

ÇANKAYA UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
MATHEMATICS AND COMPUTER SCIENCE

MASTER THESIS

OPINION MINING WITH TEXT OPERATIONS AND EXTRACTING DATA
FROM USER REVIEWS FOR E-COMMERCE APPLICATIONS

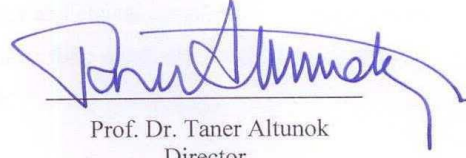
MUHAMMED BURAK UYTUN

FEBRUARY, 2013

Title of the Thesis :Opinion Mining With Text Operations And Extracting Data
From User Reviews For E-Commerce Applications

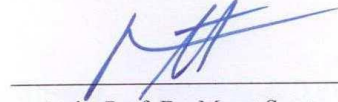
Submitted by MUHAMMED BURAK UYTUN

Approval of the Graduate School of Natural and Applied Sciences, Çankaya
University



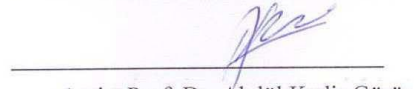
Prof. Dr. Taner Altunok
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of
Master of Science.



Assist.Prof. Dr. Murat Saran
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully
adequate, in scope and quality, as a thesis for the degree Master of Science.



Assist.Prof. Dr. Abdül Kadir Görür
Supervisor

Examination Date : 08.02.2013

Examining Committee Members

Asst. Prof. Dr. Abdül Kadir Görür (Çankaya Univ.)

Asst. Prof. Dr. Bülent Gürsel Emiroğlu (Başkent Univ.)

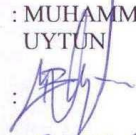
Asst. Prof. Dr. Fahd Jarad (Çankaya Univ.)



STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : MUHAMMED BURAK
UYTUN

Signature : 

Date : 26.02.2013

ABSTRACT

OPINION MINING WITH TEXT OPERATIONS AND EXTRACTING DATA FROM USER REVIEWS FOR E-COMMERCE APPLICATIONS

UYTUN, Muhammed Burak

M.Sc., Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Abdül Kadir Görür

FEBRUARY 2013, 61 pages

This thesis tries to make an understanding of using opinion mining and sentiment analysis and applying these methods for extracting data from user feedbacks. Before the data extraction step, opinion mining is examined in detail to understand better the goals to be achieved.

To extract the data, www.kobiform.com e-commerce web site is used as pilot platform. www.kobiform.com is furniture selling e-commerce web site originated in Ankara which will be mentioned in details later. Beside user reviews, multiple choice and text based surveys are given to the users while they browse thorough products for the data gathering.

The algorithm used to evaluate the data which is based on the classic view of generating word sets. The application used to gather data is developed with .net framework and Visual Studio 2008 is used as IDE. Microsoft SQL Server is used as database.

Keywords: e-commerce, www.kobiform.com, opinion mining and sentiment analysis, user feedbacks

ÖZ

DUYGU ANALİZİ VE E-TİCARET SİSTEMLERİ İÇİN KULLANICI HAREKETLERİNDEN VERİ TOPLANMASI

UYTUN, Muhammed Burak

Yükseklisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi : Yrd.Doç.Dr. Abdül Kadir Görür

ŞUBAT 2013, 61 sayfa

Bu tezde duygu madenciliği ve doğal dil işleme kullanılarak kullanıcı girdilerinden anlamlı ve işlevsel çıkarımlar yapılması konusunda bir uygulama yapılacaktır. Konunun iyi anlaşılması için uygulama detaylarından önce duygu madenciliği konusunda detaylı bilgi verilmiştir.

Tezde kullanılan veriler 2010 yılında T.C. Bilim, Sanayi ve Teknoloji Bakanlığı tarafından kabul edilen Teknogirişim programı bünyesinde kullanılan www.kobiform.com e-ticaret sistemi tarafından sağlanmıştır. Sisteme, kullanıcı yorumlarının yanısıra ürün odaklı anketler de kullanılarak veri girdisi sağlanmıştır.

Kullanılan algoritma genel yaklaşım ile belirlenen, hazır kelime ve cümle setleri üzerinden eşleştirmeye dayanmaktadır. Uygulama .net mimarisi ile Visual Studio 2008 kullanılarak geliştirilmiştir. Veritabanı olarak ise MSSQL Server kullanılmıştır.

Keywords: e-ticaret, www.kobiform.com, düşünce madenciliği ve anlamsal analiz, kullanıcı beslemeleri

ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Assist.Prof.Dr. Abdül Kadir Görür for his guidance, advice, criticism, encouragements and insight throughout the research.

TABLE OF CONTENTS

STATEMENT OF NON-PLAGIARISM PAGE	iii
ABSTRACT	iv
ÖZ	v
ACKNOWLEDGMENTS	vivi
TABLE OF CONTENTS	viii
LIST OF TABLES	ix
LIST OF FIGURES	x
INTRODUCTION	1
CHAPTER I	
1.1 A BRIEF LOOK AT OPINION MINING AND UNDERSTANDING THE WEB SEARCH	3
1.1.1 Search Engines	4
1.1.2 Content Collectors	5
1.1.3 Linguistic Operations	6
1.1.4 Indexes.....	7
1.1.5 Listing for Search Algorithm	8
1.1.6 Information Extraction	8
CHAPTER II	
2.1 DATA MINING PROCESS	10
2.1.1 Association Rule Learning	10
2.1.2 Applying Data Mining to Web.....	12
2.1.3 Opinion Mining	13
CHAPTER III	
3.1 OPINION MINING AND SENTIMENT ANALYSIS	15
3.1.1 Components of Opinion Mining.....	15

3.1.2	Architecture of the Opinion Mining System	17
3.1.2.1	POST (Part Of Speech Tagging).....	19
3.1.3	Feature Identification	19
3.1.3.1	Frequent Features Identification	20
3.1.3.2	Infrequent Features Identification.....	20
3.1.4	Opinion Sentiment Analysis.....	21
3.1.4.1	Description of Sentiment	22
3.1.4.2	Finding the Sentiment at the Feature Level	22
CHAPTER IV		
4.1	OPINION MINING AND SENTIMENT ANALYSIS	25
4.1.1	User Profile.....	26
4.1.2	Use Case Analysis	26
4.1.3	System Requirements	29
4.1.3.1	General Operations	29
4.1.3.2	Supportive Requirements.....	29
4.1.4	Implementation.....	30
4.1.4.1	Technologies	30
4.1.4.1.1	Database	30
4.1.4.1.2	Product Analyzer Software	31
4.1.4.2	Architecture.....	31
4.1.4.3	Text Process Module.....	33
4.1.4.3	Opinion Mining Module	34
4.1.4.5	Reporting Of the Analyzes and Evaluation.....	40
SUMMARY AND CONCLUSION.....		47
REFERENCES.....		49
BIOGRAPHY		50

LIST OF TABLES

Table 1. Web mining categories.....	12
Table 2. User(Customer) role description.....	27
Table 3. Database Job(Data Transfer) role description.....	28
Table 4. OMS Analyzer role description	28
Table 5. OMS Summarize role description.....	29

LIST OF FIGURES

Figure 1.1: Search engine system concept.	5
Figure 1.2: Web crawler working mechanism in detail.	6
Figure 3.1: Opinions give positive feedback about car in the figure. Hence user references the attribute it may also directly refer to object.....	16
Figure 3.2: Opinion Mining Architecture	18
Figure 3.3: Infrequent feature extraction method.	20
Figure 3.4: User opinions for a chair	21
Figure 3.5: Calculating the orientation of opinions on product features [6] Hata! Yer işareti tanımlanmamış.	24
Figure 4.1: Workflow of www.kobiform.com e-commerce system	25 Hata! Yer işareti tanımlanmamış.
Figure 4.2: Use case of opinion mining	27 Hata! Yer işareti tanımlanmamış.
Figure 4.3: Database architecture.....	31 Hata! Yer işareti tanımlanmamış.
Figure 4.4: Architecture of OMS system	34 Hata! Yer işareti tanımlanmamış.
Figure 4.5: UML Class Diagram of Opinion Mining System..	35 Hata! Yer işareti tanımlanmamış.
Figure 4.6: Feature extraction function.....	36 Hata! Yer işareti tanımlanmamış.
Figure 4.7: Selecting the reviews and opinion collection	37
Figure 4.8: An approach for calculating the negate words	37 Hata! Yer işareti tanımlanmamış.
Figure 4.9: Complexity of negate words in examples.....	38
Figure 4.10: Pseudocode of Opinion Mining Module	39
Figure 4.11: Frequent expression analysis on a product.....	41
Figure 4.12: Opinion analysis on a product	42
Figure 4.13: Frequent expression analysis on a product	42
Figure 4.14: Infrequent expression analysis on a product.....	43
Figure 4.15: Negate words analysis on a product	43
Figure 4.16: Opinion analysis on a product	44
Figure 4.17: Polarity analysis on a product.....	44

Figure 4.18: Polarity analysis on a product.....	45
Figure 4.19: Total analyze calculations	45
Figure 4.20: Frequent/Infrequent sentence comparison.....	46
Figure 4.21: Frequent/Infrequent sentence numbers with polarity	46

INTRODUCTION

In recent years with the inevitable increase at the use of internet technologies, people use the web for many reasons like entertainment, personal communication, online shopping and so on. To increase the efficiency and reliability, most E-commerce providers encourage users to comment on products or process transactions. Through these comments, users express their opinions in whatever way they feel and these reviews provided by e-commerce companies are widely read. Hence it becomes difficult for the customers to read all the reviews to make a decision and also fills databases with the meaningless words. Also e-commerce companies need to have a long-term relationship strategy to keep customers satisfied and considering customer reviews and experiences at purchasing process is one of the best ways to ensure the product quality and usability for potential customers.

Based on mined opinions it's up to system to proceed on user oriented or product oriented. With proper mining outcomes, system gains the ability to recommend user the items of could be in interest or can grade the product with multiple perspectives like beauty, usefulness, durability etc.

➤ Motivation

The amount of easily accessed information concerns the system providers because of the information overload and how to handle it. They are aware of the most important point; user should have to access the best and efficient sources with the least effort. Nowadays, especially with the spreading of interactive Web 2.0 concept, user-generated data (in our case we can name it user opinions) increased dramatically.

Companies realized the value of the information represented by these opinions. Thus, this study emphasizes the need of special mechanisms that aims to provide the community better ways to take full advantage from this data.

According to the consumer perspective, considering other consumers' opinions before purchasing a product is a common fashion even before the Internet. In favor of consumers, there is a big difference is that, a consumer has access to thousands of opinions in which simplifies the process of decision making. Essentially, consumers want to acquire the best product with the minimum price.

According to the e-commerce providers' perspective, receiving consumer's comments and feedback of products can greatly affect their marketing strategies. For example, an online e-commerce site can place smart ads in order to measure the satisfaction level of consumers for a given product. For instance, if a product has a low level of satisfaction, a smart strategy would be placing a competitor advertisement inside this page.

This thesis tries to present the ways for finding, gathering, extracting, classifying and summarizing consumer opinions or feedbacks on the e-commerce sites (In this case former www.kobiform.com). The proposed framework will examine several techniques for extracting data out of plain language text (consumer provided content), in order to generate useful information from available content in a more valuable and organized form.

CHAPTER I

1.1 A BRIEF LOOK AT OPINION MINING AND UNDERSTANDING THE WEB SEARCH

First of all, the meaning of "Opinion" must be defined. Opinion stands for a person's view, attitude or appraisal. As it shown opinion is strictly attached with a subject and the subject is the main source of the concept. Widely common opinions almost affect the definition of objects so it is significantly important for producers and sellers.

In terms of marketing by keeping the subjectivity, opinion stands for a person's both emotional and intellectual extractions of a product. These opinions could be composed before or after the buying/using of the product but for evaluation of the quality of a product, it is best for us to consider the after usage opinions. This extracted subjective information is used as source for defining a product, creating a user profile, making the product suggestion etc.

Opinion mining depends on Information Retrieval, Information Extraction and Data Mining. Information Retrieval and Information Extraction are vital for locating and gathering of valuable information out of user produced irregular data. Also, with the Web development increasingly shifts to the orientation towards the importance of semantics and integration of information, these areas of study become very important to address the new future trends of the Web.

This section gives and overview on Search Engines for a better understanding of efficient Information Retrieval and an approach to Information Extraction techniques.

1.1.1. Search Engines

Information Retrieval (from now on, called IR) is a process that manages the retrieval of a data set from a collection of data sets, usually based on keyword searches. With the unbounded expansion of the Internet, search engines have become almost the solitary instrument to find required information on the web.

Following interactions are performed during a search operation

- User submits the search query
- Query is processed and prepared to be used by the retrieval system (spellchecking, reducing to root etc.).
- Query is checked for the available indexes to retrieve the results that contains some of the terms.
- Afterwards, a ranking algorithm is applied to the result set which are finally presented to the user to decide the order of the items in result set.
- User receives the result set as response and has access to the matching results from a result list. (Fig 1.1)

The steps that we mentioned up to now show how the search engine works after a user request. In fact, the main task has to be performed as crawling web pages and their contents for indexing and ranking operations. Web search engines and the web search concept will be examined in details at following sections.

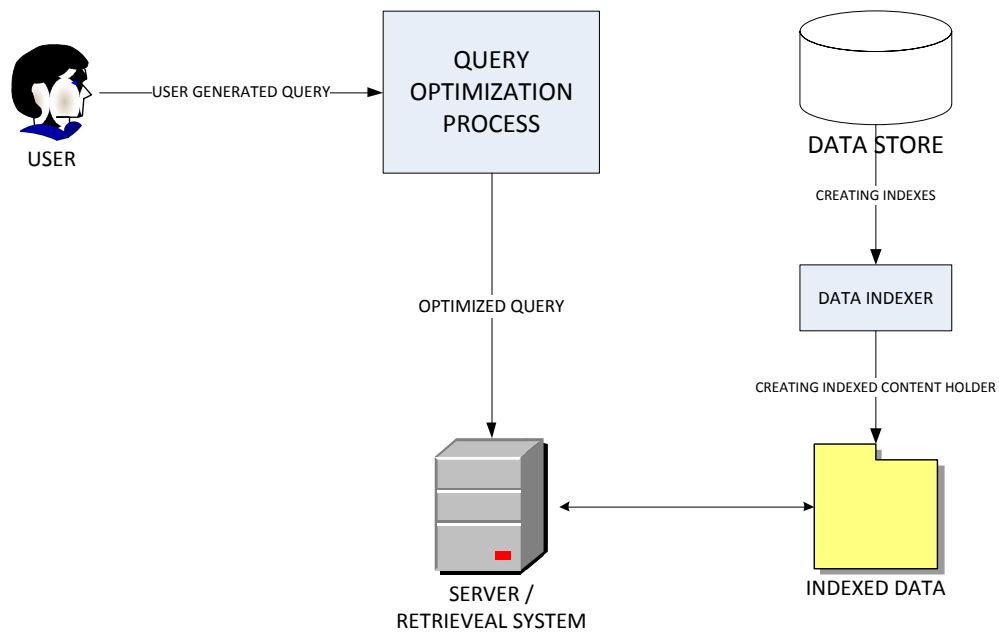


Figure 1.1: Search engine system concept

1.1.2. Content Collectors

Search engines feed on content collectors which are lightweight programs that follow links and gather contents for indexing and tuning the search operations. Since we are studying on data mining, it is best to mention about content collectors briefly. Simple explanation of a web crawler life cycle is shown at figure 1.2

Content collectors (also called as Web Crawlers, Web robots or Web spiders) are one of the most vital component of an online search engine. Content collectors have two main attributes that have importance. One of them is to have a good visiting algorithm. Hence there are many different algorithms, some common algorithms are explained.

- **Focused Crawling:** According to Menczer [1] focused crawler or crawler that focuses on topics which a web crawler that tries to gather only web pages connected to a set of predefined topics. “Topical crawling generally assumes that only the topic is given, while focused crawling also assumes that some labeled examples of relevant and non-relevant pages are available.”

- Works, working, worked → work
- Dreamy, dreaming, dreamt → dream

➤ *Lemmatization:*

Gathering the different shapes of any word together makes them marked as one single element and this procedure is known as lemmatization. It works similar to stemming as concept but lemmatization catches the words that stemming misses.

As mentioned above stemming is applicable to the words that have suffixes. But in some cases different words may map to a same root. At this point lemmatization takes places and hits the correct root.

An example that contains words which have stems and getting the roots:

- better → good
- went → go

➤ *Stop Words:*

These words are meaningless yet essential elements for a sentence that could be ignored. Even if all tools are used, there is still not one specific list of stop words could be extracted which in turn makes them undefined. However there are some common stop words such as (in English) the, by, when, with, a, an, on, of, this, that, these, those etc.

1.1.4. Indexes

Search engines work through massive amount of pages or items (products, comments, extractions etc.) to find a specific term that is supplied by user. Under normal conditions, these operations take a great deal of time. To decrease the searching time and improve the search engine performance, most common approach is to use the retrieval systems that use indexes to find structured data. An index is a data structure delegated by a term or specification that defines the item contains the index.

An example for indexes:

- Item 1: A fast car
- Item 2: A slow car
- Item 3: A car with a good design

Keywords that delegates the sentences are

- fast (for item 1)
- slow (for item 2)
- car (for item 1,2,3)
- good (for item 3)
- design (for item 3)

1.1.5. Listing for search algorithms

Listing the results after the user queried searching process is one of the main steps of a search process. Depending on the query result set, it could be massive for user to check every data and find the desired result. Therefore a proper listing of results for users' search keywords relevancy is necessary for the best user experience. Ranking algorithms are used to avoid this issue. Also with the ranking algorithms, search engines can flush the spamming words, recognize the word noises and improve the search result efficiency.

1.1.6. Information Extraction

Information Extraction (from now on, called IE) is a primitive application of artificial intelligence. Main aim of IE is to extract valuable information out of irregular and noisy data. IE system focused on data entities or objects (products, geographic objects, etc.). Irregular data can have audio, video, image form as well as text form which we are interested in this study. For this reason, the examples and further explanations on this section will be focused on texts.

A raw data (irregular and noisy) set is difficult to query and process. Main object of IE is to identify useful parts out of raw data and extract them to more valuable shaped information with semantic attributes. As it seems that IE and IR has

similarities, it is a mistake to consider both as the same. The difference between IE and IR is well described with the following quotation.

“Information Extraction is not Information Retrieval: Information Extraction differs from traditional techniques in that it does not recover from a collection a subset of documents which are hopefully relevant to a query, based on key-word searching (perhaps augmented by a thesaurus). Instead, the goal is to extract from the documents (which may be in a variety of languages) salient facts about pre-specified types of events, entities or relationships. These facts are then usually entered automatically into a database, which may then be used to analyze the data for trends, to give a natural language summary, or simply to serve for on-line access.” [3]

IE can be applied to data (as mentioned before the form of the data is going to be considered as text for following examples) and stores the targeted information explicitly. If the data is stored implicitly (cannot be read within the data set), IE generates errant results.

In summary IE systems analyses disordered text-blocks to gather data of entities, relations and events which are predefined. To put it another way, formed real information through unformed text are derived by knowledge discovery. I am going to mention these IE systems later in detail.

CHAPTER II

2.1. DATA MINING PROCESSES

As mentioned before; IE works with the information stored explicitly in texts. What if information stored in text implicitly?

Data mining (also known as Knowledge Discovery) discovers information implicitly stored in text, builds correlations between data sets and creates knowledge of it. Data mining is a powerful application when it is impractical for a human being to analyze massive amount of data which has to be analyzed.

Data mining applications were available long before the internet era. The applications were used for knowledge discovery, data or pattern recognition in large databases and data warehouses. Databases process data in a structured form, where data is kept in meaningful pieces that are responsible for building certain information of a specific form. Data mining meant to address especially databases, but recently the field has evolved to satisfy needs of the Internet

With the importance of rapidly increasing data mining; more techniques are developed lately. Two of the main data mining methods are described below.

2.1.1. Association Rule Learning

Association rule learning is one of the most popular and well researched methods for discovering relations/correlations between variables in large databases. The classic explanation of Association Rule is defined by Agrawal [4] as: Let I be a set of items $i_1, i_2, i_3 \dots i_n$. Let T be a set of transactions called to database $t_1, t_2, t_3 \dots t_n$. Each transaction in T has a unique transaction ID and contains a subset of the items in I . A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$

Association rule applications are mostly used in marketing, medical areas, web applications and security domain. To illustrate the concept, next example uses marketing:

Let's say the set of items are $I = \{\text{milk, bread, butter}\}$ and a database containing these items. An example rule could be $\{\text{butter, bread}\} \rightarrow \{\text{milk}\}$ meaning that if a customer is interested in butter and bread, customer also could buy milk. As the example shows association rule makes extractions from data sets. Hence these extractions are not facts two terms takes place to define the certainty of the statement. If we change the rule and say that; $\{\text{butter, bread}\} \rightarrow \{\text{milk}\}$ (*support* = 30%, *confidence* = 80%) , now the rule means 30% of customers are interested in butter and bread together also interested in milk 80% of the time. These two major measurements are described as;

➤ *Support:*

“The support of a rule, $X \Rightarrow Y$, is the percentage of transactions in T that contains XUY, and can be seen as an estimate of the probability, $\text{Pr}(XUY)$. The rule support thus determines how frequent the rule is applicable in the transaction set T. Let n be the number of transactions in T.” [4] The support of the rule $X \Rightarrow Y$ is shown below:

$$\text{support} = \frac{\text{Number of transactions containing } X \cup Y}{\text{Total number of transactions in } T}$$

➤ *Confidence:*

“The confidence of a rule, $X \Rightarrow Y$, is the percentage of transactions in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, $\text{Pr}(X|Y)$.”[4] Confidence is shown below:

$$\text{confidence} = \frac{\text{Number of transactions containing } X \cup Y}{\text{Number of transactions containing } X}$$

Following example processes a better and complex approach for a better understanding of association rules:

Let's define 3 sets of items which are bought from supermarkets, each T means a transaction:

T1: {cheese, bread, salt}

T2: {apple, salt, bread}

T3: {cheese, salt, sugar},

and consider minimum support (from now on, called $supp$) = 30%, minimum confidence (from now on, called $conf$) = 60%

For Bread \rightarrow Cheese ($supp = 1/3$ (33%), $conf = 1/2$ (50%))

The above statement is valid for minimum support but fails at minimum confidence

For Salt \rightarrow Apple ($supp = 1/3$ (33%), $conf = 1/3$ (33%))

The above statement is valid for minimum support but fails at minimum confidence

2.1.2. Applying Data Mining to Web

Using data mining techniques on the Internet platform is recently named as Web Mining. Although Web Mining is similar to data mining, it has some important differences from data mining approach, especially with the way of data collection. The main difference between data mining and web mining is, hence in data mining the data is stored in data base, in web mining special methods (such as IR and IE) are required to prepare data which is ready to be mined. There are three main categories of Web Mining associated with the goals and filters of the mining task (Table 1) [5].

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View of Data	- Unstructured - Semi structured	- Semi structured - Web site as DB	- Links structure	- Interactivity
Main Data	- Text documents - Hypertext documents	- Hypertext documents	- Links structure	- Server logs - Browser logs
Representation	- Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph (OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine learning - Statistical (including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine Learning - Statistical - (Modified) association rules
Application Categories	- Categorization - Clustering - Finding extraction rules - Finding patterns in text - User modeling	- Finding frequent sub-structures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaptation, and management - Marketing - User modeling

Table 1 – Web mining categories

➤ *Web Usage Mining:*

Web usage mining method based on methods which anticipate user behavior from users' behavior and interaction pattern. Most of the usage mining applications are fed from web logs. These logs mostly contain user navigation maps, clicks and similar traceable interactions. With the interpretation of these logs, users' aims are predicted.

➤ *Web Structure Mining:*

Web structure mining focuses on the structure and relations of the hyperlinks within the web itself. With the discovered information from hyperlinks, the relations between web pages are obtained.

➤ *Web Content Mining:*

Web content mining is very important as it deals directly with information. The goal is to mine content from web documents in order to build knowledge from it. This knowledge can be either hidden or somehow simply difficult to be analyzed in a straightforward way. Opinion Mining is one of the most important sub-studies of Web Content Mining.

2.1.3. Opinion Mining

As mentioned before, Opinion Mining is a field of Web Content Mining that concerns to extract information out of users' opinions. Opinion mining especially lately has become much more important because of the rapid growth rate of e-commerce.

Gathering feedbacks based on users' experiences about a product is a common behavior for the potential customer. A major problem however, is finding the desired information on them. It is not difficult to find user reviews from hundreds of e-commerce sites but most of them are disordered and difficult to extract valuable information out of them.

Also e-commerce companies can gather feedbacks from user reviews, create a decision making system for advertisements. For example, if the majority of customers express negative opinions about a given product, an alternative product

from a competitor could be placed as an alternative instead. Also, manufacturers can get the feedback of their products to improve their products or the required services.

Opinions may branch as expert opinions and user opinions (review, sharing experience etc.) Expert opinion is created by someone with a privileged status or higher knowledge on the subject while user opinion refers to opinions given by common users. An expert opinion is usually far from superior in quality, richer in technical details, and goes through all the most relevant aspects of a product. Customers usually give opinions with less importance and common sense. In this work, users' opinions, users' reviews and product opinions are used. This work will not deal with product reviews given by experts, thus it will focus mining only on ordinary customers opinions on the e-commerce sites.

CHAPTER III

3.1. OPINION MINING AND SENTIMENT ANALYSIS

This study is based on opinion mining within the e-commerce context. The first reason is to clarify importance for the marketing business and second is to help customers decide whether products fit their needs.

Most of the researches in Opinion Mining have been placing efforts on product features identification and finding opinion sentiment/orientation. There are many methods which are developed for opinion mining. In this study a classic approach *Mining and Summarizing Customer Reviews* is applied. Also a new approach *A Holistic Lexicon-Based Approach to Opinion Mining* is going to be exposed. These methods will identify features much more easily and produce sentiment analysis at considerable level. Also, both methods are considered as two best applications for e-commerce context.

3.2. Components of Opinion Mining

➤ *Object Model Definition:*

Opinions contain user wise claimed strengths and weaknesses of an object. Objects may vary such as a product or a company etc.

An object can be defined as a tree which may contains sub-components to create a new object. We can define an object as:

$O: (T, A)$ where T stands for taxonomy of components (parts that creates the object) and A is a set off attributes of object (O) Such as the hierarchy of a tree, the components can also have their own set organized. [7]

If we put this to an example; the user generated sentence, “This modem has a long range of wi-fi signal”, here wi-fi signal is an attribute of the object modem which is root object in this example. In figure 3.1, opinions and their reference to objects are illustrated

The next components are going to use this model to reference opinions as well as objects. Attribute used in the $Q: (T, A)$ statement stands for both components and attributes, which will also simplify the model by omitting the hierarchy.



Figure 3.1. Opinions give us positive feedback about car in the figure. Hence user references the attribute that may also directly refer to object.

➤ *Explicit and Implicit Features:*

As mentioned previously if a feature is readably available in an object review, it is an *explicit feature*, if it is not readably available in review and could be gathered after a process, it is called an *implicit feature*.

Example:

- “The process speed of laptop is too fast”
- “The laptop is heavy”

In the first sentence of example, *process speed* is an explicit feature and it’s supported with a positive feedback.

In the second sentence of example, the weight of the laptop is reviewed but there are no keywords such as weight so it is an implicit feature. Luckily it is easy to understand, sentence indicates a negative feature of the implicit weight attribute.

➤ *Explicit and Implicit Opinions:*

This Same concept is applied but it is applied for the opinions. An explicit opinion on a feature directly expresses the positivity or negativity but an implicit opinion does not have any judgments just implies opinions.

Example:

- “The resolution of the screen is impressive”
- “The mouse started to jam”
- “The screen of my phone is big”

In the first sentence of example there is an explicit positive opinion about *resolution of the screen*.

In the second sentence of example, there is no clear opinion about mouse but we can assume it is negative.

In the third sentence of example, there is also no clear opinion about phone screen size. We cannot decide that the *big size of the phone* is whether good or bad for the customer.

3.1.2. Architecture of the Opinion Mining System

With the approach of *Mining and Summarizing Customer Reviews* and *A Holistic Lexicon-Based Approach to Opinion Mining* following opinion mining architectural system can be illustrated. The system has influenced from both approaches.

The architecture is illustrated in figure 3.2.

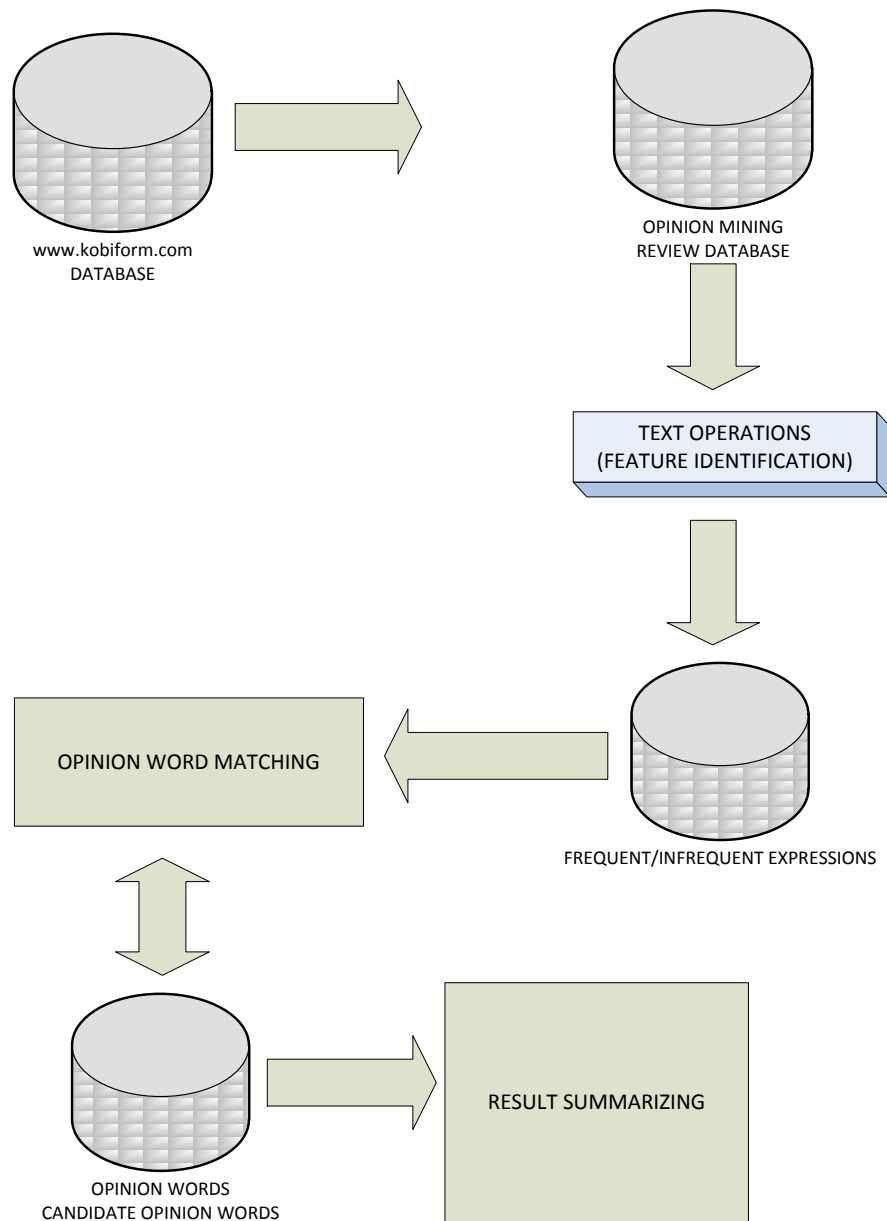


Figure 3.2. Opinion Mining Architecture

System gathers information from user reviews with a crawling module and inserts the collected data to the database. After that reviews are tagged (POST) [8]. These POST tagging is done because it will work as hooks for finding frequent features. After tagging is done and features are identified, opinion words are extracted and their semantic orientations are identified. After the opinion words are extracted and also identified, system flags the infrequent features. After this step, all the opinions are identified and processed so summarization becomes available.

3.1.2.1 POST (Part Of Speech Tagging)

“Part of speech tagging is the process of marking up the words in a text with their corresponding parts of speech reflecting their syntactic category.” [8]

To perform POST, opinions should be split into sentences to achieve a good level of granularity.

Data mining systems will depend on the noun or noun phrases in feature identification. Also, the classification of sentiment will depend on the words which are classified both as adjectives and adverbs in this step to produce a set of possible opinion words.

➤ *Opinion Sentence:*

Opinion sentence is a sentence that includes at least one reference to the object or any attribute of the object.

Example:

I have used this laptop for a year. I am really satisfied with the product performance.

There are no opinion words in the first sentence so it is discarded. The second sentence however satisfies the requirements for an opinion sentence. *Product performance* is a feature and *satisfied* is an opinion word.

3.1.3. Feature Identification

The process used for deduction of product features out of the tagged text is called *Feature Identification*.

Feature identification is the process used to deduce possible product features out of the tagged texts generated by the last step. Normally, POST gives nouns as names of the entities. In this case a noun gives name to the product and its features. We will be defining two categories of features which are named by POST, *frequent features* and *infrequent features*.

3.1.3.1. Frequent Features Identification

Our proposed system extracts only explicit nouns or noun phrases from the text. In the first step, the extracted nouns are labeled as candidate features. Then a proper association mining algorithm finds all of the frequent features' frequent item sets. The idea behind this technique is that features that appear on many opinions have more chance to be relevant, and therefore, more likely to be actually a real product feature. The Apriority algorithm [4] was used to generate the set of frequent item sets and minimum support is considered as 1%.

3.1.3.2. Infrequent Features Identification

Infrequent features identification is a method applied to discover possible infrequent (occurred a small number of time) features where the association mining is unable to identify.

Example:

- “The laptop is great”
- “The operation system on laptop is great”
- “The games installed on laptop are great”

The example shows us that an opinion word could be used more than one object. To mark this as an infrequent feature, these opinion words cannot be found in frequent features. The extraction method is described in figure 3.3 [9].

```
foreach(sentence in reviewdatabase.senteces)
{
    if(it contains no frequent but one or more opinion words)
    {
        find the nearest noun/noun phrase around opinion
        word. The noun/noun phrase is stored in the feature
        set as an infrequent feature
    }
}
```

Figure 3.3. Infrequent feature extraction method [9].

3.1.4. Opinion Sentiment Analysis

The word sentiment is polarity of meaning which is widely used to describe the orientation of texts, sentences and words. This work will deal with sentiment classification of texts represented by users' opinions, therefore the name opinion sentiment. Sentiment analysis classifies the sentiment encoded by texts.

Example:

- This is a delicious meal → Positive meaning
- This cake tastes bad → Negative meaning

This example shows the polarity of sentences with positivity and negativity.

The analysis of sentiment can be applied on different levels of simplicity (words, sentences, texts). For many applications, classifying the sentiment of documents as a whole is sufficient, for others a finer level of granularity might be necessary.

Sentiment classification can be applied to the whole opinion, to sentence or to the each feature in opinion. We prefer and apply feature sentiment classification because it is easy to observe and apply. To give an example to understand why it is preferable;

Example:



Figure 3.4. User opinions for a chair.

In example two reviews are gathered from customers. These opinions have both positive and negative aspects of product in same text. Splitting the opinion into pieces that may construct a sentimental extraction is not enough. For example “*The chair is really good but need softer cushions*” sentence has more positive meaning than negative however it still contains a negative aspect for the chair so this has to be extracted too. If we ignore the negative meaning, we may lose important information. Because of these reasons, the method proposed by *A holistic lexicon based approach to opinion mining* [6] is applied to acquire the optimum granularity level of sentences.

3.1.4.1 Description of Sentiment

Opinions hold positive or negative polarity as meaning. These positive and negative oriented expressions define opinions purpose due to understanding the sentence.

To find the orientation of the opinion words, we are going to use a database which includes the keywords. The database has a recursive relationship status so unidentified words are going to match as synonyms or inserted to database as a new keyword.

3.1.4.2 Finding the Sentiment at the Feature Level

To process an opinion, the sentiment (polarity) of the opinion should be extracted. As mentioned before, opinions can be analyzed at different levels of granularity. Analyzing the sentiment of opinions at feature level is applied in this study.

To find the opinion polarity, all the opinion words for the feature should be extracted. Therefore these rules should be considered before processing:

➤ *Negation:*

Negation words change the orientation of the following word and may affect the whole opinion polarity. As the simple mathematical rule, negation words work as following:

- Negation followed by negative word → Positive meaning
 - “not bad” contains positive meaning
- Negation followed by positive word → Negative meaning
 - “not good” contains negative meaning

➤ *Using word TOO:*

“Too” is often used to give a negative excess to the word. Therefore the when we find word too followed by an opinion, we assume that the orientation becomes negative.

Equation 4.4.2.1 Calculation of the opinion orientation [6]

- S is the sentence which contains the features
- W_i is an opinion word
- V is the set of opinion words
- $W_i.SO$ is the sentiment orientation for W_i
- $d(f, w_i)$ is the distance between feature f and opinion word w_i

For a sentence s that contains a set of features and for each feature f the above orientation score is computed. A positive word is assigned the orientation plus one (+1), and a negative one is assigned minus one (-1). The reason for the multiplicative inverse in the formula is to give low weights to opinion words that are far away from the feature f . The pseudocode of figure 3.5 [6] was used to find the opinion orientation at the feature level.

```

Algorithm OpinionOrientation()
for each sentence  $s_i$  that contains a set of features do
   $features$  = features contained in  $s_i$ ;
  for each feature  $f_j$  in  $features$  do
     $orientation = 0$ ;
    if feature  $f_j$  is in the “but” clause then
       $orientation$  = apply the “but” clause rule
    else remove “but” clause from  $s_i$  if it exists;
      for each unmarked opinion word  $ow$  in  $s_i$  do
        //  $ow$  can be a TOO word or Negation word as well
         $orientation += \mathbf{wordOrientation}(ow, f_j, s_i)$ ;
      endfor
    endif

    if  $orientation > 0$  then
       $f_j$ 's orientation in  $s_i = 1$ 
    else if  $orientation < 0$  then
       $f_j$ 's orientation in  $s_i = -1$ 
    else
       $f_j$ 's orientation in  $s_i = 0$ 
    endif
  endif

  if  $f_j$  is an adjective then
     $(f_j).orientation += f_j$ 's orientation in  $s_i$ ;
  else let  $o_{ij}$  is the nearest adjective word to  $f_j$ , in  $s_i$ ;
     $(f_j, o_{ij}).orientation += f_j$ 's orientation in  $s_i$ ;
  endif
endif
endfor;
// Context dependent opinion words handling
for each  $f_j$  with  $orientation = 0$  in sentence  $s_i$  do
  if  $f_j$  is an adjective then
     $f_j$ 's orientation in  $s_i = (f_j).orientation$ 
  else // synonym and antonym rule should be applied too
    let  $o_{ij}$  is the nearest opinion word to  $f_j$ , in  $s_i$ ;
    if  $(f_j, o_{ij})$  exists then
       $f_j$ 's orientation in  $s_i = (f_j, o_{ij}).orientation$ 
    endif
  endif
  if  $f_j$ 's orientation in  $s_i = 0$  then
     $f_j$ 's orientation in  $s_i =$  apply inter-sentence
    conjunction rule
  endif
endfor

```

Figure 3.5 Calculating the orientation of opinions on product features [6]

CHAPTER IV

4.1 APPLYING OPINION MINING TO AN E-COMMERCE SYSTEM

As mentioned before www.kobiform.com is selected as customer review data source. www.kobiform.com is an e-commerce system focuses on clustering of furnishers in Siteler/Ankara and putting their product on e-market.

Analysts made a research to supply feature database with furniture related keywords. These keywords are inserted into related tables to be used for analyzing of the sentences.

The main operational system of www.kobiform.com is shown in figure 4.1

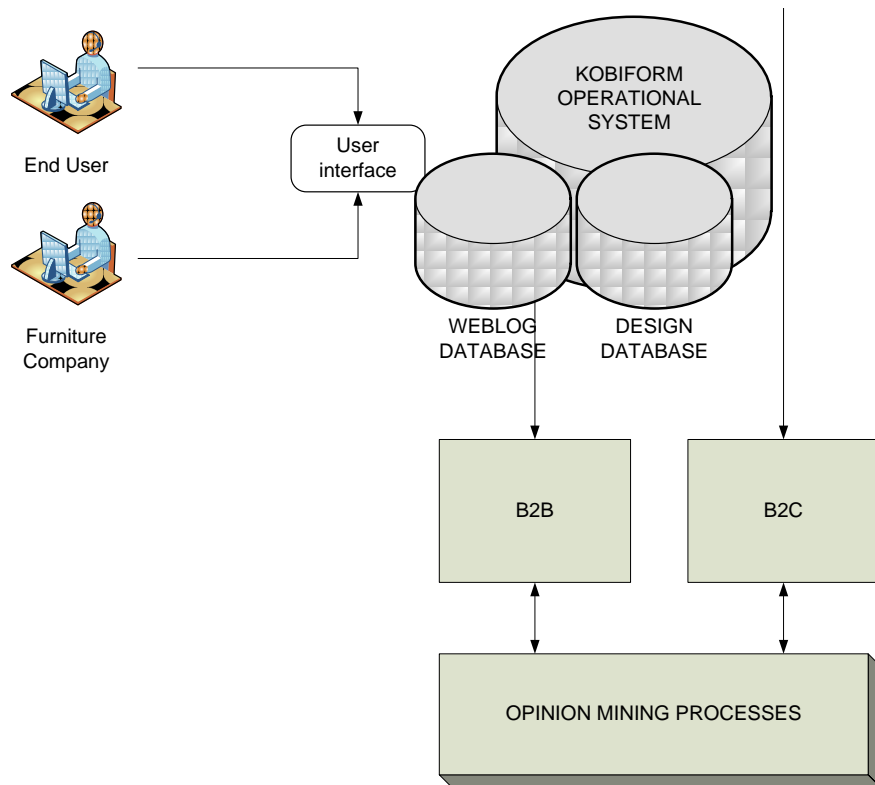


Figure 4.1 Workflow of www.kobiform.com e-commerce system

As seen on figure 4.1 two types of users provides source to the system. Furniture companies supply the system with products and end users supply the system with reviews. They both have different user interfaces.

With the provided data, system database works in sync and distributes the data. Due to the related operation, opinion mining process takes place.

4.1.2. User Profile

www.kobiform.com has wide user profile range. There are three main roles defined in the system which are administrators, product suppliers and customers. Since we focus on opinions of products, we will be focusing on customers.

Administrators are responsible for systems' stability and continuity. Administrators have direct access to feature database to keep the database up to date and can generate reports due to collected data.

Product suppliers are responsible for uploading their products (furniture) to the system. Also suppliers are obligated to send the goods to customer after a successful purchasing transaction.

Customers are the main users of system; hence it is an open system customers could be anyone who are able to purchase a product and have an internet connection.

5.2. Use Case Analysis

As mentioned before there are several steps to achieve a succesful opinion mining process. In addition to required data, system also has to transfer, analyze and schedule processes to extract opinions on products.

Following diagram shows the opinion mining process with use case and following table describes the chart in detail.

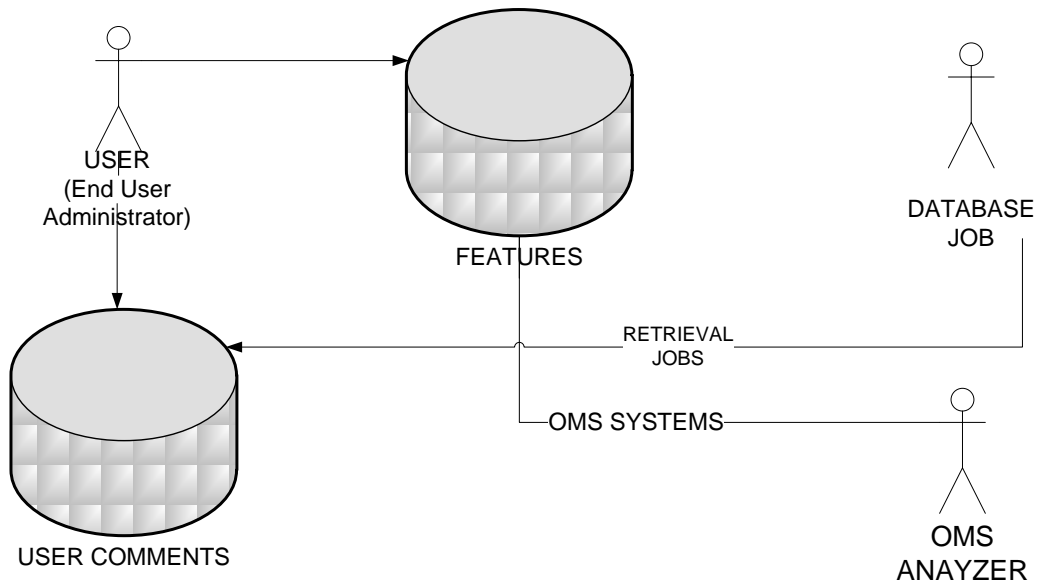


Figure 4.2 Use case of opinion mining

Description	User navigates in products and writes comments.
Primary Actor	USER (customer)
Precondition	User should be signed up to the system to write reviews
Process Flow	<ul style="list-style-type: none"> ➤ User navigates to a product page via product search or simply following links ➤ User writes a comment ➤ Comment is saved in www.kobiform.com database ➤ User is notified for success

Table 2 – User(Customer) role description

Description	User reviews transferred to the OMS databases from www.kobiform.com databases
Primary Actor	DATA CARRIER JOB (Integration Services, Database Role)
Precondition	Jobs should be implemented and services should be running
Process Flow	<ul style="list-style-type: none"> ➤ Scheduler triggers the job ➤ Reviews and related products inserted into OMS database <ul style="list-style-type: none"> ➤ If successful system logs the success message ➤ If job fails job is rescheduled

Table 3 – Database Job(Data Transfer) role description

Description	User reviews are examined
Primary Actor	OMS ANALYZER
Precondition	User reviews should be supplied by the database
Process Flow	<ul style="list-style-type: none"> ➤ User review is analyzed for frequent/infrequent/opinion/negate words ➤ If frequent sentence is matched increment the hit count and insert data to analyze result with polarity ➤ If infrequent sentence is matched increment the hit count and check for the infrequent sentence achieves to immigrate to frequent sentence ➤ If opinion word sentence is matched insert data to analyze result with polarity ➤ If negate word sentence is matched insert data to analyze result with polarity ➤ If none of the words matches insert the sentence for examination

Table 4 – OMS Analyzer role description

Description	User reviews are examined
Primary Actor	OMS SUMMARIZE
Precondition	Analyze result table should be supplied
Process Flow	➤ According to the analyze result table, opinions related to the products are summarized

Table 5 – OMS Summarize role description

4.1.3. System Requirements

www.kobiform.com needs a series of operations to perform stable. Most of these operations are implemented to perform as scheduled jobs although some of the operations require human interaction.

4.1.3.1 General Operations

General operations could be described as,

- Opinion Mining System must collect user opinions with reviews and surveys.
- Opinion Mining System must identify features in opinions.
- Opinion Mining System must log if a sentence cannot be concluded.
- Opinion Mining System must collect and store newly added feature keywords.
- Job scheduling must not fail, in case of any failure system must log the exception.

4.1.3.2 Supportive Requirements

There are some important extensions for www.kobiform.com to perform efficiently. Since these extensions do not effect on system health or operation status, can be able to improve performance and precision.

- *Performance:*

Although scheduled mining operations will run at system's least busy time periods, all the integration service, review analyzing, opinion mining summarizing processes should be handled in acceptable amount of time.

➤ *User Interface:*

Because of the system's orientation, user interface should be simple, easy to access and understand. Interface should allow users to navigate, find and interact on products with ease.

➤ *Interoperability:*

Opinion mining system integrates with www.kobiform.com. Since the system gathers the resource and opinions from same domain, interoperability becomes easy to handle.

4.1.4. Implementation

OMS requires data from www.kobiform.com. Because of this, www.kobiform.com is implemented to work compatible with OMS. Since we do not concern the way how data supplier platform works, we will be focusing on the implementation of OMS.

4.1.4.1 Technologies

Microsoft based products are used during the implementation of OMS.

4.1.4.1.1 Database

The most important part of the OMS is the database platform because of the heavy data operations weight. Beside the entire database operations and the massive amount of data, also scheduled jobs are to be executed properly for a healthy system.

MSSQL Server is used as database. Data transfer between www.kobiform.com server and OMS server is done with the SQL Server Integrated Services as scheduled job. Also the reports concerning products are generated by Analysis Services. Data Cubes are produced with these services, too.

The database architecture is shown in figure 4.3 for a better understanding.

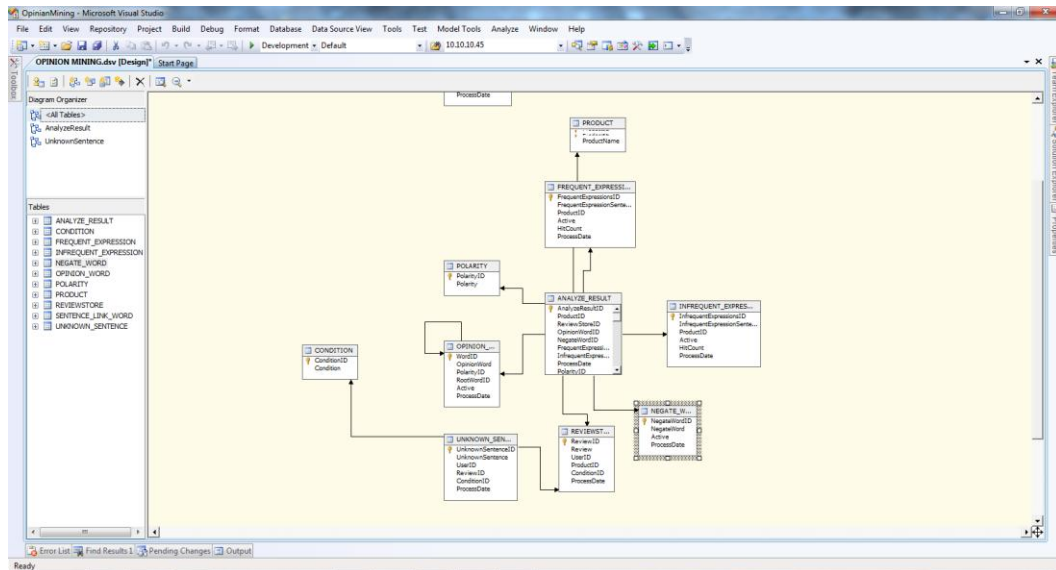


Figure 4.3 Database architecture

4.1.4.1.2 Product Analyzer Software

The analysis of products based on the user reviews are made with a scheduled running service implemented by .Net framework 3.0. All the operations are handled with the built-in libraries; also new libraries developed which inherits built-in libraries.

4.1.4.2 Architecture

Architecture of OMS is composed by two modules called *Opinion Mining Module* and *Text Process Module*. Since the data is retrieved from www.kobiform.com production database, information retrieval is handled with SSIS (SQL Services Integration Services) jobs. These services transfers user reviews from www.kobiform.com database to OMS database. This batch data transfer process is a scheduled job and due to performance concerns it is handled at most idle hours. All the review data entered by users are transferred with the product relations for OMS analyze.

After the data immigration, the reviews are analyzed with the Text Analyzer module. Due to orientation of the sentence, reviews get classified and inserted to the database for summarization.

As shown in figure 4.4, Text Process Module is responsible for classifying the sentences which contains opinions. After sentences are processed and classified, they are ready to be summarized.

Opinion mining module identifies the features of products out of tagged opinions of reviews. All features are classified and an output which consists of each discovered feature and its sentiment related analysis is generated. With the classification, the polarization of the sentence is also discovered. The discovered results are inserted into Analyze table for reporting the opinions on products.

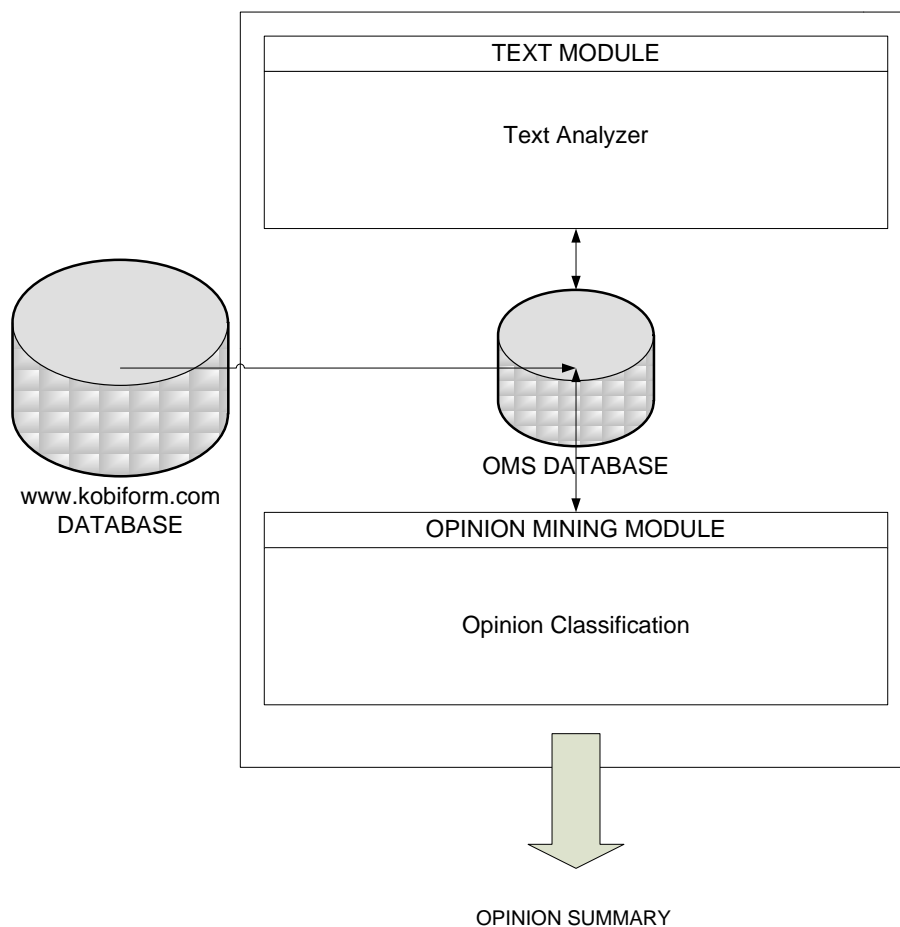


Figure 4.4 Architecture of OMS system

4.1.4.3 Text Process Module

This module handles the tagging of opinions with respect to their part-of-speech. This module works as scheduled due to performance issues, since the job activity must be logged and observed for any failures so this module requires human interaction at some point.

The extraction of sentimental opinions from review sentences is a heavy weight process. First of all the sentence is checked whether it is a frequent or infrequent sentence. Most of the time if the sentence is not a survey; result sentence is not included in the frequent/infrequent sentence set. If the sentence is not a member of these sets, sentence is checked for possible opinion words. If the sentence includes opinion words we would be checking for the negate words.

Since the inserting all sentences into infrequent sentence database results as massive and possibly unnecessary data, the data in this database cleaned periodically. Each infrequent sentence has a hit count and if the counter doesn't increment in time data is marked as unused.

Some of the reviews may not contain any of the above.

Example:

Bu ürünü aldım. (No polarity)

Bu ürünü aldığıma sevindim. (Polarity but not included in database)

If we assume that the sentences on examples are not in the database, both sentences are inserted into database for analyzing. This analyze requires human interaction because of the difficulty of extracting the meaning of the sentence. First sentence is ignored due to no polarity and added as neutral. Second sentence has a positive meaning and added into infrequent sentence database for future use.

4.1.4.4 Opinion Mining Module

This module is the last step before summarization. The outputs of the opinion mining module are associated with the products and actual reports are generated. Before the summary is generated, features should be identified and classified due to their sentimental value.

➤ *Opinion Model:*

After querying the opinions from database after processing with text analyze module, opinion model divides the review into sentences. Each sentence may contain an opinion and usually matching the keywords with the opinion word database is sufficient to polarize the sentence.

After the division of sentence to words, each word is checked for their category. Words can be opinion words or negate words. Opinion words and negate words are predefined words in database and their polarity is preset. Negate word would change the polarity of the sentence to the opposite.

Example:

Bu ürün güzel. (No negate words and positive meaning, polarity +1)

Bu ürün güzel değil. (Includes negate word and positive meaning changes to negative, polarity -1)

Bu ürün kötü. (No negate words and negative meaning, polarity -1)

Bu ürün kötü değil. (Includes negate word and negative meaning changes to positive, polarity +1)

As seen on examples, negate words change the polarity of the sentence so after the examination of the opinion words, checking for negate words is a necessary process. If the opinion and negate word operations concludes a result, the polarity and the sentence is inserted into infrequent sentence database for future reuse.

One of the most difficult parts of the opinion mining is linked sentences. Although some of the sentences are easy to divide and understand, some of them are tricky.

Example:

Ürünü çok güzel fakat biraz ağır. (Two opinions and two polarities in the sentence, +1 and -1, concludes as neutral)

Ürünü aldığımda çok beğeniyordum fakat kötü bir ürünmüş. (Two opinions and two polarities in the sentence, +1 and -1, concludes as neutral but the final opinion is negative so in fact the extracted polarity should be negative)

As seen on sentences, linked sentences change the total output of the meaning. In first example there are two opinions, one is positive and one is negative so total outcome is neutral. In second example also has a neutral polarity outcome but the actual meaning is negative due to final opinion of user. It is not an easy task to process and understand the total meaning of sentence and it is planned as future work. After matching the polarity of sentence is summarized.

If any of the words do not outcome a reasonable output sentence is stored in database to be analyzed properly with a human interaction.

The opinion analyze step is followed by the reporting step.

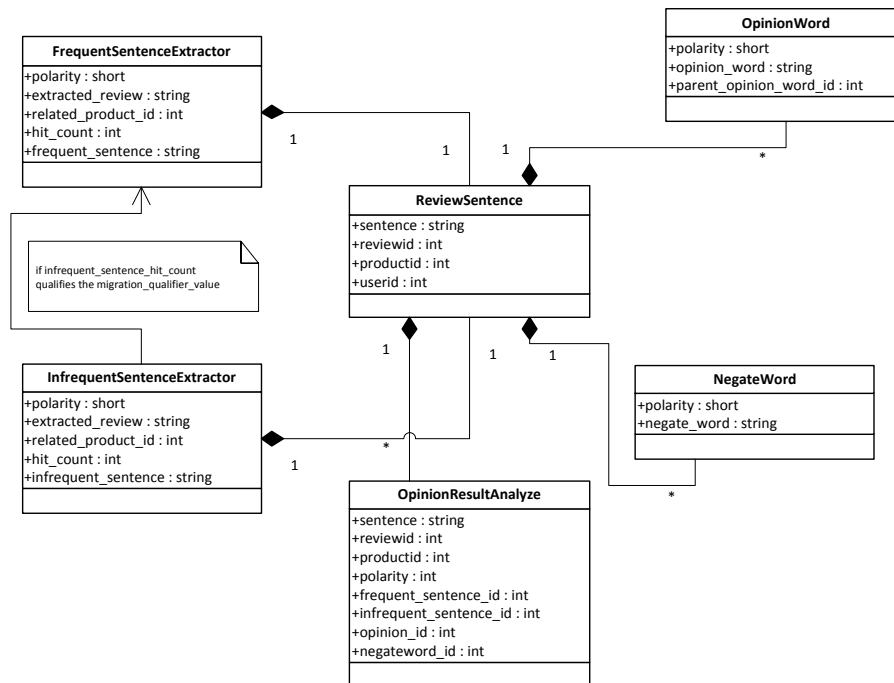


Figure 4.5 UML Class Diagram of Opinion Mining System

For generating the summary main function works with all other functions and objects. The function searches for available sentimental extraction from all of the collections. If opinion is available then the polarity of the sentence is calculated.

The sentence may contain features of the product. For calculating the polarity score all of these features must be calculated, too. In this study we will skip the feature extraction step and calculate the polarity with the containing opinion words but dividing the features from sentence and calculating scores for each feature is a better practice. To accomplish this we should find the sentiment of all opinion words, calculate the score for each feature and generate the feature based summary. A simple process is shown in figure 4.6.

```
def generate_summary
  @features.each do |frequent_feature|
    @opinions.sentences.each do |sentence|
      sentence.words.each do |feature|
        if feature.name.remove_tags == frequent_feature && feature.instance_of?(
          Feature)
          sentence.words.each do |word|
            if word.instance_of?(OpinionWord)
              word.orientation = @word_orienter.get_opinion_word_orientation(
                word, sentence)
              if word.orientation == 0
                word.orientation = self.get_orientation_from_sentence_context(
                  sentence, word)
              end
              feature.score = feature.score + word.orientation / (feature.position
                - word.position).to_f.abs
            end
          end
          @summary.add_item_to_summary(feature.name.remove_tags, feature.score,
            sentence.text.remove_tags)
        end
      end
    end
  end
  @word_orienter.save_to_seed_list
end
```

Figure 4.6 Feature extraction function [6]

As mentioned before OMS searches for the opinion words in sentence for a successful match. Whenever a word is found on collection, algorithm calculates the polarity of the sentence. The opinion may change with the negate words; hence after finding the opinion word checking for negate words is necessary. After including negate words for calculation of sentence the analysis is almost complete [6]. The summary collection is treated as the final stage of OMS and results are inserted into database. If there is no match, sentences inserted for better analyzing. Following figures illustrates the mentioned processes.

```

get_sentence_from_reviewdatabase
  while(sentences_to_be_analyzed_cap not reached)
    if(sentence is not analyzed)
      insert sentence into analyze_collection
    else
      skip the sentence
  end

get_opinion_collection_from_reviewdatabase
  foreach(sentence in collection)
    check for succesfull match for analysis
    if(succesfull)
      calculate polarity
    else
      insert for a better analyze
  end

```

Figure 4.7 Selecting the reviews and opinion collection

```

# Apply negation rules if negation words are found nearby
# an opinion word
def apply_negation_rules?(word, sentence)
  sentence.words.each do |sword|
    if sword.instance_of?(NegationWord)
      if (word.position - sword.position) > 0 && (word.position - sword.
        position) <= negation_word_range
        return true
      end
    end
  end
  return false
end

# "Too" words when togheter with adjectives usually denotes
# negative sentiment "This screen is too small"
def apply_too_rules?(word, sentence)
  if sentence.words[word.position - 1].instance_of?(TooWord)
    return true
  else
    return false
  end
end

```

Figure 4.8 an approach for calculating the negate words. [6]



Figure 4.9 Complexity of negation words in examples

```

foreach sentence si
  if sentence == linked_sentence
    divide the sentence into meaningful pieces;
    foreach meaningful granule of sentence after division
      if sentence == frequent_feature_collection
        find polarity value of sentence from frequent_feature_collection;
        process polarity calculation;
        increase frequent_collection_hit_count;
      else if sentence == infrequent_feature_collection
        find polarity value of sentence from infrequent_feature_collection;
        process polarity calculation;
        increase infrequent_collection_hit_count;
        check for the infrequent_collection_hit_count to be updated to frequent_collection;
      else if sentence contains opinion words
        find polarity value of sentence from opinion_words_collection;
        process polarity calculation;
        if sentence contains negate words
          invert the polarity value for negation;
        process polarity calculation;
      else
        insert sentence as unknown_sentence for analyze;
        mark polarity as neutral;
        summary_collection.additem(polarity_of_sentence)
      end
    end
  else
    if sentence == frequent_feature_collection
      find polarity value of sentence;
      process polarity calculation;
      increase frequent_collection_hit_count;
    else if sentence == infrequent_feature_collection
      find polarity value of sentence;
      process polarity calculation;
      increase infrequent_collection_hit_count;
      check for the infrequent_collection_hit_count to be updated to frequent_collection;
    else if sentence contains opinion words
      find polarity value of sentence from opinion_words_collection;
      process polarity calculation;
      if sentence contains negate words
        invert the polarity value for negation;
      process polarity calculation;
    else
      insert sentence as unknown_sentence for analyze;
      mark polarity as neutral;
      summary_collection.additem(polarity_of_sentence)
    end
  end

  foreach summary_item in summary_collection
    associate summary with product;
    insert summary into analyze_result;
  end
end

```

Figure 4.10 Pseudocode of Opinion Mining Module

Most of the applications decide the word orientation from accepted word dictionaries for that language but there is none in Turkish. Hence, all the words should be stored and analyzed within the database operations which make human interaction is obligatory.

Since searching opinion words in sentences is a heavyweight operation, proposed frequent and infrequent sentence concept tries to help with the process. User comments divided into smallest meaningful sentences and their polarity is calculated and stored. Due to the frequency of these sentences they are labeled as frequent or infrequent opinion sentences.

- *Infrequent Opinion Sentence:* If the processed user review is not stored in database but is eligible for opinion mining (contains opinion words and has a polarity), it is considered as infrequent opinion sentence. If the sentence could not be marked as infrequent sentence it is considered as a sentence to be analyzed. In this case system assumes the sentence whether has unknown opinion words or no adjective. System cleans the long lingering sentences from infrequent opinion sentence database.
- *Frequent Opinion Sentence:* Each frequent opinion sentence is derived from infrequent opinion sentence. After the cap hit count of infrequent sentence is hit the sentence migrates to frequent opinion sentence database and considered as frequent opinion sentence.

The frequent sentences are also considered as common opinions of users upon an object or product. These sentences could be used in surveys, product definitions, finding up and downsides of products and defines the product relations with its attributes.

4.1.4.5 Reporting Of the Analyzes and Evaluation

This study is based on extracting the user reviews on products. Since we are applying the OMS application on www.kobiform.com e-commerce system, user reviews are highly important for producers as well as the suppliers. The results have been handed to the furniture producers in Siteler and the positive effects of the recommendations are observed.

The effectiveness of the system is not %100 due to complexity of opinion mining and some of the precise data gathered from surveys however results came out better than expected. By design, first phase of the algorithm do not aim to cover all the corners of the mining process, like the unordered sentences, linguistic differences or specific word combinations. Statistical result table (table 5.4.5.10) shows the output for the efficiency of the algorithm in numbers. Results show instead of the performance benefits, creating frequent and infrequent sentence sets require a lot of data and comments are not common as anticipated. Producers benefit on the results of the OMS reports and Siteler Trade Corporation awarded the project with a plaque for precise outputs.

Some of the result reports gathered from MSSQL Analysis Services on products are shown below. These reports are created with the smart data cubes and the measurements. Also these smart cubes may be queried with the required dimensions; since it is generic it is possible to create any kind of report necessary.

Product ID	Polarity ID	Frequent Expressions ID	ANALYZE RESULT Count
Massa			89
Sandalye			87
Çocuk Takım			83
Öturma Takım			87
Mutfak Takım			84
Koltuk			87
Celvyat			87
Büroist			77
Bilgisayar Masaası	Positive	Ürün çok kullanışlı	1
		Ürün kullanımı çok rahat	1
		Gerçekten pratik bir ürün	1
		Parayla alınabilecek en iyi ürün	1
		Ürünün malzemeleri kolay bulunuyor	1
		Kullanıldığı belli ayrılmıyor	1
		Ürün tam belediğim gibi	1
		Üründen çok memnunuz	1
		Ürün tüm ihtiyaçları karşılıyor	1
		Ürünü keyifle kullanıyorum	1
		Ürünü kurulumu çok kolay	1
		Güzel bir ürün	24
		Total	35
	Negative	Ürün kullanışsız	1
		Ürünü beğenmedim	1
		Ürünün paraman değerini vermiyor	1
		Ürün çabuk eskidi	1
		Ürünün kurulumu çok zor	1
		Ürün garantisi daha uzun olmalı	1
		Görünüşü güzel değil	30
		Total	36
	Total		71
Çardap			83
Yataak			71
Baza			78
Avize			71
Decoratif Tablo			76
Abajur			68
Puf Oturak			82
Sifonyer			66
Bisiklet retro yataak			72
Tek kişilik genis oda takım			68

Figure 4.11 frequent expression analysis on a product

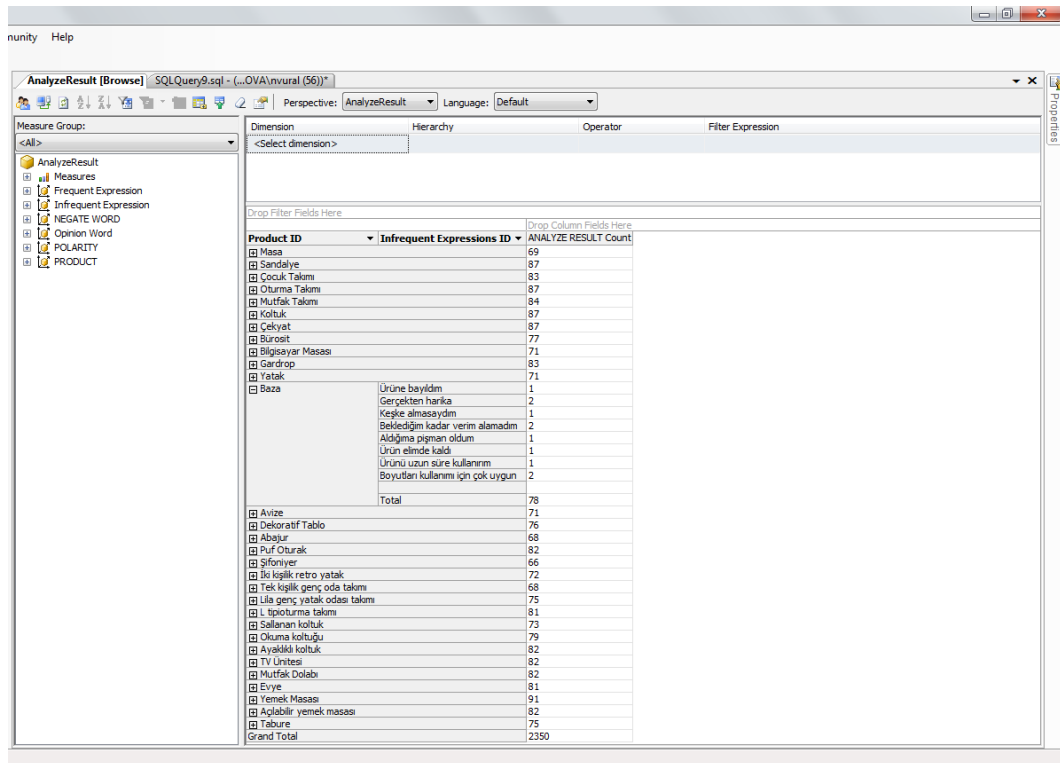


Figure 4.14 infrequent expression analysis on a product

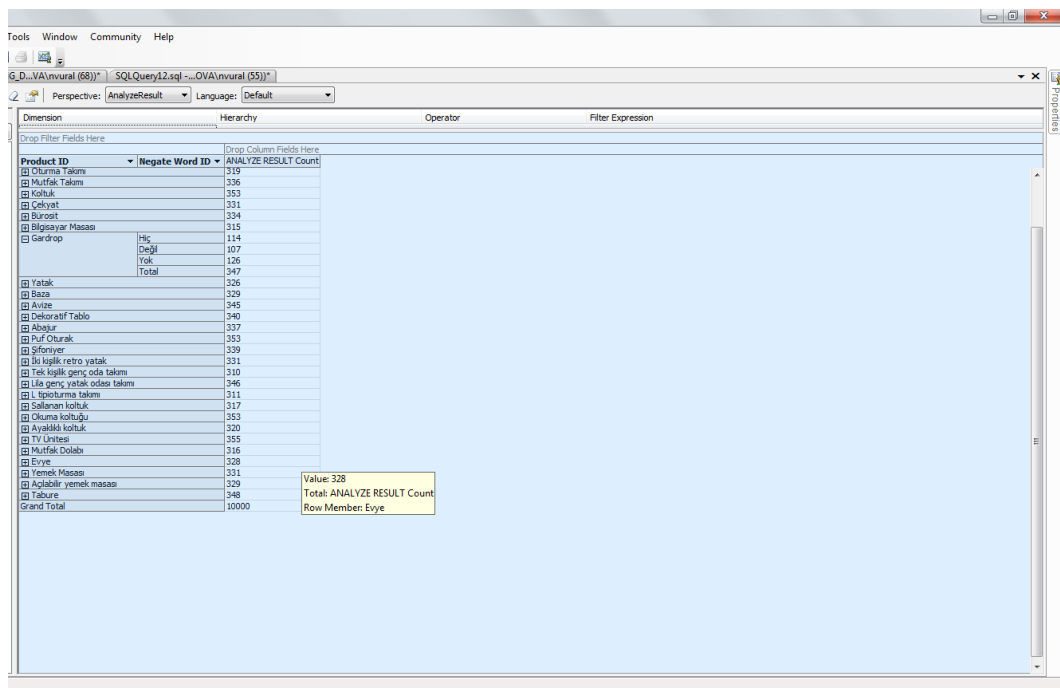


Figure 4.15 negate words analysis on a product

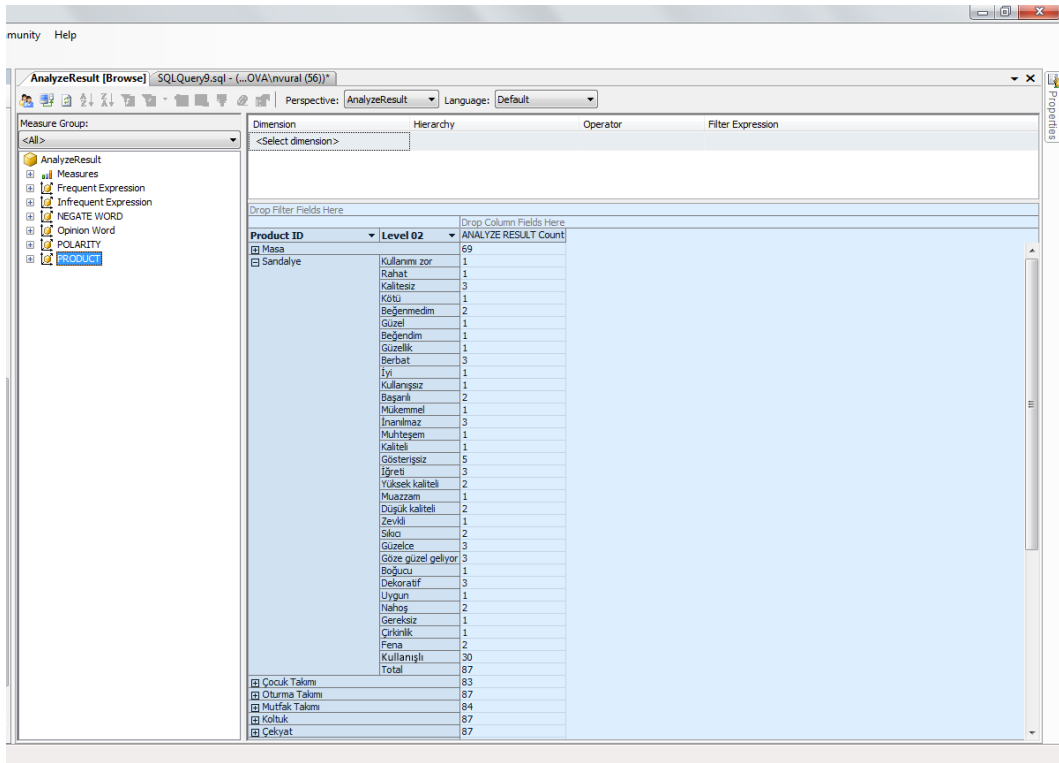


Figure 4.16 Opinion analysis on a product

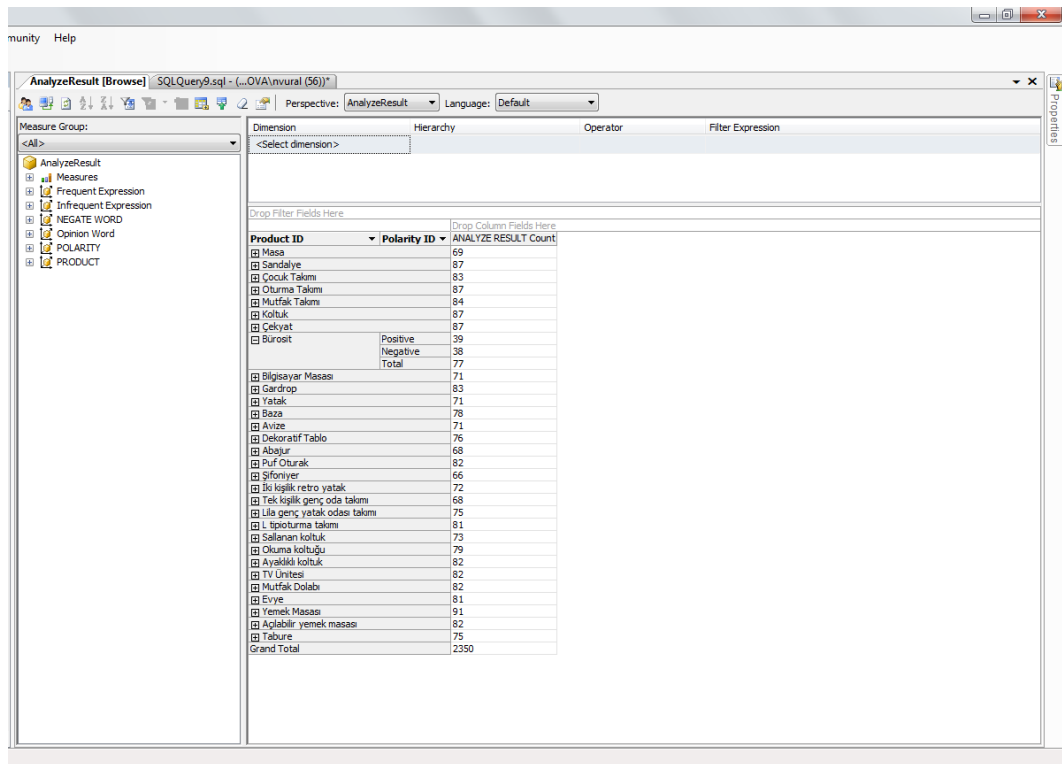


Figure 4.17 Polarity analysis on a product

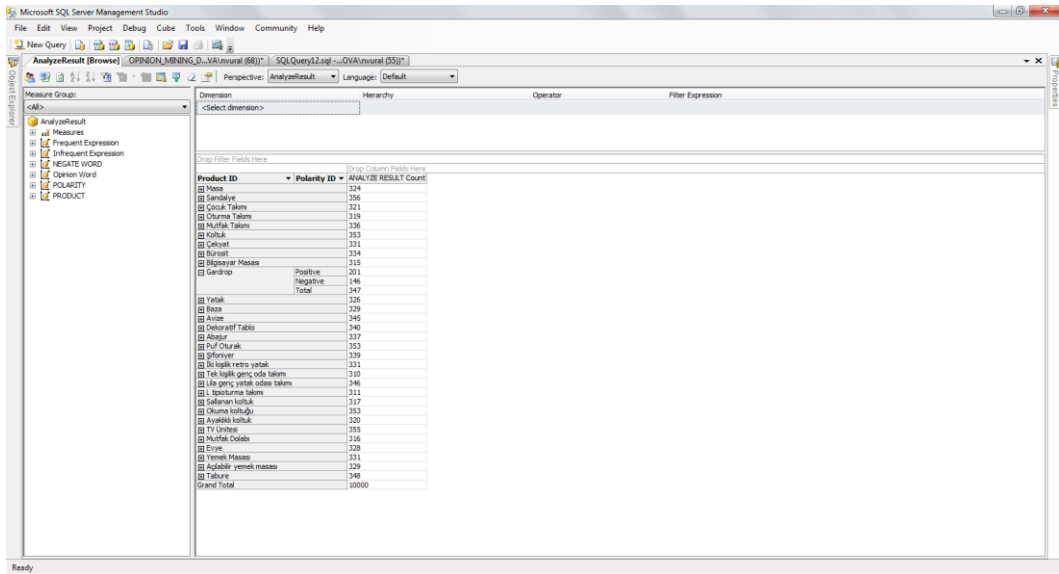


Figure 4.18 Polarity analysis on a product

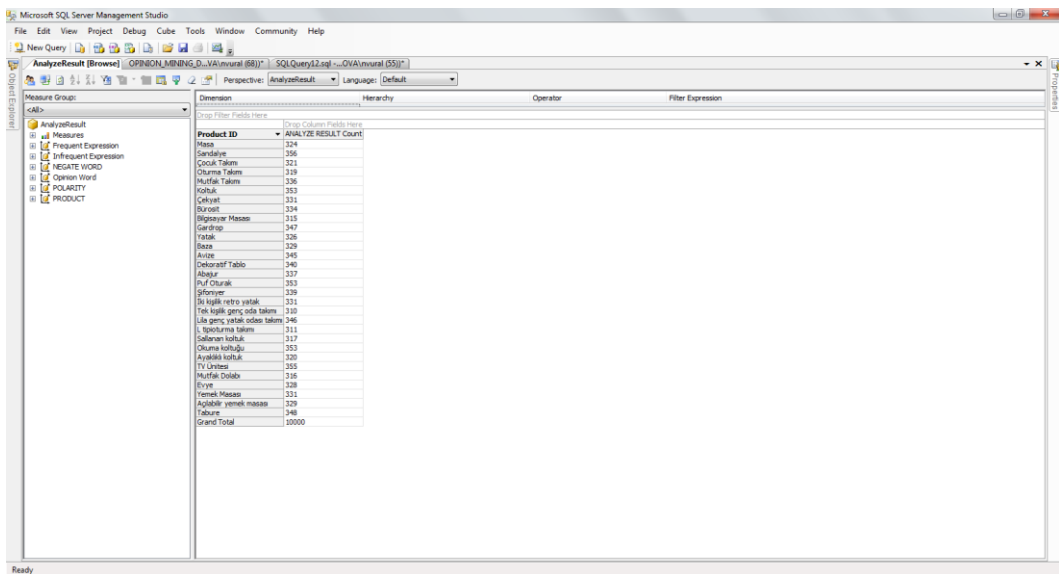


Figure 4.19 Total analyze calculations

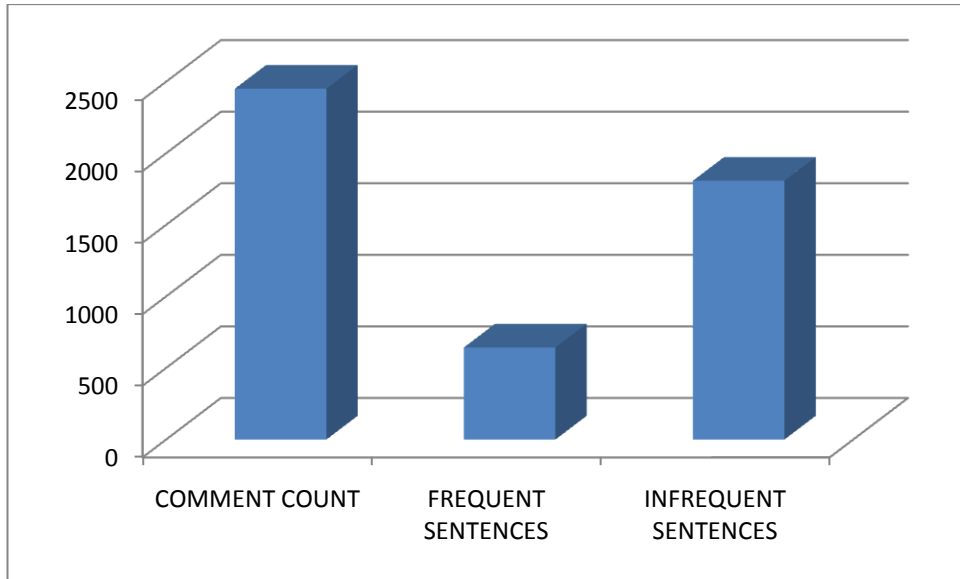


Figure 4.20 Frequent/Infrequent sentence comparison (1808 infrequent sentences, 643 frequent sentences)

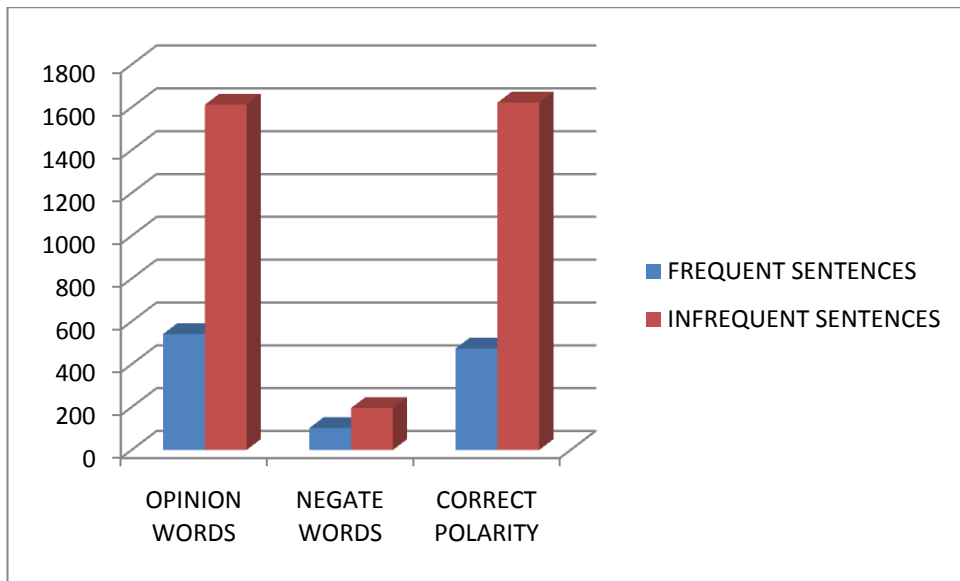


Figure 4.21 Frequent/Infrequent sentence numbers of opinion and negate words (541 frequent opinion sentences, 102 frequent negate sentences; 1612 infrequent opinion sentences, 196 infrequent negate words) with correct polarity calculation (474 out of 643 frequent sentences and 1622 out of 1808 infrequent sentences)

SUMMARY AND CONCLUSION

Hence opinion mining is a newly advancing field of study; its importance is rapidly increasing. Positive and concrete benefits gathered from user reviews increase the popularity of opinion mining. Application fields of opinion mining expand too. Although it seems to fit to e-commerce field best, opinion mining also applies to business, government intelligence and medical researches too.

This study is a basic yet useful application of opinion mining practicing the fundamentals of data mining.

In this thesis, a simple application of opinion mining is applied to an e-commerce site.

The purpose of this study is to give readers a new perspective on e-commerce applications by creating frequent and infrequent opinion sentences concept and applying opinion mining to e-commerce applications.

Also this study aims to enhance both opinion mining process performance and product definition efficiency by using the user generated structured opinion sentences.

As mentioned before, the study applied in this thesis is also submitted and won the Teknogirişim Fund which is held by Ministry of Science, Industry and Technology. Proposed project plan is divided into two sections, first section includes applying a simple and fundamental opinion mining application into the developed e-commerce clustering system and second section includes improving the mining algorithm and expands its usage.

After the analysis and comparisons this thesis showed that creating opinion container sentences (frequent and infrequent opinion sentences) dramatically improves the mining performance hence it increases the complexity and decreases the efficiency.

The idea of creating common structured user opinions might work for products but it should be noted that each user has his/her own unique way of expressing their opinions which makes grouping these under common opinion sentences quite difficult and definitely not a good practice. In other words, each person to be analyzed uses different words to express their ideas and this outcome makes this concept difficult to use.

As future work more detailed and expanded approach to analyzing the sentences and opinion extraction is aimed. However complexity of the language and linguistic operations is not trivial. Words have different meanings or they form a new meaning when come together. Solving this problem will achieve an important improvement for opinion mining.

REFERENCES

- [1] Menczer, F. (1997). ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery. In D. Fisher, ed., Proceedings of the 14th International Conference on Machine Learning (ICML97). Morgan Kaufmann.
- [2] Chen Ding et al. (2010). Network and Parallel Computing: IFIP International Conference NPC 2010. p. 91.
- [3] <http://gate.ac.uk/ie/>.
- [4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. 1994.
- [5] Raymond Kosala, Hendrik Blockeel; *Web Mining Research: A Survey*, volume 2, issue 1 - page 4
- [6] Philip S. Yu Ding Xiaowen, Liu Bing. A holistic lexicon based approach to opinion mining. WSDM'08, 2008.
- [7] Bing Liu Department of Computer Science, University of Illinois at Chicago (On-line available) <http://www.cs.uic.edu/~liub/FBS/opinion-mining.pdf>
- [8] A Composite Approach for Part of Speech Tagging in Turkish
Levent Altunyurt, Zihni Orhan, Tunga Güngör Boğaziçi University, Computer Engineering Dept., Istanbul, Turkey
- [9] Mining Opinion Features in Customer Reviews, Mining Hu and Bing Liu
Department of Computer Science University of Illinois at Chicago
(On-line available) <http://tinyurl.com/98e46ww>

BIOGRAPHY

PERSONAL INFORMATION

Surname, Name : UYTUN, Muhammed Burak

Nationality : T.C

Birthdate and birthplace : 13.03.1983

Mobile phone : +90 532 3606392

E-Mail : burakuytun@gmail.com

EDUCATION

DEGREE	SCHOOL	GRADUATION DATE
University	Çankaya University	01.03.2007
Highschool	Mimar Kemal Anadolu Lisesi	15.06.2001

JOB EXPERIENCES

YIL	YER	POZİSYON
2007 - 2008	CCNET	Software Developer/ Project Manager
2008 - 2009	Noe Grup	Partner / Project Manager
2009 - 2011	Kale Yazılım	Software Developer
2011 - 2012	Luna ICT	Partner / Product Manager
2012 - ...	Kale Yazılım	Senior Software Developer

LANGUAGES

English – Advances

French - Beginner