MEF UNIVERSITY

SENTIMENT ANALYSIS OF HURRIYET EMLAK

Capstone Project

Alev Korkmaz

İSTANBUL, 2017

MEF UNIVERSITY

SENTIMENT ANALYSIS OF HURRIYET EMLAK

Capstone Project

Alev Korkmaz

Prof. Dr. Özgür Özlük

İSTANBUL, 2017

ACKNOWLEDGEMENTS

I wish to acknowledge my adviser and teacher Prof. Dr. Özgür Özlük for his endless support and valuable feedbacks during what has been a long and challenging process of research and writing. He has been patient with me over this long process. The freedom he has afforded me to explore various paths has perhaps made my process far longer, but correspondingly far richer for all the studies made. I know more now than I did when I started, and more now than if my thesis experience had been typical. I sincerely hope I continue to have opportunities to interact with Prof. Dr. Özgür Özlük for the rest of my career.

I would like to thank also my teacher, Tuna Çakar, a talented teacher and passionate researcher. He was always so helpful and encouraging us to be independent thinkers, and having confidence in our abilities to go after new things that inspired us.

I want to thank the members of Hürriyet Emlak for their time and valuable attention. With data they provided to me, this project became a real practical way to learn and a goal to me.

EXECUTIVE SUMMARY

Sentiment Analysis of Hürriyet Emlak
Alev Korkmaz

Prof. Dr. Özgür Özlük

SEPTEMBER, 2017, 31 PAGES

Sentiment analysis refers to the task of natural language processing to determine whether a piece of text contains some subjective information and what subjective information it expresses, whether the attitude behind a text is positive, negative or neutral. Understanding the opinions behind user-generated content automatically is of great help for commercial and political use, among others. The task can be conducted on different levels, classifying the polarity of words, sentences or entire documents.

This study proposes sentiment analysis methods for evaluating and developing some prediction models in order to accurately estimate the users' opinion of Hürriyet Emlak which provides real estate services over the web.

Key Words: SA: Sentiment Analysis, ML: Machine Learning, DT: Decision Tree, Big Data: BD

ÖZET

Hürriyet Emlak'ın Duygu Analizi
Alev Korkmaz

Prof. Dr. Özgür Özlük

EYLÜL, 2017, 31 SAYFA

İnsanlar şirketimizi, ürünlerimizi, kampanyalarımızı, hizmetlerimizi sever mi, yoksa bunun için bizden talep ettikleri fazlası mı var? Bu, pazarlama için çok önemli olan bir sorudur, çünkü kim sevmediği bir şirketin müşterisi olmak ister ki? Günümüzde memnuniyet veya hoşnutsuzluk sorusunu cevaplamak için duyarlılık analizi çok önemlidir. Öyle ki, stratejideki eksiklerin tespiti ve iyileştirilmesinde harika bir bilgi kaynağı olarak kullanılabilir.

Bu çalışma web üzerinden emlak hizmeti veren Hürriyet Emlak şirketinin kullanıcılarının duygularını doğru bir şekilde tahmin etmek için bazı tahmin modellerinin değerlendirilmesi ve geliştirilmesine yönelik duygu analizi yöntemleri sunmaktadır.

Anahtar Kelimeler: DA: Duygu Analizi, MÖ: Makine Öğrenmesi, BV: Büyük Veri

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. (Thomas K., Clickworker, March 2017, [21])

In today's business world, sentiment analysis is critical because helps you see what customers like and dislike about you and your brand. Customer feedback—from social media, your website, your call center agents, or any other source—contains a treasure trove of useful business information. But, it isn't enough to know what customers are talking about. You must also know how they feel. Sentiment analysis is one way to uncover those feelings.

Opinion mining allows us to improve campaign success, product messaging and customer service consequently test KPIs and generate leads. This is an excellent source of information to make your strategy much better. But it is not a once and done effort. By reviewing your customer's feedback on your business regularly you can be more proactive regarding the changing dynamics in the market place.

1.1. Define Problem

Today's technological advances allow to any kind of business to survive online so to real estate. Internet real estate surfaced around 1999 when technology advanced and statistics prove that more than 1 million homes were sold by the owners themselves in just America, in 2000 and in Turkey early 2001.(Wikipedia, January 2016, [8])

Hurriyet Emlak is the real estate platform of the Hürriyet newspaper, which keeps the pulse of the real estate sector in Turkey since 2006.

The primary goal of Hurriyetemlak.com is to provide services to real estate offices, both visually and technically, to its individual users and corporate partners, and to bring all actors in the real estate sector together within one roof. Its purpose is to bring real buyers and sellers together for accurate and up-to-the-minute ads and to ensure they reach their goals quickly.

This project aim to help Hurriyet Emlak for its purposes in the sector and see what people really think about the company and its strategy, thus what the most convenient action could be taken to keep users in HurriyetEmlak.

1.2. Literature Survey

Machine learning approach relies on the famous ML algorithms to solve the Sentiment Analysis (SA) as a regular text classification problem that makes use of syntactic and/or linguistic features.

Text Classification Problem Definition: We have a set of training records D = {X1, X2, …, Xn} where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it. (Lee Stott, April 2016, Machine Learning, [2])

The brief details of some of the most frequently used classifiers in SA in the following subsections.

1.2.1 Probabilistic Classifiers

Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers.

Three of the most famous probabilistic classifiers are Naïve Bayes Classifier (NB), Bayesian Network (BN), and Maximum Entropy Classifier (ME).

Naïve Bayes Classifier (NB):

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label. (Bruno Stecanella, 2017, Engineer and author in Monkey Learn, [16])

Bayesian Network (BN):

The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables is specified for a model. In Text

mining, the computation complexity of BN is very expensive; that is why, it is not frequently used. (2017, Bruno Stecanella, Engineer and author in Monkey Learn, [16])

Maximum Entropy Classifier (ME):

The best explanation I've found is this: "The Maximum Entropy (MaxEnt) classifier is closely related to a Naive Bayes classifier, except that, rather than allowing each feature to have its say independently, the model uses search-based optimization to find weights for the features that maximize the likelihood of the training data." (Christopher Potts, 2015, Sentiment Symposium Tutorial: Classifiers, [17])

1.2.2 Linear classifiers

In the field of machine learning, the goal of statistical classification is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics.

A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

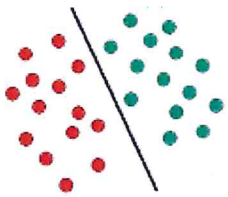

Fig 1: Linear classification

One of the most commonly used linear classifiers is SVM thanks to its usability of both classification and regression. (Greg Lamp, 2012, Why use SVM, [15])

Support Vector Machines Classifiers (SVM):

Compared to the previous schematic, sometimes full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.

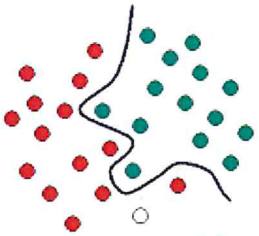

Fig2: SVM classification

1.2.3 Decision tree classifiers

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification.

1.3. Main Steps to Complete a BigData Project

According to the researches in websites such as datasienceplus.com, r-bloggers.com, Naïve Bayes is apparently the most convenient and simplest model for sentiment analysis. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document.

We will train the naive Bayes model at the first place. However, the naive bayes method is not included into RTextTools. The e1071 package did a good job of implementing the naive bayes method in R. There are also other methods which are seen in section 2.1 giving good accuracy such as MAXENT, SVM etc. So all possible methods will be seen and trained.

In this part, the steps are taken by this research to prepare the data for analysis. The standard machine learning process is flexible depending on the problem defined. (machinelearningmastery.com)

To achieve success in this project main steps are followed as below:

- Understand the business
- Discover & Prepare Data
- Evaluate Algorithms
- Improve the Method
- Present Final Results

Please see the flowchart below made by help of blogs.msdn.microsoft.com, quora, machinelearningmastery.com:

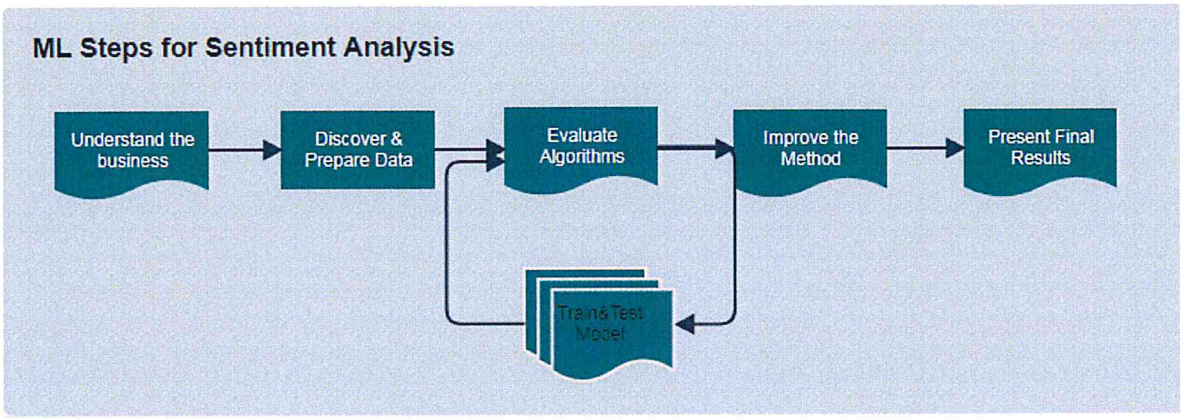

Fig3: Machine Learning Steps for Sentiment Analysis

The first step was to understand the context of the research objectives, to improve company strategy the situation of existing services of Hurriyet Emlak. The second step is to discover data followed by cleaning the data to prepare it for analysis. Text preparation is nothing but filtering the extracted data before analysis. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study from the data which takes the largest amount of time during the project. Third step is where the ML algorithms which are seen in section 1.2 are applied and tested. Sentiments can be broadly classified into two groups, positive and negative. At this stage of sentiment analysis methodology, each subjective sentence detected is classified into two groups "positive" and "negative". This is followed by improving the method applied. Finally, results are visualized and conclusions are represented.

1.3.1 Understanding Business

In a real estate company e-business model differentiation affects company success just like all other sectors. Model components of success, such as distribution channels, business partnerships, attracting customers, revenue streams, etc., are important. So, Hurriyet Emlak business strategy is clearly to improve the efficiency of the company website (Hurriyetemlak.com) consequently the marketing strategy. So that customer experience will develop, then the gaps in the current strategy will be detected and fixed, so revenue consequently will rise.(Hürriyet Emlak, 2017, [22])

In our case we have 2 main areas in which sentiment analysis will be used for:

Business: In marketing of Real estate, to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their new applications or service launches and why consumers don't prefer to use their website. For example; Analyzing reviews are important to the business holders as they can take business decisions according to the analysis results of users' opinions about their application for website usability. The reviews sources are mainly surveys and review sites such as twitter, Facebook.

When combined with social network analytics sentiment analysis becomes even more powerful and predictive. Hurriyet Emlak can better anticipate the sources and circumstances of serious problems. The most common predictive sentiment-driven application by far is identifying which users are likely to leave. The organization gets "early warning" on individual. The firm then has more time to decide what – if anything – it will do to retain them.

Public Actions: SA is extremely useful in monitoring the social media because it provides an overview of the public's opinion on certain topics. It is used to analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

2. DISCOVER AND PREPARE DATA OF HURRIYET EMLAK

Data of this project is provided from Hurriyet Emlak Company within an agreement to promise not to share any information with competitive companies. Data is gathered in Excel form then loaded in an Oracle DB.

Please see the main titles of data:

1- APP_STORE_COMMENTS

   Scores and comments gathered through apple devices.

2- MAILS

   Mails from Hurriyet Emlak users

3- SOCIAL_MEDIA

   Comments from twitter, Facebook, Instagram

4- SURVEY_MONKEY

   Scores and comments gathered through Survey Monkey Platform

5- ANDROID_COMMENTS

   Scores and comments gathered through android devices.

6- CHANNEL MAILS

   Requirements and Declarations from Hurriyet Emlak users

2.1. Discover Data and Define Relationships

As data has been provided from various systems with no Pkey there is no relationship between tables. Therefore data should be compared and associated manually.

First of all we look at the detailed data information:

APP_STORE_COMMENTS; There are A total of 50 obs., all Scores and comments gathered through apple devices. Columns are respectively; "Title": the subject of comment, "Author": nick name of user, "Version_lbl": IOS version of apple device, "Rating_lbl": rating score over 5, "Content_lbl": comment of user

MAILS; A total of 298 obs., these are received Mails from Hurriyet Emlak users. Columns are respectively; "Send_date": Date of sent mail, "Customer_type": type of customer (Kurumsal Üye / N/A), "Main_text": Mail text

SOCIAL_MEDIA: A total of 57 obs., Comments from twitter, Facebook, Instagram. Columns are; "Title": the subject of comment, "Channel": Social Media Channels (twitter, Facebook, Instagram), "Send_date": Date of Comment, "Customer_type": Consumer/Business, "Main_text": Comment of user

SURVEY_MONKEY: A total of 19K rows, Scores and comments gathered through Survey Monkey Platform. Columns are; "Send_date": Date of Commenting/Scoring, "conseil_score": Recommendation Score over 10. How much you recommend visiting Hurriyet Emlak.com to your friend, "page_usage_score": Usability Score over 100.How easy Using Hurriyet Emlak.com to use, "main_text": Comment of user

ANDROID_COMMENTS: A total of 1, 5 K row, all scores and comments gathered through android devices. Columns are, "star_rating": User score over 5, "Review_title": the subject of comment, "Review_text": Comment of user, "review_last_upd_date_time": Last updated date of comment, "Developer_reply_date_and_time": Date of Reply of Hurriyet Emlak, "Developer_reply_epoch": Serial Number of Service Request, "developer_reply_text": Comment of Hurriyet Emlak Reply," review_link": Link of Service Request

CHANNEL MAILS: A total of 194 obs., Requirements and Declarations from Hurriyet Emlak users. "Customer_type": Consumer/Business, "Main_text": Comment of User, "Numero": id of Comment, "Category": Subject of Comment, "Answer": Hurriyet Emlak Answer

2.2. Prepare Data

Machine learning algorithms learn from input data. It is highly critical that your input is the right data for the problem you want to solve. Even if you have good data, you need to make sure that it is in a useful scale, format and even those meaningful features are included.

So, in this step, we should know what data is telling us. The more noisy the data, the more difficult it will be to see the important points. A particular subset of data will be selected to be used as input consequently to be able to define a clear strategy in ML.

The actual data preparation process is three steps as follows:

Step 1: Data Selection; we will consider what data is available, what data is missing and what data can be removed. For example, in App_store_comments, "Author" has no meaning for sentiment analysis as it cannot be used as an identifier.

Step 2: Data Preprocessing; Organize your selected data by formatting, cleaning and sampling from it. For example: Internalization of Turkish character that will be explained in the following section.

Step 3: Data Transformation; Transform preprocessed data ready for machine learning by engineering features using scaling, attribute decomposition and attribute aggregation. As data attributes vary, sometimes classifications are crucial to make information more visible. For example: Time of comment consists of 24h, with a simple hour grouping, we can see all comments by day periods. (You can see time groping logic in following section)

Step 4: Final Metadata Creation; selecting, transforming and cleaning data create a clearer picture for a better understanding. With a good visualization data may give better clues for analysis as ML method should consider only the most powerful point of data.

2.2.1 Data Selection

In this part, some irrelevant information will be eliminated. For a SA, the basic needs are specially review and opinion sentences and user scores in which will focus mostly. But on the other hand, some other areas such as "date" and "application versions" that will be used for visualization are important for a better understanding of data. (Romain Paulus, Stanford Machine Learning laboratory, June 2014, What-is-the-step-by-step-procedure-of-sentiment-analysis, [4])

Useful columns are identified as below:

APP_STORE_COMMENTS: Rating_lbl (User Score) and Content_lbl (User Comment)

MAILS: Send_date (date of mail) and Main_text (mail content)

SOCIAL_MEDIA: "Send_date": Date of Comment, "main_text": Comment of user

SURVEY_MONKEY: All columns are important for this project

ANDROID_COMMENTS: "Review_text": Comment of user,
"review_last_upd_date_time": Last upd date of comment. Developer comments are not considered important because they have no contribution for data understanding due to repetition. For example: Developer_reply_text contains generally the same text in all rows.
CHANNEL MAILS: "Main_text": Comment of User. Other columns have no significant effect on data.

2.2.2 Data Preprocessing

In this step, we will make our data more convenient and easy for analysis. As mentioned in previous stage, in SA, text preparation is very important. When we looked at our text data, as Jeff Atwood said in "blog.codinghorror.com" in 2008, Turkish characters which create many problems in programming languages should be handled by internalization it. Therefore all Turkish characters such as 'İÇĞÖŞÜ' are translated to 'ICGOSU' by using PLSQL.

The next thing to do is cleansing data because it has some corrupt or inaccurate records. To detect and correct to incomplete, incorrect, inaccurate or irrelevant parts of the data PLSQL is used in the first place.

It is also should be considered that people use normally positives words in negative reviews, but the word is preceded by "not" (or some other negative word), such as "guzel değil". And since the classifier uses the bag of words model, which assumes every word is independent, it cannot learn that "guzel değil" is a negative.
Now we clean disordered data;
For PLSQL Code (see the file SCRIPTS_SURVEY_MONKEY.SQL)
Unreasonable records are detected to be removed.
For example, PAGE_USAGE_SCORE>=85 and CONSEIL_SCORE <5 is not reasonable because if a user scores very well page usability it would also recommend it.
In data cleaning part we could catch only 997 records out of 19926 records.
In the first place we keep whole data to visualize general situation but in order to use Naïve Bayes method, all records without text content will be excluded.

2.2.3 Data Transformation

In this step we will transform our data to have a standard view and meaning in overall data.

For example: The score of Page Usability (PAGE_USAGE_SCORE) is out of 100 while CONSEIL_SCORE (Recommendation Score) is out of 10. So both score are standardized to be out of 10 and new variable is named as PAGE_USAGE_SCORE_10. (Results are checked and validated. Please see data sheet in Survey_data.xlsx)

So, finally we will standardize all scores and combine them. Then we will get a set of scores called NPS.

NPS measures customer experience and predicts business growth. This proven metric transformed the business world and now provides the core measurement for customer experience management programs the world round.

Table1: Classification for NPS

Classification	Score
PROMOTERS	8,9,10
PASSIVE	7,6
DETRACTORS	5,4,3,2,1

Respondents are grouped regarding Medallia standard.

Medallia :Hundreds of the world's best-loved brands trust Medallia Experience Cloud™, a software-as-a-service platform, to help them capture feedback everywhere their customers are (on the phone, in store, online, mobile), understand it in real-time, and deliver insights and action—from the C-suite to the frontline—to improve their performance. This methodology empowers your employees to make smarter, more informed daily decisions so they can deliver better customer experiences. Through our software, companies can build more loyal customer relationships, grow faster, reduce costs, and improve corporate culture. (Net Promoter Score, June2017,[10])

With a slight flexibility application the group definitions are as below:

Promoters (score 9-10) are loyal enthusiasts who will keep buying and refer others, fueling growth.

Passives (score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.

Detractors (score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.

As it is not a clear visualization to present data by hour, a new category is made for timing with a slight flexibility of Turkey Socio-cultural structure. (Jane Mairs, Director of English Language Learning Publishing, [25])

Table2: Parts of the Day

Classification	Hours
Morning	05:00-12:00
Afternoon	12:00-17:00
Evening	17:00-21:00
Early-Night	21:00-24:00
Night	00:00-05:00

The classification below is done to visualize text contents by grouping key words. These keywords are found by examining text and user's problems. So there is no exact rule to classify comments.

Table3: Classification for comment/mail content

WORDS	CLASS
%ARAMA%', '%FILTRE%', '%HATA%', '%FILTRE%', '%KRITER%', '%ZOR%', '%KOLAY%', '%GORSEL%', 'KULLAN%','%FOTO%'	SAYFA
'%YAVAS%', '%HIZ%', '%EKRAN%', '%YUKLE%', '%INTERNET%', '%AGIR%', '%DON%', '%KAPA%', '%DURD%'	PERFORMANS
%ILAN%', '%GUNCEL%', '%ESKI%', '%GERCEK%DEGIL%', '%RESIM%', '%HARITA%','%YAYIN%'	ILAN
'%YOK%', '%EKSIK%', '%MIYOR%', '%MUYOR%', '%ERIS%'	ERISIM
'%ANKET%'	ANKET
'%TOKEN%'	TOKEN
%SAHIBIN%'	KARSILASTIRMA
'%SISTEM%', '%KAPA%', '%DURAK%', '%ACILM%', '%ACAMI%', '%ACILS%', '%CALISM%', '%ERISIM%'	HATA
'%DUZELT%', '%DEGIST%'	DUZELTME
'%ODEME%', '%BEDEL%', '%KART%', '%TUTAR%'	ODEME

2.4. Final Metadata Creation

Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier. As data is prepared for a better analysis, in this section a final look at the data will give us better ideas.

2.4.1 Understand Data - What Is The Data Telling Us?

Visualize and interpret Data (by R): R Studio is very rich for visualization use. It supports four different graphics systems: base graphics, grid graphics, lattice graphics, and ggplot2. Base graphics is the default graphics system in R, the easiest of the four systems to learn to use, and provides a wide variety of useful tools, especially for exploratory graphics where we wish to learn what is in an unfamiliar dataset.

We start by Survey Monkey as it has the largest data;

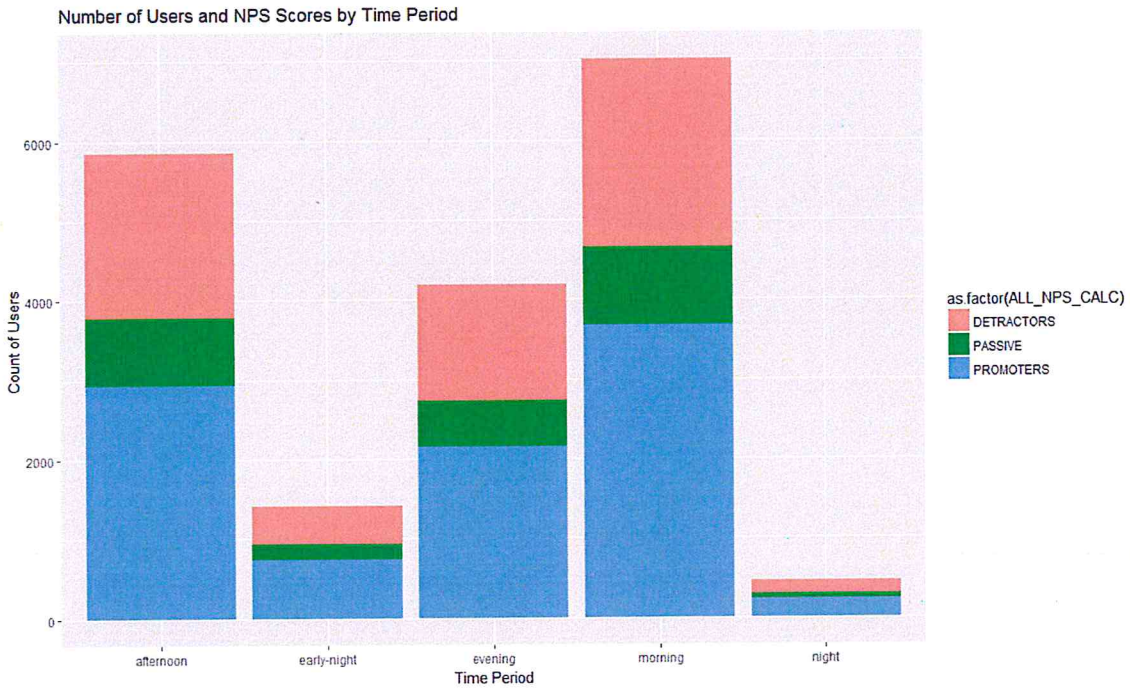

Fig4: Number of Users by NPS category and Time Period

As you can see the majority of rating is done in the morning, this is followed by afternoon. This might be because of daily period of smartphone use.

The point I want to focus on is DETRACTORS. This is where we can find all points that users get error from, dislike or have some website usability issues.

When we look at data closely we can detect rate problems:

For Example is the gplot below; "HATALI SCORE" rate is pretty high. This means potential PROMOTERS are seen like DETRACTORS because of incorrect rating of users. In comment, user writes "Harika" but the score is 10 over 100. This might be because of misunderstanding rate promotion of scoring. In all rating surveys, scores should be over the same level, but the problem of Hurriyet Emlak is that they ask for "recommendation

rate" over 10 however for "usability of webpage" it is over 100. So it is easy to get confused about when rating. We will try to detect these incorrect scores and exclude from our dataset.

Second point is that the majority of low scores are about WebPage usability (SAYFA) and access problems (ERISIM) as seen in the bar chart below : (to examine closely please see "survey_monkey.csv")

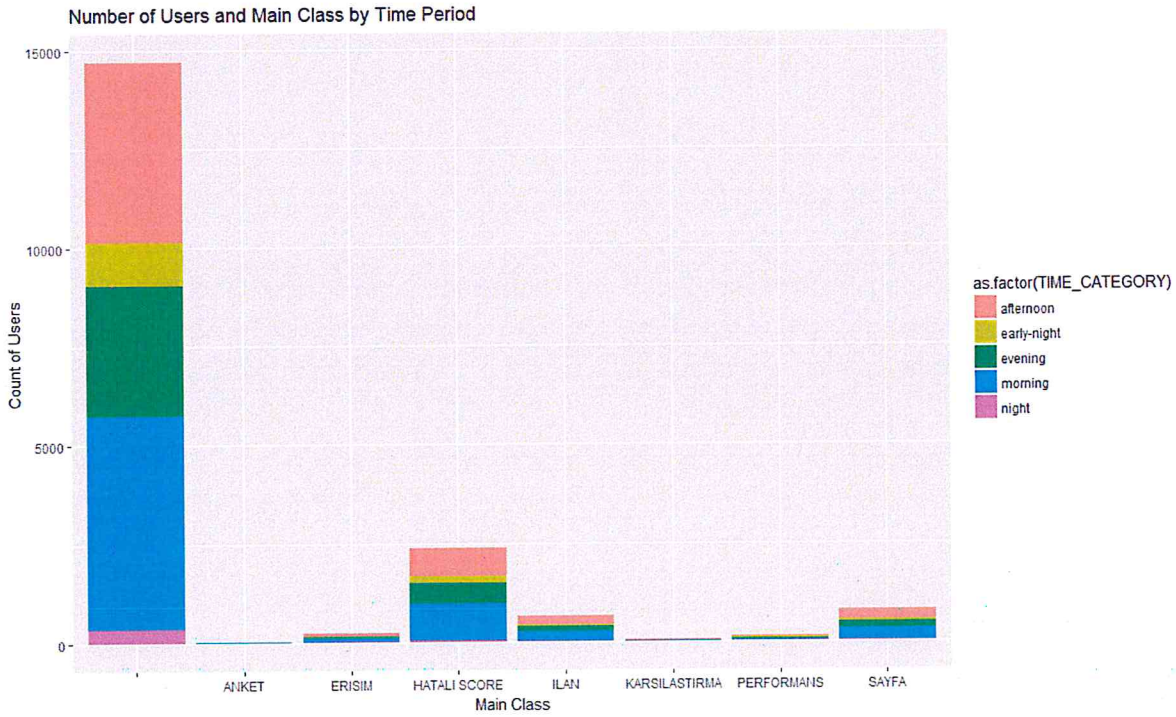

Fig5: Detractor Score Classification

ANKET: Users who don't like survey which appears when openning application

ERISIM: Users who have access issues to webpage via application

HATALI SCORE: Scores which don't match with comment For example : User score is less than 5 over 10 despite a comment like "HARIKA" , "SUPER" , "TESEKKURLER" etc.

ILAN: Users who have problems about publications

KARISLASTIRMA: comparison with competitor companies such as "sahibinden"

PERFORMANS: users facing performance problems

SAYFA: usability problems in webpage

We will continue by Android Comments which shows the level of application version use :

In this data we see all applications for HurriyetEmlak webpage use. As it can be seen clearly the most popular Application is 204800. This means users prefered to use mostly this version to reach HurriyetEmlak.com. So the version named "204800" is apparenlty a succesful version.

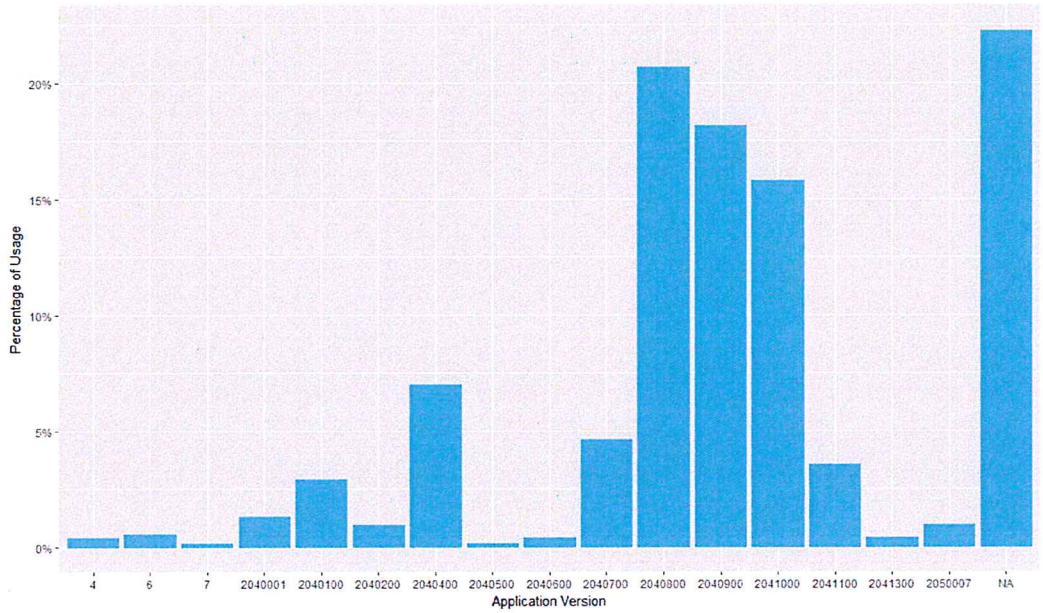

Fig6: Application usage by user count

After seeing details, we can now create our standart data from our dataset. For standardization, matching rules should be defined. After rules have been established and productionized, there will be attempts to measure the quality of each data set in regular intervals.

ALL NPS : Our final NPS data consists of "REVIEW_TEXT", "NPS","STAR_RATING" and "MAIN_CLASS" information which are standardized in all tables as following Combination of all scored data :

1- SURVEY_MONKEY : Scores and comments gathered through Survey Monkey Platform. 4 variables and approx. 20K obs. → Scored Data

2- ANDROID_COMMENTS : Scores and comments gathered through android devices. 18 variables and approx. 1.5 K obs. → Scored Data

3- APP_STORE_COMMENTS: Scores and comments gathered through apple devices. 5 variables and approx. 50 obs. → Scored Data

To simplify data, all records having score are combined and produced a new dataset : ALL_NPS (6910 obs.)

Columns:

REVIEW_TEXT : User comment

NPS: if STAR_RATING>=4 then 1, if STAR_RATING<4 then 0

STAR_RATING :User Score over 5

MAIN_CLASS : REVIEW_TEXT class that we created in the previous step.

We created NPS column by making a propotion score over 5 for all type of scoring such as over 10 and 100. For Example: As you have seen in section 2.2.3, in page usabilty scoring was over 100 while it was in "recommendation score" over 10. With a logical proportion like 60/100 = 6/10=3/5 we combined and standardized all different kind of scoring.

Finally ALL_NPS consist from combined scores over 5. NPS rule is defined according to article of Robert Biswas-Diener from University of Illinois at Urbana-Champaign Psychology as below:

Promoters (score 9-10) are loyal enthusiasts who will keep buying and refer others, fueling growth = from (9 / 2) to (10 / 2) = from 4,5 to 5.

Passives (score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings = from (7/2) to (8/2) = from 3,5 to 4

Detractors (score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth =from 0 to (6/2)= from 0 to 3

According to descriptions above, we classified NPS as if STAR_RATING>=4 then 1, if STAR_RATING<4 then 0

Table 4: Example data of ALL_NPS

REVIEW_TEXT	NPS	STAR_RATING	MAIN_CLASS
AYILDIM	1	5	
1.FAVORILERI SILEMIYORUM. 2.ARAMA YAPTIKTAN SONRA FILTRELEME YAPTIGINFIYAT ARALIGINA SADECE RAKKAM EKLEYEBILIYORSUN. SILIP YENIDEN GIREMIYORSUN. 3. SON 3 GUN ILANI ISTESEMDE TARIH SINIRLAMASI YAPAMIYOR. 1 AY EVVELILANGELIYOR.	0	2	SAYFA
HARITA KISMINYAKINLASTIRMA YAPTIKTAN SONRA HANGI MANTIKLA KENDILILIGIDEN UZAKLASIYOR ANLAMADIM... BOYLE OZELLIK KOYDUNUZ KI.	0	2	ILAN
KAYITLI ARAMALAR SEKMESIBASARKEN UYGULAMA DONUP KAPANIYOR!!!	0	1	SAYFA
BASIT	0	1	
SUREKLI ZOOM OUT OLUYOR, SORUN YOKTU YAHU DAHA ONCEDEN..	0	1	ERISIM

So the final view of the data is as below ; SAYFA and ILAN are most commented areas. They are followed by "performans" subject. This could be because of performance problems faced by users using Hurriyet Emlak Applications. More precisely, when user have a "ILAN" problem, user with 0 NPS, comment about the problem such as "...HARITA KISMINYAKINLASTIRMA YAPTIKTAN SONRA HANGI MANTIKLA KENDILILIGIDEN UZAKLASIYOR ANLAMADIM... BOYLE OZELLIK KOYDUNUZ KI...." When HurriyetEmlak employees get involed ASAP in user problem and solve it, the same user put another comment such as ""HARITALI ARARKEN AZ SONUC GELIYOR, SUREKLI HARITA KENDI KENDIOLCEKLENIYOR." DEMISTIM SURUMDE DUZELTILMIS GUZEL OLMUS, TESEKKURLER." with scoring NPS=1 .

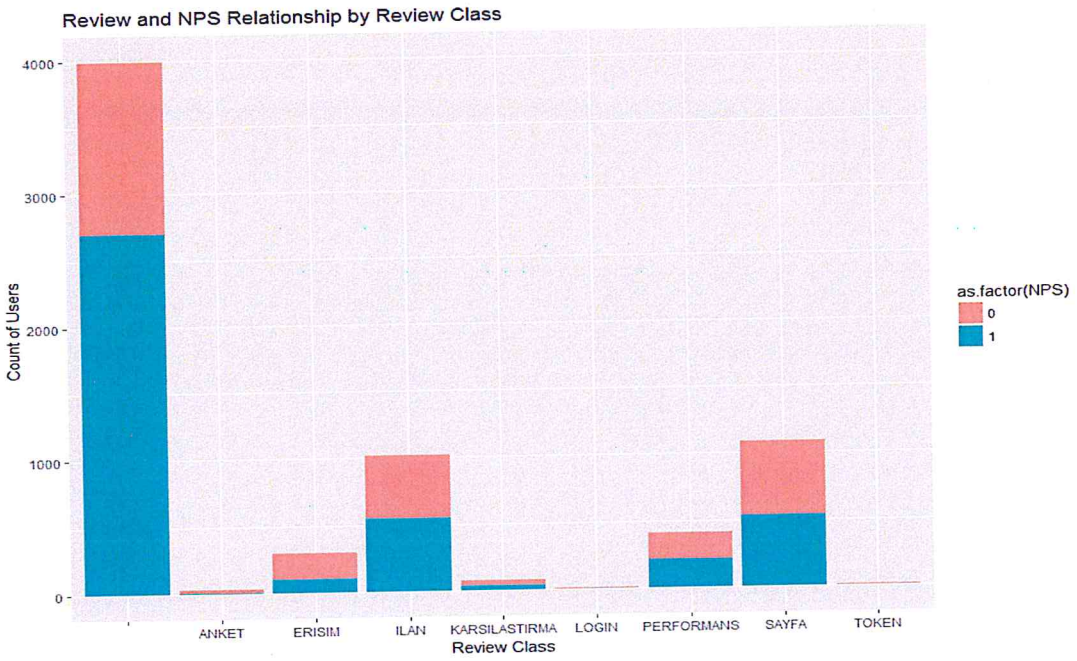

Fig 7: User Rate and Review Subject by NPS

To prevent user webpage usability issues, the graph below could be helpfull to see main areas that visitors face problems, For example: in NPS=0, 20% of users face SAYFA issues :

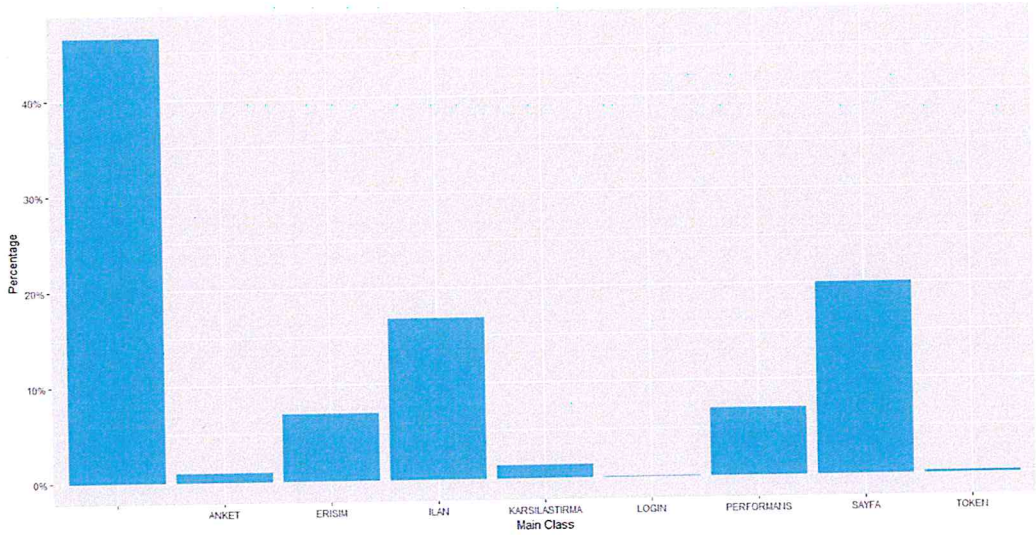

Fig 8: Percentage of Main Class for NPS=0 users

We've seen details of ALL_NPS data which is the scored part of final metada to interpret for sentiment analysis.

ALL TEXT CONTENTS :

The other part of metadata "ALL TEXT CONTENTS" is combined from all mails and comments which consists of text without score. As we have only comments and mails in this data, this part wikk definatelly a part of train dataset for analysis :

1. SOCIAL_MEDIA : Comments from twitter, facebook, instagram
    a. 5 variables and approx. 57 obs.→ TEXT_CONTENT column is used
2. MAILS: Mails from Hurriyet Emlak users
    a. 3 variables and approx. 298 obs. → TEXT_CONTENT column is used
3. CHANNEL_MAILS : Requirements and Declarations from Hurriyet Emlak users
    a. 6 variables and approx. 194 obs. → TEXT_CONTENT column is useds

The same method as ALL_NPS is applied to Mails and ALL TEXT CONTENTS data (550 obs.) is created. All comments and mails are categoritized by using main_class table given previously in Section 2.2.3.

Columns of ALL TEXT CONTENTS:

Main Text : User mail text

Main Class: Main_text class

Table 5: Example data of ALL TEXT CONTENTS

MAIN_TEXT	MAIN_CLASS
43038-396 no'lu ilan bana ait.	ILAN
10970-25550 no'lu ilan uygunsuzdur.	ILAN
Sitede mükerrer ilanlarınız var.	ILAN
4613-2057 no'lu ilan izinsiz yayınlanıyor.	ILAN
69803-654 ve 69803-653 Nolu ilanlar ile ilgili sözleşme ektedir. Kaldırılmasını rica ederim.	ILAN
21413-1198 nolu ilanda yer alan gayrimenkulun sahibiyim. Tapu ektedir.	ILAN
Mükerrer ilanlarınız bulunmaktadır.	ILAN
32618-2490 nolu ilanı aradığımda satıldığını söylüyorlar.	ILAN

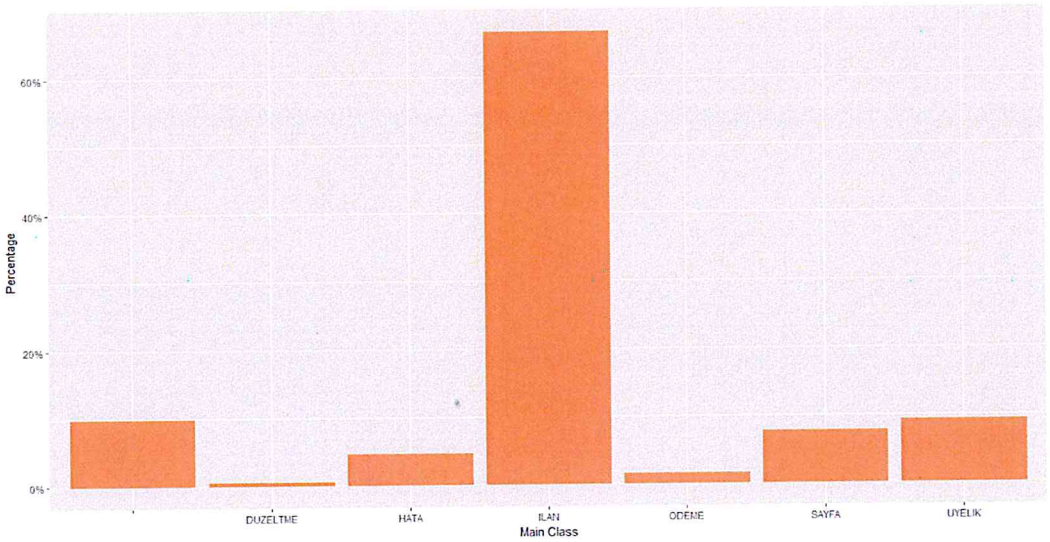

Fig 9: Main Classification of Mail contents

The graph above shows that mailS subjected "ILAN" create the majority in ALL_TEXT. This might be for that customers have problem or request mostly about publications on Hurriyet Emlak.

ILAN : all mails about publications

ÜYELIK : all mails and declaration about subscription issues

SAYFA : all mails and declarations about webpage

HATA : all mails and declarations about application error

DÜZELTME: all mails and declarations about correction request

ÖDEME : all issues about payment or fraud issues

4. EVALUATE ALGORITHMS

Once we have defined our problem and prepared our data we need to apply machine learning algorithms to the data in order to solve your problem. It is possible to spend a lot of time choosing, running and tuning algorithms. We want to make sure we are using our time effectively to get closer to our goal. So that, we discover all possible methods to get best results for sentiment analysis.

3.1. Basic Model Set Up

Naive Bayes classifier: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Infrastructure choices for sentiment analysis are Python and R. But we should first determine which one is the best for this study.

Here is a very clear comparison of both infrastructure that Jose A Dianes, Data Analytics & Visualisation - SW Engineer PhD made in 2015, [12], in summary:
R has a much bigger library of statistical packages. (R packages cover more techniques. You can find R packages for a wide variety of statistical tasks using the CRAN task view)

1- Python is better for building analytics tools (R and Python are equally good if you want to find outliers in a dataset, but if you want to create a web service to enable other people to upload datasets and find outliers, Python is better. Python is a general purpose programming language, which means that people have built modules to create websites, interact with a variety of databases, and manage users.)
2- R builds in data analysis functionality by default, whereas Python relies on packages (Because Python is a general purpose language, most data analysis functionality is available through packages like NumPy and pandas. However, R was built with statistics and data analysis in mind, so many tools that have been added to Python through packages are built into base R.)
3- Python is better for deep learning.(In our case, we are not doing deep learning)
4- Python relies on a few main packages, whereas R has hundreds

5- R is better for data visualization (Visualization is crucial to make important point visible)

We'll use Naive Bayes at the first place, because it's effective and simple to implement for a problem of this type. ML Lib also supports multi-class Decision Trees, which take more parameters than Naive Bayes classifiers. Some other models such as SVM, MAXENT, DT are compatible too for sentiment analysis. According the type of data, it might be also better to use other methods using RTextTools. All methods will be seen and trained in R Studio.

Test the model:

a. We will train the model

b. Test the accuracy

c. Summarize the results

d. Cross validation

3.2. Train the Model

Now it's time to split our data into two sets: one for training the model, and a smaller, non-overlapping data set for testing how well the model works on new data.

Firstly, we create 75:25 partitions of the data frame, 25% for test and 75% for training part. To accomplish this, we'll randomize data which is very important not to get memorized results.

Now our training data is ready to be fed into a machine learning model.

3.2.1 Naïve Bayes

We first prepare a corpus of all the documents in the data frame. This means Tokenization of the text data. So we represent each word in a document as a token (or feature) and each document as a vector of features. In addition, for simplicity, we disregard word order and focus only on the number of occurrences of each word i.e., we represent each document as a multi-set 'bag' of words.

Next, we clean up the corpus by eliminating numbers, punctuation, white space, and by converting to lower case. In addition, we discard common stop words such as "ben", "biz", "benim", yani", etc. by using PLSQL in data Cleaning part.

Third step is to find most frequent words by using findFreqTerms function in R. This is very useful and easy to use. We start by finding most frequent words with at least 6 times used rule.

Next step is creation of a matrix, The Document Term Matrix. We represent the bag of words tokens with a document term matrix (DTM). The rows of the DTM correspond to documents in the collection, columns correspond to terms, and its elements are the term frequencies. We use a built-in function from the 'tm' package to create the DTM.

Finally we apply naïve bayes model to our train dataset (5000obs.), then test dataset (1910 obs.)

You can find confusion matrix results, accuracy rate rises with the freqTerms number as seen below:

Table 6: Confusion matrix & Accuracy by FreqTerms

		PREDICTED	
	ACTUAL	0	1
FindFreqTerms=6 ACCURACY: 0.611	0	161	628
	1	114	1007
FindFreqTerms=7 ACCURACY: 0.619	0	173	597
	1	129	1011
FindFreqTerms=8 ACCURACY: 0.626	0	197	573
	1	141	999
FindFreqTerms=10 ACCURACY: 0.646	0	274	496
	1	179	961
FindFreqTerms=15 ACCURACY: 0.668	0	316	454
	1	180	960
FindFreqTerms=18 ACCURACY: 0.672	0	333	437
	1	189	951
FindFreqTerms=22 ACCURACY: 0.672	0	356	425
	1	212	941
FindFreqTerms=20 ACCURACY: 0.673	0	345	425
	1	199	941

As seen above, the best freqTerm level is 20 and the best accuracy possible with this data using Naive Bayes is 0.6732.

3.2.2 Other Models

To create a model, you need to pass it a container. I had used ALL NINE (9) algorithms "GLMNET, SLDA, BAGGING, RF, NNET, TREE, MAXENT, SVM, TREE" initially, but due to the memory limitations of R (32-bit) project, I was forced to create models ONLY for THREE (3) algorithms:

- SVM – Support Vector Machines
- MAXENT – Maximum Entrophy.
- Boosted Regression Trees

Support Vector Machines (SVMs; Vapnik, 2014, What is the step by step procedure of sentiment analysis, [4]) apply a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space, but in practice, it does not involve any computations in that high-dimensional space. This simplicity combined with state of the art performance on many learning problems (classification, regression, and novelty detection) has contributed to the popularity of the SVM. (See section 2.1)

MaxEnt (Phillips et al., 2006, Maximum Entropy, [26]) uses mostly environmental data for locations of known presence and for a large number of 'background' locations. But in our case we will use it for Word classification.

Boosted Regression Trees (BRT) is, unfortunately, known by a large number of different names. It was developed by Friedman (2001), who referred to it as a "Gradient Boosting Machine" (GBM). It is also known as "Gradient Boost", "Stochastic Gradient Boosting", and "Gradient Tree Boosting". The method is implemented in the gbm package in R. Due to memory limitations it takes too long to run it but we will try for once to see results.

The confusion matrixes for 3 algorithms are respectively as below:

Table 7: Other Methods Results

		PREDICTED	
	ACTUAL	0	1
SVM ACCURACY: 0.70	0	235	325
	1	96	754
MAXENT ACCURACY: 0.697	0	362	198
	1	228	622
BOOSTING ACCURACY: 0.697	0	362	198
	1	228	622

Please find Recall Accuracies for 3 algorithms as below:

Table 8: Ensemble Summary

n-ENSEMBLE	COVERAGE	n-ENSEMBLE RECALL
n >= 1	1.00	0.67
n >= 2	0.68	0.79

Table9: Algorithm performance

SVM_PRECISION	SVM_RECALL	SVM_FSCORE
0.705	0.655	0.655
MAXENTROPY_PRECISION	MAXENTROPY_RECALL	MAXENTROPY_FSCORE
0.685	0.690	0.685
LOGITBOOST_PRECISION	LOGITBOOST_RECALL	LOGITBOOST_FSCORE
0.685	0.690	0.685

Results above are produced to see;

Speed, we should get results fast and use small samples of your data and simple estimates for algorithm parameters. Turn around should be minutes to an hour.

Diversity, we should use a diverse selection of algorithms including representations and different learning algorithms for the same type of representation.

Scale-up, we shouldn't be afraid to schedule follow-up spot-check experiments with larger data samples. These can be run overnight or on larger computers and can be good to flush out those algorithms that only do well with larger samples (e.g. trees).

5. IMPROVE THE METHOD

When tuning algorithms you must have a high confidence in the results given by your test harness. This means that you should be using techniques that reduce the variance of the performance measure you are using to assess algorithm runs. This is why we recheck result by doing Cross Validation. (Sunil Ray, November 2015, Improve Your Model Performance using Cross Validation, [23])

Cross Validation: Standard machine learning assessment technique used for assessing how the results of a statistical analysis will generalize to an independent data set.

See the CV results with N=3 as below:

> cross_SVM = cross_validate (container, N,"SVM")

Fold 1 Out of Sample Accuracy = 0.6796537

Fold 2 Out of Sample Accuracy = 0.6505608

Fold 3 Out of Sample Accuracy = 0.6468011

> cross_MAXENT = cross_validate (container, N,"MAXENT")

Fold 1 Out of Sample Accuracy = 0.05012744

Fold 2 Out of Sample Accuracy = 0.05477651

Fold 3 Out of Sample Accuracy = 0.04573439

When we check the cross validation above, SVM algorithm looks much more confident than MAXENT.

Having one or two algorithms that perform reasonably well on a problem is a good start, but sometimes you need to go further to get better accuracy.

In this part, we will improve data quality to get better results of classification. One possible improvement is word correction in data. For Example: "Teşekkürler" is meant with many different words such as "TSK, TESEK, TESEKKKUR, TESEKKURLER" etc. So I've tried to FIX this word to "TESEKKUR" and the new results are as below:

MAXENT: The new result is almost the same as before

SVM: the new result is slightly better compared the previous one.

Table 10: Maxent, SVM & Boosting Error Matrix

		PREDICTED	
	ACTUAL	0	1
SVM	0	242	318
ACCURACY: 0.70	1	94	756
MAXENT	0	362	198
ACCURACY: 0.697	1	229	621
BOOSTING	0	362	198
ACCURACY: 0.697	1	228	622

In conclusion, our case is compatible to apply Extreme Feature Engineering where the attribute decomposition and aggregation seen in data preparation is pushed to the limits. We applied it for "TESEKKUR" word, but not to all words because Turkish is a complicated language as there is increasingly complex expressions are constructed by adding suffixes to a base word. As R Studio has no Turkish library, text analytics gets harder with Turkish texts. The important missing thing was a good Turkish word DB in this project so I could do better comparison of word.

6. FINAL RESULTS & CONCLUSION

In this study, a process of prediction modeling was developed and tested with a sample of an online real estate company "Hurriyet Emlak", Probabilistic Classifiers (Naïve Bayes, Maximum Entropy Classifier); linear classifiers (Support Vector Machines Classifiers) methods were tested on Hurriyet Emlak Survey & Mail Data to produce Sentiment Analysis of hurriyetemlak.com users.

Naïve Bayes method is the simplest and best method for a sentiment analysis according to online researchers. But with this project it is proven that some other methods such as SVM with 0, 70 accuracy could produce better results compared to Naïve Bayes having 0, 67 accuracy.

Cross validation is very important to see the confidential level of applied model. So the main idea is that we want to minimize the generalization error. The generalization error is essentially the average error for data we have never seen. In this study MaxEnt lost accuracy in cross validation part while SVM keeps it well.

However, sentiment analysis in R in Turkish language is not quite well performed compared to English or Spanish as R library contains them. So the most obvious focus should be a good Turkish word DB to correct and standardize all words to go further to boost the accuracy in this subject.

According to data analysis made when preparing data to ML model, it is seen that people Review comes mostly in the morning(05:00-12:00) and afternoon(12:00-17:00), also Promoters (65%) are more than Detractors (35%), in application versions the Most successful android version was 2040800, in user problems SAYFA & ILAN are top categories in NPS rates.

In conclusion, this study attempted to detect HurriyetEmlak's main comments coming from users and measure NPS of social network comment. Most Sentiment Analysis machine learning algorithms are near 80% (Stream Hacker- MAY 10, 2010, [20]). So the accuracy at 0, 7 of this study turned out to be a reasonably successful result for an introductory level sentiment analysis.

APPENDIX A

Table11: Label summary 1

	NUM_MANUALLY_CODED	NUM_CONSENSUS_CODED	NUM_PROBABILITY_CODED	PCT_CONSENSUS_CODED
0	560	395	395	71
1	850	1015	1015	119

Table12: Label summary 2

	PCT_PROBABILITY_CODED	PCT_CORRECTLY_CODED_CONSENSUS	PCT_CORRECTLY_CODED _PROBABILITY
0	70,5	44,1	44,1
1	119,4	82,58	82,58

Table 13: Document summary

	MAXENTROPY_LABEL	MAXENTROPY _PROB	SVM_LABEL	SVM_PROB	MANUAL_ CODE	CONSENSUS_ CODE	CONSENSUS_ AGREE
1	1	0.9755926	0	0.6569044	0	1	1
2	1	1	1	0.6604988	1	1	2
3	1	0.9997938	0	0.6884557	0	1	1
4	1	0.9999986	1	0.8193627	1	1	2
5	0	0.9873691	1	0.7884026	1	0	1
6	0	1	0	0.7953102	0	0	2

Table13: Ensemble Agreement

	n-ENSEMBLE COVERAGE	n-ENSEMBLE RECALL
N>=1	1	0,67
N>=2	0,68	0,79

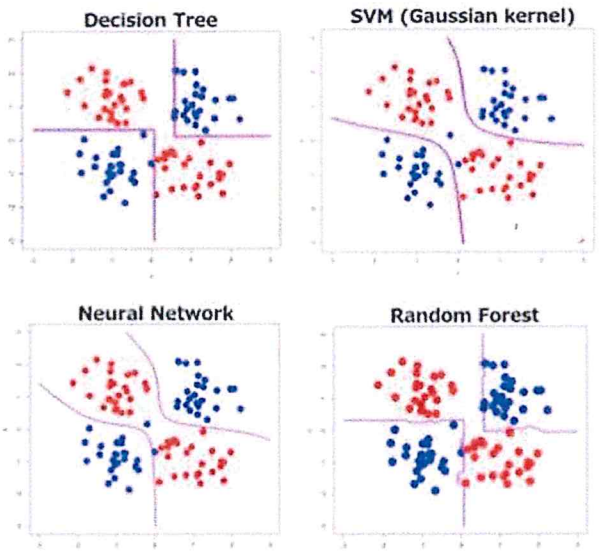

Fig10: SVM, DT, NN, RF Comparison

REFERENCES

[1] March 16, 2017 by Robert Stanley, What is Sentiment Analysis? Examples, Best Practices, & More
https://callminer.com/blog/sentiment-analysis-examples-best-practices/

[2] April 13, 2016 by Lee Stott, Machine Learning – Hands on getting started teaching Machine Learning
https://blogs.msdn.microsoft.com/uk_faculty_connection/2016/04/13/machine-learning-hands-on-getting-started-teaching-machine-learning/

[3] February 10, 2016, Tavish Srivastava, Step by step guide to building sentiment analysis model using graphlab
https://www.analyticsvidhya.com/blog/2016/02/step-step-guide-building-sentiment-analysis-model-graphlab/

[4] Jun 19, 2014, Vapnik, What is the step by step procedure of sentiment analysis?
https://www.quora.com/What-is-the-step-by-step-procedure-of-sentiment-analysis

[5] Jan 23, 2014, Gaurav Shankhdhar. Sentiment Analysis Methodology
https://www.edureka.co/blog/sentiment-analysis-methodology/

[6] January 21, 2016, Rajeev Ranjan Ishwar, 10 STEPS TO BEGIN SENTIMENT ANALYSIS TO GLEAN CUSTOMER INSIGHT
http://www.infogix.com/blog/10-steps-to-begin-sentiment-analysis-to-glean-customer-insight/

[7] November 25, 2013, Jason Brownlee, A Tour of Machine Learning Algorithms
https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[8] January, 2016 , Wikipedia : https://en.wikipedia.org/wiki/Real_estate

[9] 13 Mar 2008, Jeff Atwood, What's Wrong With Turkey?
https://blog.codinghorror.com/whats-wrong-with-turkey/

[10] 19June2017, Net Promoter Score
http://www.medallia.com/net-promoter-score/

[11] May 2006 Craig Ball, the Role of Metadata in Machine Learning for Technology Assisted Review
https://www.umiacs.umd.edu/~oard/desi6/papers/JonesFinal.pdf

[12] Aug 10, 2015, Jose A Dianes
https://www.codementor.io/jadianes/data-science-python-r-sentiment-classification-machine-learning-du107otfg

[13] 01 Nov 2012, F. Javier Alba, Basic Sentiment Analysis with Python http://fjavieralba.com/basic-sentiment-analysis-with-python.html

[14] January 10, 2016, Cheng-Jun Wang, Sentiment analysis with machine learning in R
https://www.r-bloggers.com/sentiment-analysis-with-machine-learning-in-r/

[15] December 23, 2012 by Greg Lamp, Why use SVM? http://www.yaksis.com/posts/why-use-svm.html

[16] May 25th, 2017, Bruno Stecanella, A practical explanation of a Naive Bayes classifier https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

[17] 2011,Christopher Potts, Sentiment Symposium Tutorial: Classifiers http://sentiment.christopherpotts.net/classifiers.html#nb

[18] 2010, Seth Grimes,Information management, Expert Analysis: Is Sentiment Analysis an 80% Solution? https://www.informationweek.com/software/information-management/expert-analysis-is-sentiment-analysis-an-80--solution/d/d-id/1087919

[19] Dec, 2015, Amanda Sivaraj ,Sentiment HQ Sets a New Accuracy Standard https://indico.io/blog/sentimenthq-new-accuracy-standard/

[20] May 10, 2010, Stream Hacker-Text classification for sentiment analysis – naive bayes classifier https://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/

[21] March 2017, Thomas K., Clickworker, Sentiment Analysis – What is it for? https://www.clickworker.com/2017/03/14/sentiment-analysis-what-is-it-for/

[22]Hürriyet Emlak web site: http://www.hurriyetemlak.com/

[23] November 2015, Sunil Ray, Improve Your Model Performance using Cross Validation (in Python and R) https://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-r/

[24] January 2009, Robert Biswas-Diener. Scale of Positive and Negative https://internal.psychology.illinois.edu/~ediener/Documents/Scale%20of%20Positive%20and%20Negative%20Experience.pdf

[25] February 2012, Jane Mairs, Director of English language Learning Publishing http://www.learnersdictionary.com/qa/ways-to-improve-your-english

[26] 2006, Phillips et al., Maximum Entropy https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779976/