positive.



- **Recall** is defined as the ratio of true positives to sum of true positive and false negatives. Recall also have a range between 0 to 1 as precision. Recall increases while the model correctly classifies positive cases.



- **F1 score** is defined as weighted average of the precision and recall. Since our target is to achieve better classification without interruption of imbalance F1 score would be useful since it recognizes both precision and recall.

$$\boldsymbol{F1\ Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

5. COMPUTATIONAL RESULTS

The performance of each model is evaluated with the performance metrics as defined in the previous chapter. In order to reduce the chance of overfitting and to have a better parameter tuning, grid search with 10-fold cross validation is applied to all model deployments. The best parameters that yield highest mean accuracy of each 10-iteration is selected. 99% confidence interval is also calculated to observe the margin of error in order to select the most consistent model.

Table 5 includes all performance metrics for each model deployment. The highest scores for each metric are presented with bold characters.

Table 5 Performance Metrics

Performance Metrics		TRAIN				TEST			
		10-Fold Mean Accuracy	Margin of Error (99% C.I)	C.I Upper Limit	C.I Lower Limit	Accuracy	Precision	Recall	F1 Score
Over Sampled	SVM	**86%**	10%	**96%**	75%	**78%**	**84%**	45%	58%
	DT	68%	6%	74%	62%	69%	55%	**67%**	**60%**
	RF	84%	11%	94%	73%	69%	55%	43%	**60%**
	GBM	85%	8%	93%	**76%**	76%	72%	49%	59%
Under Sampled	SVM	69%	5%	74%	65%	72%	59%	61%	60%
	DT	57%	8%	65%	49%	65%	51%	32%	39%
	RF	67%	7%	73%	60%	68%	54%	54%	54%
	GBM	66%	**4%**	70%	62%	69%	54%	65%	59%

Table 6 shows the parameters for each model which produced the results of Table 5.

Table 6 Best Parameters

Parameter Tuning		Selected Parameters
Over Sampled	SVM	Kernel Type: RBF, C: 2, Gamma: 0.5
	DT	Max Depth: 16, Min Sample Split: 3
	RF	Max Depth: 18, Min Sample Split: 5, Min Sample Leaf:1, Estimators:200
	GBM	Max Depth:4, Min Sample Split:2, Min Sample Leaf:1, Estimators:1500, Learning Rate:0.1
Under Sampled	SVM	Kernel Type: Linear, C: 0.5, Gamma: 0.5
	DT	Max Depth: 5, Min Sample Split: 6
	RF	Max Depth: 9, Min Sample Split: 19, Min Sample Leaf:1, Estimators:200
	GBM	Max Depth:5, Min Sample Split:4, Min Sample Leaf:1, Estimators:750, Learning Rate:0.1

Performance metrics show models trained with over sampled train subset yielded better results but on the other hand models trained with under sampled train set have lower margin of error. Also, decision tree models have the lowest overall scores and SVM and GBM models trained with over sampled data resulted with overall best scores.

SVM and GBM comparison on train subset shows SVM have higher upper band in terms of accuracy and GBM have higher lower band. SVM's better performance on test subset might suggest SVM would yield better results on unlabelled original dataset which is expected to be imbalanced as test subset. Considering simplistic approach and much lower processing power requirements compare to GBM predictive model for this paper defined as SVM trained with over sampled dataset.

5.1 Average Daily Sentiment Score Predictions

As targeted at the beginning of the project, two different predictions had been made with lexicon and machine learning based models. Both approaches are presented with similar trends in macro terms. The mean score for lexicon-based approach is 0.20 and 0.18 for machine learning based approach. Standard deviations are 0.13 for both approaches. Although mean scores are very close to each other, predictions based on machine learning model seems slightly more negative.

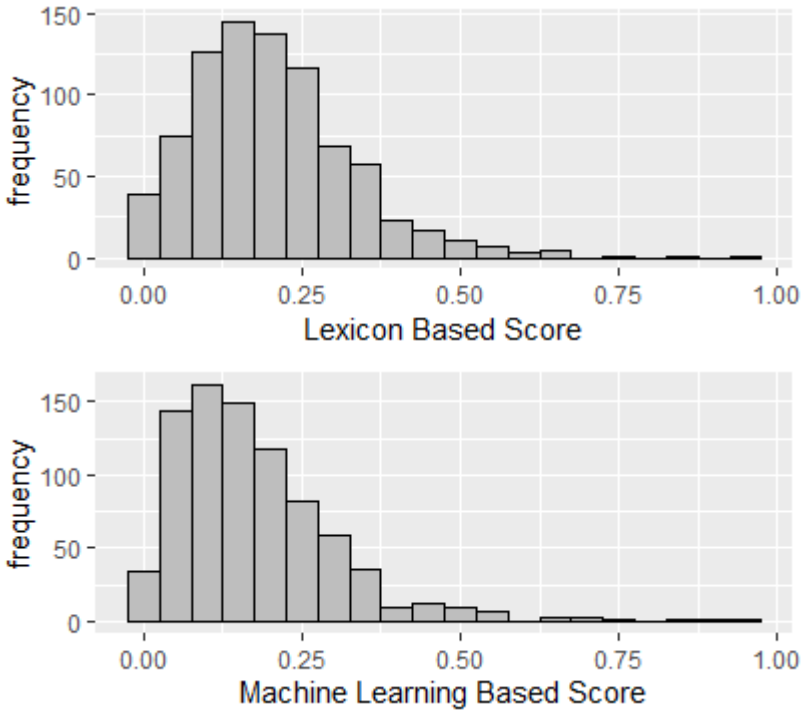
Table 7 includes summary statistics for sentiment scores based on two approaches and tweet counts.

Table 7 Summary Statistics for Daily Average Sentiment Scores

Summary Statistics	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Lexicon Based	0	0.116	0.1883	0.2033	0.2628	0.9479
M.L Based	0	0.08597	0.15346	0.17752	0.2325	0.95455
Number of Tweets	2	23	34	54.82	54	1,431

Figure 2 shows the distribution of average daily sentiment scores based on two approaches.

Figure 2 Histograms of Daily Average Sentiment Scores

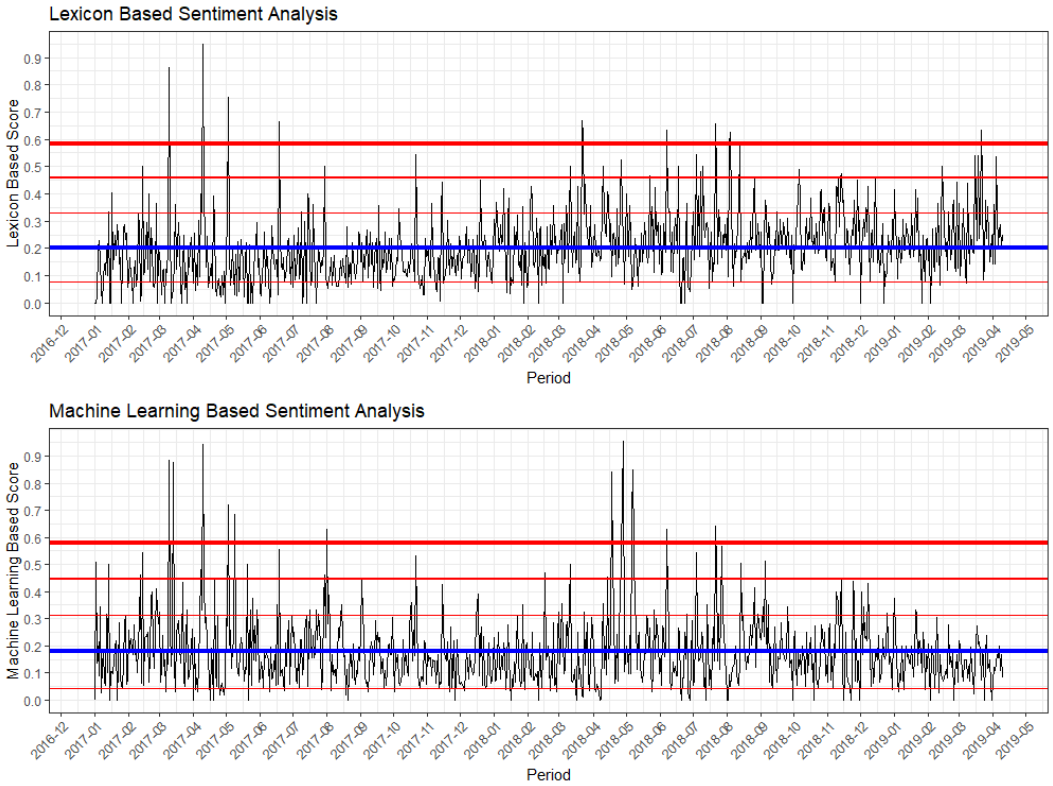


As above histograms show, predictions according to both models are highly right skewed to the negative side of the scale. However, skewness on machine learning based prediction is higher with 1.96 compare to 1.18 lexicon-based skewness.

As stated in previous chapters, target of this project was to answer the question of what can be done to improve customer satisfaction. Thus, extreme points need to be further analysed. Days where average score is below or above mean score by three standard deviation were defined as extreme points for this study.

Figure 3 presents daily average sentiment scores, predicted with both lexicon and machine learning based approaches where blue line indicates the mean and red lines are standard deviation distance to mean. The thickest red line is the distinctive line of extreme where cases are farther than three standard deviations of the mean. There are 10 extreme days according to lexicon-based predictions and 13 according to machine learning based predictions. 5 days were commonly predicted as extremes by both models. That means there were 13 extreme days that was predicted by only one of the approaches which might be indicating misclassification. Major deviations between two models would provide valuable information in order to improve either model.

Figure 3 Average Daily Sentiment Scores



There are also days with high activity where tweet counts reach extreme points. Figure 4 shows there are 17 extreme days where at least 314 tweets with @YapiKredi mention. These tweets might be reaction to an event or participation to a competition where company does sometimes in order to boost public interest to a new product or facility.

Figure 4 Number of Tweets per Day

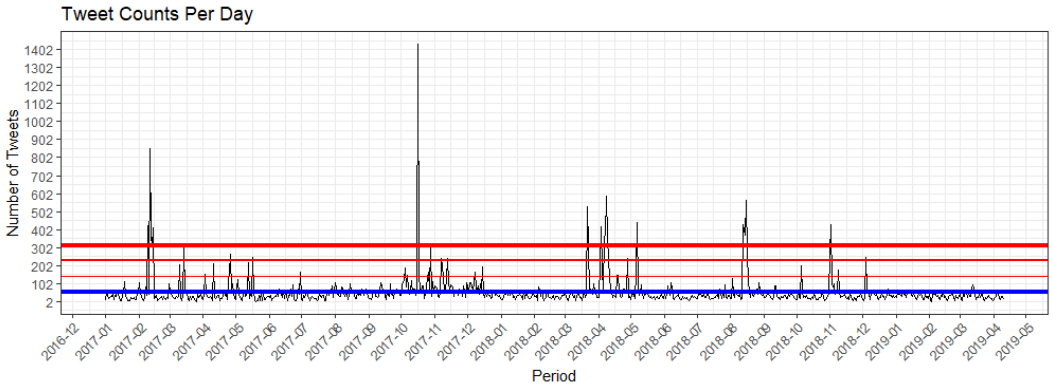


Table 8 shows all the extreme days in detail. There were total 32 days which has at least one extreme value amongst three indicators. In order to understand the drivers of

positive sentiment a further event-based analysis was carried on. Analysis with two sub segments including extreme days according both approaches and extreme days with high variance between approaches would provide better structure.

Table 8 Extreme Days

Date	Lexicon Based Score	Machine Learning Based Score	Number of Tweets
09/02/2017	0.106	0.101	**426**
11/02/2017	0.286	0.302	**849**
12/02/2017	0.005	0.462	**364**
14/02/2017	0.500	0.544	**410**
10/03/2017	**0.865**	**0.884**	207
14/03/2017	0.046	**0.877**	302
10/04/2017	**0.948**	**0.943**	211
03/05/2017	**0.754**	**0.721**	122
09/05/2017	0.030	**0.687**	67
18/06/2017	**0.667**	0.556	9
01/08/2017	0.075	**0.632**	106
16/10/2017	0.170	0.335	**1431**
22/03/2018	**0.669**	0.017	**529**
23/03/2018	**0.623**	0.013	**318**
03/04/2018	0.175	0.029	**417**
06/04/2018	0.189	0.028	**359**
07/04/2018	0.155	0.000	**368**
08/04/2018	0.166	0.003	**589**
09/04/2018	0.353	0.221	**357**
18/04/2018	0.097	**0.841**	145
28/04/2018	0.333	**0.754**	240
29/04/2018	0.227	**0.955**	22
06/05/2018	0.049	**0.689**	164
07/05/2018	0.066	**0.848**	**441**
07/06/2018	**0.632**	**0.632**	106
22/07/2018	**0.657**	**0.643**	70
04/08/2018	**0.625**	0.050	40
13/08/2018	0.574	0.139	**432**
14/08/2018	0.081	0.505	**370**
15/08/2018	0.159	0.293	**567**
02/11/2018	0.366	0.255	**432**
21/03/2019	**0.634**	0.073	41

5.2 Extreme Days According to Both Approaches

- 10.03.2017 have a lexicon-based score of 0.865 and machine learning based score of 0.884 with 207 tweets which is above the 54.8 mean tweets. Unfortunately, out of 207 tweets 178 of them belongs to a retweet campaign. Original tweet was not typed directly but shared as a picture in order to bypass the character limit of Twitter format. Due to deleted original tweet, cause of positive sentiment cannot be identified. However, the word “lütfen” which means “please” in English typed in original tweet created a positive sentiment.
- 10.04.2017 have a lexicon-based score of 0.948 and machine learning based score of 0.943 with 211 tweets. One of the banks latest and humorous advertisement campaigns first aired as of these days called “Gary & Metin”. Two famous comedian actors took part in this campaign.
- 03.05.2017 have a lexicon-based score of 0.754 and machine learning based score of 0.721 with 122 tweets. Another part of ad campaign of “Gary & Metin” seemed to have successful returns. But this time there was specific twitter leg of the campaign with collaboration of famous @incicaps account. This account is linked to another social media domain called “inci sözlük” a forum like web site where focus is also humorous.
- 07.06.2018 have 0.632 sentiment score according to both lexicon-based and machine learning based approaches. On this day there were rumours and expectations for long waited cash payment option instead of military obligation. Some Twitter users with high followers tried to create a public opinion by asking banks “Are you ready to provide loans for paid military obligations?” and lots of emojis were used.
- 22.07.2018 have 0.657 lexicon-based score and 0.643 machine learning based score. There was a minor spam attack to the banks promoting a block-chain company who was after getting a contract or improve interest. Positive classification of these tweets should be an area of improvement for both models.

5.3 Extreme Days with High Variance Between Approaches

- 14.03.2017 have 0.877 machine learning based score but lexicon-based score is only 0.046. There was a chain of retweets with the topic focused on gratitude to the company. Apparently, the bank made some adjustment on a branch located in east part

of the country (Siirt) to improve handicap access. Machine learning based model successfully labelled these tweets as positive.

- 09.05.2017 have 0.687 machine learning based score and only 0.030 lexicon-based score. Another ad part of ad campaign of "Gary & Metin" performed with @incicaps user. Positive sentiment successfully labelled by machine learning based model. However, lexicon-based model could not succeed for this occurrence of advertisement campaign.
- 01.08.2017 have 0.632 machine learning based score and 0.075 lexicon-based score. A twitter user with 180k followers wrote a message and inform the bank about a phishing attempt and got a lot of tweet. Although the tweet written in a good manner positivity of these message is subjective.
- 22.03.2018 and 23.03.2018 have lexicon-based score of 0.669 & 0.623 and machine learning based score is only 0.017 & 0.013 respectively. There is again a retweet chain labelled as positive by lexicon-based approach. However, original tweet was about a harsh complaint. This means lexicon-based approach made a misclassification on this case.
- 18.04.2018, 28.04.2018 and 29.04.2018 all have high machine learning based scores. These three dates also a part humorous questions about possible loan opportunities in order to not oblige military service as the same on 07.06.2018. However, this time only machine learning based method labelled those tweets as positive.
- 06.05.2018 and 07.05.2018 have high machine learning based score. Both days users are responding to another ad campaign collaborated with popular TV show "Jet Sosyete". There are many retweets and many humorous responses to the campaign successfully caught by the machine learning based model.
- 04.08.2018 have lexicon-based score of 0.625 and machine learning based score of 0.050. A Twitter user with 5k followers sent a complaint tweet about a transaction and requested support from his followers. Around 40 retweets achieved and due to words linked to support request, lexicon-based model overlooked the words linked to actual complaint.
- 21.03.2019 have 0.634 lexicon-based score. On the other hand, machine learning based score is only 0.073. This is particularly interesting case due to sarcastic nature of original tweet which was retweeted 20 times. A user was complaining about receiving

many calls for marketing activities, but tweet was written in sarcastic way including two positive emojis. Thus, lexicon-based approach classified those tweets as positive where machine learning based model did not.

5.4 Extreme Days According to Tweet Counts

There were 17 days where number of tweets reached extreme points. Three of those days also produced extreme scores and analysed in previous chapters. That means 14 days despite having high number of tweets did not have extreme results. In order to keep the analysis brief as much as possible only 5 highest tweet counts selected.

- 11.02.2017 have 849 tweets. There was online campaign where people suggesting playlist of songs that are suitable for approaching valentine days. Those tweets did not create any significant sentiment score according to both models.
- 16.10.2017 is the day where the famous "Gökay 425" incident happened. Due to performing test on production environment all mobile application users of the bank received a pop-up message containing the text of "Gökay 425". This was an honest mistake that triggered mixed responses amongst twitter users. Some of the responses were sarcastic or humorous and some of them worried about security breaches and was negative. Total number of tweets was 1,431 and this was the most extreme day in terms of tweet counts.
- 08.04.2018 had 589 tweets with relatively normal lexicon-based sentiment score. However, machine learning based sentiment score is 0.003 which is indicating one of the most negative days in project scope. The hashtag "#adioslareziloluyos" was circulated as of this day. People were complaining very roughly about unsuccessful campaign about banks credit card product called "adios". Slang like hashtag of "adioslareziloluyos" is actually a combination of three separate words, "adios" is the name of credit card product, "rezil" and "olmak" means when used together means to be in an infamous situation. Complaints were asking the bank to fulfil its commitment about the campaign. Since this hashtag is also one of the most frequent words in our dataset it was also represented in labelled data that machine learning based models were trained. Thus, machine learning based model successfully caught all the tweets, but lexicon-based model overlooked due to some of the other words offsetting the impact.

- 15.08.2018 had 567 tweets and both models produced sentiment scores within normal margins. This was the day where a large currency shock impacted Turkey's macro-economic environment. A significant portion of the population had the opinion of this situation induced by USA in order to punish Turkey. Previous two days were also extreme days by tweet counts where multiple hashtags were circulating and requesting companies to boycott USA based companies and products.

5.5 Most Frequent Words

Most frequent words could provide hints about what people are talking about. However, the context is important considering in Turkish language a word can have multiple means.

Twitter format itself is a challenge also it increases importance of word frequency in the same time. Twitter users usually prefer a sarcastic expression. This can be linked to people's choice, in general sarcastic tweets collects more interest and retweets or likes. Thus, a common language is built to attract more interest with high sarcasm levels. This situation creates a challenge for lexicon-based model more than the machine learning based one due to tokenization of words. On the other hand, character limit of each tweet, lead people to express themselves in less words where people create hashtags such as "#adioslareziloluyos" which is not actually a word but combination of three words. However, this slogan like hashtag represented as a feature in both lexicon and machine learning based models. Naturally sentiment libraries do not have any sentiment score for such combined words built out of general grammar of the language. Thus, they do not yield any sentiment by lexicon-based models but machine learning based models can easily detect such words if they have been provided with a suitable training dataset.

Figure 5 as a word cloud contains examples of all cases mentioned above. The three most frequent words were "kredi" (loan), "banka" (bank), and "para" (money). All these three words can lead positive or negative sentiments according to the context they have been used in. There are also observable hashtags such as "adioslareziloluyos" – a slang like combined word which is used to complain about a specific product of the bank- or "bedelliaskerlik" -another combined word for payment to avoid obligatory military service-.

There are also words like “merhaba” meaning “hi” in Turkish. Although this is a stop word that lexicon-based model is not evaluating, human intuition suggest it is a nice way to start a conversation and following words might be positive as well. Considering also machine learning based models was trained with a data labelled by actual human beings, we might expect this word is associated with a positive sentiment in some cases.

Figure 5 Most Frequent Words

6. CONCLUSION

The main conclusion of this study is both lexicon-based and machine learning based approaches can predict sentiment changes. However, there is still room for improvement for both approaches.

The main challenge for both models were lack of the resources for Turkish language, such as stop word library, word stemmer and most importantly word by word sentiment scores. Stop words and stemming performed with public R libraries such as SnowballC. Although there is a better performing stemmer available on Turkish language called “Zemberek”, due to lack of knowledge on Java programming language it was not used for this project. According to official document, Zemberek is also provides a typo correction module which is expected to improve model performance due to language dynamics on tweets.

Detailed analysis on extreme scores showed machine learning based model performing better than lexicon-based model. This is mainly linked to lexicon-based model’s exclusion of context. It is also clear that there is much more potential to be discovered on machine learning approach. Better models can be built if larger training sets are available with more processing power.

Conclusion on predictive results can be summarised under three topics such as complaint management, campaign management and influence of popular accounts.

The most striking hashtag of the dataset “adioslareziloluyos” is a great example of complaint management. Twitter users invented a single word by combining three which was also a reference to companies own campaign about the product and used it very effectively to show their dissatisfaction. Users affected by the same instance immediately responded to hashtag by retweeting it. This is a perfect example of collective wisdom that is accumulating with the social media. People with similar interest meet under an invented word and their collective voice sounded stronger than any individual. The lack of positive score on the following days of this hashtag usage suggests the company could not handle this problem very well.

“Gary & Metin” campaign unfortunately ended badly due to actor’s personnel life and negative public sentiment towards him. However, it was collecting very successful feedback from Twitter community. The reason behind that success was linked to the focus

point of the campaign itself. It was a humorous story of AI robot that run out of the factory and started to live together with a regular guy. This campaign checked the approval boxes for humour, tech-savviness, and popular actors amongst young people. On top of these aspect campaign also supported with different popular media such as @incicaps.

Influencers on social media are a new phenomenon. Even regular people with no obvious talent or previous fame can become an influencer. This users with high number of followers can steer the social media sentiment simply tweeting in an interesting way. Two of the extreme point that we explored in previous chapters triggered by such people. Social media provides important power to individuals. As a result of this situation companies need to treat each and every customer as VIP.

REFERENCES

[1] Kaplan, A.M. & M. Haenlein (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 53.

[2] Mohammed ALSADI, Sevinç GÜLSEÇEN, Elif KARTAL (2016). TOP 10 Turkish Universities Twitter analysis user sentiment analysis and comparison with international ones.

[3] López Barbosa, R., Sánchez-Alonso, S. and Sicilia-Urban, M. (2015), "Evaluating hotels rating prediction based on sentiment analysis services", Aslib Journal of Information Management, Vol. 67 No. 4, pp. 392-407.

[4] Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz. Predicting Depression via Social Media.

[5] Sitaram Asur, Bernardo A. Huberman (2010). Predicting the Future with Social Media.

[6] Jansen, B.J.; Zhang, M.; Sobel, K.; Chowdury, (2009). A. Twitter power: Tweets as electronic word of mouth. JASIST 2169–2188.

[7] Moghaddam, S.; Popowich, F. (2010) Opinion polarity identification through adjectives. CoRR, arXiv: 1011.4623.

[8]

Pak, A.; Paroubek, P. (2010) Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, CA, USA; pp. 436–439.

[9] Rajaraman, A., & Ullman, J. (2011). Data Mining. In Mining of Massive Datasets (pp. 1-17). Cambridge: Cambridge University Press.

[10] Edyta Rudawska (2011). Social conscience of financial institutions as a solution for retaining relationships with customers in the face of the global crisis.

[11] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques.

[12] Dehkharghani, R., Saygin, Y., Yanikoglu, B. et al. Lang Resources & Evaluation (2016) 50: 667.

[13] P. Kralj Novak, J. Smailovic, B. Sluban, I. Mozetic (2015). Sentiment of Emojis, PLoS ONE 10(12): e0144296.

[14] Jolliffe I. (2011). Principal Component Analysis. In: Lovric M. (eds) International Encyclopedia of Statistical Science.

[15] I. Mani, I. Zhang. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction, In Proceedings of workshop on learning from imbalanced datasets.

[16] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[17] Pedregosa et al.,(2011). Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.

[18] Boiy, E. & Moens, MF. Inf Retrieval (2009) 12: 526. https://ezproxy.mef.edu.tr:2109/10.1007/s10791-008-9070-z

[19] Rokach, L.; Maimon, O. (2005). "Top-down induction of decision trees classifiers-a survey". IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews. 35 (4): 476–487.

APPENDIX

R code for cleaning process

```
cleaner <- function(data){
  #url
  regex_http <- "http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+"
  data <- gsub(regex_http,"",data)
  #çoklu mention
  data <- gsub("@\\S+\\s*", "",data)
  #hashtag
  data <- gsub("#", "",data)
  #sayilar
  data <- gsub("[[:digit:]]","",data)
  #whitespace
  data <- gsub("      "," ",data)
  data <- gsub("    "," ",data)
  data <- gsub("  "," ",data)
  data <- gsub(" "," ",data)
  #trim
  data <- gsub("(^ +) | ( +$)", "",data)
  #1-2 letters
  data <- gsub(" *\\b[[:alpha:]]{1,2}\\b *", " ", data)
  data <- gsub("^ +| +$|( ) +", "\\1", data)
  #noktalama
  data <- gsub("[[:punct:]]","",data)
  processed_data <<- data
}
```

R Code for Bag of Words

```
word_list <- mydata %>%
  unnest_tokens(word,clean_tweets) %>%
  anti_join(get_stopwords("tr", source = "stopwords-iso")) %>%
  mutate(word = wordStem(word,language = "turkish") %>% unlist()) %>%
  count(word) %>%
  filter(
    !str_detect(word, pattern = "[[:digit:]]"),     # removes any words with numeric digits
    !str_detect(word, pattern = "[[:punct:]]"),     # removes any remaining punctuations
    !str_detect(word, pattern = "[^\x20-\x7E]"),    # removes non latin
    !str_detect(word, pattern = "(.)\\1{2,}"),      # removes any words with 3 or more repeated letters
    !str_detect(word, pattern = "\\b(.)\\b"),       # removes any remaining single letter words
    !str_detect(word, pattern = " *\\b[[:alpha:]]{1,2}\\b *")) %>%   #remove two letter words
  filter(n >= 5) %>% # filter for words used 5 or more times
  pull(word)


bow_features <- mydata %>%
  unnest_tokens(word,clean_tweets) %>%
  anti_join(get_stopwords("tr", source = "stopwords-iso")) %>%
  mutate(word = wordStem(word,language = "turkish") %>% unlist()) %>%
  filter(word %in% word_list) %>%     # filter for only words in the wordlist
  count(Id, word) %>%                 # count word useage by  ID
  spread(word, n) %>%                 # convert to wide format
  map_df(replace_na, 0)
```

R Code for Emoji Extraction

```
emoji <- read.csv("emoji.csv", stringsAsFactors = TRUE, sep = ";")
emoji_regex <- sprintf("(%s)", paste0(emoji$Bytes..UTF.8., collapse="|"))
compiled <- ore(emoji_regex)

found_emoji <- ore.search(compiled, mydata$Ileti, all=TRUE)
emoji_matches <- matches(found_emoji)

#loop
emoji_lines <- which(grepl(emoji_regex, mydata$Ileti, useBytes = TRUE))
mydata$emojis <- rep(0,nrow(mydata))
for (i in emoji_lines){
  mydata$emojis[i] <- list(matches(ore.search(compiled, mydata$Ileti[i], all=TRUE)))
}

emoji_scores <- mydata %>%
  select(Id,emojis) %>%
  filter(emojis != "0") %>%
  unnest(emojis)

emoji_scores$sentiment <- data_frame(
  Bytes..UTF.8. = emoji_scores$emojis %>%
    map(charToRaw) %>%
    map(as.character) %>%
    map(toupper) %>%
    map(~sprintf("\\x%s", .x)) %>%
    map_chr(paste0, collapse="")
) %>%
  left_join(emoji) %>%
  select(Sentiment)
```

Python Code for Machine Learning Part

```
import pandas as pd
import numpy as np
import rpy2.robjects as robjects
from rpy2.robjects import pandas2ri
pandas2ri.activate()
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import NearMiss
from sklearn.decomposition import PCA
from sklearn import svm
from sklearn import tree
from sklearn import ensemble
from sklearn.model_selection import cross_val_score
import seaborn as sns
```

```
##############################PCA################################################
pca = PCA(n_components=2000).fit(features)
exp_var = pca.explained_variance_ratio_.cumsum()
plt.plot(exp_var)
features_pca = pca.transform(features)
pca_data = pd.DataFrame(data=features_pca)
pca_data["Is_Positive"] = label["Is_Positive"]
pca_data["Text"] = full_data["İleti"]
####training_subset############
classification_data_pca = pca_data.dropna()
classification_data_pca = classification_data_pca.drop(["Date","Id","Text"], axis = 1)
X = classification_data_pca.drop(["Is_Positive"], axis = 1)
y = classification_data_pca[["Is_Positive"]]
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2, random_state = 123
y_train.Is_Positive.value_counts()
y_test.Is_Positive.value_counts()
################################Re-Sampling#################################
#smote
sm = SMOTE(random_state=123)
X_train_sm, y_train_sm = sm.fit_sample(X_train,y_train)
unique_elements, counts_elements = np.unique(y_train_sm, return_counts=True)
print(np.asarray((unique_elements, counts_elements)))
#nearmiss
nm = NearMiss(version=2,random_state=123)
X_train_nm, y_train_nm = nm.fit_sample(X_train,y_train)
unique_elements, counts_elements = np.unique(y_train_nm, return_counts=True)
print(np.asarray((unique_elements, counts_elements)))
```

```
grid_param_sv_o={'gamma': [0.5,1,1.5,2,2.5,3], 'C': [0.5,1,1.5,2,2.5,3],
                 "kernel":["rbf","linear","poly"],"degree":[2,3,4]}
classifier_sv=svm.SVC(random_state=123)
gd_sr_sv_o = GridSearchCV(estimator=classifier_sv,
                     param_grid=grid_param_sv_o,
                     scoring='accuracy',
                     n_jobs = -1,
                     cv = 10,
                     verbose = 0)

#fit & pred
gd_sr_sv_o.fit(X_train_sm, y_train_sm)
pred_sv_o = gd_sr_sv_o.predict(X_test)
pred_sv_o_tr = gd_sr_sv_o.predict(X_train_sm)

#stats for best fit model
best_parameters_sv_o = gd_sr_sv_o.best_params_
best_result_sv_o = gd_sr_sv_o.best_score_
con_mat_sv_o=metrics.confusion_matrix(y_test,pred_sv_o)
train_result_sv_o = metrics.accuracy_score(y_train_sm, pred_sv_o_tr)
test_result_sv_o = metrics.accuracy_score(y_test, pred_sv_o)
train_f1_sv_o = metrics.f1_score(y_train_sm, pred_sv_o_tr)
test_f1_sv_o = metrics.f1_score(y_test, pred_sv_o)
train_prec_sv_o = metrics.precision_score(y_train_sm, pred_sv_o_tr)
test_prec_sv_o = metrics.precision_score(y_test, pred_sv_o)
train_rec_sv_o = metrics.recall_score(y_train_sm, pred_sv_o_tr)
test_rec_sv_o = metrics.recall_score(y_test, pred_sv_o)
sv_stats_o = [best_result_sv_o,train_result_sv_o,train_prec_sv_o,train_rec_sv_o,train_f1_sv_
              test_result_sv_o,test_prec_sv_o,test_rec_sv_o,test_f1_sv_o,con_mat_sv_o,best_p
```