DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

WEB MINING: PATTERN DISCOVERY ON THE WORLD WIDE WEB

by
Mustafa TURAN

June, 2011
İZMİR

WEB MINING: PATTERN DISCOVERY ON THE WORLD WIDE WEB

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Computer Engineering

by
Mustafa TURAN

June, 2011
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"WEB MINING: PATTERN DISCOVERY ON THE WORLD WIDE WEB"** completed by **MUSTAFA TURAN** under supervision of **ASST. PROF. DR. DERYA BİRANT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Derya BİRANT

Supervisor

Asst. Prof. Dr. Caner Erden ATAY

(Jury Member)

Asst. Prof. Dr. Özge ŞAHİN

(Jury Member)

Prof.Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank to my supervisor, Asst. Prof. Dr. Derya Birant, for her support, supervision and useful suggestions throughout this study.

I owe my deepest gratitude to my family. This thesis would not have been possible without their unflagging love and support.

Mustafa TURAN

WEB MINING: PATTERN DISCOVERY ON THE WORLD WIDE WEB

ABSTRACT

The uncountable size of the data in the World Wide Web (WWW) nowadays makes it the largest cloud database that ever existed on Earth. The problem with data is that it is not a structured database, which makes it meaningless. To make the data usable, web mining methods are created. Web mining is the application of data mining techniques to discover patterns from the World Wide Web (WWW). Web mining is a powerful research area to gather and examine content from web pages or web services. It has methods for information retrieval from web pages and analyses the structure of gathered documents. Moreover, web mining gathers data related to the structure of a website and its users using the web-server logs and session logs.

However, although reaching data from the WWW is possible with web mining techniques, the reached data might not be sensible or meaningful without machine learning techniques. To make the data sensitive and meaningful, there exist a lot of methods depending on one's aims. Classification, which can classify web data according to its content, is one of most popular data mining methods in machine learning.

This thesis proposes the hybrid combination of web mining techniques and machine learning techniques. The developed approach can gather Turkish text data from various web pages and web services and serve it in a structured data format. The study in this thesis basically covers web content mining, web structure mining for gathering data and analyzing the structure of web pages and services. It also uses various internal and external web services for language detection, Turkish spell-checking, Turkish 'Part of Speech Tagging' (pos-tagging) and stemming operations. Moreover, the study uses two machine learning techniques, which are Naïve Bayes and 'Support Vector Machines' with weighting method of TF-IDF (Term Frequency – Inverse Document Frequency)', to sentimentally classify the data gathered from web pages.

In this work, firstly, how and where the data is gathered is given. Secondly, the operations over the text data are explained in detail. Then, finally, sentimental classification with accuracy values over the gathered data with multiple perspectives is given.

Keywords: Web Mining, Web Content Mining, Web Structure Mining, Feedback Mining, Sentimental Classification

WEB MADENCİLİĞİ: WEB SAYFALARINDA ÖRÜNTÜ KEŞFİ

ÖZ

Web sayfalarında bulunan sayılamayacak derecede verilerin çokluğu, interneti 'Dünya' gezegeninin en büyük veritabanı haline getirmiştir. Bu kadar verideki problem bu verilen düzenli bir veri yapısı içermemesidir. Bu verinin düzenli hale getirilip çeşitli amaçlar için kullanılması amacıyla web madenciliği metotları ortaya çıkmıştır. Web madenciliği veri madenciliği tekniklerinin web sayfaları üzerinde örüntü keşfi amacıyla kullanılması için kullanılan bir tekniktir. Web madenciliği web sayfalarından ve web servislerinden veri toplamak ve veriyi incelemek için güçlü bir araştırma alanıdır. Web madenciliğinin web sayfalarından ve servislerinden veriyi elde etme, elde edilen veri üzerindeki yapıyı analiz etme gibi metotları vardır. Bunların dışında, web madenciliği web sunucu kayıtlarını ve kullanıcı oturumlarından yararlanarak kullanıcılar ve web sayfasının yapısı hakkında veri elde etme özelliğine sahiptir.

Her ne kadar web madenciliği teknikleriyle ile web sayfalarında veri elde etmek mümkün olsa da bu verileri tam manada anlamlı hale getirmek için makine öğrenme teknikleriyle kullanmak gerekmektedir. Bu verileri anlamlı hale getirmek için birçok teknik vardır. Makine öğrenme teknikleri arasında sınıflandırma, metin tabanlı verileri içeriklerine göre sınıfını belirlemek için kullanılan en popular metotlardan biridir.

Bu tez web madenciliği teknikleriyle makine öğrenme tekniklerini birlikte kullanarak hibrit bir yapıyı amaçlamaktadır. Uygulama çeşitli web sayfalarından ve servislerinden Türkçe yazılmış verileri elde edip, bu verileri düzenleyerek servis halinde sunmaktadır. Bu çalışma temelde, web içerik madenciliği, web yapı madenciliği tekniklerini kullanarak web sayfalarından veri elde edip bu verileri yapısal olarak incelemektedir. Bunun dışında, yazı dili tanıma, Türkçe kelime doğrulama, Türkçe ek kök ayırma gibi metin işlemleri için çeşitli iç ve dış web servislerini kullanmaktadır. Dahası Naïve Bayes ve 'Destek Vektör Makine'lerini

TF-TDF (Terim Frekansı – Ters Doküman Frekansı) ağırlıklandırma yöntemi ile kullanarak web sayfalarından elde edilen veriler üzerinde sezgisel sınıflandırma yapmaktadır.

Çalışmada öncelikle, verinin nasıl ve nerden elde edildiği hakkında bilgi verilmekte, ikincil olarak bu veriler üzerinde yapılan metin operasyonları detaylı bir şekilde doğruluk oranları hesaplanarak açıklanmaktadır. Son olarak ise, elde edilen metin dokümanlar üzerinde birçok açıdan sezgisel sınıflandırma yapılmakta ve doğruluk değerleri verilmektedir.

Anahtar sözcükler: Web Madenciliği, Web İçerik Madenciliği, Web Yapı Madenciliği, Geri Bildirim Madenciliği, Sezgisel Sınıflandırma

CONTENTS

Page

CHAPTER ONE

INTRODUCTION

1.1 General

In today's world, companies and products are getting online; what is meant by this is that even if they are neither online product nor Internet based companies, they are getting online with enterprise, product reviews on social media as their customers are connected to the Internet with personal computers, work computers and mobile devices. The huge demand to the Internet made mobile devices become online and this brought about the need for mobile devices to include the necessary applications to get online. According to the United Nation's report, Internet users are to exceed 2 billion at 2010 (Lynn, 2010). With easy access of applications, people get used to writing on social media platforms. Today's social media platforms can be expressed but not be limited as discussion boards, Foursquare venues, Twitter tweets, Facebook profiles, community groups and blogs on the World Wide Web (WWW).

With the availability of put, get, post requests and responses through mobile applications to WWW, people are able to put their thoughts on social media platforms much faster than a desktop computer. This online word-of-mouth behavior represents new and measurable sources of information with many practical applications which makes social platforms are a feedback treasure for companies and company products. Feedbacks are online, but there exist problems about where they are in the WWW and which sentimental classes they are in. If a company takes some people to perform only this task, the employers won't be able to track each and every page on the WWW.

"Feedback" is the communication term that is used to describe any response, critique, criticism, or comment. The social media platform reviews for a special product or a company can be taken as a feedback of this product or company.

It is a fact that it’s more important to keep the customers you already have, since it’s much cheaper than acquiring new ones (Markey & friends, 2009). The best way of keeping customers is their valuable feedbacks.

Customer feedbacks are very important resources for keeping customers, re-shaping companies and company products. A feedback can be neutral, positive and negative in sentimental classification. Positive and negative feedbacks includes very essential information about keeping in touch with customers. While a positive feedback gives an idea about how well things are going, a negative feedback gives an idea about how it should be changed to make the customer happier. Negative feedbacks help the clients to see what is wrong with a company, a company agency office or a company product, etc...

In this thesis, an approach, namely "Sentimental Feedback Miner" (SFM) was created. This is a web mining approach that finds Turkish reviews, comments and blog posts on social media platforms and makes sentimental classification over them. SFM uses web mining techniques for mining the data and uses machine learning techniques for sentimental classification.

SFM's main aim is discovering text patterns on web sites for user defined keywords and gather them in realtime and analyze them using machine learning algorithms. SFM is language independent at basic; however Turkish language for Turkish text data has been focused on the WWW for sentimental classification.

1.2 Purpose

The purpose of this thesis is discovering patterns over Turkish text data related to user given keyword or keywords using web content mining and web structure mining techniques on WWW. Then this thesis aims to analyze this content sentimentally for determining the polarity of the text data gathered. Moreover, the purpose covers the comparisons of cleaned and pruned test data results with pure test data results in sentimental classification over multiple categorized keyword queries to web sites.

The aim in sentimental classification is increasing the accuracy of polarity results for Turkish text data by applying pruning, cleaning, spell-checking and pos-tagging methods to the text and gathering an efficient way to do it.

This study mainly differs from others with real-time web mining on WWW and applying machine learning based sentimental classification over Turkish text data.

Users of the proposed approach will find the answers of the following questions by using this thesis study:

1. How many posts are written about my search keyword in a day, in a week, in a month, in a year and even in a specific time periods?
2. What are the posts about my search keyword in social media?
3. In which platforms my keyword is written?
4. Who are the people writing about my search keyword? How can I reach these posters (writers)?
5. Are the posts negative or positive?

1.3 Organization of the Thesis

The thesis consists of six chapters. Remaning parts of the thesis are organized as follows in five more chapters.

In Chapter 2, web mining techniques and classification techniques are explained without literature reviews.

In Chapter 3, previous approaches and studies are shared in a categorized way. Firstly, web mining studies are explained. In web mining studies, approaches are divided into four sub-categories based on web content mining studies, web usage mining studies, web structure mining studies and hybrid web mining studies. Then sentimental classification studies are explained. Furthermore, sentimental

classification studies are divided into two sub-categories based on the methods in which they are implied.

In Chapter 4, proposed approach is defined with system architecture, database model, flow diagram and technologies behind the system. Furhermore, each module of the approach is explained in detail.

In Chapter 5, the application is given with thumbnails of the system. Moreover, the experiment results over two different training and test datasets both with and without text cleaning, pruning operations are shared.

Finally, Chapter 6 presents conclusion of tests and gives future research directions related to the study.

CHAPTER TWO

WEB MINING WITH MACHINE LEARNING

2.1 Web Mining

The uncountable size of the data in WWW nowadays makes it the largest cloud database that ever existed on the Planet. According to Garruzzo & friends (2007) the application of data mining techniques in order to extract useful information that implicitly lay among web data is a very essential task. Main web data includes the web pages, the web page structures, the linking structure between web pages, the surfing behavior of the users and the user profiles including demographics like age, sex, education, location, etc...

With the availability of these huge data on the WWW, data mining techniques are carried to the web. Thus, web mining is the application of data mining techniques to discover patterns from Web. According to analysis targets, web mining can be divided into three different types, which are Web Usage Mining (WUM), Web Content Mining (WCM) and Web Structure Mining (WSM).

2.1.1 Web Usage Mining

The term 'Web Usage Mining' was first introduced by Cooley & friends (1997) in which they define web usage mining as the ‘automatic discovery of user access patterns from Web Server. Web usage mining techniques are used to get web browsing information of users (Srivastava, 2000). The WUM targets web logs which are logged via web servers and typically contain information of the visitor's IP address, hostname, time stamps, exact location of visited page, proxy type, browser, user-agent, browser-language, operating system title, screen resolution, support for plugins like java and flash player. Moreover, with some web server extensions, it is possible to see the geo-location of the visitor on web logs.

Even in Web Usage Mining, it is not guaranteed that the user always visits a web page with the same IP address, hostname or browser, etc... IP addresses of visitors are assigned dynamically because the connection to the Internet is made through an Internet service provider (ISP) which does not want users to spend the system resources of its servers (Hui & friends, 1998; Cohen, 2002). Moreover, with DSL based modems, users are able to switch off/on their devices any time. That causes a new IP address assignment from ISP side to user side. The main purpose of WUM is to discover useful information from WWW users' browsing data in order to fulfill business goals by addressing of strategies at customer relationship management or services and marketing (Hui, 2008).

WUM techniques need some processes to gain relevant information. WUM consists of three phases (Omeri, 2009): Data pre-processing, pattern discovery and pattern analysis. Logging process (pre-processing) can be done via web server, web application or third party service providers (Google Analytics, Quant-cast, etc...). The important thing, while choosing logging process is which kind of data will be used for data mining process. Also the accessibility of the data is important too. In server side web logs you have a lot of options to see about users' behavior and you have options to choose which data should be stored. Moreover, it is possible to combine server logs and application logs together. In that case the data may become more relevant to process.

2.1.2 Web Content Mining

Web Content Mining (WCM) uses the ideas and principles of data mining and knowledge discovery to monitor specific data from WWW. The data in the Web is more complex than a static database (Xu, Zhang & Li, 2010). The documents on the WWW belong to MIME types (Content-Type) and each of these MIME types has their own templates for Web pages. However, they are usually semi-structured documents like HTML pages; on the other hand some web data like database generated JSON (JavaScript Object Notation) Data, XML (Extensible Markup

Language) data are structured. Web pages consist of unstructured free text data which makes it more difficult to extract information from them (Gupta, 2006).

According to Liu and Chang, there is a classification of mining tasks (Masseglia & friends, 2008 and Velásquez, 2010):

Structured Data Extraction: Structured data is easy to handle, it contains tidy data for mining and it's faster to extract data from unstructured data. XML Data, JSON Data, Site maps can be given as an example for structured data.

Unstructured Text Extraction: However, there exist several MIME-types on WWW, where most of the data is in text format. This research is closely related to text mining, natural language processing and information retrieval.

Web Information Integration: Web sites may service same, smilar data or related information using different template systems. In order to make use of multiple web pages to provide value added services, it is needed to integrate information semantically from multiple web resources.

Building Concept Hierarchies: For instance, a linear list of pages ranked and produced by search engines.

Segmenting Web Pages: By taking a general look on source code of a web page it can be seen that a Web page includes many divisions like Meta tags, navigation menu, content area, footer, etc... Separating these segments will help web mining application to clarify the source.

Mining Web Opinion Sources: Opinions can be thought as reviews or feedbacks on several web pages or web services. This is the most generally used mining technique, in this thesis. The opinions from various web services and web pages are mined in this thesis.

2.1.3 Web Structure Mining

Web Structure Mining (WSM) is gathering structural summary of a web page and also discovering the link structure of the hyperlinks for navigation purposes (Markov & Larose, 2007). Usually, a webmaster generates pages according to a logical structure because most of the professional systems are using web frameworks and these allow routing between pages related to a coder defined automated link. Thus, WSM is used to evaluate web sites to provide efficient linking structure and grouping similar resources together. Sometimes, WSM could not be enough to reach relevant information from a web page. To deal the desired data inside the underlying templating system of a web page, WSM technique and WCM can be used together.

2.2 Text Classification with Machine Learning

Classification (Japkowicz & Shah, 2011) uses training data to generate a mining model and then uses this mining model with new data for predicting the class of data as shown in Figure 2.1. Some of the most widely used classification methods for text-classification methods based on machine learning are described in the following.

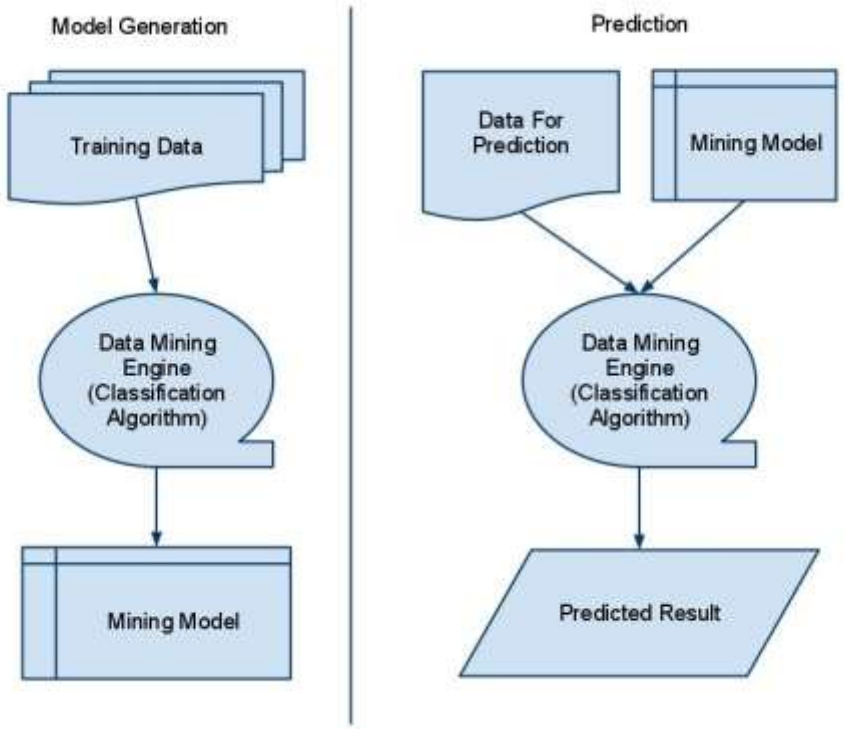

Figure 2.1 Classifications in two steps

2.2.1 Decision Tree

Decision Tree (DT), which uses 'If-Then-Else' rules to make classifications, is extensively used method in machine learning. Several decision tree algorithms are published, the first one is CHAID which was published by Kass in 1980 and the most popular one is C 4.5 which is an extension ID3 algorithm and was published by Quinlan in 1993 (Friedman, 1996).

DT has several advantages such as easy interpretation, implementation, fast results and reasonable time for training. With basic 'If-Then-Else' rules decision trees can be interpreted, implemented easily and with its top down architecture it works fast. Thus enables it to handle large number of nodes in small amount of time.

Although, DT comes with several advantages, it has some drawbacks too. In DT, classification goes to only one output node. Thus, due its top down model, DT can not handle relations between nodes.

According to the decision tree in Figure 2.2, a person who works as a civil servant gets a vehicle loan directly from the X Bank. However, other job applicators have the chance to get credit too. The bank worker will firstly look at the applicant's salary and then the applicant's age to give credit.

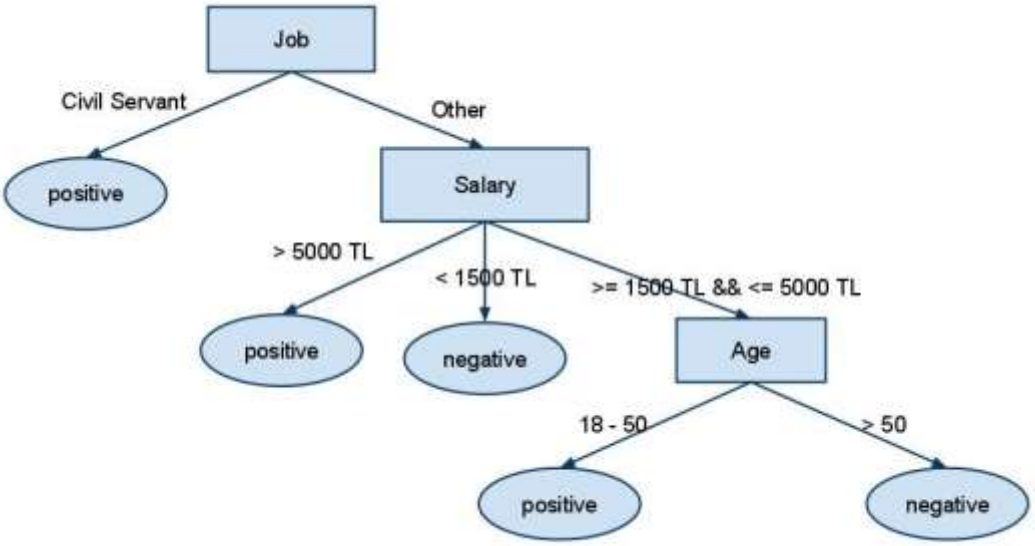

Figure 2.2 Decision Tree for vehicle loan application at X Bank

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANM) is based on human brain's computing technology (Braspenning & others, 1995). Thus, its working style simply looks like the biological nervous system. It learns by examples just like a human. ANM has several learning algorithms; the common approach is that there exist three main layers in neural networks which are the input layer, the hidden layer and the output layer as shown in Figure 2.3. Each neuron in a layer which has a weight associated with each one of the neurons in the next layer. In the input layer, every neuron acts as a predictor variable for the output nodes. The second layer in Figure 2.3 corresponds to the hidden layer which carries the necessary functions to calculate output nodes. Lastly, the output layer neurons can be thought as results of input layer neurons computed in hidden layer neurons.

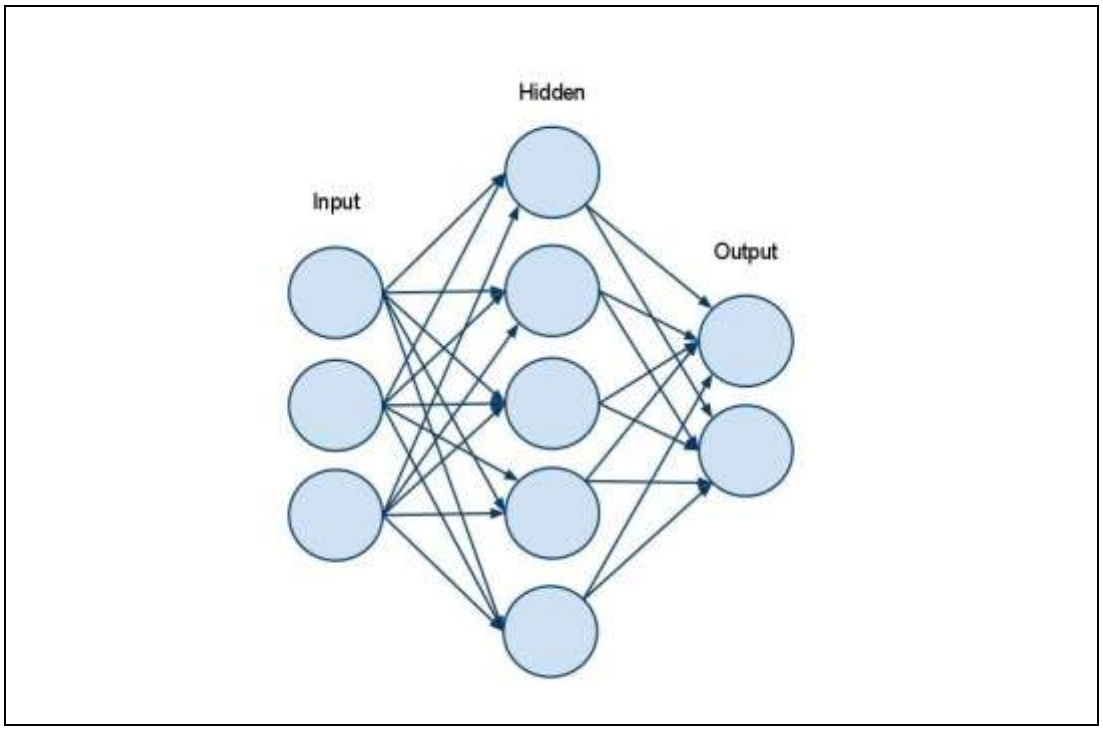

Figure 2.3 A neural network with input, output and one hidden layer

2.2.3 Bayesian Classification

Bayesian classification is a statical classification method based on Bayes Theorem which uses probability to make predictions (Han & Kamber, 2006). Suppose that there are n classes like C_1, C_2, C_3 ,…, C_n. And that there is also an example data X, which has not been classified yet. In such a case; Naïve Bayes classifier sets the class of X to a class C_i which has the highest probability values from given classes. Every data X is shown as feature vectors like X = (X_1, X_2, X_3 ,…, X_m). In Bayes classification every feature has the same importance and each feature is independent from the others. A value of any feature does not contain information about another feature. As an example, the probability of X in class C_i is like in formula (1).

$$P(Ci\ |X) = \frac{P(X|Ci)\ P(Ci)}{P(X)} \tag{1}$$

If P(X) is static for all classes, the probability of X in class C_i can be obtained using the equation $P(X|C_i)\ P(C_i)$ where $P(C_i)$ is the propability of each class as in formula (2).

$$P(Ci) = \frac{Ti}{T} \tag{2}$$

In equation (2), T_i is the number of examples trained in class C_i and T is the total number of training examples. If the priority of class is unknown, then it is assumed that all the classes are equal in the way that $P(C_1) = P(C_2) = P(C_3) = \ldots = P(C_n)$ and for this reason the statement $P(X|C_i)$, is used for finding the probality of X in class C_i. The probabilities $P(X1|Ci)$, $P(X_2|C_i)$, ..., $P(X_m|C_i)$ can be guessed from training samples,

$$P(Xk|Ci) = \frac{Tik}{Ti} \tag{3}$$

in equation (3), T_{ik} is number of tranining data that has the value X_k in class Ci and T_i is the number of training data in C_i. To classify an unknown X, each C_i class is calculated with $P(X|C_i)P(C_i)$ like in formula (4).

$$P(X|Ci) = \prod_{k=1}^{m} P(Xk|Ci) \tag{4}$$

For instance, the class of sample DX can be predicted by the Naïve Bayes classification method using the training documents in Table 2.2 and training dataset in Table 2.3 according to the word list in Table 2.1. Firstly, the training documents in Table 2.2 are converted to training sets as in Table 2.3 in accordance with the word list in Table 2.1.

Table 2.1 Polarity based training word list

bad	beautiful	Confortable	Smart	expensive	nice
antic	uggly	Wordless	Unconfortable	cheap	

Table 2.2 Sentimental classification training documents

Id	**Documents**
D1	It's a very smart idea to buy that cheap and comfortable car.
D2	It is a very beautiful and comfortable car.
*D*3	It is such a bad idea to drive that uncomfortable and wordless car.
D4	My dad will not like that to pay for that antic, ugly and expensive car.
D5	It is a nice car.

Table 2.3 Sentimental classification training set with polarity words

Class	**Document Id**	**W1**	**W2**	**W3**
Positive	D1	Smart	confortable	cheap
Positive	D2	Beautiful	confortable	-
Negative	D3	Bad	wordless	uncomfortable
Negative	D4	Ugly	antic	expensive
Positive	D5	Nice	-	-

Table 2.4 Sentimental classification test document

Id	**Documents**
DX	I do not like wordless, uggly cars.

Table 2.5 Sentimental classification test set with polarity words

Class	**Document Id**	**W1**	**W2**	**W3**
?	DX	wordless	uggly	-

To find the class of DX in Table 2.4 with given testing data set in Table 2.3, the value of $P(X|C_i)P(C_i)$ needs to be maximized. Firstly, the class 'Positive' contains 3 documents and the 'Negative' class contains 2 documents. For each class, the probalities can be calculated as P('Positive') = 3/5 = 0.6 and P('Negative') = 2/5 = 0.4.

Thus, the probability of DX in class ‘Positive’ and ‘Negative’ is calculated as in Figure 2.4.

Positive:

$$P(\text{'uggly'}|\text{'Positive'}) = \frac{(\text{Count of words 'uggly' in class 'Positive'} + 1)}{(\text{Number of words in 'Positive'} + \text{Number of words in all documents})} = \frac{0+1}{6+12} = \frac{1}{18}$$

$$P(\text{'wordless'}|\text{'Positive'}) = \frac{(\text{Count of words 'wordless' in class 'Positive'} + 1)}{(\text{Number of words in 'Positive'} + \text{Number of words in all documents})} = \frac{0+1}{6+12} =$$

$$P(DX|\text{'Positive'}) = P(\text{'Positive'})\left(P(\text{'uggly'}|\text{'Positive'})P(\text{'wordless'}|\text{'Positive'})\right) = \frac{3}{5} * \frac{1}{18} * \frac{1}{18} = 0.0018$$

Negative:

$$P(\text{'uggly'}|\text{'Negative'}) = \frac{(\text{Count of words 'uggly' in class 'Negative'} + 1)}{(\text{Number of words in 'Negative'} + \text{Number of words in all documents})} = \frac{1+1}{6+12} =$$

$$P(\text{'wordless'}|\text{'Negative'}) = \frac{(\text{Count of words 'wordless' in class 'Negative'} + 1)}{(\text{Number of words in 'Negative'} + \text{Number of words in all documents})} = \frac{1+1}{6+12}$$

$$P(DX|\text{'Negative'}) = P(\text{'Negative'})\left(P(\text{'uggly'}|\text{'Negative'})P(\text{'wordless'}|\text{'Positive'})\right) = \frac{2}{5} * \frac{1}{9} * \frac{1}{9} = 0.004$$

Figure 2.4 Probability calculation for document DX

When P(DX | ’Negative’) is greater than P(DX | ’Positive), then the test document DX is in class ‘Negative’.

2.2.4 Support Vector Machines with TF-IDF Values

‘Support Vector Machines’ was first published in 1995 by Cortes and Vapnik (1995). SVM is constructed for compromising between the model's complexities and learning ability with limited sampled data according to risk minimization principle (Vapnik, 1995). The aim of support vectors is to find the best line which divides the

classes into two other sides. Sampled data for Support Vectors (SV) contains the important information for classification.

In SVM, there exists a hyperplane like in Figure 2.5 which divides the samples of each class from each other. The dots on this hyperplane will support the equlity of w.x+b=0, where w is normal to the hyperplane and |b|/||w|| is the vertical distance to the origin. Support vector machine method tries to find the highest distance (margin witdh) between the nearest positive and negative samples to the line. The margin can be calculated like in formula (5).

$$\begin{cases} wx^{+} + b = +1 \\ wx^{-} + b = -1 \end{cases} \rightarrow w(x^{+} - x^{-}) = 2 \qquad (5)$$

$$M = \frac{(x^{+} - x^{-})w}{|w|} = \frac{2}{|w|}$$

When ‘w’ value in the formula (5) decreases, margin width increases with inverse proportion. Every point in hyperplane is shown as x_i. If the training set of two classes are $x = (x_1, x_2, \ldots, x_n)$ and the class values y = (-1, + 1) can be represented as formula (6).

$$\begin{cases} if\ y = +1\ then\ wx_i + b \geq 1 \\ if\ y = -1\ then\ wx_i + b < 1 \end{cases} \rightarrow \text{foreach i } y_i(wx_i + b) \geq 1 \qquad (6)$$

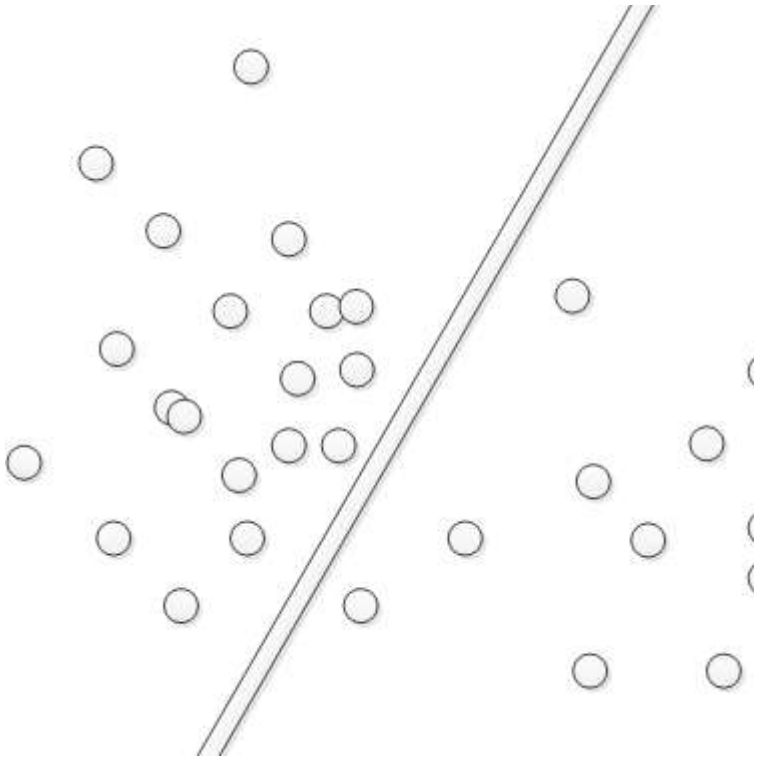

Figure 2.5 SVM sample hyperplane

To classify sample X, firstly the most suitable hyperplane should be found. One side of this hyperplane refers to the positive side and the other side refers to the negative side. Sample X is formalized with support vector machines method and if f(x) function is greater than zero then it is automatically assigned to the positive class and if it is less than zero, it is automatically assigned to the negative class:

- f(x) = sign(wx+b),
- if f(x) >= 0 then assign it to the positive class,
- f(x) < 0 then assign it to the negative class.

To use SVM to classify documents, a vector space style model is used to give each term in a document an identifier id as a dimension and a weight based on it is importance to the document. To calculate weights, the TF-IDF model is used in this thesis.

TF means ‘Term Frequency’ which can be sampled as the count of a word in a document and IDF is ‘Inverse Document Frequency’. TF-IDF can be represented as formula (7).

$$TF = Number\ of\ the\ times\ the\ term\ appears\ in\ the\ document \quad (7)$$

$$IDF = \log\left(\frac{Total\ Number\ of\ Documents}{Term\ Frequency\ in\ All\ Documents}, 2\right)$$

$$TFIDF = TF * IDF$$

After the TF-IDF values were calculated, they were used as vector values in SVM method and the classification was done in light of this method.

CHAPTER THREE

WEB MINING AND SENTIMENTAL CLASSIFICATION: STUDIES AND ISSUES

3.1 Web Mining Studies

Medelyan & others (2009) made a study and an application which mines meaning data from Wikipedia that contains 18GB of English written text data. In the study, they firstly analyzed the structure of Wikipedia and they figured out some basics about Wikipedia as follows:

- Wikipedia uses WikiMedia software which can provide data in XML format,
- Wikipedia contains information as an encyclopedia, a thesaurus, a database, an ontology, and a network structure,
- Wikipedia contains several parts for explanation of a topic which are articles, disambiguation pages, hyperlinks, category structure, templates and info boxes, discussion pages and edit histories.
- Unlike WordNet, it is not fully lexicon resource; it has behavior of human language too.

After analyzing the Wikipedia content structurally and ontologically, they reviewed a lot of literature about various topics. The studies done on Wikipedia content are following the direction of a semantic web mining approach.

Alvarez & others (2007) studied for an approach based on web content mining and web structure mining to gather information on HTML based text data. And they presented a page creation model method to automatically detect underlying web structure of HTML pages and extracting meaningful data from web pages. Their method requires a single page with inputs which then can automatically extract the attribute values of each data record using multiple-string algorithm. The approach was validated with real web page data.

In another study of web mining approaches, a music information system was derived using web content mining (Schedl, Widmer, Knees, & Pohle, 2008). The system contained more than 600,000 information about music artists. They built relations between artists according to similarity and prototypically. For defining this similarity they used co-occurrence analysis. The study also covers auto-tagging of artists based on DF, TF, and TF-IDF vectors.

3.2 Sentimental Classification Studies

Sentimental classification studies were divided into two sub titles which are ‘Lexicon Enhanced Sentiment Classification’ and ‘Machine Learning Enhanced Sentiment Classification’.

3.2.1 Lexicon Enhanced Sentiment Classification

Turney (2002) introduced a simple unsupervised learning algorithm to rate comments like thumbs up or down. Firstly, he extracted adjectives and adverbs in phrase, then calculated semantic orientation of each phrase and classifies them according to the average semantic orientation phrases. In this study, the average accuracy was 74%. However, in movie reviews he got 66% accuracy, for the bank and automobile reviews he got accuracies between 80% and 83%.

An approach was created by Dang, Zhang, & Chen (2010) at the University of Arizona. In this approach, they combined machine learning and semantic-orientation approaches into one framework in order to achieve an improvement in the performance of classification. They firstly took data from a U.K. product price comparison website where people gave votes for online summaries of products and companies. They also took data from a U.K. based camera review website which has a rating system from one to five for each product. In their research, they used some HTML parsing software to mine and parse data. Then, they added gathered data into a relational database (RDB). In the study, three different features, which were content free, content specific and sentiment features, were used. After getting the

documents into RDB, they used a POS tagger (Toutanova & friends, 2003) to find adjectives, adverbs and verbs from the documents. Having tagged these words, SentiWordNet was used to figure out the sentiment scores. Then, they calculated the score of each word and compared the positive negative score of each word to conclude their experiments. As a result, the combination of three features passed all other accuracy scores and in digital camera reviews they got 83.3% accuracy.

Another research by Yessenalina, Choi & Cardie (2010) at Cornell University generates annotator rationales for replacing human resources. Research was done at document level sentiment classification to improve performance using automating annotator rationales. They used movie review data to compare no annotator rationales, human annotator rationales and automatic annotator rationales. At the end of this study, they got an accuracy of 92.5% which was significantly better than a human based one which had an accuracy of 91.61%.

In a paper by Lu, Kong, Quan, Liu, & Xu (2010) from China, the sentiment analysis method was used to explore the strength of sentiments. In the study, they firstly extracted the opinion phrase using POS tagging and they took noun words nearby adjectives with adverbs as sentiment features. Secondly, sentiment strength was calculated with multiplication of adjectives' strength and adverbs' strength in the document. Finally, they divided the dataset into five parts by using a five-fold cross validation method. As an experiment they used a hotel five-scale review dataset from a Chinese Hotel Review website. Then, they compared the cross validation of results with sentiment strengths and they achieved a 71.65% precision in their approach. According to the results, their experiment reached a better performance based on efficiency and precision values without considering the adjective sentiment strength.

3.2.2 Machine Learning Enhanced Sentiment Classification

Pang & friends (2002) studied on three different machine learning algorithms to classify reviews. The algorithms are ‘Naïve Bayes’, ‘Support Vector Machines’ and ‘Maximum Entropy’. They made their test on a movie reviews dataset which

contained 752 negative and 1301 positive dataset based on 144 reviewers. The results of the study showed that machine learning based methods are better than human generated baselines. When looked at the machine learning perspective, SVM approach gave the best results and the Naïve Bayes the worst results among these three methods in terms of relative performance.

Another work by Pang & Lee (2004) proposed that the machine learning based text categorization techniques be applied only to the subjective part of document. They gave the relation between polarity and subjectivity on documents. They showed subjectivity part of review still contains polarity data. With the study, they proved that including only subjective portion of documents gave more efficient results for sentiment analyses. Moreover, by using minimal-cut-framework, the accuracy of sentimental classification was improved.

A paper by Agarwall & friends (2008) from Indian Institute of Technology Kanpur and General Motors Technical Centre India uses linguistic resources to determine sentiments on sentence level. The paper used movie, car and book reviews to examine their results with support vector machines. They firstly, separated the sentences as either subjective or objective. Then, the subjective sentences were classified as positive or negative. They tested both unigrams and N-grams (bigrams and trigrams) as feature. In the experiments, N-grams gave a better accuracy, though it had a complexity drawback. This experiment was a bit different from the other ones because it did not put away the stop-words; it was using stop-words to understand the strength of the adjectives and adverbs. In the final experiment for sentiment analysis they combined conjunctions with n-grams as a feature and they got a 92.7% precision and a 92.5% recall using car reviews data.

Thet & friends (2008) made a study using SVM to classify polarity of movie reviews. They used multiple perspectives on review dataset and compared all the results and found the best accuracy states on three different datasets. They got the best accuracy when they applied term weighing, stemming, negation and removal of

stop words methods together. They got the worst accuracy when they used adjective terms only.

CHAPTER FOUR

PROPOSED APPROACH

4.1 General Structure of Proposed Approach

Proposed approach uses its internal components and external components to complete web mining and classification tasks. The proposed approach contains nine basic components. These components are database server, web server, crawling web service, 'Title Miner' web service, Turkish spell-checker web service, sentimental classification web service, Google language detecter API web service, Zemberek post-tagger, stemmer web service for Turkish language, and lastly the clients as shown in Figure 4.1. Although, each component has a different task in the system, all components except database server are connected to internet to complete their tasks. The definitions of the components, where six of the components are internal components and the rest three components are external components of the system, are given in the subtitles of Chapter 4.1. Moreover, in Chapter 4.3, Algorithms and Flow Charts, how these components works systematically to complete their tasks interoperability with others is explained.

4.1.1 Internal Components

First of all, each internal web service has the ability to serve data in XML, JSON data and multi-line text formats; thus, making the proposed approach accessible globally from near all internal and external requests. With this feature, each component of the system is able to communicate with each other in the desired data format.

Database server is the component to keep the data in secure and organized way to respond to queries to access stored data when needed. In this approach every customer has one dedicated database. For security reasons the database server is only connected to the web server and completes requests over the web server.

The web server component is the core component for connecting the database server, sending queries to the database server and getting results from the database server to serve them to the clients and the crawler web service.

Crawling web service is the most important component as it makes the major web content mining and web structure mining operations. It simply crawls the web pages; it uses other components of the proposed approach to prune clean data and then serves the data in the desired text-data formats like XML, JSON data, and plain text.

'Title Miner' web service is just used for getting the exact title of the document on web pages. It uses both web content mining and web structure mining to complete its task. This component is used to gather the document title of blog pages from various blog web sites. It simply learns the exact title of the web page and responds to it in a desired format.

Turkish spell-checker web service, which is another internal component of the system, reads the Turkish text data and makes spell-checking on it and if any of the words is misspelled, corrects the word and responses it in desired formats as in the crawling web service.

Sentimental classification web service is the second most important component of the proposed system. It gets training data and the new document to make a prediction about the new documents sentimental class. It uses two different algorithms to define the class of a document.

4.1.2 External Components

Google language detection API web service which is an external component of the system is used to check whether the text data is in Turkish language or not. Google language detection web service simply gets the text query and responses the result language prediction in JSON data format.

Zemberek web service, which is distributed as open source by TUBITAK, is a Turkish pos-tagger and stemmer (Akın, 2006). Like all stemmers, Zemberek offers finding the root of the word given. By default, it comes with a Java server and a Java console options. In this approach, Java web server version is used with a small modification in code to get responses in JSON data format.

Since this thesis focuses on the classification techniques based on the supervised learning, the clients of the proposed approach are the supervisor of their data. Clients are also the watcher for feedbacks using the web server component of the proposed approach.

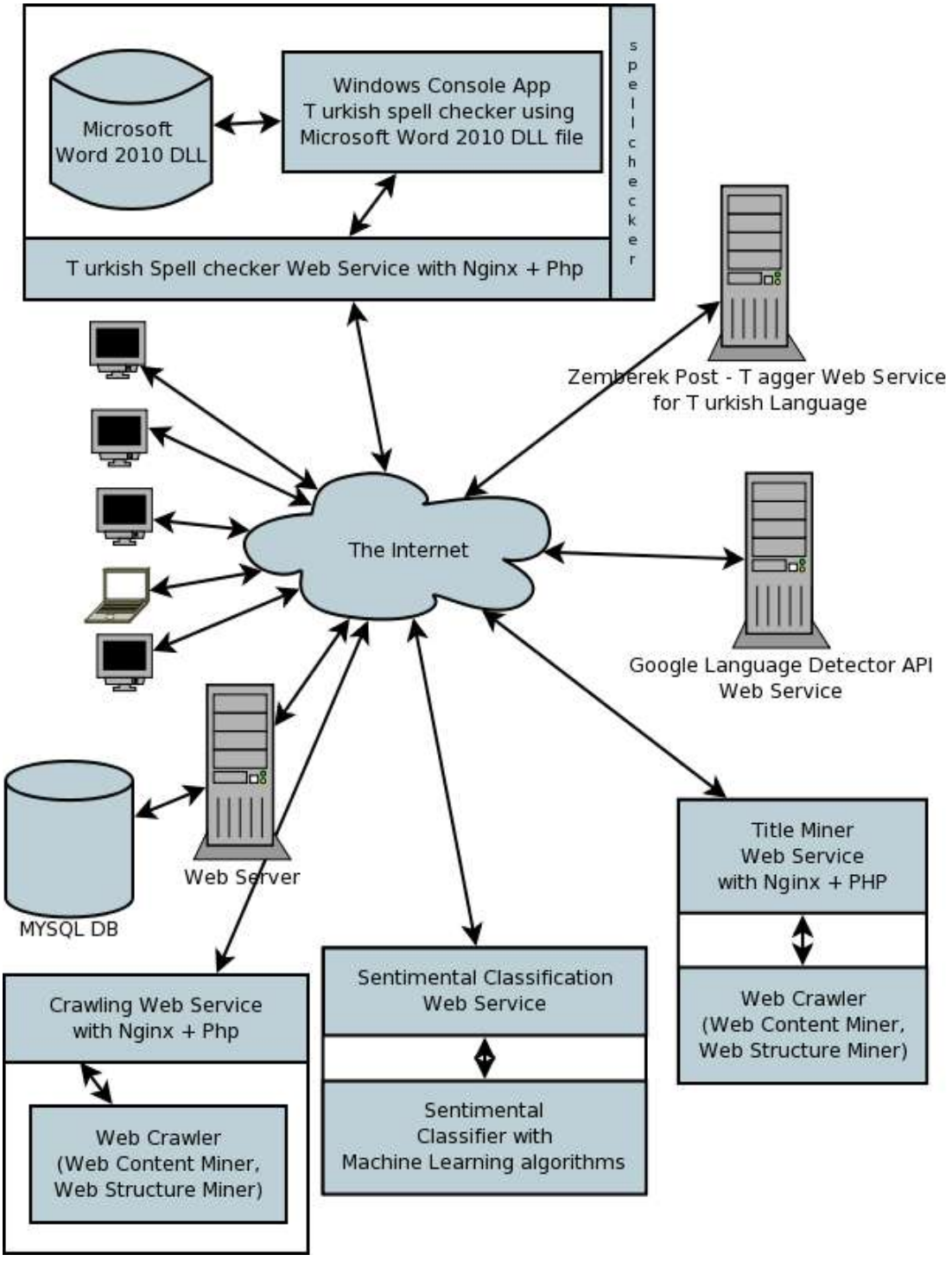

Figure 4.1 General structure of proposed approach

4.2 Database Model

The database model is designed for a dedicated usage which means that for every new customer the database should be created at like in Figure 4.2. The dedicated model contains six tables. The web mining and classification results are directly stored to the 'docs' table. Other tables are generally used for moderating and administrating storages. All tables in the schema are explained with definitions of columns in Table 4.1 – 4.6.

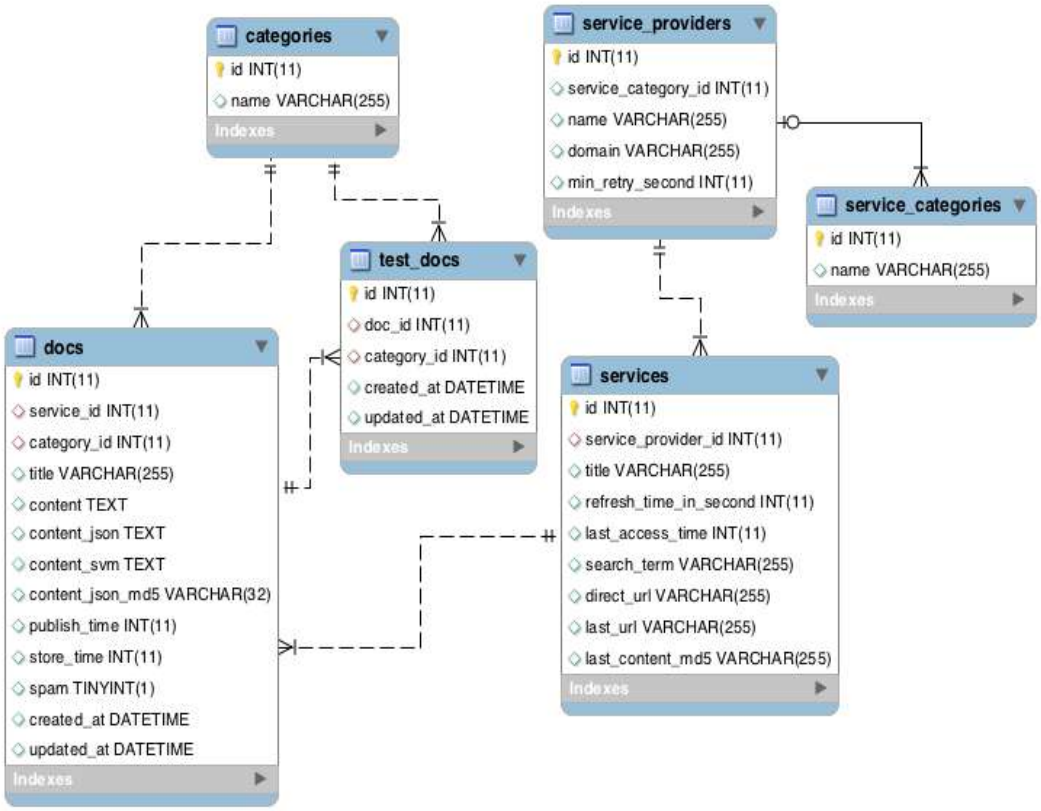

Figure 4.2 Database schema

Table 4.1 Definitions for database table ‘service_categories’ columns

Column	Definition
Id	Primary Key (unique idenfier for service category)
name	Shown name for service category

Table 4.2 Definitions for database table ‘service_providers’ columns

Column	Definition
Id	Primary Key (unique idenfier for service provider)
service_category_id	Foreign Key for ‘service_categories id’
name	Shown name for service provider
domain	The name of the service provider forexample: facebook.com
min_retry_second	The minimum time in seconds to refresh service usage

Table 4.3 Definitions for database table ‘services’ columns

Column	Definition
Id	Primary Key (unique idenfier for service)
service_provider_id	Foreign Key for ‘service_providers id’
Title	User defined title for service
refresh_time_in_second	User defined service access time; must be greater or equal to service provider’s min_retry_second
last_access_time	Last access time for the service in unix time format
search_term	User defined keyword for quering the service
direct_url	Direct access page link for the service content. For instance, it is considered for static links like a Wikipedia link about search_term
last_url	Last access page link for the service.
last_content_md5	It is used as an identifier for the latest document gathered wih this service. It helps not to make a duplication of the same content from this service.

Table 4.4 Definitions for database table ‘categories’ columns

Column	Definition
Id	Primary Key (unique idenfier for catgeory)
name	Sentimental classification category name: ‘pos’ and ‘neg’

Table 4.5 Definitions for database table ‘docs’ columns

Column	Definition
Id	Primary Key (unique idenfier for service)
service_id	Foreign Key for ‘services id’
category_id	Foreign Key for ‘categories id’
Title	Document title
content	Document content without html tags
content_json	The untouched gathered content from a service in JSON data format
content_svm	Clean content for machine learning algorithms
content_json_md5	An identifier to prevent spam from same users post
publis_time	Document publish time in unix time format
store_time	The unix time when this document fetched from service
spam	Boolean idenfier for checking spam content
created_at	Datetime stamp for keeping when this entry was created
updated_at	Datetime stamp for keeping when this entry was updated

Table 4.6 Definitions for database table ‘test_docs’ columns

Column	Definition
Id	Primary Key (unique idenfier for service)
doc_id	Foreign Key for ‘docs id’
category_id	Foreign Key for ‘categories id’
created_at	Datetime stamp for keeping when this entry was created
updated_at	Datetime stamp for keeping when this entry was updated

4.3 Flowcharts of Proposed Approach

As described in Section 4.1, the proposed approach has four internal components and three external compenents that work in cooperation with each other. In this section, the flowcharts and algorithms behind the scene are shown. First of all, in Figure 4.3, a very general flowchart, which shows how the system works properly, is

given. Then, the other flowcharts and algorithms are given for a deep investigation of the proposed approach.

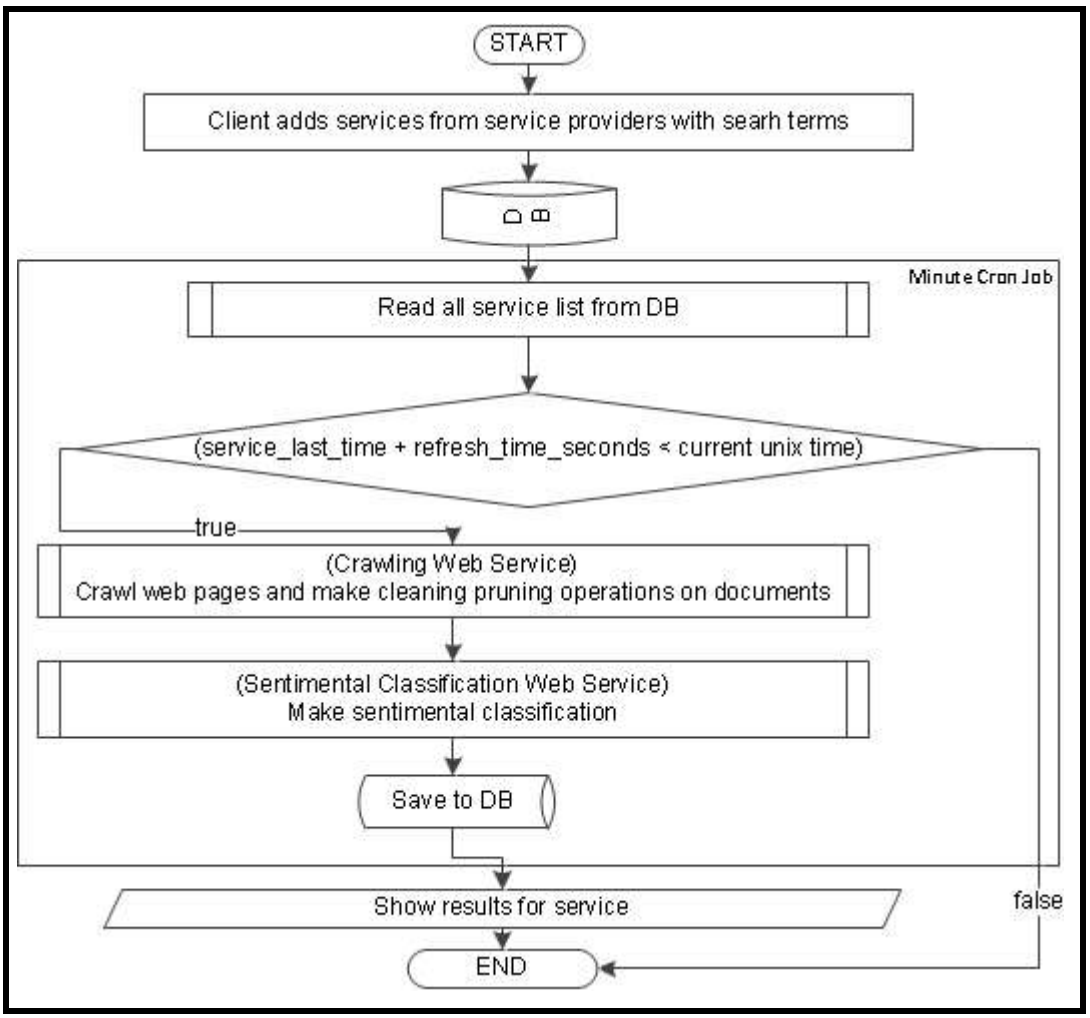

Figure 4.3 General flow chart of the proposed approach with association of client with cron service

In the next flowcharts, every internal service is deeply investigated with the order given in the Figure 4.3. First of all, 'Crawling Web Service' is shown with details in Figure 4.4. Then, in Figure 4.5, flowchart of 'Title Miner' web service, in Figure 4.6, pseudo code of 'Title Miner' web service, and in Figure 4.7, flowchart of Turkish spell-checker web service are given. After those, in Figure 4.8, generating model with training data and sentimental classification with Support Vector Machine and in Figure 4.9, sentimental classification with Naïve Bayes are given separately.

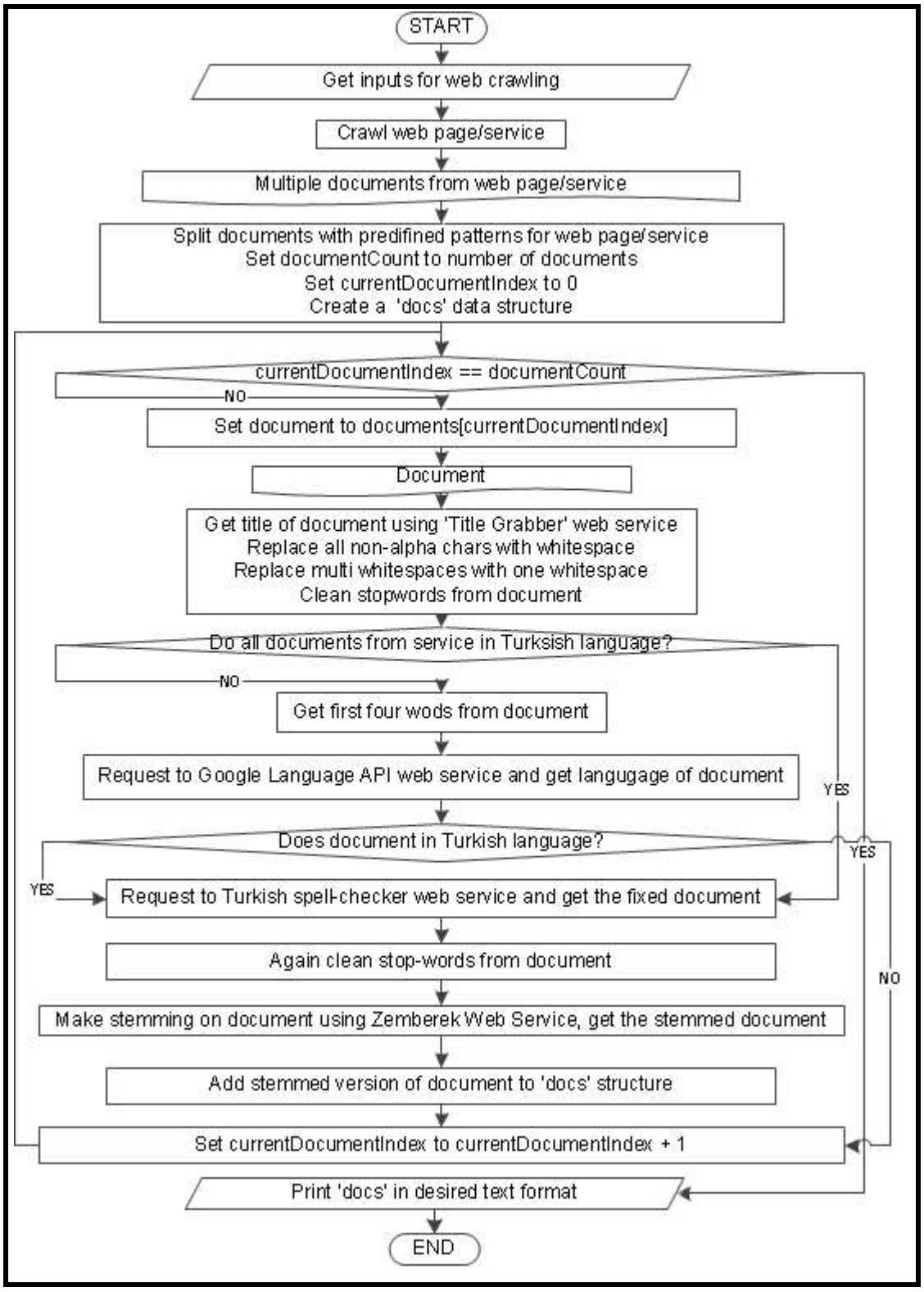

Figure 4.4 Flowchart of web crawling web service

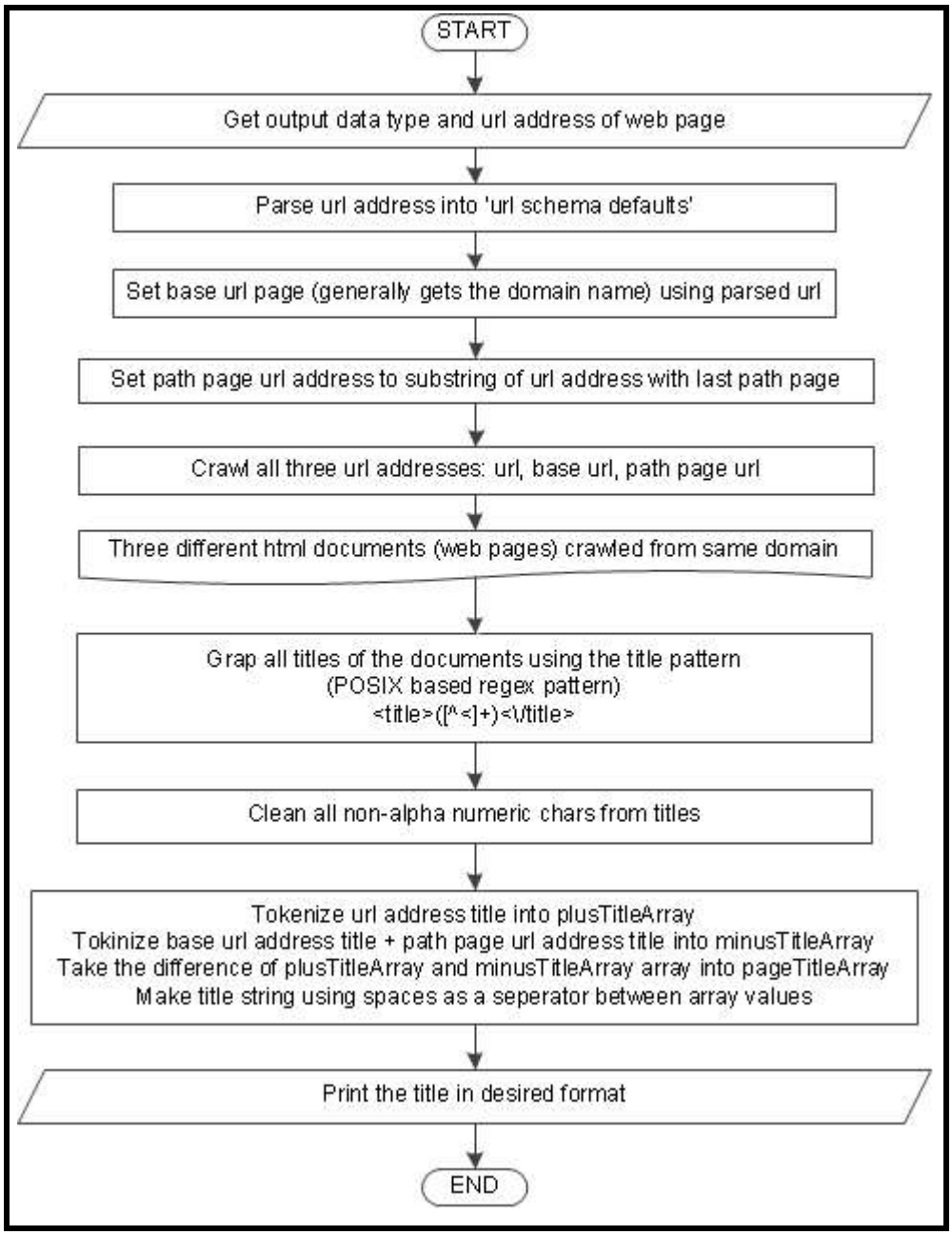

Figure 4.5 Flowchart of 'Title Miner' web service

The pseudo code of 'Title Miner' web service is given in Figure 4.6, whereas the flowchart of 'Title Miner' web service is given in Figure 4.5. The pseudo code will help to implement a similar application in a desired programming language using web mining and web structure mining techniques.

```
START
  SET dataType = url_decode( HTTP_REQUEST['dataType'] )
  SET url = url_decode( HTTP_REQUEST['ur'] )
  SET parsedUrl = url_parse ( url )
  SET baseUrl = parsedUrl[schema] + '://'+parsedUrl[host]
  SET pathPageUrl = substring ( url, 0, strrpos (url, '/') )
  IF pathPageUrl + '/' == url THEN
    SET pathPageUrl = substring ( url, 0, strrpos (pathPageUrl,
'/') )
  END IF

  SET minusTags = ARRAY()
  SET plusTags = ARRAY()
  IF baseUrl != url THEN
    // fetch url content
    SET baseUrlContent = get_url_content (baseUrl)
    SET baseUrlTitle= pattern_match(baseUrlContent)
    minusTags.append ( tokenize (baseUrlTitle) )
  END IF

  IF pathPageUrl != url THEN
    // fetch url content
    SET pathPageUrlContent = get_url_content (pathPageUrl)
    SET pathPageUrlTitle = pattern_match(pathPageUrlContent)
    minusTags.append ( tokenize (pathPageUrlTitle) )
  END IF
  // fetch contents
  SET urlContent = get_url_content (url)
  SET urlTitle = pattern_match(urlContent)
  plusTags.append ( tokenize (urlTitle) )
  SET newTags = plusTags - minusTags
  print_format( un_tokenize (newTags) , dataType)
END
```

Figure 4.6 Detailed pseudo code of 'Title Miner' web service

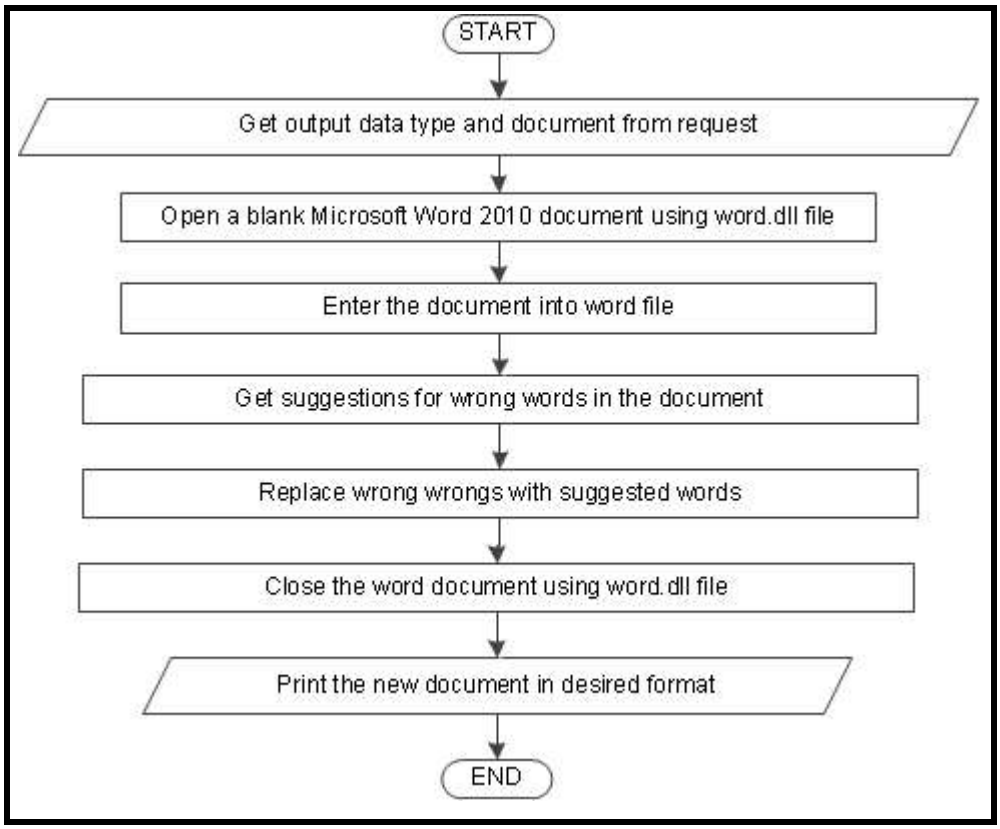

Figure 4.7 Flowchart of 'Turkish spell-checker' web service

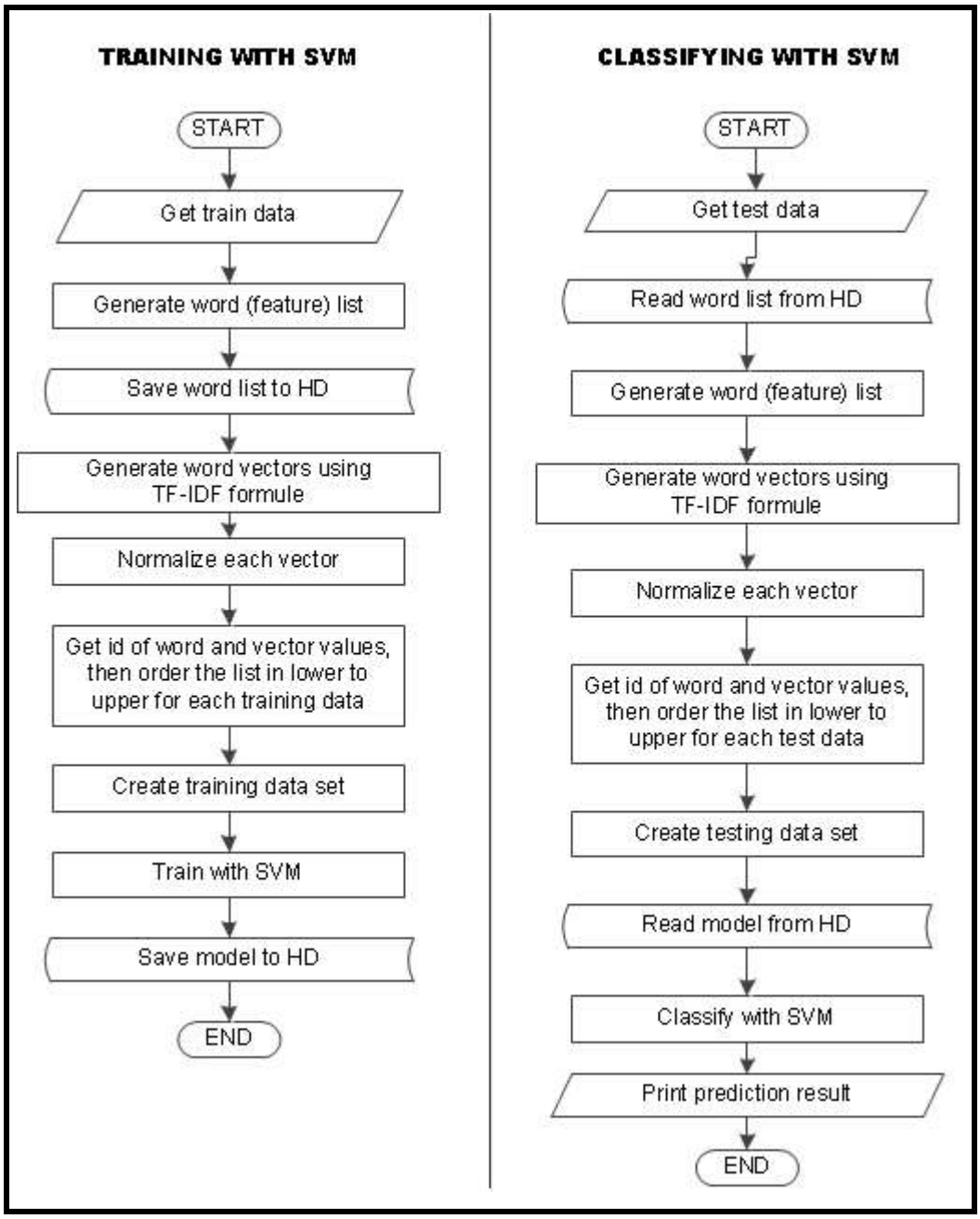

Figure 4.8 Flowchart of training and sentimental classification with Support Vector Machines (SVM)

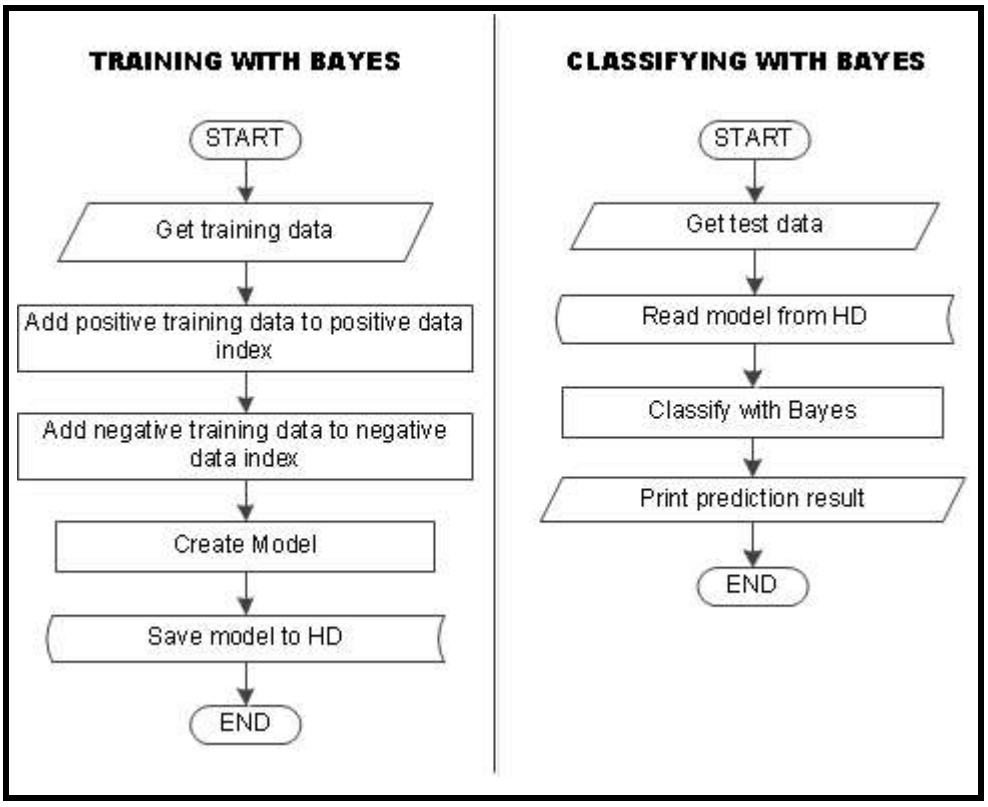

Figure 4.9 Flowchart of training and sentimental classification with Naïve Bayes

4.4 Technologies behind the System

Although the complexity of the proposed approach makes it hard to implement, the technology behind the system makes it easier for implementing and applying it almost costlessly for educational purposes. The choices from components are generally done in opensource ways. The technologies behind the system are as follows.

Eventhough, the majority of tasks can be done at any UNIX based operating system; Windows OS is needed for spell-checking operations. Basically, at UNIX based OS, the system has Nginx web server, Mysql Database Server, PHP programming CLI and PHP-CGI for serving the dynamic content, Java for Zemberek stemmer and SVM-light for support vector machine calculation. At Windows side, Nginx web server, PHP-CGI and C# programming language for interacting with Microsoft Office 2010 Word library are used. For data serializing, de-serializing and

interconnections of each module with other modules, JSON data, XML data and multi-line text formats are used.

For the web server side, open source web server Nginx, which is a new age, light-weight, high performance web server that handles large amounts of traffics with small amounts of RAM, is chosen. It can be used in both Windows OS and UNIX based OS to serve content over the Internet.

For the database operation, Mysql Database Server is choosen. It is an opensource and it supports relational databases with a high performance.

For web based programming, PHP language is used as it has C language like syntax web based language. Furthermore, it offers easy content fetching function with one line code and supports POSIX based pattern maching with regular expressions.

Java programming language is used to add JSON based output functionality to Zemberek pos-tagger and stemmer for Turkish language. With the benefit of Java framework, the stemmer software both works at Windows and UNIX based OSes without any problems.

SVM-light (Joachims & others, 1999) is C language based implementation for support vector machine trainer and classifier which is opensource and freely available for scientific usage purposes.

From Windows side, it is used Windows Server 2003 OS at a virtual machine and Microsoft Word 2010 Home and Student edition trial version. .Net framework based C# programming language is choosen for communicating with the Word library. On Windows operating system for serving data on the WWW, Nginx web server and PHP-CGI interpreter are used.

For serializing, deserializing and communicating between components, JSON data and XML data are used. Both JSON and XML data include a set of rules for encoding data in machine-readable formats. Moreover, they also support Unicode encoding for serving data which makes all characters are more readable for Non-Ascii character based languages. As the Turkish language has six non-ascii characters (ç, ğ, ı, ö, ş, ü), it becomes problematic when it is not used in Unicode chars headers like UTF-8. Thus, both XML and JSON formats handle this problem and give a loseless communication between modules.

CHAPTER FIVE

SENTIMENTAL FEEDBACK MINER APPLICATION & EXPERIMENTS

5.1 Sentimental Feedback Miner

Sentimental Feedback Miner (SFM) is a hybrid application that uses web mining technologies for discovering patterns on WWW and uses machine learning classification techniques for analyzing the discovered documents semantically to find out the polarity of a document. SFM has the ability to gather data from most common Internet data files like XML data, JSON data and HTML data. Moreover, for functionality it can convert XML and JSON data formats into dynamic array structures for fast data access purposes. Also, for HTML data types, it has pattern discovery and gathering mechanism to push data from HTML text to dynamic array structures.

SFM enables its clients to watch their valuable feedbacks from various internet resources. Thus, it supplies semantic social media monitoring opportunity to its clients. Furthermore, it answers the five questions in Chapter 1.2 for its clients as listed below:

1. Clients are able to see how many posts are written in the given time periods like today, yesterday, this month, and also handles listing for choosen date ranges with given search keywords,
2. Clients can see the content of posts,
3. With SFM's categorized service providers, clients can see the service providers with content,
4. If the main content of the page containes information about the author of the cotent, SFM provides this information to clients,
5. If a sufficient number of test docs are supplied by the clients using the documents provided by the system, SFM can automatically classify the documents sentimentally by using machine learning algorithms.

In addition to these five answers, SFM can handle multiple categorized searches and sentimental classifications over the gathered documents. Every client is responsible for creating their own training data for classification. Thus, it supplies an efficient training-set for client-based classification. With the dynamic mechanism of SFM, a new database for each client is created. Thus, polarity training data does not confuse other clients' training data and enable clients to train their data for themselves.

SFM has two control panels which are the admin panel and the user panel. The admin panel is basically used for manipulating datas whereas; the user panel is generally for displaying data.

By using the admin panel with administrative privileges, users are able to see / add / delete operations over services, test documents and documents. However, they are not able change the predifined service providers. Only the super admin user can add / delete service providers and sentimental classification categories. Preview of the admin panel is given in Figure 5.1 with super admin rights and in Figure 5.2 with admin rights.

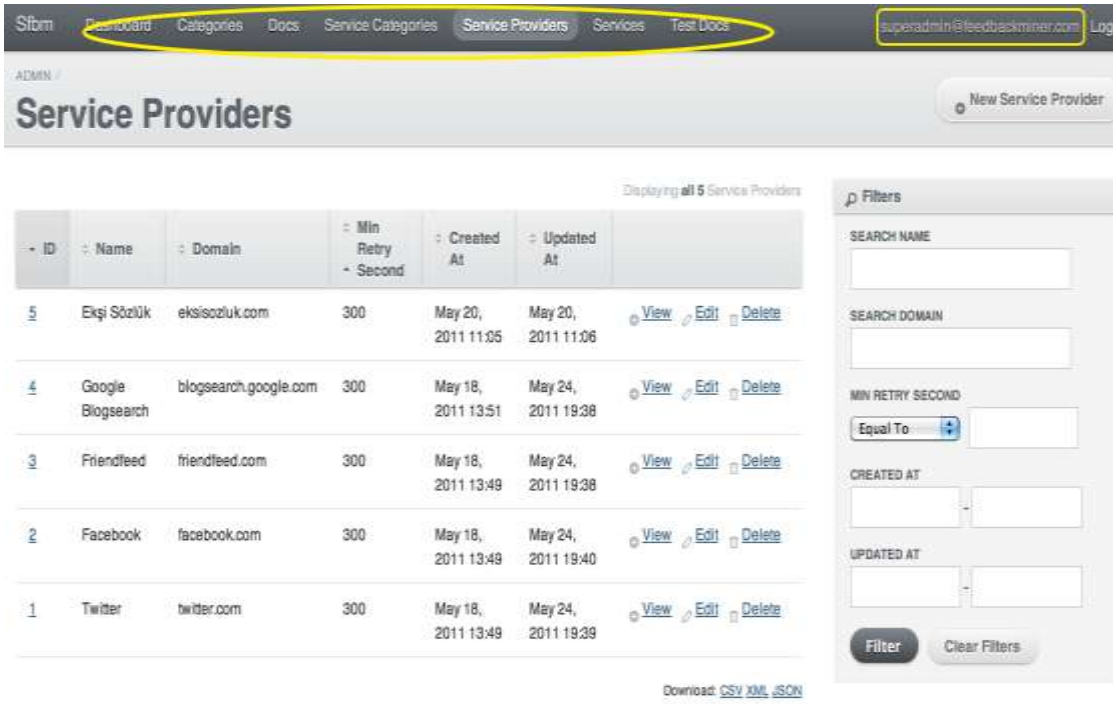

Figure 5.1 SFM preview from the admin panel with super admin privileges

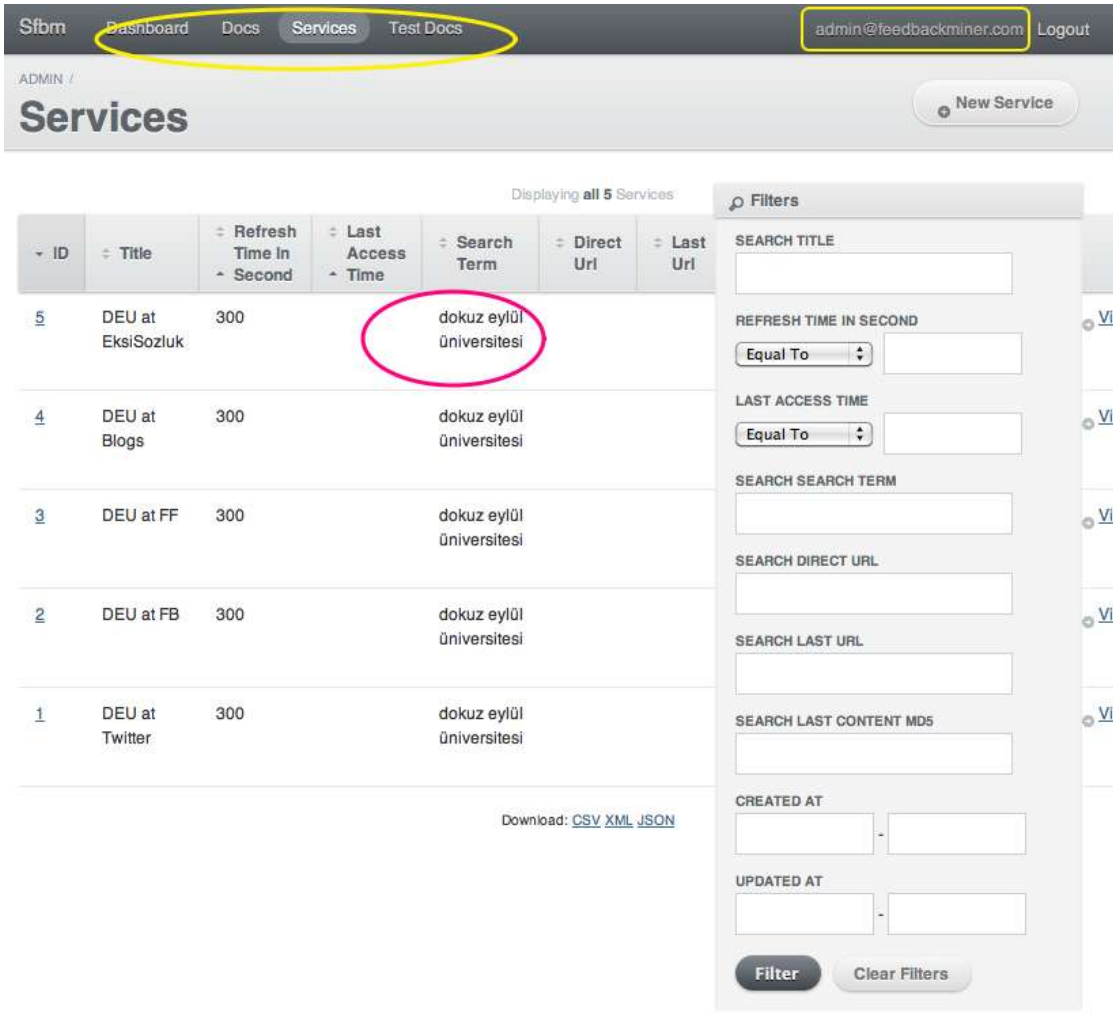

Figure 5.2 SFM preview from the admin panel with admin privileges

The user panel is just for displaying data and adding test documents automatically for classification. Other options such as adding/deleting/editing service and deleting documents are disabled for the user panel to prevent system from end-users. In Figure 5.3, a preview from the user panel is given.

Figure 5.3 SFM preview from the user panel

5.2 Experiments and Results

Experiments are divided into three sub titles which are 'Web Mining Experiments', 'Text Manupulation Experiments' and 'Sentimental Text Classification Experiments'.

5.5.1 Web Mining Experiments

Web mining experiments starts by getting content from a web page and goes on with the pattern discovery processes on the gathered documents. For the experiments, the five most popular text-based social media platforms in Turkey according to Alexa (2011) records were chosen. Facebook, Friendfeed, and Twitter were chosen for the micro-blogging category. Google Blogsearch was chosen for finding

documents from Turkish written blogs and lastly, a popular discussion board from Turkey Ekşisözlük was choosen for the general discussion category. From Table 5.1 to 5.5, the definition and ontologic answer for web mining from chosen web sites are given.

Table 5.1 Definition and ontologic answers for Facebook

Question	**Answer**
Service supplier?	Facebook.com (http://www.facebook.com/)
Meta tag definition?	'Facebook is a social utility that connects people with friends and others who work, study and live around them. People use Facebook to keep up with friends, upload an unlimited number of photos, post links and videos and learn more about the people they meet.'
Web service API support?	It has API support.
What is API url for any search term?	https://graph.facebook.com/search?q={QUERY}
Supported output formats	JSON

Table 5.2 Definition and ontologic answers for Friendfeed

Question	**Answer**
Service supplier?	Friendfeed.com (http://www.friendfeed.com/)
Meta tag definition?	'FriendFeed enables you to discover and discuss the interesting stuff your friends find on the web.'
Web service API support?	It has API support.
What is API url for any search term?	http://friendfeed-api.com/v2/search?q={QUERY}
Supported output formats?	JSON, XML(ATOM)

Table 5.3 Definition and ontologic answers for Twitter

Question	**Answer**
Service supplier?	Twitter.com (http://www.twitter.com/)
Meta tag definition?	'Instant updates from your friends, industry experts, favorite celebrities, and what's happening around the world.'
Web service API support?	It has API support.
What is API url for any search term?	http://search.twitter.com/search.{FORMAT}?q={QUERY}
Supported output formats?	JSON, XML(ATOM)

Table 5.4 Definition and ontologic answers for Google Blogsearch

Question	**Answer**
Service supplier?	Google.com (http://blogsearch.google.com/)
Meta tag definition?	'Google Blog Search provides fresh, relevant search results from millions of feed-enabled blogs. Users can search for blogs or blog posts, and can narrow their searches by dates and more.'
Web service API support?	It has API support.
What is API url for any search term?	http://blogsearch.google.com/blogsearch_feeds?q={QUERY}&output=atom&hl=tr
Supported output formats?	HTML, XML(ATOM,RSS)

Table 5.5 Definition and ontologic answers for Ekşi Sözlük

Question	**Answer**
Service supplier?	Eksisozluk.com (http://www.eksisozluk.com/)
Meta tag definition?	'Sözcük ve terimler konusunda kullanıcıların subjektif sunumlarıyla genişletilen sözlük. (A dictionary with users' subjective posts for keywords and terms.)'
Web service API support?	It does NOT have API support.
What is HTML url for any search term?	http://www.eksisozluk.com/show.asp?t={QUERY} &kw=&a=&all=&v=&fd=&td=&au=&g=&p=1
Supported output formats?	HTML

As seen on Table 5.1 to 5.5, four of the service providers are supporting at least XML or JSON formats for response which makes available data in structured array for PHP programming languge in easy way. But for the last one 'Ekşi Sözlük' only responses in HTML fornat which makes it a bit hard to gather data into a structured array format. If a service supports both XML and JSON, for this application it is choosen response data in JSON format, because PHP has its own one line deserializer function to get data into structure array for JSON. Before going on with the pattern discovery, a few lines of partial response from each service in Figure 5.4 for the search query 'dokuz eylül üniversitesi' or ASCII formatted one 'dokuz eylul universitesi' depending on service provider are shown.

Figure 5.4 Partial contents from five service providers for ‘dokuz eylul universitesi’ keyword group

As seen in Figure 5.4, the contents are not generally human readable, without browsers like Mozilla Firefox, Internet Explorer, Google Chrome, Safari, etc… It is not easy to see the content in a tidy way. However, for computers, after discovering

patterns with web content mining and web structure mining techniques, it is easy to make this various formatted data-list into one well-structured data format and use it at various services. A sample of output provided by the web crawler of SFM can be seen in Figure 5.5 with three data formats. All the data coming from any source are shown in the same data structure, which makes it the machine communicator for other services. The ‘item’ nodes are hidden because they all have more than one node, but for the preview, a sample ‘item’ node is shown in Figure 5.6.

Figure 5.5 Web crawler responses from five services in XML format

Figure 5.6 An example item node from web crawler service responses

'Title Miner' is another web mining approach that finds the exact document title of a web page document. An example of title attribute from html tags is given in Figure 5.7. However, Meta title contains both the document title and the web page information which is unrelated with the document. In the example, 'Türk Telekom’dan denizci öğrencilere jest' is the exact document title and the rest '| B KAPISI …Dergisi' contains information about the web page.

```
<!DOCTYPE html>
<html>
<head>
<meta charset="UTF-8" />
<title>Türk Telekom&amp;#8217;dan denizci öğrencilere jest | B KAPISI
&amp;#8211; Sivil Havacılığın Online Dergisi</title>
```

Figure 5.7 A partial html response contains title of web page

'Title Miner' web service simply crawls different pages from the given the URL address and tries to predict the given URL address' exact title. It uses web content mining to get title of each crawled web page and uses web structure mining to analyze the linking structure of the given URL address.

To test the accuracy of the ‘Title Miner’, fifty blog pages from various blog posts were used. After the titles were gathered with ‘Title Miner’ web service, they were controlled manually and the results were compared. The results were very satisfactory; from forty-seven of the fifty pages given, the desired document title was achieved. Accuracy is calculated with formula (8) where the accuracy (A) for ‘Title Miner’ web service is 47/50 = 0.94, which can be represented as 94%.

$$A = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ test\ documents} \tag{8}$$

5.5.2 Text Manupulating Experiments

For text manipulating experiments, one document from each service provider, which was defined in Chapter 5.5.1, was taken. The original five documents with document identifiers can be seen in Table 5.6. After collecting the documents in text data format, four more steps were applied to these five documents to make them efficient for classification techniques as follows:

1. Removing non-alpha characters and Turkish stop words from the documents. After the pruning operation the text data was formatted like in Table 5.7.
2. Turkish spell-checking and correction with suggestions were applied to the documents using ‘Turkish spell-checker web service’ as seen in Table 5.8.
3. Stemming operation was applied using ‘Zemberek Pos-tagger, Stemmer Web Service’. The results can be seen in Table 5.9. For some words, the ‘-n’ tag after the word was added, this means it has a negative suffix in Turkish language.
4. Lastly, to be sure about deleting the stemmed stop-words, a second stop word removing operation was done over the documents that can be seen in Table 5.10.

Table 5.6 Original - Five example documents from web crawler responses for experiments

Id	Document
D1	Dokuz Eylül Üniversitesi Kamu Yönetimi Başkanlığı'nı yürüttü. Hak-İş ile Müsiad'ın siyasi danışmanlığını yaptı. Değişik gazete ve dergilerde, Türkiye'nin siyasi ve idari sorunlarına ilişkin yüzlerce araştırması yayınlandı.'
D2	'ÇAMAŞIR SUYUNDAKİ TEHLİKE Kadınların sık kullandığı çamaşır suyu da çok zararlı. Dokuz Eylül Üniversitesi Çevre Mühendisliği Bölümü Öğretim Üyesi Doç. Dr. Mustafa Odabaşı, bunların kanser riskini önemli ölçüde arttırabileceğini söylüyor. Çamaşır suyu içeren ürünlerin, amonyaklı veya asidik (tuz ruhu, kireç çözücü gibi) temizlik maddeleriyle karıştırılması zehirli gazların (klor gazı ve klor aminlerin) açığa çıkmasını sağlıyor, ortamdaki oksijeni durduruyor ve insanları nefes alamaz hâle getiriyor. Bu tür zararlı maddelerin aşırı teneffüs edilmesi hâlinde solunum yolları ve akciğerde tahribata yol açıyor. Odabaşı araştırma sonuçlarını şöyle değerlendirdi: "Piyasadaki çamaşır suyu içeren temizlik ürünlerinin sayısı gün geçtikçe artıyor. Katkısız, parfümlü, deterjan katkılı koyu kıvamlı sıvı, jel, ovma tozu, sprey gibi bir çok ürün Türkiye'de ve dünyada yaygın olarak kullanılıyor. Geçtiğimiz yıl yapılan bir araştırma, bu ürünlerin ülkemizdeki her 100 evden 85'inde kullanıldığını, hane başına yıllık tüketimin ise 3 kilograma ulaştığını gösteriyor. Çamaşır suyu içeren temizlik ürünlerinin kullanımında dikkatli olunması gerekiyor.
D3	Dokuz Eylül Üniversitesi akademik üstünlüğünü yayınlarıyla başarılarıyla kanıtlamış Türkiye'nin en iyi üniversitelerinden biridir.
D4	tip fakultesi hastanesi gordugm en guzel (ozel olmayan) hastane.
D5	1982'de tıp fakültesine girdigimde (amanin yasim belli oldu) o siralar yeni yeni kurulmakta oldugundan her yerinde insaatlarin hüküm sürdügü, bu nedenle beni ziyaret eden bir arkadasimin, oolum seni kandirmislar, burasi tıp degil insaat fakültesi diye espiri yapmasina vesile olan üniversite.

Table 5.7 O1 - The five documents, after applying stop-words and non-alpha characrter removing operations on example documents of Table 5.6 (Original)

Id	Document
D1	eylül üniversitesi kamu yönetimi başkanlığı nı yürüttü hak iş müsiad ın siyasi danışmanlığını yaptı değişik gazete dergilerde türkiye nin siyasi idari sorunlarına ilişkin yüzlerce araştırması yayınlandı
D2	çamaşir suyundaki tehlike kadınların sık kullandığı çamaşır suyu çok zararlı eylül üniversitesi çevre mühendisliği bölümü öğretim üyesi doç dr mustafa odabaşı bunların kanser riskini önemli ölçüde arttırabileceğini söylüyor çamaşır suyu içeren ürünlerin amonyaklı asidik tuz ruhu kireç çözücü temizlik maddeleriyle karıştırılması zehirli gazların klor gazı klor aminlerin açığa çıkmasını sağlıyor ortamdaki oksijeni durduruyor insanları nefes alamaz hâle getiriyor tür zararlı maddelerin aşırı teneffüs edilmesi hâlinde solunum yolları akciğerde tahribata yol açıyor odabaşı araştırma sonuçlarını değerlendirdi piyasadaki çamaşır suyu içeren temizlik ürünlerinin sayısı gün geçtikçe artıyor katkısız parfümlü deterjan katkılı koyu kıvamlı sıvı jel ovma tozu sprey çok ürün türkiye dünyada yaygın olarak kullanılıyor geçtiğimiz yıl yapılan araştırma ürünlerin ülkemizdeki evden inde kullanıldığını hane başına yıllık tüketimin kilograma ulaştığını gösteriyor çamaşır suyu içeren temizlik ürünlerinin kullanımında dikkatli olunması gerekiyor
D3	eylül üniversitesi akademik üstünlüğünü yayınlarıyla başarılarıyla kanıtlamış türkiye nin iyi üniversitelerinden biridir
D4	tip fakultesi hastanesi gordugm guzel ozel olmayan hastane
D5	tıp fakültesine girdigimde amanin yasim belli oldu siralar yeni yeni kurulmakta oldugundan yerinde insaatlarin hüküm sürdügü nedenle ziyaret eden arkadasimin oolum kandirmislar burasi tıp degil insaat fakültesi espiri yapmasina vesile olan üniversite

Table 5.8 O2 - The five documents, after Turkish spell-checking operations on O1 documents

Id	Document
D1	eylül üniversitesi kamu yönetimi başkanlığı mı yürüttü hak iş müsait ıh siyasi danışmanlığını yaptı değişik gazete dergilerde Türkiye nin siyasi idari sorunlarına ilişkin yüzlerce araştırması yayınlandı
D2	çamaşır suyundaki tehlike kadınların sık kullandığı çamaşır suyu çok zararlı eylül üniversitesi çevre mühendisliği bölümü öğretim üyesi Doç. Dr. Mustafa odabaşı bunların kanser riskini önemli ölçüde arttırabileceğini söylüyor çamaşır suyu içeren ürünlerin amonyaklı asidik tuz ruhu kireç çözücü temizlik maddeleriyle karıştırılması zehirli gazların klor gazı klor aminlerin açığa çıkmasını sağlıyor ortamdaki oksijeni durduruyor insanları nefes alamaz hâle getiriyor tür zararlı maddelerin aşırı teneffüs edilmesi hâlinde solunum yolları akciğerde tahribata yol açıyor odabaşı araştırma sonuçlarını değerlendirdi piyasadaki çamaşır suyu içeren temizlik ürünlerinin sayısı gün geçtikçe artıyor katkısız parfümlü deterjan katkılı koyu kıvamlı sıvı jel ovma tozu sprey çok ürün Türkiye dünyada yaygın olarak kullanılıyor geçtiğimiz yıl yapılan araştırma ürünlerin ülkemizdeki evden inde kullanıldığını hane başına yıllık tüketimin kilograma ulaştığını gösteriyor çamaşır suyu içeren temizlik ürünlerinin kullanımında dikkatli olunması gerekiyor
D3	eylül üniversitesi akademik üstünlüğünü başarılarıyla yayınlarıyla kanıtlamış Türkiye nin iyi üniversitelerinden biridir
D4	tip fakültesi hastanesi gördüğüm güzel özel olmayan hastane
D5	tıp fakültesine girdiğimde amanın yaşım belli oldu sıralar yeni yeni kurulmakta olduğundan yerinde inşaatların hüküm sürdüğü nedenle ziyaret eden arkadaşımın oğlum kandırmışlar burası tıp değil inşaat fakültesi espri yapmasına vesile olan üniversite

Table 5.9 O3 - The five documents after ‘Modified Turkish Stemming’ operations on O2 documents

Id	Document
D1	eylul universite kamu yonetim baskan mi yurut hak is musait ih siyasi danisman yap degisik gazete dergi siyasi idari sorun iliskin yuzlerce arastirma yayin
D2	camasir su tehlike kadin sik kullan camasir su cok zarar eylul universite cevre muhendis bolum ogretim uye odabasi bu kanser risk onem olcu art soyle camasir su icer urun amonyak tuz ruh kirec coz temiz madde karis zehir gaz klor gaz klor amin acik cikma sagla ortam oksijen durdur insan nefes al-n hale getir tur zarar madde asiri teneffus et hal solunum yol akciger tahribat yol ac odabasi arastirma sonuc degerlen piyasa camasir su icer temiz urun say gun gec art katki parfum deterjan katki koy kivam sivi jel ov-n tozu sprey cok urun dunya yaygin ol kullan gec yil yapi arastirma urun ulke ev in kullan hane bas yil tuket kilogram ulas goster camasir su icer temiz urun kullan dikkat ol gerek
D3	eylul universite akademik ustun basari yayin kanitla iyi universite biri
D4	tip fakulte hastane gor guzel ozel ol-n hastane
D5	tip fakulte gir amanin yas belli ol sirala yeni yeni kur ol yer insaat hukum sur neden ziyaret ede arkadas ogul kandir bura tip degil insaat fakulte espri yap vesile ol universite

Table 5.10 O4 - The five documents, after applying stop-words removing operations on O3 documents

Id	Document
D1	eylul universite kamu yonetim baskan yurut hak is musait ih siyasi danisman yap degisik gazete dergi siyasi idari sorun iliskin yuzlerce arastirma yayin
D2	camasir tehlike kadin sik kullan camasir cok zarar eylul universite cevre muhendis bolum ogretim uye odabasi kanser risk onem olcu art camasir icer urun amonyak tuz ruh kirec coz temiz madde karis zehir gaz klor gaz klor amin acik cikma sagla ortam oksijen durdur insan nefes al-n hale getir tur zarar madde asiri teneffus et hal solunum yol akciger tahribat yol ac odabasi arastirma sonuc degerlen piyasa camasir icer temiz urun say gun gec art katki parfum deterjan katki koy kivam sivi jel ov-n tozu sprey cok urun dunya yaygin ol kullan gec yil yapi arastirma urun ulke ev in kullan hane bas yil tuket kilogram ulas goster camasir icer temiz urun kullan dikkat ol gerek
D3	eylul universite akademik ustun basari yayin kanitla iyi universite
D4	tip fakulte hastane gor guzel ozel ol-n hastane
D5	tip fakulte gir amanin yas belli ol sirala yeni yeni kur ol yer insaat hukum sur ziyaret ede arkadas ogul kandir bura tip degil insaat fakulte espri yap vesile ol universite

As seen in Tables 5.6 - 5.10, word counts were decreased from each operation, because pruning, cleaning, spell-checking, and stemming and then again cleaning operations were applied over documents. In Figure 5.8, a chart showing the word counts versus the operations over these five documents are given.

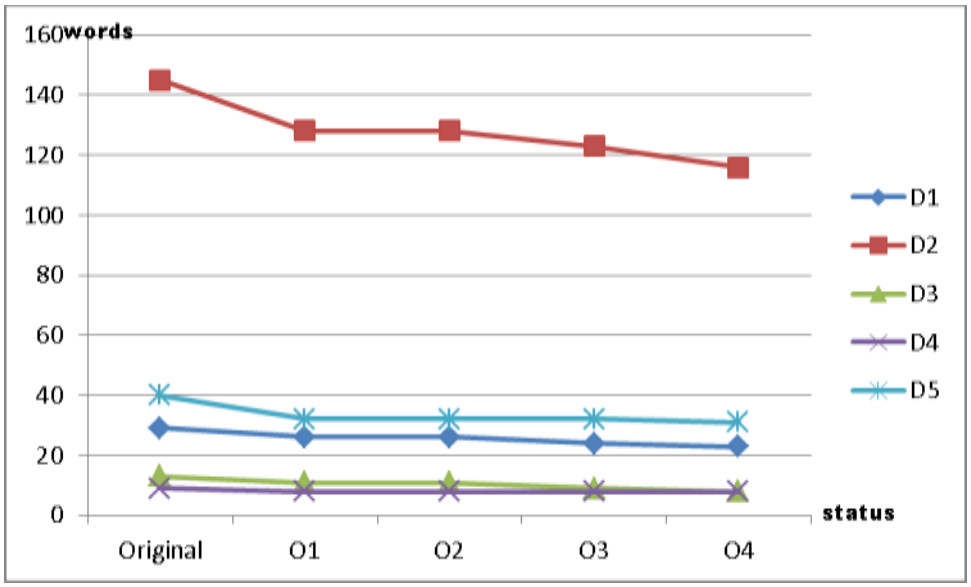

Figure 5.8 Word counts versus operation statuses with charts over five documents

Text manipulation is very important on Turkish language because on web pages most of the user comments were written in misspelled collection of words. Moreover, Turkish language has many suffixes for a single word. Those reasons make it harder to push data to classification algorithms. Thus, spell-checking and stemming operations were used to prevent a word from being perceived as a completely different word.

5.5.3 Sentimental Text Classification Experiments

The supervised learning methods, which can be called inductive learning in machine learning or classification, were used. For testing purposes, polarity dataset were gathered from services sepecified in Chapter 5.5.1 and assigned to 'test_docs' database table manually. After gathering 200 positive, 200 negative samples, the steps in Chapter 5.5.2 were applied to manipulate the text data. To supply how useful it is to manipulate the text, the experiments were made over un-touched comments and the text data was cleaned. To make classification, firstly used Naïve Bayes classifier was used and then for comparing results TF-IDF based SVM was used.

For classifying the first 175 positive and 175 negative samples were used as training data and the remaining 25 positive and 25 negative data were used for testing purposes.

Accuracy is calculated according to formula (8) in Chapter 5.5.1, and the results are listed in Table 5.11. As seen on experiments, the manipulated (processed) data, gives more accurate results in both algorithms. Although SVM needs more operations to calculate weights for SVM classification, it gives more accurate results for sentimental classification.

Table 5.11 Accuracy results sentimental classification with Turkish dataset

Dataset	**SVM with TF-IDF Accuracy**	**Naïve Bayes Accuracy**
Processed data	72.00% (18 correct, 7 incorrect)	68.00% (17 correct, 8 incorrect)
Un-touched data	60.00% (15 correct, 9 incorrect)	60.00% (15 correct, 10 incorrect)

For testing purposes, another ready-made dataset was used from Cornell University which was gathered from imdb.com (Internet Movie Database) movie reviews. The dataset contains ~5300 positive and ~5300 negative reviews (Pang & Lee, 2005). To test the accuracy the first 5000 of both positive and negative reviews were used for training and the remaining reviews were used for testing purposes. Again the accuracy is calculated according to formula (8) and results are shown in Table 5.12. Because the data was in the English language, no cleaning, pruning and stemming operations were performed on it.

Table 5.12 Accuracy results of sentimental classification with the ready-made dataset

Dataset	**SVM with TF-IDF Accuracy**	**Naïve Bayes Accuracy**
Processed data	82.01%	80.00%

CHAPTER SIX

CONCLUSION & FUTURE WORK

6.1 Conclusion

A very useful application was developed during this study for the clients who want to monitor companies or products from social media platforms and track the positive, and negative feedbacks about them if compared with other approaches that only use web content mining. It is a very smart web mining application that can find the exact title of a web document and surf on web pages using the web structure of pages.

For web mining purposes, the approach provides data from various resources and various text data types from the WWW. It discovers patterns and analyzes them as tidy outputs in three different text representation formats. This feature helps the communication of internal and external components with each other.

In the studies, web mining techniques were used and these techniques were combined with text classification techniques. Although, there are some studies on various languages about sentimental classification over web mining, the study of sentimental classification over web mining is a new topic for the Turkish language.

The study comes with revolutionary features such as a sentimental classification over the Turkish language and the unique title mining approach. Good results were achieved over title extracting from various web pages. It is hoped that the study will guide for new research areas for the Turkish language, like spam detecting over web based data on Turkish language and full sentimental classification including neutral results in addition to positive and negative ones.

6.2 Future Works

Sentimental Feedback Miner application were developed which can gather information, discover patterns on web sites and web services and semantically analyse the incoming text data. However, it is a very good approach for polarity based documents. Furthermore, it needs some improvements like spam checking on documents. It also needs to be able to classify documents into one more category, namely ‘neutral’.

For future work, with the data gathered from clients, SFM can also be used with web usage mining data to make suggestions to visitors. With the usage of web usage mining, it might be good to recommend new keywords to new clients with experience of old clients.

REFERENCES

Agarwal, R., Prabhakar, T. V., & Chakrabarty, S. (2008). "I Know What You Feel": Analyzing the role of conjunctions in automatic sentiment analysis. *Advances in natural language processing: 6th international conference, GoTAL 2008* (28–39). Gothenburg, Sweden: Springer.

Akın, A., & Akın, A. A. (2006). *Zemberek, an open source NLP framework for Turkic Languages*.

Alexa – Top sites in Turkey. Retreived May 2, 2011, from http://www.alexa.com/topsites/countries;0/TR

Alvarez, M., Pan, A., Raposo, J., Bellas, F., & Cacheda, F. (2008). Extracting lists of data records from semi-structured web pages. *Data & Knowledge Engineering*, 64, 491–509.

Braspenning, P. J., Thuijsman, F., & Weijters, A. J. M. M. (1995). *Artificial neural networks: an introduction to ANN theory and practice*. Germany: Springer.

Cohen, F. (2002). How to get around your ISP? *Managing network security*. Retrieved February 25, 2011, from http://fredcohen.net/Analyst/netsec/2002-02.html.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning* 20, 273-297.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. *Proceedings of IEEE International Conference Tools with AI*, 558–567.

Dang, Y., Zhang, Y., & Chen H. (2010). *A lexicon-enhanced method for sentiment classification: An experiment on online product reviews* (46-53). University of Arizona.

Friedman, K. A. (1996). *The Decision Tree*. Washington, USA: Heart Publishing.

Garruzzo, S., Rosaci, D., & Sarne, G. M. L. (2007). MARS: An agent based recommender system for the semantic web. *Distributed Applications and Interoperable Systems*, (181-194). Berlin: Springer

Gupta, G. K. (2006). *Introduction to data mining with case studies*. India: Prentice-Hall of India Pvt.Ltd.

Han, J., Kamber, M., Kaufmann, M. (2006). *Data mining: concepts and techniques*. San Francisco, USA: Elsevier.

Hui, S-C., & Foe, S. (1998). A dynamic IP addressing system for Internet telephony applications. *Computer Communications*, (21, 3), 254-266.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. USA: Cambridge University Press.

Joachims, T. (1999). *Making large-scale SVM learning practical.* B. Schölkopf , C. Burges & A. Smola (Ed.), *Advances in kernel methods - Support vector learning*. USA: MIT-Press.

Lu, Y., Kong, X., Quan, X., Liu, W., & Xu, Y. (2010). Exploring the Sentiment Strength of user reviews. *WAIM'10 Proceedings of the 11th international conference on Web-age information management*, 471–482.

Lynn, J. (October 19, 2010). *Internet users to exceed 2 billion this year*. Retrieved February 12, 2011, from http://www.reuters.com/article/2010/10/19/us-telecoms-internet-idUSTRE69I24720101019

Markey, R., Reichheld, F., & Dullweber, A. (2009). Closing the customer feedback loop. *December 2009 Harvard Business Review*.

Markov, Z., & Larose, D. T. (2007). *Data mining the web: Uncovering patterns in web content, structure, and usage* (1st ed.) (1-59). USA: Wiley-Interscience.

Masseglia, F., Poncelet, P., Teisseire, & M., Marascu, A. (2008). Web usage mining: extracting unexpected periods from web logs. *Data Mining, Knowledge, & Discovery,*
16(1), 39-65.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). *Mining meaning from Wikipedia*. New Zealand: University of Waikato.

Omeri, A. (2009). Data mining for improved web site design and enhanced marketing. *Data Mining for Desing and Marketting*, 96-107.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the ACL,* 271-278.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL 2005,* 115-124.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (79-86).

Schedl, M., Widmer, G., Knees, P., & Pohle, T. (2008). *A music information system automatically generated via Web content mining techniques*. Linz, Austria: Department of Computational Perception, Johannes Kepler University.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations*, 12–23.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of Speech tagging with a cyclic dependency network. *In Proceedings of HLT-NAACL*, 252-259.

Thet, T. T., Na J-C., & Khoo, C. S. G. (2008). Sentiment classification of movie reviews using multiple perspectives. *ICADL 2008*, LNCS 5362, 184–193.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 417-424.

Xu, G., Zhang, Y., & Li, L. (2010). *Web mining and social networking: Techniques and applications* (1st ed.). New York: Springer.

Yessenalina, A., Choi, Y., & Cardie, C. (2010). Automatically generating annotator rationales to improve sentiment classification. *Proceedings of the ACL 2010 Conference Short Papers*, 336–341.