Integration of Opinion Mining into Customer Analysis Model

A THESIS SUBMITTED TO
THE FACULTY SCIENCE AND ENGINEERING
OF QUEENSLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mohd Ridzwan Yaakub

Faculty Science and Engineering
Queensland University of Technology

June 2015
Copyright in Relation to This Thesis

© Copyright 2015 by Mohd Ridzwan Yaakub. All rights reserved.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

Signature:

Date: 23 June 2015
To my family
Abstract

In today’s information era, ample data are easily available through many media such as the Internet, books, journal papers, newspapers, radio and also television. The collection of data sometimes complicates the job in information technology area because of huge data pool to be scrutinized and extracted. Data mining is an area, most affected by this situation because a lot of data are extracted for selection before any decision is made or do any tasks. The emergence of data warehouse and opinion mining in information technology helps data mining in term of choosing the best data for the right tasks. In doing this, products’ ontology model for database in data warehouse will need to be developed to get the best data for data mining tasks.

In the field of Information Technology (IT), Customer Relationship Management (CRM) system can be used for dealing with structured data in a database. It can establish contacts and manage communications with customers, analyse information about customers and make campaigns to attract new customers. However, it is very hard to integrate customers’ comments and feedback into the CRM system because most comments and feedback are described in text opinions (unstructured data). Currently, Opinion Mining is increasingly important than ever before, especially in customer preference analysis and prediction. This work proposes a novel, multi-dimensional model for Opinion Mining, which integrates customers’ characteristics and their opinions about products (or services). Gaining access to customer opinions helps to deliver the right products and services to them, and thus, help organisations and companies to make more profit. Opinionated comments are transferred to a fact table and represented in multi-dimensions such as customer, product, time and location. Based on the fact table, customer opinions are then analysed by exploiting data warehouse techniques including Online Analytical Processing or OLAP and data cubes. This research presents a comprehensive way to evaluate customers’ orientation for any possible attribute of a products.

Our new architecture of CRM combines personal record of customer, product record and
feedback from customers regarding a particular product that they have already used. This new architecture is very significant for companies and manufacturers to obtain the best comprehensive summary of opinions from customers regarding a particular product, especially for future product improvement. Our testing and evaluation also showed better results compared to our baseline models, with a precision of 0.890 and a recall of 0.872. In addition, a report on the orientation of user opinion based on customer group \( (ogc) \) can help customers to make a decision while they are still undecided about a particular product. The research discovered that feature selection based on ontology technique and synonym produced very significant finding for every feature and its sentiment word. Furthermore, our \( ogc \) formula shows a good calculation for every single pair of feature and its sentiment word since the system’s report, which is generated by SQL Server shows the polarity of every product feature commented by others customers.
Keywords

opinion mining, sentiment analysis, customer analysis model, feature ontology, data warehouse, structured data, unstructured data, OLAP, Data Cube
Acknowledgments

I wish to thank the following people who have contributed their ideas, time and energy towards making a successful completion of my thesis.

- The utmost thanks goes to my supervisory team, led by Professor Dr Yuefeng Li as my principal supervisor and Professor Dr Yanming Feng and Dr Jinglan Zhang as my associate supervisors.
- My team from e-Discovery Lab
- My mum
- My lovely children
- My supportive wife
# Table of Contents

Abstract v

Keywords ix

Acknowledgments xi

Nomenclature xvii

List of Figures xx

List of Tables xxi

1 Introduction 1

1.1 Background ........................................ 3

1.2 Aim and Research Questions .................................. 6

1.3 Research Objectives ..................................... 9

1.4 Methodology .......................................... 10

1.5 Significance of the Research .................................. 12

1.6 Publications ........................................... 13

1.6.1 Journal Paper .......................................... 13

1.6.2 Conference Papers ....................................... 13

1.6.3 Plenary Speaker ........................................ 14

1.7 Thesis Structure ......................................... 14

1.8 Summary ................................................ 15
# Literature Review

2.1 Introduction .................................................. 17
2.2 Term Definition ................................................ 18
2.3 Customer Relationship Management ........................... 23
   2.3.1 Customer Relationship Management ...................... 23
2.4 Ontology ......................................................... 26
2.5 Data Mining (DM) .............................................. 27
   2.5.1 Data Warehouse ........................................ 29
   2.5.2 Opinion Mining and Polarity ............................ 31
2.6 Feature Extraction ............................................. 32
2.7 Opinion Mining Methods ...................................... 35
2.8 Levels of Opinion Mining .................................... 36
   2.8.1 Association Rules ....................................... 37
   2.8.2 Manual .................................................. 37
   2.8.3 Lexicon Approach ...................................... 38
   2.8.4 Machine Learning ...................................... 39
2.9 Summary ......................................................... 39

# Customer Relational Model

3.1 Introduction .................................................. 41
3.2 Customer Relationship Management ........................... 42
3.3 Data Mining in CRM ........................................... 52
   3.3.1 Performing Customer Analysis .......................... 57
3.4 Problem in CRM ................................................ 59
3.5 New Idea to Enhance CRM .................................... 61
3.6 Integration of Customer Comment with CRM system ......... 63
3.7 Summary ......................................................... 65
# Opinion Integration Model

4.1 Introduction ........................................ 67
4.2 Customer Data Integration \((CDI)\) ...................... 69
4.3 Research Process ........................................ 72
4.4 Integration of Opinion in \(CRM\) ......................... 76
4.5 Summary .............................................. 83

# Algorithms of Opinion Integration Model

5.1 Introduction ........................................ 85
5.2 Customer Relationship Management \((CRM)\) ............. 86
5.3 Model Development and Implementation .................. 89
  5.3.1 Extraction Method .................................. 91
  5.3.2 Transformation Method ............................. 96
  5.3.3 Opinion Extraction ................................ 100
  5.3.4 Attribute Polarity ................................ 102
5.4 Opinion Integration .................................... 103
  5.4.1 Data Warehouse Based Design ....................... 104
  5.4.2 Orientation ........................................ 104
  5.4.3 Report ............................................ 107
5.5 Summary .............................................. 110

# Evaluation and Discussion

6.1 Introduction ........................................ 113
  6.1.1 Testing Environment ............................... 115
  6.1.2 Dataset ........................................... 116
  6.1.3 Baseline Models .................................. 117
  6.1.4 Evaluation Measures ............................... 118
  6.1.5 The Proposed Model ............................... 119
## Nomenclature

### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Attribute Polarity</td>
</tr>
<tr>
<td>CD</td>
<td>Customer Detail</td>
</tr>
<tr>
<td>CDI</td>
<td>Customer Data Integration</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>DM</td>
<td>Data Mining</td>
</tr>
<tr>
<td>EC</td>
<td>Extracted Comment</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract, Transform, Load</td>
</tr>
<tr>
<td>FeN</td>
<td>Feature Noun</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FO</td>
<td>Feature Ontology</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FS</td>
<td>Feature Synonym</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>NN</td>
<td>Noun</td>
</tr>
<tr>
<td>OGC</td>
<td>Opinion Group of Customer</td>
</tr>
<tr>
<td>o(gc)</td>
<td>Orientation Group of Customer</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Processing</td>
</tr>
<tr>
<td>OM</td>
<td>Opinion Mining</td>
</tr>
<tr>
<td>OS</td>
<td>Opinion Sentence</td>
</tr>
</tbody>
</table>

xvii
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>Part of Speech</td>
</tr>
<tr>
<td>PT</td>
<td>Proposed Technique</td>
</tr>
<tr>
<td>SCD</td>
<td>Slow Changing Dimension</td>
</tr>
<tr>
<td>SCV</td>
<td>Single Customer Value</td>
</tr>
<tr>
<td>SMS</td>
<td>Short Message Service</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SSRS</td>
<td>SQL Server Reporting Service</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>WiFi</td>
<td>Wireless Fidelity</td>
</tr>
<tr>
<td>WWW</td>
<td>World Wide Web</td>
</tr>
<tr>
<td>3G</td>
<td>Third Generation</td>
</tr>
<tr>
<td>4G</td>
<td>Fourth Generation</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Four Important Steps in Opinion Mining</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>Research Methodology</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>Changing Marketing Strategies (Chen and Popovich [2003])</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Product Ontology by Somprsertsri and Lalitrojwong [2010]</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>Data Warehouse</td>
<td>30</td>
</tr>
<tr>
<td>2.4</td>
<td>Opinion Mining Framework</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>Levels of Opinion Mining</td>
<td>36</td>
</tr>
<tr>
<td>3.1</td>
<td>Customer Relational Model based on People, Process and Technology</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>Customer Relational Model</td>
<td>46</td>
</tr>
<tr>
<td>3.3</td>
<td>Customer Interaction Frequency</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>Segmentation Management Strategy</td>
<td>54</td>
</tr>
<tr>
<td>3.5</td>
<td>Customer Value and Customer Lifecycle Management</td>
<td>56</td>
</tr>
<tr>
<td>3.6</td>
<td>Cross Industry Standard Process for Data Mining</td>
<td>59</td>
</tr>
<tr>
<td>3.7</td>
<td>Relation between CRM’s Entities</td>
<td>62</td>
</tr>
<tr>
<td>3.8</td>
<td>New CRM Model Structure</td>
<td>63</td>
</tr>
<tr>
<td>3.9</td>
<td>Processes in New Model</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Customer Data Integration</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>A Data Cube for Customers’ Opinion on Product Categories</td>
<td>71</td>
</tr>
<tr>
<td>4.3</td>
<td>Research Plan</td>
<td>73</td>
</tr>
<tr>
<td>4.4</td>
<td>Basic Ontology</td>
<td>74</td>
</tr>
</tbody>
</table>
4.5 Integration Business Model .................................................. 77
4.6 Opinion Polarity .................................................................. 82

5.1 Ontology of Mobile Phone .................................................. 87
5.2 The Model for our Sentiment Analysis System ......................... 90
5.3 The Diagram of ETL ........................................................... 92
5.4 The Data source .................................................................. 93
5.5 Detail Process of Extraction .................................................. 94
5.6 Process of searching noun, adjective and adverb ....................... 96
5.7 Detail of Transform Process .................................................. 97
5.8 The Main Dimension in Data Warehouse ................................. 105
5.9 Summary of Polarity and Orientation based on Product Feature .... 108
5.10 Details of Polarity and Orientation based on Connectivity Attributes 109

6.1 Our Testing Environment ...................................................... 115
6.2 Results for Precision using Our Ontology for All Products ............ 121
6.3 Results for Recall using Our Ontology for All Products ............... 122
6.4 Results for Precision using Our Feature Ontology and Synonym for All Products 122
6.5 Results for Recall using Our Feature Ontology and Synonym for All Products 123
6.6 General Polarity for Mobile Phone ........................................ 128
6.7 Detail of Customer’s Comments based on Product Attributes ........ 129
6.8 Detail of Customer’s Comments based on Product Attributes’ Feature 130
6.9 Synonym of Product Ontology ............................................. 132

C.1 MP3 Ontology ................................................................... 152

D.1 DVD Player Ontology ......................................................... 154

E.1 Camera Digital Ontology ...................................................... 156

F.1 Mobile Phone Ontology ....................................................... 158
List of Tables

2.1 Example of POS Tagger ........................................... 20

3.1 Segmentation Types and Business Tasks .......................... 53

4.1 Initial Idea for Data Structure ....................................... 71
4.2 Comment from One Customer Group ............................... 78
4.3 Features Polarity .................................................... 80
4.4 Attribute Polarity .................................................... 81
4.5 Feature Synonym .................................................... 81

5.1 Seven Level Polarity System ........................................ 102
5.2 Attribute Polarity .................................................... 103

6.1 The Confusion Matrix ............................................... 119
6.2 Result for Recall. .................................................... 124
6.3 Result for Precision. .................................................. 125
6.4 Result for Precision between Pattern Knowledge and Our Techniques ........ 126
6.5 Result for Recall between Pattern Knowledge and Our Techniques .......... 126
6.6 Result for F Score. .................................................... 127
Chapter 1

Introduction

The emergence of the Internet in this information age has changed the business trend between an organisation or a company and their respective customers. The increasing Internet penetration per household in most part of the world has made online transaction more essential than before as it is easier and it can be done anytime and anywhere, which results in customers sharing their views and opinions about a service or a product with other customers in online community-based social media, such as blogs, product review websites, Facebook, and Twitter among others. Information that customers review or share on these social media are invaluable information needed not only by other customers but also organisations. For the customers, the information gives them more information about a particular product or service that they are interested in buying or using, based of which they can gain more knowledge about the product or the service for comparison purposes before they proceed to buy the product. Besides that, organisations can capture a lot of information from online reviews to improve their product, and also plan new strategies for their marketing segment.

Research on big data or Data Mining (DM) is more important than before as the analysis process becomes more critical in helping end users such as customers and organisations in making decision for their future task(s). The rise of the Internet gives opportunity for customer to gather as much information as they need about a particular product (s) or service (s) from other customers’ review online. On the other hand, it also allows organisation or business that sells the products or services to improve their products or services based on customer’s review online.

Opinion Mining (OM) or Sentiment Analysis (SA) is a respective field that combines many
techniques from data mining and natural language processing (NLP) from research areas of Information Retrieval (IR) and Knowledge Discovery (KD) based on text. Meanwhile, opinion mining emphasises on the analysis of users’ opinion, comment, feeling, and characteristic among others based on features from products, services, individual, organisation, and event (Liu [2012]). In the year 2003, Nasukawa and Yi [2003] introduced the term sentiment analysis, while Dave et al. [2003] used opinion mining in their research about user sentiment based on text or customer review.

NLP becomes an important part in OM for its ability to extract a huge volume of human opinion based on text comment. Research on NLP is essential for opinion mining as it can monitor positive or negative human feelings based on the sentence or feature extraction in a customer’s comments. Opinion mining research is also known as an exclusive type of text classification as it classifies user based on groups such as personal information, demographic, and education, among others.

Opinion mining is implemented in various areas such as election, regulation, business, broadcasting, marketing, education, research and development, policies, and health care (Liu [2012] and Talele and Badgujar [2012]). The main contribution of opinion mining and sentiment analysis are their ability to extract human opinion into the orientation of human sentiment, either positive or negative, based on sentiment words found in texts. Good, excellent, amazing, okay, and perfect are among the example of positive words from a user’s comment. Meanwhile, negative sentiment words among others are terrible, worst, bad, poor and horrible.

This research study viewed current practices of CRM in the business field and how technology, especially information technology, can enhance the ability of CRM -based on the combination of customers’ background and their opinionated comments from a particular website. This chapter first outlines the background of the study (Section 1.1) and the aim and the questions of the research (Section 1.2). Next, the chapter discusses about our FOUR research objectives (Section 1.3). The next section briefly describes the significance of the research (Section 1.4), followed by the publications produced for this research (Section 1.5). The thesis structure and the outline of the remaining chapters of the thesis are presented in Section 1.6, and a summary of this first chapter is presented Section 1.7.
1.1 Background

One of the toughest jobs in this information age is gathering relevant and useful information from many resources such as the Internet, databases, and webpages, which change rapidly every second, and also from traditional resources for instance papers and books. Many research areas such as data mining need large data to produce better results. Normally, data mining with plentitude information or data produces better results, but sometimes, it can generate inaccurate results due to mismatch and useless information (noisy data). In fact, very often, huge amount of patterns (or rules) extracted from data are redundant or useless. It is also difficult for human to understand the discovered patterns (rules) due to two obstacles: the overwhelmingly large volume of discovered patterns (or rules) and the lack of semantic information in them.

Nowadays, data is more important than before. The collection of data, sometimes making job in information technology field becomes more complex and difficult to understand because of its huge data pool to be extracted. One of the important areas in Data Mining research is Data Mining with incomplete data. The emergence of Knowledge Management in information technology has improved the ability of data mining, especially in choosing the best data for the right tasks. As shown in a study by Misra et al. [2003], the main objective of Knowledge Management is to make sure that the best decision making is produced by the right knowledge at the right time as it is needed. In doing this, ontology-based model for knowledge management has been developed to get the best data for data mining tasks. Ontology is a study about what already exists and what we want the system to do in achieving a cogent description of reality (Vickery [1997]). Currently, it is hard to find research in data mining, which uses knowledge management to increase the confidence level of data collection.

Nowadays, a lot of data are stored in different formats, places, and schemas, making the integrating process difficult before they can be analysed for the best decision making by an organization. The variety of systems in the market also gives a tough time for system developers to merge or integrate the systems to analyse data on customers’ characteristics. Currently, the relational model is one of the popular models used in most companies or organizations. In the relational model, data are stored in rows and columns (Rainardi [2008]), record-oriented (Sukumaran and Sureka [2006]), and assigned to dedicated fields, known as structured data (Baars and Kemper [2008]). However, 80 percent of the data in an organization or a company are unstructured data as mentioned by Sukumaran and Sureka [2006] and Losee [2006], which
CHAPTER 1. INTRODUCTION

shows that it is inaccurate to use only structured data (Yaakub et al. [2011], Yaakub et al. [2012]) for customers analysis because it only covers 20 percent of the entire data for an enterprise. The majority of the data such as web pages, text comments, e-mails, images, video, telephone conversations, music, and presentations are unstructured data (Rainardi [2008], Losee [2006]).

In this research, we integrate both structured and unstructured data in order to cover a broader spectrum of information for an enterprise. The question here is how do we make sure that the best possible data will be chosen for customer analysis model in order to get the best result.

Normally, data mining is knowledge discovery at the data level, and it is very difficult for human to understand the meaning of discovered knowledge in databases, whilst knowledge management describes any activity involving knowledge performed by human being. However, there is a big gap between data mining and knowledge management. In this research, we present a new method for people to understand data mining results, especially for opinion mining.

Opinions and comments from customers regarding products and services are more significant today because the information can be obtained easily from the Internet. However, the customer relationship management (CRM) system today cannot benefit this large amount of information because it focuses more on the pattern of customers’ purchasing behaviours and how to retain the current customers with various loyalty campaigns. CRM itself consists of three main components, which are people (customer and company’s staff), process and technology (Chen and Popovich [2003]). Our research emerged from the components of technology and people. We used current CRM data for customer records and we enhanced the technology using data mining to analyse the orientation of customer’s opinions regarding a respective product.

Pang and Lee [2004] describes the main function of opinion mining is to extract the opinion or human sentiment based on texts from a document analysis process. Most people in today’s community post their opinions or reviews about a particular product on media social spectrum such as in personal blog, forum, product’s website, and Facebook. Consequently, this brings forth the importance of sentiment analysis or opinion mining technology in analysing customers’ reviews. Both customers and organisations receive positive impact from the opinion mining result. On the customers’ side, they have a lot of information about a product or a service that they are interested to buy or use. With the right information, they can compare prices among suppliers and, product’s quality, and analyse the percentage of customers satisfied
with the product that they want to buy or the service that they plan to use.

Meanwhile, organisations or companies can gather a lot of important information from customers about their product or service. As a result, they can improve their product or service based on customers’ comments (Agarwal and Mittal [2013]). They can also plan new marketing strategies to stay competent with their competitors in a particular business.

Figure 1.1 shows four important steps in opinion mining process. The first process is data pre-processing, followed by feature selection. The next process is sentiment word selection and finally, the last step is pairing the feature word with the right sentiment word. Every single step in opinion mining is important to produce an accurate sentiment classification between feature and sentiment word.

The main problem in opinion mining is how to classify text based on human sentiment. This is different with text mining, which only identifies the text about a topic or a title. Accordingly, opinion mining does not only identify the text about feature, recognises the right sentiment word to pair with a particular feature word.

Opinion mining classification consists of three main types, document level, sentence level
CHAPTER 1. INTRODUCTION

and feature level. Document level classification concerns the document’s orientation, either as a positive document or a negative document, based on the topic discussed by user or author (Pang and Lee [2004], Turney [2002]). In this level, the sentiment orientation for a document is based on the conclusion of the sentiment word, where more positive words in a document mean that the particular document is regarded as a positive document. Research by Kim and Hovy [2004] and Riloff et al. [2003] discussed about the classification in sentence level. The difference between document and sentence levels classification is that the document level emphasises on the document’s title, while at sentence level it highlights the sentiment orientation for every sentence in the document.

Next, feature level classification emphasises on classifying a sentiment word based on a product’s entity or feature (Hu and Liu [2004], Popescu and Etzioni [2005], Liu et al. [2005] and Ding et al. [2008]). The main objective of feature level classification is to identify a feature and to pair it with the right sentiment word, which is not an easy task because the same sentiment word can have different meanings in different contexts. Liu [2010] claimed that feature level classification is the best technique compared to document or sentence level, as sentiment classification based on feature level is far more accurate and thorough.

In this research, we introduced an ontology model and a summary of polarity in report form to help customers find the right product(s) for their need and also for companies and manufacturers to improve their services and products.

1.2 Aim and Research Questions

Data mining plays a pivotal role in today’s business trend as a lot of data are available easily through many media such as the Internet, books, journal papers, newspaper, and also radio and television. This kind of information gives opportunity for both customers and companies to analyse a product and a service using data mining technology. Customers can gain valuable insights about a product and a service, which they are interested in buying or renting from other customers that write a review about particular product or service. Meanwhile, a company can gather a lot of information about their product or service, which in return, can be used to improve the quality of products or services and make it more competitive.

Big data gathered from social media technologies sometimes contain noisy data. Data
mining technique(s) can be used to analyse the data based on a particular subject needed either by the customer or the company. The combination of data warehouse and opinion mining, which is a subset of data mining area can help user in choosing the best result. In order to achieve this, an ontology model for database in data warehouse will need to be developed to get the best data for data mining tasks.

The major issue in this research was to integrate structured data and unstructured data from customers’ comments. Kantardzic [2003] described that structured and unstructured data are imperative to provide valid insights into current business developments. Meanwhile, Sukumaran and Sureka [2006] used text tagging and annotation techniques to integrate structured and unstructured data. Their researches are quite similar to this research, but our research technique used factual data and descriptor to integrate the unstructured and structured data. Baars and Kemper [2008] used three approaches to integrate structured and unstructured data. The first approach was using a portal add-on to build a bridge between the access and the logic layers. Secondly, tools were used to integrate components from the data layer. The last approach was positioning the middleware hub at the logic layer. The above opinion mining methods either adopted a context-free sentiment classification approach or relied on a large number of manually annotated training examples to perform context-sensitive sentiment classification. Guided by the design of science research methodology, the works in Lau et al. [2009b] and Lau et al. [2009a] illustrated the design, development, and evaluation of a novel fuzzy domain ontology based context-sensitive opinion mining system. The ontology extraction mechanism was underpinned by a variant of Kullback – Leibler divergence, which can automatically acquire contextual sentiment knowledge across various product domains to improve the sentiment analysis processes.

The research problem above was selected based on some reasons especially, the importance of using appropriate data to help in producing a good product or service by companies and also manufacturers. Our research shows the importance of using both structured and unstructured data for customer analysis, instead of using either one individually to generate a more accurate result. We classified the knowledge (or ontology) about the related information into multiple data dimensions, including customer, date, product and opinion. For each dimension, we also considered its concept hierarchies.

The Customer Relationship Management (CRM) is more important than before. Structured
data is the main type of data stored in CRM. This research proposes a new architecture for Opinion Mining, which uses a multidimensional data model to integrate customers’ characteristics (structured data in CRM) and their comments about products/services (unstructured data) from a particular website.

Usually, a CRM system can be used for dealing with structured data in a database. Information such as customer’s contact and personal information make it easy for companies to communicate with them, to analyse about any particular customer and eventually plan the best strategic campaign to retain current customers as well as to attract new customers. However, the biggest challenge for a CRM system is to integrate customers’ comments with customer data in the system because most comments are unstructured data, while the data in CRM system are structured. Normally, customers’ opinion covers a broad spectrum of issues about products. Some comments are about specific technical issues, but some comments may be very general. Also, some comments are positive, but others maybe negative or neutral.

The main research activity in this study was to integrate the structured and unstructured data (comments) into a multidimensional data model. The key step to achieve this objective was to transfer comments (opinions) to a fact table, which has several dimensions, such as customers, products, time and locations. This research presents a comprehensive way to calculate customers’ orientation for all possible products’ attributes or services.

Besides that, the creation of ontology in multidimensional database also plays a vital role in this research. A good database design will help the CRM gathers the information more accurately and consistent compared to badly designed database. Ontology is a database with information about what categories or concepts exist in the world (domain), what properties they have, and how they relate to one another (Vickery [1997]).

In this research, we present a new approach to integrate costumers’ opinion into a traditional CRM system. This new approach firstly identifies the entities and opinion descriptors in customers’ comments. The approach also maps the entities to the corresponding concepts in the product concept hierarchies. In addition, it transfers customers’ comments into a fact table to show the associations between customers’ comments and other data dimensions.

The other reason for selecting this research problem was the possibility to complete the project within three years, which is the time frame given for a PhD course. The development of ontology-based Data Warehouse is the most important part in this proposed research. So,
1.3 RESEARCH OBJECTIVES

The evaluation of this particular ontology needs to be done thoroughly to make sure that the end result is useful in Data Mining field for the next stage of research. In doing these, some new techniques were introduced to support the research.

1.3 Research Objectives

This thesis has two main outcomes. The first is to extract customers’ comments from the Internet or a particular company’s website. In this research, we used the data from Amazon website (www.amazon.com). The importance of extracting customers’ comments is because comment is unstructured data format, so it cannot be integrated with other data from CRM. The second outcome is product’s polarity report based on the combination of customer comments and customer record from CRM system. Details about the research processes for this work also mentioned in chapter 4 (subchapter 4.3 Research Process).

The achievement of these main goals depends on the following objectives:

- Acquire Users’ Comments

   It is very important in this research to acquire the right data from respective users’ comments. The data were selected from the Bing Liu and Minqin’s project (Hu and Liu [2004]). These data were chosen because they have a large amount of customer’s comments based on a particular product. Besides that, the performance of our model was compared with the five baseline models from (Hu and Liu [2004]). The performance for this research was measured by recall ($R$), precision ($P$), and $F$ test.

- Develop Ontology in Data Warehouse for Opinion Mining

   The development of ontology was done in two phases, which were prototype in the early stage of the research and the main ontology framework after the first phase of prototype was completed (after testing and evaluation were done). Higher and lower levels ontology designs are based on their characteristics. The higher level contains entities divided based on their uniqueness, such as the name of the product in general and basic information about the product. Ontology then finds the relationship between
CHAPTER 1. INTRODUCTION

the entities. The second or lower level contains instances, which describe the usages of entities in some applications and more detailed than in the first level. The instances are divided into groups based on entities. The structure of two-tier ontology was also assessed to find out its usefulness or efficacy for opinion mining. These involved the development of schema design, table design, and also the implementation of the table in the database.

- Establishment of Data Cube for Analysing Customer Opinion

This involved text analysis using the research model. The Extract, Transform, and Load (ETL) process design was created to implement data transfer and integration of the table in the database, which means that the integration of structured and unstructured data was completed in this phase. Besides ontology development, this phase was also one of the main research developments in the study. Data extraction and integration between unstructured data (opinionated comments) and other CRM data (structured data) were investigated in this phase.

- Evaluation on the Significance of Opinion Mining Research based on Polarity

This research proved the importance of integrating unstructured and structured data in helping Data Mining, especially in Opinion Mining area to increase the accuracy of data selection based on the polarity of a product. The evaluation method design with the baseline were used to prove the significance of the research. This research showed tremendous improvement compared to the six baseline models used in this research.

1.4 Methodology

Figure 1.2 shows the research methodology of our work. The methodology consists of four major stages. First, the theoretical study refers to the literature review of past and current research. Next, the development phase involved pre-processing data, and identifying subjective sentences among others. Then, the implementation and testing phase, in which the newly developed techniques were tested. Finally, the evaluation phase studied the testing result to determine the best solution.
1.4. METHODOLOGY

The literature review on sentiment analysis for current and past research was done in theoretical study. In this first phase, the author identified current issue(s) on sentiment analysis and opinion mining based on methods used by current and past researchers. The best method from the literature review chosen and used for testing and as baseline models on third phase, testing and evaluation. Past processes were studied in the test to improvise previous research.

The next phase was development, where pre-processing data was done. This phase also involved identifying feature and sentiment word. A new technique proposed in this phase to improve processes on feature selection, and also a process of pairing sentiment word and its feature word (sentiment classification).

Then, was the implementation and testing phase. In this phase, the implementation and the testing of the proposed method was done with dataset. The proposed method was later compared with the baseline model(s) based on precision, recall, and F-measure.

Lastly is the evaluation phase where the results from the testing phase were analysed.
1.5 Significance of the Research

For the given topic, the objective of opinion mining is to decide the opinions of a text document, including positive support, neutral or negative support. It is easy for human being to do so; however, how to use data mining to solve this problem is still an open problem. This research model has proved the importance of Opinion Mining’s research in producing better results using the proposed ontology and synonym. Ontology is a catalogue that makes everything in real world is possible with certain terms, such as how the entities are put together and how they work. Ontology works on the concept level (a higher level), rather than the data level (like data mining), which means concept level is easy to be used and understood by human being. The used of data cube to integrate unstructured and structured data is another interesting chapter in data mining research. The main contributions of this research are as below:

- Ontology Model for Product
  
  This research produces a new ontology model for a particular product based on high level and low level. The information on the product is grouped into several levels based on entities, attributes, instances and features of the particular product. After grouping the data, every single group item then is involved in synonym process. This synonym process is important in integrating between unstructured and structured data because a lot of customers used different terms to define the same item. For example, message is also mentioned by another customer as msg.

- Algorithms to Integrate Unstructured and Structured Data
  
  New algorithms have been built in this research to enhance the performance of the existing model in integrating two different format data namely as structured and unstructured data. These algorithms have been tested and evaluated, and the result shows that its performance is far better than six baselines used in this research.

- New Formula to Calculate the Product Orientation
One of the distinctive discoveries in this research is a new formula to calculate product orientation based on customers’ opinionated comment (Yaakub et al. [2011], Yaakub et al. [2012]). This new formula can support CRM application in helping customer to find a product based on other customers’ comments. This model produces report based on this formula. A customer or company can easily use it, especially in searching for details in comment or a summary about the product and its features.

Full detail about this research contributions are discussed in the last chapter, Conclusion.

1.6 Publications

1.6.1 Journal Paper


1.6.2 Conference Papers


1.6.3 Plenary Speaker


1.7 Thesis Structure

The rest of this thesis is summarised as follows:

Chapter 2: This chapter is a literature review of related research topic and work done by other researchers. Details about past and current work regarding our topic are discussed in this part. Customer Relationship Model (*CRM*), ontology, data mining, data warehouse, and opinion mining and polarity are among the topics that we summarised in this research work. A simple summary about our work is mentioned in every sub-topic or work.

Chapter 3: This chapter is about *CRM* in detail. We discuss about the importance of *CRM* in business world today and what it can do to increase company profit and at the same time improve customer satisfaction toward the company. In addition, we elaborate the constraints in current *CRM* and how our model can be suited in the *CRM*.

Chapter 4: This chapter is about our idea in developing our *OpinionIntegrationModel*. During the early part of this chapter we mentioned about customer data integration (*CDI*) from *CRM* and how we enhance the *CDI* process in *OpinionIntegrationModel*. The research tasks are discussed in details to show steps of our work processes. Also, in a sub topic, we outlined our research questions based on hypothesis and basic idea of our ontology. In the last part of this chapter, we discuss about our basic idea of integrating structured and unstructured data.

Chapter 5: This chapter is about the algorithms of Opinion Integration Model of our model.
We discuss in detail about our model development including technical processes such as extraction model and transformation model. The two main key points in our research, feature ontology and synonyms are also discussed in particular before we show the report about polarity of customer’s sentiment comments for a particular product.

Chapter 6: This chapter is an evaluation and discussion. We outline and describe the dataset that we used and the six baseline models that we compared with our work. Precision, recall and F-test were used to evaluate our model performance. Comparison of results between the current research model and baseline models are discussed thoroughly in this sub topic. In the last part of the chapter, we show a detailed report regarding the polarity of customer comments on a particular product.

Chapter 7: This last chapter is conclusion. In this chapter, we summarise the importance of our model to enhance CRM especially in customer’s satisfaction on a particular product. We also discuss the future direction of our model and what improvement can be made to increase its performance.

1.8 Summary

The integration of unstructured and structured data is an ongoing project by data mining researchers worldwide and it is difficult to be completed in just three years. The proposed research uses database as the main tool to develop product’s ontology. This means the ETL and data cube are the main research process in finding the final outcome (polarity of product). The proposed research only emphasised to find the summary of one product at one time based on Opinion Mining. Other problem in Data Mining will not be done in this proposed research.
Chapter 2

Literature Review

This chapter investigates the past and current topics related to the research project. Literature review elaborates more about related works in sentiment analysis (opinion mining) and describes the definition of sentiment analysis (opinion mining) and the common terms used in this thesis. Also included here are the subtopics that have been used in the current research such as feature selection, sentiment word and sentiment classification. Current and past techniques by other researchers are described and analysed to get more information and understanding to improve the current technique in opinion mining or sentiment analysis.

2.1 Introduction

Opinion mining is a research, which comprises of two important parts. First, to identify an object of entity that gives important meaning to sentiment word and second, to identify sentiment word, which has relation (meaning) with the object from the same document. In other word, opinion mining is a pair of object and sentiment word, which determines the semantic orientation of human sentiment, either positive or negative, based on particular data.

Online social media have changed the way human share and exchange their opinions, ideas and views with their colleagues, friends and others (Yin et al. [2012]). A good understanding on others’ opinions and views is an advantage to any interested party in gathering or collecting data from a group of users from the Internet for future use. For example, a film director can get inputs from movie-goers about their preferences for future movie genre based on the sentiment from their intended audience.
Text mining is an important subset in opinion mining research as its techniques such as Natural Language Processing (NLP) and Information Retrieval (IR) are popular techniques used, especially for feature selection process. In this research, the challenging process is to integrate structured data from the CRM system with unstructured data from customers’ reviews. Two machine learning theories, supervised and unsupervised are also described in this chapter.

2.2 Term Definition

Extract, Transform, Load (ETL)

Extract, transform and load is a process in database, where extraction is a process of reading dataset from the system’s database, while transform is a process to change the initial state of data to another required form. Load is a process of writing the data into the new targeted database (database new format). ETL has three tasks:

- Extract

  Extraction of data from source (dataset). The most challenging task as a result from this task will influence other tasks, hence accuracy in this task is important.

- Transform

  Transformation involves lot of rules or functions such as columns, code values and formula for new calculation.

- Load

  The last phase is load, where the system loads the data into the targeted destination (data warehouse).

Opinion Mining (OP)

Opinion mining is also known as Sentiment Analysis (SA) by some researchers (Liu [2012]). Liu [2008] described opinion mining as $d \in D$, where $d$ is a particular document that contains
sentiment word and object, while \( D \) is a set of text document, which consists of opinion or sentiment about a particular object. Opinion mining can be defined as human sentiment on specific entity based on the characteristics or features of that particular entity. For example, political analysis can collect a lot of public opinion about a particular political party based on the party’s leadership, performance and candidates for future election. The main objective of sentiment analysis or data mining is to extract opinionated word and its object from any text document with a particular title as discussed by users.

**Data Pre-Processing**

Data pre-processing plays a pivotal roles in text mining. Furthermore, data pre-processing is also important in opinion mining technique as well. In opinion mining, this process is used to identify a targeted word for example *play* as a verb, to remove unnecessary words, and to filter meaningless words from particular sentences (Bagheri *et al.* [2013] and Mouthami *et al.* [2013]).

Based on a research by Hu and Liu [2004], they identified five phases involved in the data pre-processing task :-

- **Tokenization**

  A character in text has block(s) known as token. This token is divided by space, comma and full stop or other punctuation marks. This separation of character gives different meaning to the system in pre-processing data.

- **Stemming**

  This phase is common in data pre-processing, which transforms words into root words, such as *caring* into *care*.

- **Stop Words Removal**

  Many reviews, comments and opinions have a lot of meaningless words in the sentence known as stop word, which is not important in supporting word or giving meaningful sentence such as *a, the, is, are, was and were* among others.
• Part-of-Speech Tagging

Part-of-Speech (POS) tagging is important to divide every single word in sentence to the type of word such as noun, verb, adverb, and so on. This phase is important for selection of feature words and sentiment words in sentences for opinion mining processes. This tagger was first built by Toutanova and Manning [2000] and improved by others in Toutanova et al. [2003].

Table 2.1 shows an example of POS tagger used from the sentence: "heat water in a large vessel". All categories of words are defined in simple notation such as NN for noun, VBZ for verb and JJ for adjective. This extraction process makes the opinion mining process easier compared to dividing every single word manually.

<table>
<thead>
<tr>
<th>Word</th>
<th>tag</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>heat</td>
<td>VBZ</td>
<td>verb</td>
</tr>
<tr>
<td>water</td>
<td>NN</td>
<td>noun</td>
</tr>
<tr>
<td>in</td>
<td>CC</td>
<td>Coordinating Conjunction</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>large</td>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>vessel</td>
<td>NN</td>
<td>Noun</td>
</tr>
</tbody>
</table>

• Normalization

Normalization phase is more of maintenance works such as correcting misspelled words, and removing repeated characters.

Object

Object is referred to as an entity $e$ in a sentence, where it can be a product, a topic, an issue, a service or an organisation, based on the main topic discussed by reviewer. Most $e$ are defined or supported by its feature $f$, which defines its characteristics. In a document $d$, we can define an object as $e(f)$. 
2.2. **TERM DEFINITION**

**Opinion Holder**

Bethard *et al.* [2004] acknowledged the importance of knowing the source of information or opinion holder from online review such as a blog’s author. In another word, opinion holder is a person or individual that gives an opinion or a review about a particular topic. Product reviews by customers, for example can be traced based on owner’s *id* or log in details.

**Feature Extraction (FE)**

Feature extraction is a process to extract feature word from comments and opinion sentence, which is the most frequent noun used in a sentence. All features recognised by the system are extracted from the sentences. Social media such as blog and online forum have too many words for the system to identify the target item. The target item could be a title, a feature and an attribute. Thus, this is not an easy task in opinion mining because a targeted item or feature word cannot be recognised without doing any tagger’s process first. Some researchers such as Popescu and Etzioni [2005] and Hu and Liu [2004] stressed that the identification of target item or feature extraction as an additional task in opinion mining research.

**Product Feature (PF)**

Product feature in this research is words that are noun in a sentence derived from feature ontology. In this context, product feature consists of the product itself, its attributes, and its characteristics, which is based on Sony’s product website.

**Feature Selection (FS)**

Feature selection is a processed in opinion mining to identify feature(s) that can be paired with sentiment word in a document. At this level, polarity for the sentiment word is not counted as this particular work was done based on sentiment classification task. Jain and Zongker [1997] defined feature selection as a list of selected feature from a document to produce the best feature subset for future used. Nicholls and Song [2010] and Liu and Yu [2005] explained that the main task for feature selection is to identify a list of selected feature based on some criteria from the original feature set (from feature extraction process). They also mentioned that irrelevant feature
is removed from this new list of selected feature. Agarwal and Mittal [2013] also supported the definition by Nicholls and Song [2010] and Liu and Yu [2005] by informing that feature selection is selecting important features and removing irrelevant feature from the feature set. Meanwhile, other researchers including Guyon and Elisseeff [2003] defined feature selection as selecting appropriate feature’s subset to develop a new and better forecasting model.

Based on the definitions given by other researchers above, feature selection is an important process to remove irrelevant features from the list of selected feature. This process only identifies suitable features based on conditions from the opinion mining of a particular project.

**Feature Synonym**\( (fs) \)

Feature synonym\( (fs) \) refers to words of the same meaning or same group based on synonym word from Microsoft Word’s thesaurus.

**Sentiment Classification**

Sentiment classification is a combination of product features with opinionated word, which determines the polarity of this pair as either positive or negative sentiment.

**Attribute Polarity** \( (AP) \)

Attribute polarity is the strength of attribute based on feature ontology level. Level 0, which is the highest level of feature ontology is counted as 1. Level 1 of feature ontology, is considered as 0.5 and level 2 is considered as 0.25\( (Yaakub \textit{et al.} [2012]) \).

\[ ogc(g,c) \]

\( ogc(g, c) \) is a formula to calculate product features’ orientation based on set of customer, product and user’s opinion on a particular product feature(s) \( (Yaakub \textit{et al.} [2012]) \).
o(g,c)

o(g, c) refers to the orientation of sentiment based on calculation of sentiment and attribute polarity from feature ontology level (Yaakub et al. [2012]).

2.3 Customer Relationship Management

2.3.1 Customer Relationship Management

Business is a transaction that happens everyday between customers and companies or organizations. The transaction in business is either a product or a service offered by a company to customers, involving a small amount of money to a very large amount. Every company has many tasks in order to attract customer for business transaction such as marketing to find the right customer at the right time and producing a product or service that customers really need. At the same time, a company must be competitive to compete with their competitor in producing good product, to retain their loyal customers and at the same time to attract potential customers. Before the emergence of the Internet, companies increased the production of any products to cope with customers’ demand based on the analysis from mass marketing and mass production techniques (Chen and Popovich [2003]). These techniques decreased the relationship between customers and companies because only one side (company) is interested in keeping the transaction record without knowing what the customers really want and for customers, they did not receive any appreciation as loyal customers such as discount program (Chen and Popovich [2003]).

In modern era, every company must be creative in their marketing strategy and understand what exactly does the customers need to retain their loyal customers and at the same time attract new customers because the competition between companies and their competitors are really stiff and getting tougher compared to previous era, even in the same market area. Drastic changes are needed in business strategy especially in marketing paradigm to make sure the companies stay relevant in the business world. Good relationship between customers and companies is a must to improve customer satisfaction that leads to increased profit in the future. Mass or broad marketing and target marketing are not enough to attract customers in this Internet era as customers prefer one-to-one or personalised marketing. This new marketing mechanism must
be applied in every company to survive the current trend of business market, which requires the company to be more efficient and effective to meet customers’ need (Chen and Popovich [2003], Prabhaker [2001]). Figure 2.1 shows the importance of good relationship between customer and marketing focusing on company.

Most companies have many tangible assets to support their businesses such as knowledge worker and business premise but the greatest asset that a company has is the customers, as expressed by Campbell and Cunningahm [1983]. As the greatest asset for a company, customers’ satisfaction is the main target for every company to achieve their goal in business. Ittner and Larcker [1998] acknowledged the importance of intangible assets (satisfaction of customer) that needs to be highlighted in predicting future financial performance and not historical accounting measure that is currently used by many companies. Furthermore, every company must gather as much information as possible with meticulous knowledge about the customers’ need (Achilladelis et al. [1971]) to make sure that they produce or market the product that customers are looking for. Thus, the ability of any company to understand their customers’ needs, preferences, buying behaviors, and price sensitivity is a major advantage against their competitors (Chowdhury [2009]). It is very important for every company to retain their loyal customers as Reichheld and Teal [2001] stated that a five percent increase in customer retention would result in 35 to 95 percent increment in average customer lifetime value, which is an easy profit for a company compared to investing to find new customers, whose loyalty are uncertain.

Customer analysis, an analysis about customer behaviors and activities while transacting, is increasingly important among companies today. As a result, a lot of research on this topic has been done (Campbell and Cunningahm [1983], Achilladelis et al. [1971], Parsons [2003], Ittner and Larcker [1998]). Rainardi [2008] defined customer analysis as an analysis of customer activities and behaviors when doing a transaction with an organization or a company. Customer analysis is an analysis performed with sufficient information produces better results, especially the relationship between a company and the customers. Campbell and Cunningahm [1983] emphasized the importance of any company that has good relationships with customers to oversee the marketing and purchasing strategies. Meanwhile, Peppers and Rogers [2001], and Chen and Popovich [2003] also stressed that company would move up or down depending on their capability to produce good one-to-one relationships with their customers. This statement is supported by Parsons [2003] as his finding showed that sales by companies and visits by customers are the two main attributes that would generate optimum customer behavior on
Figure 2.1: Changing Marketing Strategies (Chen and Popovich [2003])
purchasing items.

As mentioned earlier by Campbell and Cunningham [1983], customers are the greatest asset that a company has, but it is not easy to maintain the same customers for a prolonged period without specific research on customers’ behaviors. Every company must have a really good knowledge about their customers needs (Achilladelis et al. [1971]) in order to make sure that the new product is relevant for the targeted customers. Marketing and business strategies are the two main areas, in which every company needs to emphasize to thrive and remain competitive in the current business trend. Today, many companies are looking to establish connections with customers for longer term as they know that customer’s loyalty means a lot, especially in reducing cost and increasing company profit (Chen and Popovich [2003], Campbell and Cunningham [1983]). In achieving these goals, companies must find a new tool to improve their businesses. One of it is known as Customer Relationship Management or *CRM*.

### 2.4 Ontology

One of the popular terms in computer science currently is Ontology. Even though, this term was first used in the medical field for ”leukaemia” or ”terminal illness”, Swartout and Tate [1999] stated that this term become more popular in other research filed, including information technology and computer science itself. In 1999, Swartout and Tate [1999] also cited that the emergence of ontology in computer science was because more researchers were interested in reusing or sharing knowledge across the systems. Vickery [1997] cited that ontology is a catalogue of everything to make the possible world with emphasis on how it is put together and how it works. However, Gruber [1993] described ontology as an explicit specification of a conceptualisation. The main idea of this definition is derived from conceptualisation but this researcher emphasized that ontology is not limited to taxonomic hierarchies and its relation only, but wider in terms of its definition. This proposed research also built its ontology based on this definition because it was more suitable to develop ontology using database as the main development tool. This idea is also supported by Poli [1996]. He also mentioned that ontology looks like a database because of the usage of terms such as categories, the properties that describe each of the categories, and how the categories relate to each other. Furthermore, the most relevant finding byPoli [1996] that makes it relevant to this proposed research is the usage of ontology as a unifying framework to integrate the systems. Almeida and Barbosa [2009]
2.5. DATA MINING (DM)

also cited ontology as a hierarchical structure based on concepts and relations, but they divided ontology into two categories known as Real Ontology ($R$ – Ontology) and Epistemological Ontological ($E$ – Ontology). $E$ – Ontology is the task of conceptualizing a domain, which is very relevant with the proposed research project. Meanwhile, some researchers defined ontology using different statements, but still evolved around representing the concept about the real world. Guarini [1998] cited ontology as representing a specific view of the world by explaining the meaning of the symbols in Information System (IS). Fonseca [2007] also defined ontology as a component in IS, but with the purpose of supporting the creation of conceptual schemes. He also differentiated ontology between ontologies of IS and ontologies for IS.

In opinion mining, ontology is the ability to share knowledge, exchange information and minimize ambiguity (Xue et al. [2009]). Meanwhile, Bhatt et al. [2006] highlighted that ontology plays important roles in sharing sources and defining terms precisely for future uses such as meta-data. Figure 2.2 shows product ontology proposed by Somprasertsri and Lalitrojwong [2010]. They used ontology to differentiate product features terms based on the level of the product’s structure name. The purpose of this idea is to normalise terminologies used by different product features. This product ontology is quite similar with our model of product ontology, especially in grouping the products into four levels as product name, product feature, product attribute, and product instance. Our product ontology, and Somprasertsri and Lalitrojwong [2010]’s product ontology are used a tree-hierarchy concept. To differentiate this product ontology with our model of product ontology, in our new model, we integrate every single term of ontology with product feature synonym file from thesaurus.com website and manually done for popular terminologies such as msg and txt. Meanwhile, in earlier product ontology, Somprasertsri and Lalitrojwong [2010] stated that product features were manually inserted based on the same group of features.

2.5 Data Mining (DM)

Knowledge is very important in our life today, especially after the emergence of the Internet. Everyone can access information from all around the world without barrier. Misra et al. [2003] defined knowledge as a precious resource. They stressed that knowledge has three important characteristic, which are:
CHAPTER 2. LITERATURE REVIEW

Figure 2.2: Product Ontology by Somprsertsri and Lalitrojwong [2010]

- Information lending the meaning from the knowledge
- The producers and consumers of knowledge are people
- The knowledge can be specified in two types; the tacit or explicit

Data mining is a very popular and well known research area. Therefore, it has multiple definitions by DM researchers. Wang and Wang (2008) defined DM as the process of trawling through data to find previously unknown relationships within the data, which are interesting to the user of the data. However, Binali et al. [2009] defined it as a process to discover knowledge in databases. Tsiptsis and Chorianopoulos [2010] mentioned that the aim of DM is to extract knowledge and insight by gathering the large amount of data, which analysed by using sophisticated modeling techniques. This definition is also supported by Rygielski et al. [2002] who explained that data mining as analogy which gather and extract hidden information in data warehouse or websites such as gold or coal mining.

Another researcher mentioned that DM is an analytical process designed to search consistent patterns or systematic relationships between variables by using large amounts of data, and then validate the findings by applying the detected patterns to new subsets of data (Schoech et al. [2000]). Han and Kamber [2006] state data mining is a result from the process of discovering interesting pattern and knowledge by a large amount of data. Chowdhury [2009]
described that DM has five steps in its development, which are preparing data, defining a study, reading data and building a model, understanding the model that has been built, and predicting the outcomes of the model. This research emphasised on opinion mining. Binali et al. [2009] mentioned that the aim of opinion mining is to extract knowledge based on user’s opinion from certain information to present it back to users in user-friendly manner. Based on the definitions of data mining by many researchers above, the term data mining refers to the processes or techniques of extracting a large number of data with one objective, which is to produce meaningful information for future use or prediction by analysing the data first. This definition is important for our research work throughout the phase of development of database and analyses of data.

2.5.1 Data Warehouse

IBM researchers, Berry Devlin and Paul Murphy in late 1980s introduced the concept of Business Data Warehouse. From that day, the usage of data warehouse concept had spread all over the world, especially in a system containing large database. The emergence of data warehouse benefited a large section of business sector such as in banking, retails, and airlines. Data warehouse is recognised as a database for report and analysis. Silvers [2008] described data warehouse as a valuable concept for the entire enterprises because its existence has benefited the whole enterprises instead of a single entity. Rainardi [2008] mentioned that data warehouse is a system that retrieves and consolidates data periodically from the source system into a dimensional or normalized data store, which means that it updates the database by batches, not every time the transaction happen. Data warehouse can be categorized as subject-oriented, integrated, time-variant, and non-volatile (Han and Kamber [2006]). Malinowski and Zimanyi [2008]) also agreed with Han and Kamber regarding the data warehouse categories as they defined the data warehouse as a collection of subject oriented, integrated, non-volatile, and time varying data to support management decisions.

Figure 2.3 below shows the data warehouse architecture used in our research (Malinowski and Zimanyi [2008]).

Malinowski and Zimanyi [2008] proposed the data warehouse architecture as shown above, which consists of data sources segment and four tiers, known as back-end, data warehouse, OLAP, and front-end. Malinowski and Zimanyi [2008] stated that data source is the data store,
from where data warehouse must retrieve the data. This segment contains information that provides the connectivity interfaces to a data store. Among the information in data store is the location of server, login and password, a method to retrieve data, and lastly the security permission.

*ETL* (Extraction, Transformation, and Loading) tools, which are some of the most important processes in data warehouse, are situated in the back-end tier. *ETL* is used to extract all formats of data into the intermediate database before they are transformed or loaded into the data warehouse. *ETL* is one of the processes used in our research to extract unstructured data from customers’ comment into structured data. Other data format handled by *ETL* is structured data, operational, or other sources. The second tier is data warehouse tier, which is composed of an enterprise data warehouse, several data marts, and meta-data repository. The storage of related data and information is the repository. Meanwhile, the *OLAP* (Online Analytical Processing) tier is regarded as an *OLAP* server to support multidimensional database and operations. The last tier, front end tier provides the function of dealing with data analysis and invalidation. *OLAP* tools, reporting tools, statistical tools, and data mining tools are among the several

![Figure 2.3: Data Warehouse](image-url)
client tools contained in this tier. Our research heavily depended on this architecture, especially when performing the data warehouse parts such as ETL process, OLAP, and CUBE to extract unstructured data into structured data.

2.5.2 Opinion Mining and Polarity

One of the popular works in data mining regarding customers’ behaviour is opinion mining. Opinion on web is expressed in many different ways such as image, audio, video and also text form (Binali et al. [2009]). In data mining research, opinion mining is also known as sentiment analysis, sentiment extraction, and affective rating (Binali et al. [2009], Esuli and Sebastiani [2005]). Opinion mining is defined as the determination to find subjectivity in sentence and orientation of it either positive or negative based on the strength of polarity (Binali et al. [2009]) and Mishra and Jha [2012] added that what feature is represented by the opinion. Meanwhile, Kim and Hovy [2004] used seeding expansion of adjectives for positive and negative terms with score method for example, stronger negative than positive. Synonyms and antonyms’ were used to enrich the meaning of every term in adjective. The problem with this method is that it cannot detect any new adjective because it only detect adjectives based on its seed and synonym terms as this method also without learning process in its method. Binali et al. [2009] mentioned that the aim of opinion mining is to extract opinions knowledge from information sources to present it back to users in a user-friendly manner. Binali et al. [2009] acknowledged that opinion mining has few related fields in data mining research such as:

- Information Extraction (IE)
  Unstructured data is transformed into structured data format in database for future data mining uses based on machine language learning. Indexing and querying databases are two techniques in information extraction that improves information retrieval.

- Information Retrieval (IR)
  Basically, information retrieval is searching for information based on a query. The major different between information retrieval and information extraction is information retrieval has better result in data precision based on a querying databases for a topic.

- Natural Language Processing (NLP)
NLP is a process used by computer to translate human language into language that computer understands and used this computer language to communicate with other computers.

- Machine Language Learning (ML)
  
  ML is processes used by computer to understand and duplicate human thoughts and behaviours based on data evolution from its initially stored databases.

- Web Data Mining
  
  Web data mining is a process to discover informative knowledge based webs and databases.

Figure 2.4 shows the opinion mining framework based on Binali et al. [2009]. Item in this figure referred to entity and object, and item extraction is representing object that referred in review such as mobile phone. Binali et al. [2009] highlighted that this item extraction is very important because this entity normally has opinion in the review that referred to them.

![Figure 2.4: Opinion Mining Framework](image)

2.6 Feature Extraction

Feature extraction is the extraction processes for lower level than item extraction as this process targeting item features. The basic idea for this process as mentioned by Binali et al. [2009] is
that a negative comment on product does not mean that the customer gives a negative opinion on its product features. This idea was initially highlighted by Hu and Liu [2004]. Opinions expressed by customers in any review refer to product features known as feature sentiment. Binali et al. [2009] stated that currently, there is no research in this particular area that has algorithms with the ability to differentiate between the opinions for a product and its features. However, item sentiment is a collection of overall sentiment captured from customer’s review of a product. This item sentiment is divided into positive, negative and neutral sentiment orientation as it supports the idea of Turney [2002]. Turney [2002] divided the sentiment into three categories as he emphasized that all items in product review is subjective. Item comparison and feature comparison compare between more than one product and its features, respectively. These comparison processes help customers who do not have much time to gather information about a product and its features in detail before making any online purchase. Adjective extraction is Turney [2002]’s first step in his works. This adjective came with a noun described by an adjective. Then, semantic orientation used by using distance between noun and adjective. Then, average semantic orientation used for a pair of words (noun and adjective) on review orientation.

Opinion mining framework based on Binali et al. [2009] work is:

- **Extraction Process**
  Research works used extraction process to transform unstructured data from customer reviews. This process is essential to capture the product and its features.

- **Sentiment Detection**
  This framework uses three categories of sentiment as proposed by Turney [2002] as neutral and has its part in improving the accuracy of results. Binali et al. [2009] did not discuss how they extract the sentiment word either for feature or item.

- **Report Presentation**

  Binali et al. [2009] final output is report on item and feature comparison.
CHAPTER 2. LITERATURE REVIEW

Meanwhile Zhou and Chaovali [2008] mentioned that the aim of opinion mining is to discover common patterns in user opinions from their textual statements automatically or semi-automatically. This research covered the polarity review of users (positive, negative, or neutral), which is one of the main issues in Opinion Mining as proposed by Turney [2002] and also used by Binali et al. [2009]. Zhou and Chaovali [2008] defined polarity mining as a task of determining the positive and negative orientations of textual information. Other research on this topic is by Pang and Lee [2004]. Their research was regarding sentence-level subjectivity summarization based on minimum cuts by implementing Naive Bayes support vector machines technique. Hu and Liu [2004] used the technique of opinion word extraction and aggregation enhanced with WordNet to get the polarity from users’ reviews. The major issue in this research is to integrate structured and unstructured data from customers’ comments.

Some researchers named feature extraction as feature selection. Feature selection is a process after the feature extraction process. Feature selection is a process of selecting a feature subset based on an original set of features. Saeyes et al. [2007] mentioned feature selection can be done using supervised and unsupervised learning techniques. Furthermore, Saeyes et al. [2007] highlighted three advantages of feature selection in opinion mining as follows:

- To prevent over fitting and increase model performance
- To produce faster and more effective model
- To get a clearer picture for the process of producing data

Meanwhile, sentiment classification is a process to classify sentiment word (opinionated word) with feature to identify the polarity of a sentence, either positive, negative or neutral. Sentiment classification can be divided into three types, which includes document-level sentiment classification, sentence-level sentiment classification and feature-level sentiment classification. Details about this level will be discussed in the section ”Level of Opinion Mining”, later.
2.7 Opinion Mining Methods

Kantardzic [2003] described that structured and unstructured data are imperative to provide valid insights for current business developments. Meanwhile, Sukumaran and Sureka [2006] used the text tagging and annotation technique to integrate structured and unstructured data. Baars and Kemper [2008] used three approaches to integrate structured and unstructured data. The first approach was using portal add-on to build a bridge between the access and the logic layer. Second, tools were used to integrate components from the data layer. The last approach was residing the middleware hub at the logic layer. The above opinion mining methods either adopted a context-free sentiment classification approach or relied on a large number of manually annotated training examples to perform context-sensitive sentiment classification. Guided by the design science research methodology, the works in Lau et al. [2009b], Lau et al. [2009a] illustrated the design, development, and evaluation of a novel fuzzy domain ontology based context-sensitive opinion mining system. The ontology extraction mechanism underpinned by a variant of Kullback-Leibler divergence automatically acquired contextual sentiment knowledge across various product domains to improve the sentiment analysis processes. Kumar et al. [2012] also used context-sensitive approach in their work using score scales from a modified SentiWordNet system (using average score) on adjectives, then based on the polarity system, they find negative and positive documents. The context-sensitive approach in this system used three rules which were intra-sentence conjunction rule, intra sentence comma rule and inter sentence similarity rule. For the first and the second rules, they mentioned that an opinion between a conjunction, such as *and* and *notonly* and a comma is the same, but an opinion between conjunction, such as *but* and *yet* has opposite meaning. However, for inter sentence similarity rule, they expressed that sentences in review should have the same opinion, except that the sentence has words such as *but* and *however*.

The difference between opinion mining and information retrieval is the data format (Binali et al. [2009], Mishra and Jha [2012]); subjective data is involved with opinion mining while factual data works with information retrieval (Mishra and Jha [2012]). Mishra and Jha [2012] classified opinion mining into three different levels, namely sentence, document and feature. Opinion mining at the sentence level identifies opinionated sentence and then classifies it into positive, negative or neutral. While, opinion mining at the document level identifies a whole review either as positive, negative or neutral. In our research, we did not discuss about these
two levels because we were only interested on feature.

Most opinion mining research used adjective as the main term to determine the orientation. Early research in this area mostly used dictionary based approach such as the works of Hu and Liu [2004], Kim and Hovy [2004], Esuli and Sebastiani [2005] and Kanayama and Nasukawa [2006]. The problem with dictionary based approach is the difficulty in finding domain specific opinion words. Hatzivassiloglou and McKeown [1997] proposed a corpus-based approach to solve this problem, in which they used adjective to find the orientation of a sentence. Clustering algorithm is a technique used by them to find a pair of adjectives, than combined by conjunctions such as and and but which and referred to same orientation adjective and but referred to opposite orientation objective. Kanayama and Nasukawa [2006] improved this method by using adjective and verb (the idea of coherency).

### 2.8 Levels of Opinion Mining

Opinion mining can be divided into three levels. Figure 2.5 shows the levels of opinion mining.
Document-level opinion mining presumes the whole opinion or sentiment in a document as one unit representing a document based on the concept of a single opinion holder relating to a single object. Pang and Lee [2004] and Turney [2002] argued for this level by classifying the entire opinion from a reviewer or a comment in a document represents only one object to identify the polarity of a document either as positive, negative or neutral.

Meanwhile, sentence-level opinion mining is based on subjective classification (Hatzivas-siloglou and McKeown [1997]). In this level, every sentence is identified either as objective sentence (factual sentence) or subjective sentence (contains opinionated word). A sentence recognised as subjective sentence, then identified either its opinionated word either as positive, negative or neutral that will determined the sentence’ polarity (Liu [2010]).

The document and the sentence levels have failed to identify individual’s preference (Hu and Liu [2004]). Thus, based on Liu [2012], the main tasks in opinion mining at feature level are to identify and to extract object features from users comments, and then to determine the opinion from the comment as positive, negative or neutral. After that, synonym word for feature word is used, and then classified in set of group based on positive, negative or neutral.

Feature-level opinion mining is getting popular in data mining research areas. A lot of techniques used by researchers in this particular level are discussed in next subsection.

### 2.8.1 Association Rules

One of the popular techniques in this area uses the association mining rule to find the product feature from frequent noun as infrequent noun is hardly referred to as product feature proposed by Hu and Liu [2004] and this technique is improved by Popescu and Etzioni [2005] with the introduction of part-of-relations that removes the frequent noun, which is not a feature. But this technique is time consuming as it needs to use query on web for finding the product features.

### 2.8.2 Manual

Liu [2012] emphasised that the human approach is the best technique to identify feature and its sentiment as human know exactly the meaning of every sentence. Normally, this technique is the most expensive and time consuming as it takes a long time to analyse the sentence. Currently, this technique is used to evaluate automatic technique as manual technique is the most accurate
technique.

2.8.3 Lexicon Approach

Opinion lexicon is a set of opinion words compiled by a system such as adjective, adverb, verbs and noun. Lexicon based approached was introduced by Ding et al. [2008]. Meanwhile, Hu and Liu [2006] used pros and cons format to identify opinion words. NLP’s technique was used to extract a review based on POS’s tag. Then, the identification of frequent features or nouns was done by the association rules technique and supervised mining method. Frequent nouns were used to find the nearest opinion word based on adjective. Three methods were used to identify an opinion’s orientation; opinion words(adjective), semantic orientation (positive or negative polarity) and dominant orientation. The advantage of using the lexicon is model or system easily can identify opinion polarity in sentences or a product review. The approach can be grouped into:

Dictionary-based Selection

This concept is also known as semantic orientation. This technique does not need any training to be developed, thus it also considered as an unsupervised method. Initially this technique only uses small lexicon developed manually based on root word. Synonym and antonym are then used to increase the size of original words (Liu [2012]). Kim and Hovy [2004] used this concept to develop their list of two seeds (positive and negative) for verb and adjective. Hu and Liu [2004] also used this technique in their early work for feature identification. Two popular applications in opinion mining research area WordNet and SentiWordNet were developed based on this technique. The weakness of this technique is that it depends on domain.

Corpus-based Selection

Based on the weakness of dictionary-based technique, corpus-based was created. This technique is domain independent, where a set of sentiment word is enhanced based on documents that has large corpus from particular domain. Turney [2002] used this technique to calculate Pointwise Mutual Information (PMI) for sentiment classification.
2.8.4 Machine Learning

There are two approaches in machine learning technique:

**Supervised Machine Learning and Unsupervised Machine Learning**

Most of research in opinion mining used the supervised technique. The tasks of supervised machine learning is to train a function, to make the system capable to identify sentiment orientation and also understand to use a list of sentiment’s corpus in document. *Support Vector Machines (SVM)*, *Naive Bayes (NB)*, *Maximum Entropy Classification (MES)* and *Neural Networks (NN)* are among popular techniques in supervised machine learning. Based on Pang and Lee [2004] SVM is the best technique as this technique has the most accurate classification. *Minimum cuts’s* method used by Pang and Lee [2004] to classify a sentence into objective or subjective, where objective sentence is removed to improve the accuracy of classification. They also proved that their technique is better than manual technique in classifying process. Agarwal and Mittal [2013] and Pang and Lee [2004] are among researchers who used SVM technique in their work due to good empirical efficiency. Meanwhile, the advantage of NB technique is because it is easy and simple to use as it does not require big data set to classify words. However, unsupervised method is only used in dictionary-based selection.

2.9 Summary

The literature review is one of the important parts before and during doing the research task. Previous works by researchers have helped this proposed research in giving new ideas, guidelines and references. Based on the baseline selected in this research, many techniques already have been done to closely achieve the understanding of opinion by the system. The usage of automatic features creation by the system based on the frequent term or nouns in the sentences is not the best solution because not everyone is thinking the same thing regarding a particular product. Pre-defining features or entities regarding the product through the creation of ontology in Multi Dimension Database is the main difference between the previous researches and our research in this area.

In our research a new architecture for Opinion Mining, which uses a multidimensional
model to integrate customers’ characteristics and their comments about products (or services) was proposed. The key step to achieve this objective was to transfer comments (opinions) to a fact table that includes several dimensions, such as, customers, products, time and locations. Our model used ontology and synonym of product and product features to group the product features, than we used the combination of frequent noun (adjective and adverb) and polarity lexicon to capture pair product features and opinionated noun (Yaakub et al. [2013]). For summarization of opinionated comments, we used seven levels polarity to calculate the product orientation.
Chapter 3

Customer Relational Model

3.1 Introduction

In modern era, every company must be creative and know what exactly customers need to attract new customers and to retain their current customers. This is very crucial for them to be competitive and thriving against their competitor. The changes of marketing paradigm from broad marketing target to customer-centric, one-to-one marketing need new mechanism for every company to survive the current trend of business market. Campbell and Cunnigham [1983] said the greatest asset that a company has are their customers. Furthermore, every company must gather as much information as possible to gain knowledge about their customers’ need(Achilladelis et al. [1971]). This is to ensure they produce or market the product that customers are looking for.

Customer analysis, an analysis about customer behaviors and activities while doing transactions is getting more important among companies today. As a result, a lot of research on this topic have been done (Campbell and Cunnigham [1983], Achilladelis et al. [1971], Parsons [2003], Ittner and Larcker [1998]). Rainardi [2008] defined customer analysis as an analysis of customer activities and behaviors when doing a transaction with organization or company. Customer analysis that analyzed with sufficient information produced a good result especially relationship between company and customers. Campbell and Cunnigham [1983] emphasized the importance of any company that has good relationships with customers to oversee the marketing and purchasing strategy. Meanwhile, Peppers and Rogers [2001] also stressed that company would be up or down depending on their capability to produce a good one-to-one
relationships with their customers. This statement was supported by Parsons [2003] as his finding shows that sales by companies and visits by customers were the two main attributes that generate optimum customer behavior on purchasing items.

Customers’ satisfaction is the main target for every company to achieve their goal in business. Ittner and Larcker [1998] acknowledged the importance of intangible assets (satisfaction of customer) that need to be highlighted in predicting future financial performance and not historical accounting measure that currently used by many companies. Thus, the ability of any company to understand their customers’ needs, preferences, buying behaviors, and price sensitivity posses major advantage against their competitor (Chowdhury [2009]. Reichheld and Teal [2001] stated that a five percent increase in customer retention would resulted in 35 to 95 percent increment in average of customer lifetime value, that means an easy profit to company when compared with investing to find a new customer.

3.2 Customer Relationship Management

Customers are the greatest asset that a company has (Campbell and Cunnigham [1983] but it is not easy to maintain the same customer for a prolonged period without specific research on customers’ behaviors. Every company must have meticulous knowledge about the customers’ needs (Achilladelis et al. [1971]) to make sure the new product is relevant for the targeted customers. The changing landscape of marketing and business strategies among companies and their competitors has changed the way company manage their company. Many companies today are looking towards establishing their connection with customer for longer term as they know that loyalty of customer means a lot for them especially in reducing cost and increasing the company profit (Chen and Popovich [2003]). In achieving these goals, companies must find a new tool to improve their businesses. One of it is known as Customer Relationship Management or CRM.

Customer Relationship Management (CRM) is an important tool for companies to communicate with their customers or target customers. The large amounts of websites in the World Wide Web (WWW), especially for business segment have made the CRM more important than before. Tsiptsis and Chorianopoulos [2010] and also Winer [2001] described the CRM as a strategy for companies to build, manage and strengthen the relationship with their customers for
3.2. CUSTOMER RELATIONSHIP MANAGEMENT

extended time period if not forever. However, Rainardi [2008] defines the CRM as an activity to gather customers contact and manage the customers through communication. He also mentioned that companies analyze the information about customers to attract new customers, and retained current customers through campaigning and also by providing services and support to them. In technical term, CRM is a middle ware of company’s system between front end (sales, marketing, and customer service) and back end (financial, operation, and logistic) (Eckerson and Watson [2000]. Chen and Popovich [2003] highlighted this by taking full advantages of technology from CRM, a company will boost their ability to gather more information and analyze data regarding on customer patterns, get better understanding of customer behaviors, then produce more accurate predictive model. In return, customers will get better service and information about product which helps company to increase their retention rate of loyal customer.

The main objective of CRM is to do everything for customers from tracking customer behaviors to sending the email for direct communication with them (Winer [2001]. Chen and Popovich [2003] have highlighted two main objectives of CRM, which are intended for companies and customers. For company, getting higher revenues and reduces the operational cost are the two main objectives, besides helping company to assess customers in terms of loyalty and forecast future transaction. On the customers’ side, they can get a lot of information about product and service from company in a consistent manner and this will help them to save their money and time. However Tsiptsis and Chorianopoulos [2010], claimed in their work that the main objectives of CRM are to retain customers through customer satisfactions and to build customers development through customers’ insight such as customers’ needs, behaviors, and potentials. Meanwhile, Chen and Popovich [2003] have emphasized that to be successful in implementing CRM, every company has to integrate three main components in CRM which are technology, process and people. Figure 3.1 shows the model produced by Chen and Popovich [2003].

Our new architecture of CRM combines customers’ personal record, product’s record and feedback from customers regarding the particular product that they already used.

Business strategy, processes based on technology and customer centric business, and cross functional integration are the four main contexts that Chen and Popovich [2003] highlighted as a vital part in implementing CRM to support the three components of technology, process
Figure 3.1: Customer Relational Model based on People, Process and Technology
and people. They define people as customer and staff of company. Customer needs must be the most important thing for company to understand and analyze for future transaction. Good understanding about customer needs help a company to strengthen their relationship with customers. Every staff in company must change their culture especially in producing customer-centric model, which makes CRM is one of the best tools to implement it.

The emergence of technology in this era has given a lot of opportunities for company to take advantages of it. A lot of data and information about customer and product can be retrieved easily than before. CRM application has the ability to collect and analyse data in large amount with less time. Predictive model and report are among functions that can be produced by using CRM.

CRM (technology) can’t be used properly without people (customer and staff) and process of business (Business strategy). Changes on technology also need changes in the process of business. Every company must learn on what to change in their business strategy especially in marketing and promotion’s strategy. Customer relationship marketing strategy need to focus more on customer to gain better understanding on what customer really needs, their preferences, and buying behavior (Chen and Popovich [2003]).

The traditional method used by companies or marketing in acquiring retaining customers is by training staff about customer’s behavior in purchasing product (Winer [2001]). The emergence of CRM has improved the way companies analyzing their customers’ behavior by gathering information about customers, their attitudes on making transaction to forecasting future payment that will be made by customers. The companies used the database management to manage relationship with their customers (Anderson et al. [2007]). Rogers [2005] mentioned the importance of customers to improve the value of companies. While the work by Mithas et al. [2005] mentioned that CRM has improved customers’ satisfaction by understanding the customers’ behavior. A lot of information such as customers’ shopping patterns and behaviors, promotions and sales data, and pricing data are stored in companies database (Anderson et al. [2007]).

Winer [2001] had introduced CRM Model based on 7 basic components as shown in Figure 3.2 below.

First step in CRM is creating a customer database. Customers record includes their personal information, customers’ contacts, response to marketing packages, and also descriptive
Figure 3.2: Customer Relational Model
information. Personal information is the basic background about customer but it is the most important part in segmentation customer group especially based on customer’s income. Customer contact used to make sure company has good relation with their customers in making two ways communication between company and customer. Descriptive information and marketing package are important in doing data analysis about customer.

Next step is analyses from the database. Traditionally, this task is to define the right customer segment. The goal of this step is to find the most profitable prospect that companies can give to customers. This is our main research activity for CRM. Result from the analysis phase plays a vital role in the customer selection phase to find which customers are targeted for particular products or services. In the customer targeting phase, selected tools used for targeting customers such as e-mail, short message service (SMS), papers, and telephones. After that, the companies must find the most accurate way to communicate directly with their targeted customers by using the tools that was chosen in previous step. The main objective in this phase is to deliver the best level of satisfaction to the customers. The last two steps in CRM Model are privacy issues and the measurement for the CRM’s program success. As mentioned earlier, our research only involved the analyses of the database phase which is the second phase of CRM Model. In this part, the data warehouse concept, OLAP, and data mining (opinion mining) are used to enhance the current CRM architecture.

Besides the strategy to obtain new customers, the relationship between companies and current customers is very crucial for companies’ revenue. Reichheld and Teal [2001] mentioned from his work that five percentage increase by companies to retain their current customers had impacts as high as 95 percentage on the companies’ net present value. The level of communication between companies and customers is different depends on the domain of the business. Figure 3.3 below shows the level of interaction between a company and their customer based on the business domain.

The company in domain business such as banks, telecommunication, and retails have many direct customer interaction. This gives easy task for company to construct the database because they can easily gather information from customers directly every time they do transaction with the customers. The large data collected from the transactions can give more accurate result for system to predict the customers’ interest in the future services or products.
However, businesses such as furniture and automobile may have harder job to gather customer’s information because the less interaction and also always communicate indirectly with their customers. Meanwhile, Rainardi [2008] has highlighted eight main activities of CRM such as performing campaign segmentation, managing permissions, performing customer analysis, and managing customer support. Performing customer analysis is one of our main activities in our research task. This has resulted in using data mining as the best statistical process analysis for data warehouse.

Nowadays, CRM concept receives very good reception from many companies as it help companies to improve, especially in marketing segment. It has become one of the most competitive and effective strategy for company. In defining the CRM meaning, Rainardi [2008] has emphasized four main activities which are the activities to establish contact and manage communications with customers, analyzing information about customers, campaigning to attract new customers, and performing business transactions with customers such as servicing customers, and providing support to customers. Furthermore, Rinaldi has listed the main activities when performing CRM by utilizing data warehouse as follows:

- A Single Customer View (SCV)
3.2. CUSTOMER RELATIONSHIP MANAGEMENT

- Performing Campaign Segmentation
- Managing Permissions
- Monitoring Deliverability
- Performing Customer Analysis
- Managing Customer Support
- Managing Personalization
- Administering Loyalty Schemes

Two main activities that is an interest to our research are SCV, and Performing Customer Analysis, as we mainly considered and performed a thorough research in this part. Customer analysis means the analysis of customer activities and behaviors, which categorized into two section: Descriptive analysis which requires the understanding and description of customer activities and behaviors, and predictive analysis which emphasizes on prediction and forecasting for future product and marketing (Rainardi [2008]).

Usually, CRM applications are systems that manage customer information such as customer background, and transaction history, demographic information and information about company’s business such as company’s campaign program. From these basic information, CRM then use it to analyse and predict customer behavior regarding the business transaction. Results from analysis and prediction then used to manage customer in several group based on category set-up by company to increase their sales and profits. Activities such as loyalty and rewards program to customer are among the strategy used by company either to retain loyal customer or to attract new one. In current CRM application, system only integrate customer data based on business transaction history but not the feedback or comments from customer about particular product. Hackney [2000] warns that CRM tools in market today still can’t solve 100 percent business problem, as risks still happen such as unhappy customer and unplanned company budget revisions to enhance the CRM capabilities. Furthermore, the main goal of CRM is to concentrate more on selling and less on administrative tasks (Chen and Popovich [2003]).

As mentioned earlier by Chen and Popovich [2003], to success in CRM all three main components in CRM must be integrated in real world. Our model was developed to integrate
not only about customer data but also comments about product from customer. The combination of technology (CRM, customer analysis and data warehouse), people (opinionated comments) and process (integration of customer analysis and comments) are our main tasks in this research. Information from customer’s comments and also their background are our main input in developing our model. Customer analysis is helping companies to have better understanding on their business model, especially their customer. Good understanding on what customer need, give advantages for company to forecast on what the future market segment and strengthen the loyalty of customers. In next two chapter, we elaborate details about the idea of data integration and how we implement it in real world respectively in chapter four and five.

**Single Customer View (SCV) and Customer Segmentation**

Single Customer View or SCV is an integrated from customer data from various places or departments that is stored in one central place. The main idea CRM application need SCV is to describe a customer by only one definition based on data of customer’s attributes. This is the main part in CRM that has large data about customer. In dimensional modeling, SCV used SCD’s (Slowly Changing Dimension) technique to preserve historical information especially keeping it in row’s format. The main advantage of SCV is it gives opportunity to company to perform better customer analysis because customer data is in dimensional database which less join table in database compared to normal database. Furthermore, customer data in CRM is already integrated, which gives opportunity for company to combine customer attributes easily. As a result, descriptive and predictive analysis produced better result with large information. Eckerson and Watson [2000] supported this statement by mentioned that company can create a 360 degree view of customer to get advantage of future business transaction by analysing the past data or records from customers.

Table 3.1 shows segmentation types of customer in customer markets (Tsiptsis and Chorianopoulos [2010]). Marketing is a very important element that needed by all company to promote their product or service to right customer at the right time. Thus, every company must understand customer segmentation to optimize its market strategy to achieve their business goal. Based on table 3.1 six types of segmentation types are identified as:

- Value Based
3.2. CUSTOMER RELATIONSHIP MANAGEMENT

The most important segmentation type as it used to identify the most valuable customer and analyses the changes of value all the time. For marketing strategy, this segment is used to optimise the budget for marketing initiatives and also used in deciding the best strategy for service delivery.

- Behavioral

The most useful segmentation because data collection for this segmentation is the easiest as most of the data stored in company database. Customers’ behavior and transaction patterns are collected then classified together based on company’s chosen characteristic. Loyalty program, new product development and product marketing strategies are among the end result based on this type of segmentation.

- Propensity based

This segmentation group customer based on propensity score such as churn score and cross-selling score. One main objective for this segmentation is to optimise the marketing action to all customers.

- Loyalty based

Most research about customers’ loyalty are done here. This segmentation classified the loyal customer based on value of customer, in which high value customer given more priorities than low value customer. Marketing for this group of customer is essential as loyal customer give more revenue to company compared with other customer group.

- Socio-demographic and life-stage

Customers grouped, based on their socio-demographic and life-stage such gender, education and household income. This type of customer group is suitable with specific marketing target.

- Needs
3. CUSTOMER RELATIONAL MODEL

The most difficult segmentation as customer is grouped based on data from market research. Company must identify customer needs, preferences, attitudes and the most difficult thing is what they expect from company’s future product and service. Furthermore, this segmentation is used by company to support their new product development and future research works on designing product.

Besides customer segmentation, Tsiptsis and Chorianopoulos [2010] also mentioned about segmentation management strategy. Figure 3.4 shows four steps for segmentation management strategy. The first step is to identify customer segments in database. Result from customer segmentation processes must be deployed for this purpose. Report from this segment such as customer profiling used in second step, evaluate and position the segments. Evaluation and positioning the segmentation are based on the market’s study. At this stage, management team must do lot of research survey regarding what customer needs, wants and also customer’s demographic information. Next, they must analyses segments and identify competitive advantages, opportunities and threats, set marketing and financial goals among others. Key performance index (KPI) must be defined at this stage to monitor the performance of every single segment. Third step is to perform cost-benefit analysis to prioritise actions per segment. Data mining and market research at this phase is to analyses customers behavior. The objectives for this phase are identification of the main factor of cost and profitability and also assessment on customers’ behaviors. Last phase is to build and deliver differentiated strategies. The keyword for this phase is specialisation. Therefore, every single segment must be identified by different segment management team. Each team must address their own segment needs and preferences. As a result, customer in different segment must be handle with different ways and different marketing strategies.

3.3 Data Mining in CRM

Rygielski et al. [2002] mentioned data mining as an analogy like gold or coal mining, which looks up and extracts information hidden in data warehouses or websites. Other definition for data mining is Knowledge Discovery from data. Meanwhile, Han and Kamber [2006] emphasized data mining as a sequence of processes namely as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Therefore, Han and Kamber [2006] described data mining as a process of discovering
### Table 3.1: Segmentation Types and Business Tasks

<table>
<thead>
<tr>
<th>Business Task</th>
<th>Segmentation Criteria</th>
<th>Tools and Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>New product design and development</td>
<td>Needs and behavioral</td>
<td>Combination of data mining and Market Survey</td>
</tr>
<tr>
<td>Design of customized product offering strategies</td>
<td>Behavioral</td>
<td>Data mining and Cluster analysis</td>
</tr>
<tr>
<td>Brand image</td>
<td>Need</td>
<td>Market survey and Cluster analysis</td>
</tr>
<tr>
<td>Differentiated customer service</td>
<td>Customer value with personal attributes</td>
<td>Customers’ grouping</td>
</tr>
<tr>
<td>Resource allocation and prioritisation of marketing</td>
<td>Customer value and understanding customer</td>
<td>Value tiles and market survey</td>
</tr>
<tr>
<td>Identify target group</td>
<td>Propensity scores (classification model)</td>
<td>Data mining (classification model)</td>
</tr>
</tbody>
</table>
interesting patterns and knowledge from large amounts of data. Based on the above theories, we summarized data mining as the process of extracting interesting patterns or knowledge from a large amount of data with the objective of analyzing or predicting meaningful information. Other definition for data mining is knowledge discovery in database, knowledge extraction, data/pattern analysis, data archaeology, data dredging, information harvesting, and business intelligence (Rainardi [2008]).

Data mining is a very powerful technology as it can implement different types of data such as flat files, semi-structures data, world wide web (WWW) and others. The combination between data mining and CRM is a very interesting research as CRM can integrate customer data from various data sources in a central place with the concept of Single Customer View or SCV that gives developer a holistic view of customers in term of contact management and marketing services (Rainardi [2008]). Hence, this concept gives opportunity for company to understand more about customer. The ability of data mining to extract knowledge about customer reduces company’s cost especially in finding customer’s pattern and behavior through the survey technique (Chen and Popovich [2003]). Furthermore, Chen and Popovich [2003]
have highlighted five advantages of using data mining especially data warehouse in CRM as follows:

- Better accuracy of data and faster access to information
- Better quality of data because bad and duplicate data are eliminated
- Extract and manipulate data more quickly for analysis
- Result of analysis represent in report
- Ability to calculate and estimate future value

These five advantages of data warehouse are the main reasons that we used data warehouse as our main technique to implement our research. Less time in accessing information to analyze data is a key issue in business area because company and customer’s time are very valuable. The ability of a data warehouse to generate good report from analysis gives company and customer a convenient way to understand the result from analyses based on what they need.

An effective CRM strategy can be produced by data mining technology, which performs analyses on customers’ transaction pattern with a company (Tsiptsis and Chorianopoulos [2010]) and customer product reviews from particular website. Good strategy by company will lead to better relationship with their respective customers. Furthermore, company-customer’s relationship increases customer satisfaction and profitability for company as customer promotes the image of company to their colleague either physically or by using social media web2.0 such as Internet’s forum and Facebook.

Figure 3.5 shows the relationship between customer lifecycle and its value to company (Tsiptsis and Chorianopoulos [2010]). Customer value is low in the very first relation with company, as at this time the company’s main priority is to capture new customer. In the second stage, company starts to establish relationship with customer and strengthens it with a lot of selling model based on the result from customer analysis model by data mining technology. The longer a company could retain a customer, the value of that particular customer is more valuable to the company. This figure proves the importance of loyal customer to a company.

In the current CRM, descriptive and predictive analyses are easily done by storing integrated customer data in a dimensional data warehouse to discover customers purchasing patterns
Figure 3.5: Customer Value and Customer Lifecycle Management (Tsiptsis and Chorianopoulos [2010])
and forecasting customer behaviors (Rainardi [2008]). Our research emphasized on helping CRM to forecast what product a customer really needs based on customers’ comments.

3.3.1 Performing Customer Analysis

As mentioned earlier in this chapter, analysis of customer is important for company to get a better understanding about their customers. A successful marketing by any company depends on how good reactions are received from customer on a particular product or service produced by the company. Company that has thorough knowledge about their customer has good prospect to be successful in their business. Féraud et al. [2010] mentioned besides understanding what customer needs, a company must also use a good application that can analyze data rapidly to generate predictive analysis. In doing this, company needs comprehensive tools such as CRM that has the ability to process a large data.

Customer analysis is an analysis about customer and their purchasing habit based on customer past transaction record with the company. Two types of customer analysis are descriptive analysis and predictive analysis. Descriptive analysis is an analysis about customer behaviors and activities in doing business transaction to get a better understanding about customer. Example of descriptive analysis is risk pattern identification in insurance industry. Meanwhile, predictive analysis is an analysis about predicting or forecasting customer behavior and activities to get better understanding on what customer needs from a company in their future transaction. Example of predictive analysis is the annual sales forecasting for company. However, Tsiptsis and Chorianopoulos [2010] claimed that analytical CRM is how a company analyses its customer information to get better understanding about their own customer and then give a better service or sell a product to the right customer at the right time. Hence, data mining model is critical as it is used to analyse the value of every customer by understanding their need and characteristic before predicting customer behavior. Then, this particular data is used to analyse patterns of customer in doing transaction(s) with company to extract valuable knowledge for strengthening the relation between company and customer. The result(s) from this analysis was highlighted by Tsiptsis and Chorianopoulos [2010], which must be integrated into the whole of CRM system especially OperationalCRMFront – lineSystem, as they believed that the interaction between a company and its customer can be more effective as customer can receive more information about the company’s product(s) or service(s) directly
from the company’s representative such as email and SMS notification. For companies, they can be personalized customers’ experience (one to one marketing) as they already possess details about what customer needs from them and also what customer expects from them for future product or service.

Figure 3.6 shows the phases involved in Cross Industry Standard Process for Data Mining (CRISP – DM) (Tsiptsis and Chorianopoulos [2010]). This figure shows the current practice by companies in multi industries in predicting their future business model especially with customers. Data mining involvement is essential for industry to produce the best model for their business.

This data mining project involved six phases. "Business Understanding” is the first phase, which is important to define a company’s business goal (s). Business circumstances and current company’s situation are assessed before the company’s business goal is transformed into data mining objective. A good understanding and definition about current business process is the key for success for this whole data mining project. The second phase is Data Understanding. It is compulsory for any data mining process to analyse data and its quality because the data will influence the final result of this project. Data mining’s pre-processing techniques for data is required for DataPreparation phase. Tasks such as data acquisition, data cleaning and data transformation are performed in this phase.

The fourth phase is Modeling. In this phase, some modeling techniques were chosen to run the experiments. Dataset for this phase was divided into training and testing datasets for evaluation purposes. The next phase is ModelEvaluation. The main objective in this phase was to examine if the chosen model is suitable for company’s use. The best model from the previous phase is chosen for implementation in the Deployment phase. This model’s results are integrated with company’s operational CRM system. Maintenance works are performed to accommodate company’s business goal.

Féraud et al. [2010] built a customer analysis platform based on data mining platform to build a predictive model using two orders of magnitude in more explanatory variable using naive Bayes classification technique. Féraud et al. [2010] main target was getting knowledge about customers in CRM. They used score on customer to detect churn, the inclination to subscribe a new service. The time for processing this technique was quite slow as this was one of the main problems in implementing this technique in CRM.
3.4 **Problem in CRM**

In 2001, US$20 billion was spent for CRM tools and in 2003 this spending grew up to US$46 billion. The amount that companies had spent was very high, but the result from using CRM in business was not up to what they expected. Rigby *et al.* [2002] reported that companies which used CRM not only lost 55 percent of their customers, but also damaged their loyalty program with customer. Meanwhile, Reinartz *et al.* [2004] mentioned that many companies which deployed CRM failed to use the CRM effectively and managed most of the CRM program. Rigby *et al.* [2002] revealed four reasons why CRM failed to achieve its goals. Four reasons are:

- Implementing CRM in ad-hoc
- Installing CRM technology without preparation

![Cross Industry Standard Process for Data Mining](image)
• Assuming that more CRM technology is better

• Stalking, not wooing, customer

In facing stiff competition in business market, every company must redefine their business strategy to ensure that they remain competitive in finding what customer really need from their company. Every company needs to define smaller and smaller segments in their target customer and produce the right product or service to cope with customer need and preference that keep changing from time to time (Chen and Popovich [2003]). Therefore, every company must learn faster than before to understand what the customer really need from them. As a result, company must conduct segmentation analyses and determine market goals first before implementing CRM (Rigby et al. [2002]). In this case, company needs thorough understanding about customer’s need by creating a customer strategy. Customer has different needs based on their background. Rigby et al. [2002] emphasized the importance to segment customers into sets of group depending on some range of attributes such as the most profitable, education, and income. Companies can implement the CRM based on their own business strategy. Thus, the firm needs to urgently learn how to develop and strengthen their CRM capabilities (Wang and Feng [2012]).

Installing CRM before a company really understands how they do business process is another aspect that brings the company in trouble. Organization and its processes are two main components in every company that needs to be clarified in details, especially the job scope for every staff. Chen and Popovich [2003] mentioned that CRM is a combination of business process and technology. Thus, this shows that every staff including the management level must understand about CRM thoroughly and must change their way of doing business in parallel with the CRM. Rigby et al. [2002] warned that to be successful in implementing CRM, a company must change their structures and processes, and also alter their corporate cultures too. The META Group Report (Meta [1998]) also supported the importance for a company to change their processes in implementing CRM because without any change in company’s staff attitude, it is like throwing money into a black hole.

Company must be made aware that the success in achieving their business target depends heavily on their understanding on what the customer really needs and not on what they have with CRM technology. Wang and Feng [2012] highlighted that the combination of three components that bring success to implement CRM are customer orientation, customer centric
organizational systems and CRM technology. Chen and Popovich [2003] also emphasized that to use CRM to its optimum ability, every company must combine three aspects in their work, people, process and technology. This shows that CRM technology alone is not the key factor in implementing a good CRM in company.

A good relationship with customer is the most important thing that every company must have. Customer is the most valuable asset that any company must protect from losing to competitors. Rigby et al. [2002] described that one of the failure in implementing CRM is when a company does not really understand their target customer. Most companies try to build relationship with the wrong target customer or the right customer at the wrong time. Companies must possess skills to strengthen their relation with customer by identifying and understanding the customer, and they also need the ability to acquire and retain profitable customer (Wang and Feng [2012]).

Besides that, understanding what customer needs, especially product was our main task in this research. Our research gives more information about a product in detail, including a summary of customer’s orientation on every product features. Our model enhances the current CRM model in term of technology. Wang and Feng [2012] emphasized that one of the best abilities that CRM must have is to enable a company to capture accurate product or service and timely insights pertaining to customer needs.

3.5 New Idea to Enhance CRM

This chapter is highlighted the important of opinion mining (sentiment analysis) in CRM and our idea to implement opinion mining in CRM, because this is the first time for us to combine CRM with opinion mining.

Tsiptsis and Chorianopoulos [2010] claimed that for every strategy in developing a CRM system, an organisation must always emphasise that the company must satisfies its customer. Consequently, customer must not only remain with the company, but also manages a good relationship between company and them. In further discussion, Tsiptsis and Chorianopoulos [2010] compared two different strategies used by companies to try to sell a product. The first, is to try to sell everything from their stores, while the other, the company analyses customers’ need first than gives plenty of choices based on the analysis result. As a result, customer appreciates
the second company more as the company knows what they really need.

Based on the seven components of CRM in Figure 3.2 and the underlying problem in CRM as mentioned in the previous section, there were two useful entities influencing the success of a CRM system; customer and product. Figure 3.7 shows the relation between customer, product and other entities in the CRM system. Existing CRM system stresses on customer compared to other entities because it focuses on one to one marketing. Customer’s entity contains information such as customer’s background, record of purchasing, and loyalty program. Meanwhile, other CRM entities such as marketing, customer service, sales and delivery, all have relationships with customer. In current CRM practice, product entity that contains information about the product in detail is not a CRM main entity as it emphasizes more on customer, not on product.

![Diagram of CRM's Entities](image)

**Figure 3.7: Relation between CRM’s Entities**

In this new idea, we highlighted the combination of customer and product entities together with other CRM entities. Customer records without any product or service attribute means nothing in CRM. Thus, this situation also applies to other CRM’s entities. Product entity or data is important as customer entity because without these two entities, customer analysis cannot be done. A good product means a good business. For example, *i-phone* is one of the best mobile phone in the market today, with this product Apple (company that produce *i-phone*) has increased the number of their customers. This situation shows that customer and product are important entities in the CRM system with support from other entities.

Customers’ satisfaction on a particular product is one of the requirements in business as well as the satisfaction with a company’s treatment on customer before and after sales. This shows
that company needs feedback from customers about their product. One of the feedback medium is customer’s review or comment on the company website. This feedback also needs responses from other entities, especially customer service and marketing entities. The emergence of Internet has produced a new way of obtaining feedback from customer regarding the product through the comment section on the company website. Normally, a lot of feedback or comments from customers in this comment section on company website with various terms and sentiments used by customer to express their opinion on a specific product. This information is important for company to improve their product and also their business. The problem is comments are unstructured data and other entities are structured data in *CRM* system. In the next section, we will discuss on how to solve the problem regarding the structured and unstructured data.

### 3.6 Integration of Customer Comment with *CRM* system

Comment on product review is an unstructured data, meanwhile others entities in *CRM* system are structured data. The integration between structured and unstructured data is a complex process. Figure 3.8 shows the new *CRM* model proposed for this research project.

**Figure 3.8:** New CRM Model Structure

Comments from customers’ review were collected together with other data sources such as
customer, product and marketing. However, only comments data needed to be extracted first because its format is unstructured data. Figure 3.9 shows the processes involved in developing our new model in details. The target of comments’ extraction was to capture product features and sentiment words that represent the orientation of customers on a product. Part of Speech technique was used to differentiate every single word in groups such as noun, adjective and adverb. Association rules were also used to capture product’s features. Meanwhile, information from product database was used to produce product ontology based on four tiers of product description, which are product name, product attributes, product’s attribute features and instance of the attribute. Synonyms were also used in product ontology to detect and capture infrequent nouns or product’s features.

The output from the extraction process was then used in the transform process together with product ontology for matching product feature and sentiment words. The outcomes of this process were then loaded into a data warehouse together with customer database. The combination of product, customer and opinion extraction(output from load process) were used in data warehouse for customer analysis. After the customer analysis process, data cubes were designed to present customer analysis based on Opinion of product features, customer groups and set of product known as OGC. The OGC’s report main objective was to produce a report on
3.7 Summary

In this chapter, we discussed about the state of the art for CRM, current design of CRM models and how our enhanced model can fit in existing CRM model. The idea to improve the current CRM model was based on business strategy and CRM’s current processes based on technology and customer centric business. As the CRM main target is to retain current customers and find new segment of customers, we assumed that it is important to add product or service in CRM as companies will not draw any customer without product or services. The idea of this research model was to combine product and customer entities with support from new entity (customer review) for customer analysis as product that satisfy customers can boost the profit of a company. Every customer is important for companies to improve their business models, so companies must really understand their customers and targeted customers as different customer have different needs, taste, behaviour and potential (Tsiptsis and Chorianopoulos [2010]).
Chapter 4

Opinion Integration Model

4.1 Introduction

The emergence of Internet has changed the way companies conduct their businesses. A company cannot depend on the strategic placement of their shops anymore because in today’s business trend, customers come from all over the world. Demographic issue is not relevant in this Internet era. Online businesses are important nowadays because many customers do not have much time to frequent shops at shopping mall. Online commercial website is the alternative way for customers who are always busy with their everyday activities. Many business websites today offer a variety of choices about products and services with cheaper price compared to shopping at the malls or shops. Competition in online business is getting tougher by day as customers can compare products and prices offered by companies easily on websites. The problem arises when customers have difficulties to choose the right products or services offered by companies because many business websites give a lot of information that can persuade them to buy products or service that they are not interested in buying (Yaakub et al. [2011]). Another problem is some of the information is useless or inaccurate due to misleading advertisement by some companies.

Most business websites today that market a particular product have commentary facility on their website. This comment section gives opportunity for customers to write down some opinions about the product they already purchased or to share some ideas or thoughts about that particular product with other customers. Thus, these comments from customers are diversified
Comments from customers about a particular product vary from a general term such as product’s brand name and the product in general, to specific product features, technical parts or terms. Sometimes in the comment, customers compare the product with other similar product features or function (Yaakub et al. [2011]). Normally, customer gives their opinion about the product by showing the strength of their comments by expressing terms or words such as excellent, worst or okay. The combination of opinionated sentence and product in comments were our main extracted data for this research project.

To be specific, we assumed that we have concept hierarchies (CH) for all products. We defined two concepts in this research called as:

- **Concept Hierarchy (CH)**

  A concept hierarchy defines a sequence of mappings from a set of low-level to higher level, which is a more general concept (Yaakub et al. [2011], Yaakub et al. [2012], Yaakub et al. [2013]). In our model, we used four levels of concept hierarchy. An example of four levels of concept hierarchy is address: street name < city < state < country.

- **Opinion Sentence (OS)**

  An opinion sentence is a concept consisting of a set of sentiment words and product features pairs. This concept is produced from customer comments, which have combinations of opinionated word and product terms such as the product in general, product features or product attributes (Yaakub et al. [2011], Yaakub et al. [2012], Yaakub et al. [2013]). This concept also emphasizes on the comments with opinion sentence concept, which our model used in the model development. Comments without this combination were ignored by our model.

  Details about these two concepts are discussed in chapter five.

The most difficult part in this research was to find the entities or product features and the associated sentiment words in documents. By using the Part of Speech (POS) technique, all entities were grouped in noun, while the sentiments words were grouped in either adverb and adjective. Detail discussion about the extraction process is presented in chapter five.
Sentiment orientation for every opinion sentence is hard to decide either as positive or negative or neutral. One simple idea to decide the sentiment orientation is based on the customers comments in general. Normally, a positive customer comment is more likely to give a positive sentiment-feature pair rather than a negative comment (Yaakub et al. [2011]).

Another important task in our model was to differentiate an opinion sentence for product itself. It is not accurate to discuss the orientation of an opinion sentence if the comment from customer describes the product in different levels of product features (Yaakub et al. [2011]). For example, customer may comment that a NOKIA phone is not good in general but the model NOKIA Lumia gets an excellent comment. A single comment from a customer may reveal more than a single features or entities; thus, we integrated all possible customers sentiment descriptions for a particular product feature. This was the main point that we created our own ontology for product features.

4.2 Customer Data Integration (CDI)

Our research tasks was based on five processes of customer data integration in CRM as introduced by Rainardi [2008]. Figure 4.1 shows the customer data integration’s process.

The first process of CDI is to retrieve, which models the process of extracting customer data from various resources. All information about customers such as customer’s background, demographic location, and historical data of purchasing are among the data retrieved by a model. In our model, we also retrieved the customer data, but we enhanced this process by including comments from customers regarding a particular product, which are in unstructured data format. Comments from customer normally have opinionated words or sentences. Thus, one of our tasks was to extract comments, which were unstructured data format into structured data format using Part of Speech (POS) tag. The next process is the cleaning process. Normally, CRM performs correction for incorrect customer information and removes duplicate information appeared in customer data. In our model, this processed was not involved as we only received a set of cleaned data from companies. Therefore, our task in this part was we cleaned the unused tag for POS such as word ”a”, and ”the”.

CRM’s third process is storing all customer data in a central data source. We also stored all data in a central data source, in our data warehouse. The combination of customers’ data
and comments from customer were stored in a same place, but in different dimensions. One main difference between our model and the current CRM model is that we also stored product ontology, and product features synonyms in our data warehouse model. The main function of this enhancement was for the matching process in our customer analysis process. Our customer analysis process is completely different with the current definition of customer analysis. Our model analysed the customer character and historical purchasing transaction and also opinionated comments from customers to enhance the satisfaction of customers, especially on a particular product because our model can show in detail the polarity of the comments from customer for every product feature (Yaakub et al. [2012], Yaakub et al. [2013]).

In our work we mapped all opinion sentences into a fact table, which was our main table in the model. This fact table was associated with other dimensions in the data warehouse. The initial idea for our data structure (fact table) is shown in Table 4.1, however, the complete fact table and data warehouse is discussed in chapter five. In this fact table, it is associated with a Customer ID, Product ID, Time and Opinion from their respective dimensions. These four dimensions were our main dimensions in this research development.

After all important data were in the data warehouse, we created some algorithms (Yaakub
et al. [2012], Yaakub et al. [2013]) to describe how to transfer an opinion sentence into a list of attribute polarity pairs (Yaakub et al. [2011]). In these algorithms, we introduced the seven level polarity system for any sentiment word captured by the model. Polarity 3 is stronger than polarity 2, and polarity 2 is stronger than polarity 1. These four levels of polarity show the orientation in positive sentiment. However, negative polarity means negative sentiment with −3 as worse than −2. Meanwhile, polarity with 0 means that it is a neutral comment.

Using the data warehouse, we created many data cubes based on what the company and customer need. Our important task here was to analyse the orientation of some groups of customers for a particular product in certain levels (Yaakub et al. [2011]). Figure 4.2 shows a three dimensional data cube that we created to calculate the orientation of customer, named as OGC(Opinion Group Customer). This cube was built based on the categories of product, age groups of customer and opinions in polarities (Yaakub et al. [2011]). This data cube was used by our model to calculate the orientation for a given group of customers based on a category of product and the polarity of opinion.

![Figure 4.2: A Data Cube for Customers’ Opinion on Product Categories](image)

The next two processes were updating data and distributing report to customers. As mentioned earlier, our tasks did not involved customer data, so we did not update any customer data. In our model, we updated the ontology of the product and the product features synonym.
In other word, if we receive a new product feature from companies, then we need to update the product ontology and also product features synonym in our data warehouse. The last process was we produced product reports based on comments from customers. Customer can easily access our model to find the summary of comments from customer regarding a product (general description) and its features (specific description).

The main advantage in using a data integration model is that companies and customers can use cleaner, single and reliable version of customer data. This means that customer analysis done by companies can produce better results, and hence, customer satisfaction with a company and its product will increase. This is made possible by using a cleaner and more reliable data.

4.3 Research Process

Research plan can be defined as the strategy for achieving the goal of this proposed research project. In this part, the phases of the research were defined thoroughly including the research question for each phase (hypothesis). Figure 4.3 shows the main research plan for the proposed research. It was divided into four phases, namely gathering data, designing ontology for the database and developing the multi dimensional model, implementing ETL and data warehouse, and evaluating the model.

- Gather/Analyse data

This was the first phase of the research project. Data was collected and analysed in the database. The main data for this research were users comments. Other data came from the CRM database such as customer data and product data.

H1: Is it possible to gather unstructured and structured data for the usage of data mining in later steps?

The first research question for this research project was regarding the data for the research work. This research gathered data users comments of a particular product from Hu and Liu [2004] research data because of the large amount of data from the website amazon.com and we
also used their models as our baselines in the evaluation stage. We analysed this unstructured data and divided them into subcategories such as entities and attributes.

- **Design Ontology for Database and the Development of Multi Dimensional Database**

This is the most important part of this research. We designed our model in this phase. The creation of a data warehouse from a fact table and other dimensions is important in producing a good analysis model for our opinion integration model. We designed our ontology based on product features and synonyms for the model to easily recognise the product and product features. The calculation of $OGC$ (Opinion Group Customer) was also done in this phase.

H2: Will the database design give impact on the development of a good model?

This hypothesis is about the factors that affect the designing of a database. A good database design will help data mining in producing better results. This phase is the most important
task in our research because we must integrate customer comments, which is unstructured data with the data from current CRM model, which is structured data. It is hard to extract the unstructured data, and then transfer it to structured data without a good database design. By understanding opinion analysis, the database’s cubes help the model in choosing the accurate data in opinion mining, which means that the output of this phase is significant to minimise noisy and incomplete data in data mining tasks especially opinion mining.

In achieving this objective; an ontology design was produced to group the product features from customer comments. The data from phase one, was divided in the database by categories of entities, attributes, and relationships. This research project involved high and low levels of ontology. Figure 4.4 shows our early idea for ontology design.

![Figure 4.4: Basic Ontology](image)

Normally, a two-tier ontology is defined as a tuple \((E,R,I,G)\), where \(E\) is a set of entities from product dimension, \(R\) is a set of relations that describe the relationship between entities, \(I\) is a set of instances that describes the usages of entities in a real application, and \(G\) is a mapping showing the citations of using entities in instances. The first tier is a taxonomy, which is defined by a graph \(<E,R>\). The second tier is a set of instances. The relationship between the two tiers is described by a mapping \(G\) and its reverse.

Ontology consists of two parts; entities and relationships between entities. The entities involve data such as customer, city, house, and university. The relationships describe the
relations between entities, such as Ridzwan (customer) likes Filzah (customer) very much. It is very difficult to measure the strength between entities. In this case, Ridzwan likes Filzah 90 percent or 60 percent. This problem was solved in this research using the four tier structure as shown in Figure F.1 by calculating the strength of each entity differently based on tier. In chapter five, we will discuss this issue in detail.

• The Implementation of ETL and Data Warehouse

The third phase is the implementation of ETL and data warehouse. The extraction of customers comment were made in this process.

H3: How current CRM can process unstructured data from customer comments and then transfer it into structured data.

This phase is parallel with phase two. After we designed the data warehouse and product ontology, we extracted useful features and polarity from customer comments and integrated them with the structured data such as product, customer, and other data from CRM database. Techniques such as Part of Speech (POS) and frequency noun were used in this phase. The input in this task was the raw data from customers comments and the output was the extracted data in the structured data. The process of extraction was the hardest part in this research, especially capturing the sentiment and product feature words. The ontology of product features and synonym was used in solving this problem. In the next chapter, chapter five, we explain in detail this important process. Besides that, we created numerous data cubes to produce report about customer orientation on a particular product. Our model has a unique ability to produce reports about product based on customer comments. The calculation of OGC produces different levels of report about a product based on customer sentiment orientation, positive or negative. The model can show results about a product in the general level or attributes (features) level or instances level. Results from the general level may be different in any different level because our model can calculate the customer orientation of a product in four levels, as mentioned in the previous subsection.
• Assessment of the Project to determine the Significance of Research

The last phase of this research project is the testing and evaluation of the whole model. We used five models from Hu and Liu [2004] as our baseline models. These models were used as our baseline as we deployed the same data as Hu and Liu [2004] in their research works.

H4: How significant is the integration of structured and unstructured data from customers’ comments and CRM model using the data warehouse process.

This phase is the last part of our research project. Every research that has been completed must be evaluated to determine its significance to the particular area. The evaluation in this project involved the success of data warehouse implementation in integrating structured and unstructured data as a possible solution to improve the data mining result, especially Opinion Mining. The main measures for this research are precision, recall and F-Score to determine how this model is more accurate and performs better than the baseline models. Our research shows better results compared to all six baseline models for all measurements. Our techniques of using features ontology and ontology have proven their significance in this research area.

4.4 Integration of Opinion in CRM

Every company in this era of technology must take advantage of the advanced tools produced in line with their business. In finding a new product segment and retaining existing customers, companies can emphasize the use of customer analysis tools in their businesses. Customer analysis tool plays an important role in making the CRM system as one of the best solutions for business solution software.

In our work, we produced our own feature ontology for each product that we worked with, such as the mobile phone as shown in Figure F.1. This ontology is based on a four tier ontology as we divided the features of a product based on their subcategories (characteristics) or attributes, such as general term and other technical terms. Normally, customers comment
about a particular product in different ways. For example, some customers comment on the technical features such as size, battery, color, and applications that are built in the product in specific ways but other customers only comment about the technical features in general. These situations influenced our design of the product feature ontology based on the subcategories of the product. Our fact table was also created based on these subcategories or attributes.

![Diagram of Integration Business Model]

**Figure 4.5: Integration Business Model**

We built our dimension database based on our integration business model as shown in Figure 4.5. Four main dimensions were chosen to create this integration business model to represent product, customer, time and opinion. These dimensions were chosen because they were our main dimensions used in customer analysis process. Product dimension has four important attributes, encompassing of Product Name as the general information of the product, Features as the definition of the product and Attributes to describe the details of the product features. Customer dimension has data about customers, while the Time dimension determines the tie to produce report, either daily, weekly, monthly or yearly. However, the Opinion dimension is the calculation of customers’ orientation about a product, either positive, negative or neutral.

Comments by customers are different based on a product’s subcategories such as product name, product features, and product feature attributes. Table 4.2 shows examples of comments from one group of customers on a particular product. This table shows different types of customers comments. As shown in this table, the first comment is a general comment about mobile phone, which is very good. This type of comment is easy for any model or system to determine the product orientation. The second comment is quite difficult for a model or a
Table 4.2: Comment from One Customer Group

<table>
<thead>
<tr>
<th>No</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Screen is excellent and also the speed.</td>
</tr>
<tr>
<td>2</td>
<td>Camera and video quality need to improve.</td>
</tr>
<tr>
<td>3</td>
<td>Beautiful display, fast and responsive.</td>
</tr>
<tr>
<td>4</td>
<td>Price, build quality, and battery life are bad.</td>
</tr>
<tr>
<td>5</td>
<td>It has excellences in Speed, Slimness and Sharpness.</td>
</tr>
<tr>
<td>6</td>
<td>Battery life, screen, radio, and accessories are very bad.</td>
</tr>
<tr>
<td>7</td>
<td>The phone came huge but extremely cool after putting the case.</td>
</tr>
<tr>
<td>8</td>
<td>I like it.</td>
</tr>
<tr>
<td>9</td>
<td>Everything is same as its previous model but this one is smaller and lighter.</td>
</tr>
<tr>
<td>10</td>
<td>The size is ok but the color and its applications are very bad.</td>
</tr>
<tr>
<td></td>
<td>This mobile phone is very good.</td>
</tr>
</tbody>
</table>

Note: The numbers in the table correspond to the priority level of the comments, with 10 being the highest priority.
4.4. INTEGRATION OF OPINION IN CRM

system to detect the product orientation because it contains more than one sentiment words. In our model, this raw data was extracted based on some techniques such as POS tag, level frequency of words, and matching the words with sentiment words from the sentiment words database acquired from sentiwordnet’s website.

After the process of identifying the pair of entities and descriptors from raw data was completed, our model transferred these pairs into another dimension, the Features-Polarity dimension as shown in Table 4.3. This table shows every single product or product feature mapped to their pair, sentiment word. For example, mobile phone was paired with very good. It is not easy for a model to calculate the opinion without grouping the product features into particular groups. Next, we used the ontology created in Figure F.1 and their synonyms to group the feature-polarity pairs. All product features were grouped into seven subcategories based on this ontology and its synonyms. As mentioned earlier, the main function of this product feature ontology was to capture as many features as possible for every single product feature. Ontology without synonym is not as good as ontology with synonym. Our test result in chapter six showed that feature ontology with synonym was better than feature ontology without synonym in all measures that we tested. In Table 4.3, product feature are still separated, such as size, color, mobile phone, screen, and slim.

Meanwhile, Table 5.2 shows that all the product features are grouped together in their particular subcategories. This mapping process is essential because most customers give comments which are too specific, and it is difficult to implement in the real world without grouping features with the same characteristics in one group (Yaakub et al. [2011]). One problem with this technique alone was that sometimes, customer use different terms to refer to the same product or its features. For example, a customer used phone for mobile phone, while other customer referred to it as cellphone. This means that one term of a product feature may have more than one meanings. We solved this problem using product feature synonyms (Yaakub et al. [2012], Yaakub et al. [2013]).

Table 4.5 shows an example of synonym words for phone in our model’s database. This table matched all the possible terms for phone. Even though in the real world there are more than five terms similar to the word phone, in our model we listed five words similar to the root
<table>
<thead>
<tr>
<th>No</th>
<th>Comment</th>
<th>Features Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mobile phone, very good</td>
<td>{screen, excellen}</td>
</tr>
<tr>
<td>2</td>
<td>size, ok, color, very bad, application, very bad</td>
<td>{price, bad}, {quality, bad}, {battery, bad}</td>
</tr>
<tr>
<td>3</td>
<td>smaller, like, lighter, like</td>
<td>{battery, very bad}, {screen, very bad}, {radio, very bad},</td>
</tr>
<tr>
<td>4</td>
<td>phone, nice, case, cool</td>
<td>{phone, huge}</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>{battery, very bad}, {screen, very bad}, {radio, very bad}</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>{speed, excellent}, {shimmer, excellent}, {sharpness, excellent}</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>phone, huge</td>
<td>{phone, huge}, {case, cool}</td>
</tr>
<tr>
<td>9</td>
<td>camera, need improve, video, need improve</td>
<td>{screen, excellen}</td>
</tr>
<tr>
<td>10</td>
<td>battery, very bad, phone, very good</td>
<td>{mobile phone, very good}</td>
</tr>
</tbody>
</table>

**Table 4.3: Features Polarity**
Table 4.4: Attribute Polarity

<table>
<thead>
<tr>
<th>No</th>
<th>Comment</th>
<th>Attribute Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>(general,3)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(general,1)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>(general,-3)</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>(application,-3)</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>(...)</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>(connectivity,3)</td>
</tr>
</tbody>
</table>

Table 4.5: Feature Synonym

<table>
<thead>
<tr>
<th>Feature</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone</td>
<td>phone</td>
</tr>
<tr>
<td>phone</td>
<td>mobile phone</td>
</tr>
<tr>
<td>phone</td>
<td>cell phone</td>
</tr>
<tr>
<td>phone</td>
<td>handset</td>
</tr>
<tr>
<td>phone</td>
<td>cellular phone</td>
</tr>
</tbody>
</table>

word (product’s feature). Besides that, all of the sentiment words also change the polarity of sentiment words based on the seven polarity system as mentioned earlier in this chapter.

The result from attribute polarity table is shown in Table 4.4. We then calculated opinion polarity (\( \text{ogc} \)) based on the \( \text{OGC} \) formula. Figure 4.6 shows the result of our model for early calculation of opinion polarity based on the seven level polarity system.

As shown in figure 4.6, our model was able to calculate the orientation of customers about a product based on their comments on the product’s website. We calculated this polarity using \( \text{OGC} \) formula discussed in chapter five. This result gives opportunities for companies to provide the best product for their current customers and also their potential customers. Customers also benefit from this research outcome, especially in getting information about a product based on reviews by other customer for every single product feature that they need. For example, if the customer is looking for a mobile phone that has received positive or negative comments in mobile’s application, they can use this model to make the right choice.

This chapter discussed about the basic idea of our entire research model and the final output of our research tasks. In the next chapter, we will discuss the technical details of the
Figure 4.6: Opinion Polarity

<table>
<thead>
<tr>
<th>Attribute</th>
<th>General</th>
<th>Entertainment</th>
<th>Connectivity</th>
<th>Application</th>
<th>Accessories</th>
<th>OCC Orientation</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negative 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive 1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Negativa 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negativa</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
4.5 Summary

Two main concepts behind our opinion integration concept are hierarchy and opinion sentence. The concept of hierarchy brought this research into another dimension with the creation of ontology and synonym for every product ontology terminology. Meanwhile, opinion sentence produced a set of product features and its sentiment words. Product review has sentiment words that express the orientation of the product features. In this research, ontology and synonym eased this new model to recognise a product feature than without ontology and synonym.

The technique of data warehouse, especially ETL (Extraction, Transform, Load) was the main process in this research project. The creation of many data cubes presents more reports, especially based on customer’s opinion, product and group of customer (OGC) as this formula is our new formula to calculate the polarity of product’s sentiment based on the product’s review by customers. This OGC formula gives opportunity to all customers and companies about the orientation of every product feature based on product ontology level.
Chapter 5

Algorithms of Opinion Integration Model

5.1 Introduction

Analysis on customer’s character and behaviour are the key factors for companies to achieve their business target. Thus, companies began to realize the importance of investigating and understanding customers’ characteristics, behaviours, and needs. The success in analysing the need of a customer will enhance a company’s revenue and profit. In doing this, companies must shift their focus from only product-based to customer and product-based with the support from Customer Relationship Management (CRM). CRM is basically a business activity between a company and its customer to enhance a good relationship between them.

The key of this research activity is the analysis of customers comments (unstructured data) regarding a particular product and customers background (structured data) from CRM database, which enables companies to understand what customers need, identify valuable customers, predict future customer activities or behaviours, and most importantly, to make a proactive decision on the product that customers are interested in. Therefore, companies will not only attract new customers but also successful in maintaining loyal customers to do business transaction with them.

In performing customer analysis, opinion mining becomes more common and popular than ever before. Opinion mining (OM), or also identified as sentiment analysis, sentiment classification, and opinion extraction in the literature is a key point in information retrieval and computational linguistics, which is linked with the opinion expressed in a document. The key issue in our research is integrating customers’ opinion into the customer analysis model.
We produced a new model for Opinion Mining using a multidimensional model to integrate customers’ opinions into the CRM systems. Usually data are stored in two formats: structured and unstructured. Structured data from CRM databases are stored in rows and columns; in contrast, data from customer’s comments are unstructured data. Most web contents such as web pages, text comments, emails, images, videos, music, and presentations are unstructured data.

This research involved two main parts: To extract and integrate structured and unstructured data from customers’ comments; and to produce report on polarity of product based on customers’ opinionated comments. We used Microsoft SQL Server 2008, Microsoft Visual Studio 2008 and Microsoft SQL Server 2008 Report Builder 2.0 to perform the integration process. Besides that, JAVA programming and Part-of Speech (POS) tagging were also applied in the data mining process. This research used multidimensional data warehouse and some method in Extract, Transform, Load (ETL) process for customer information and comments extraction, transformation and loading. This model transfers comments from traditional text files and load them into a multidimensional data warehouse, which involves fact table and other dimensions (customer, products, time and location). The scope of this research involved the design of multidimensional models, data transformation and loading, and data report building executed in data warehouse, as well as customer comments extraction (which is also recognized as opinion mining) using Java programming. In other words, this project involved cleaning of data, data warehouse design, data transfer and implementation, data analysis, and evaluation.

5.2 Customer Relationship Management (CRM)

CRM is important to handle transactions with customers. All records about product, supplier, and customer are saved in CRM’s database. Every company must use this record to analyse what their customers really need from them. Online transaction keeps a lot of valuable information regarding customers and the products that they really need. Most websites also give a lot of information regarding the product, however, sometimes it does not give the right information that the customers look for.

Business websites usually have the comment section, which would give the opportunity for customers who already used the particular product to make some notes regarding the product.
Figure 5.1: Ontology of Mobile Phone
As mentioned in the earlier section, 80 percent of data are unstructured data in which the CRM case refers to as opinionated comments from customers. This shows how customer comments can influence other customers in making a decision to purchase a product. To analyze opinionated comments from customers, it is crucial for CRM to understand what customers need and to forecast which product can be more profitable than other products. Normally, customers comments are diverse, for example, regarding the products in general, or very specific features or technical parts, or sometimes comparison with other products with the same functions or features. Sometimes, customers describe their opinions by showing the strength of their comments by putting scales such as excellent, very bad, or just okay/OK. Customers’ expression in commenting a particular product will give new ideas or information to other customers who are looking to buy a new product or service. The strength of comments from reviewers of a product will influence the decision of new customers to buy or not to buy any particular product. It is easy for customers to find a product, based on the features that they are eager to use such as application and connectivity in a mobile phone. It is not an easy task to integrate the comments from customers (unstructured data) with structured data such as product, customer, and place in the Customer Relationship Management (CRM) system. Our approach has been developed not only to cater this problem but also to summarize every feature of the product as positive, negative, or neutral based on comments by customers.

Customer analysis is one of the most important aspects that companies and organizations need to emphasize on, especially in finding a new customer group and retaining the loyalty of existing customers. In this research, this new model produced an ontology for each product that we worked with such as mobile phone, as shown in Figure F.1. The idea to produce a new ontology was because every ontology term in every tier comes with a synonym file for every level of ontology to capture more product features from the opinionated comments. One of the difficult parts in extracting sentences from product comments was to capture the right feature. Customers usually refer to the same feature using different terms. For example, "SMS" is also referred to as "text", "message", and "msg". We divided the features of a product depending on their subcategories (characteristics) such as general, and other technical terms. The idea was that customers usually comments on products based on technical features such as size, battery, color, and applications that are built in the product. These subcategories were our main attributes in producing the comment fact table.
One business model was produced to guide our creation of a better multidimensional database. Based on Figure 4.5 in previous chapter, we used four dimensions representing product, customer, time and opinion. These four dimensions were used to find the best solution to represent opinions in the next phase of our research, customer analysis. The comments from customers were different based on their backgrounds such as age and gender.

5.3 Model Development and Implementation

The main task in this research model was the extraction of useful comments elements from data source, then to transform and load them into the data warehouse. Data cubes were developed for analysis of comments and the results of these analyses were shown in a report. As a result, five steps of development process were done as shown in Figure 5.2, where the idea about the whole model/processes that used to integrate the customers’ comments into CRM is presented.

As shown in this figure, the research had five major steps in implementing the model. These processes are:

- **Extraction Method**
  Firstly, all data (comments) sources are inserted into the model. The extracted processes of comment file including extracted noun and sentiment word happen in this process. This process is one of the most important parts in this research. The extraction of customer comment involves a lot of techniques such as pruning, and frequent nouns.

- **Transformation Method**
  This process involves extracting data from previous processes to this process to match the extracted nouns and sentiment words to feature and polarity pair and insert them into a fact comment table with related dimension key. The combination of product ontology and synonyms is a major part in this process to capture the product and it’s features.

- **Data Warehouse**
  All successful data process are saved in this data warehouse for Customer Analysis process.
Figure 5.2: The Model for our Sentiment Analysis System
• Building Cube

Data cubes are created to perform data analyses based on customer comments.

• Report

All of research results are presented in this process based on the cube’s dimensions.

5.3.1 Extraction Method

In general, this research project involved data source (comments) as the main data input, which was extracted into nouns and sentiment words using the ETL (Extraction, Transformation, and Loading) process. The ETL process is the main part of this system. The diagram for the ETL process is shown in Figure 5.3. The extraction and transformation methods were discussed in the previous section, meanwhile the loading method or process includes the data warehouse, data cubes creation and report presentation.

Figure 5.4 shows the data source used in this research. This research used comment files in .txt format, which includes the introduction and the comments section from Bing Liu and Minqing Hu’s works (Hu and Liu [2004]). Data extraction aimed to use some methods to extract useful feature and polarity as well as data from certain tables such as product, customer and date from raw comments.

The data extraction part aimed to use some methods to extract useful features and polarity from raw comments as well as product, customer and date information from other CRM files. This part involved five main steps in the research development processes as shown in Figure 5.5:

Based on Figure 5.5, the details on the extraction processes are:

• Pre-processing Data Source

Based on Figure 5.5, data pre-processing involved three earlier processes which are addCustomer, RawFile and Pre-process. Data pre-process was done in this part to delete useful information from the input file such as document introduction and special characters, for instance the symbol $t$. Meanwhile, some information from CRM files such as product and customer were added in this early process.
Figure 5.3: The Diagram of ETL
5.3. MODEL DEVELOPMENT AND IMPLEMENTATION

- **Removal of Useful Data from Customer and Product Files**
  The second process was removing unnecessary data such as product ID, customer ID and date from the input file into another file in the CRM system. After this process, the data would be in a text file format (.txt) and ready for POS tag’s process.

- **Part of Speech Technique**
  The part of speech tagging (POS) was used to pre-process the data in the source file. It processes the customers reviews (data source) in text form and produces another text file, which has data such as product name, customer information, date, and pairs of entity descriptor based on our business model, involving main tables or entities such as product, customer, opinion, and time. Thus, the extraction process produced structured data format. The most important words in this research were noun, adverb, and adjective. Examples of outcome after POS phase is 1/CD, ./good/JJscreen/NN, ./excellent/JJservice/NN./.

- **Frequent Nouns Detection**
  Noun is an attribute that can describe the feature of a product. Adverb and adjective can express the opinion of the customers regarding a feature. In the pre-processing step, all words were considered as stop-words except those three type of words (noun, adverb and adjective). The most important process in this part was to get the frequent nouns because all entities or features are noun. Based on the normalized $tf \times idf$ in Lau et al. [2009a], to compute noun frequency ($tf$) and document frequency ($df$), we defined the support as follows:

---

**Figure 5.4:** The Data source
CHAPTER 5. IMPLEMENTATION

Figure 5.5: Detail Process of Extraction
where $tf(n)$ is the number of appearances of $n$ in a sentence and $tn$ is the total number of noun appearing in the sentence.

Most extracted nouns were not related to the product or features based on observation. Therefore, to reduce the number of extracted features from the text, a frequent noun was selected using maximum and minimum thresholds to reduce noisy data. A frequent noun can be defined as follows:

$$
\text{min}_\text{sup} \leq sup(n) \leq \text{max}_\text{sup}
$$

where $df(n)/tf(n) < tr/tn$, with $df(n)$ is the document review, $tr$ is the total sentence review, and $tn$ is the total noun appearing in the sentence. This concept of $df(n)/tf(n) < tr/tn$ was needed in this research to make sure that the total sentence reviewed by the system was more than the document because every single document reviewed contains at least 10 sentences of dataset.

The minimum support $\text{min}_\text{sup}$ and maximum support $\text{max}_\text{sup}$ are coefficient. $\text{min}_\text{sup}$ is a parameter and the default value is 1 % of total nouns appearances. Meanwhile, $\text{max}_\text{sup}$ gives the upper limit of frequent nouns and it is also a parameter with a default value of 10 % of total nouns appearance. For example, if there are 1000 total appearances of nouns, if $n$ appearance is 11 appeared in 5 reviews, then $sup(n) = 11/1000$, $1% \leq sup(n) \leq 10%$, $n$ is a frequent noun because $df(n)/tf(n)=5/11$, where $df(n)$ is the number of reviews that include $n$. These minimum support and maximum support are important in this research for capturing only accurate noun as a feature and for the system to remove unnecessary noun from the dataset.

Every noun with the closest adjective and adverb in term of distance with a product feature was saved in the output file, together with the structured data from the CRM because these types of words such as good, worst and excellent represent the sentiment of comments.

- Noun Distance
In this process, the research model searched every noun closed to the adjective or the adverb using useful sentence element with distance and put it into an output file, while linking the product, customer information, and date as the input of transformation. Figure 5.6 shows the process of the research model searching the noun and adjective or adverb in comment sentences. In this research, POS tag function divided every single word into its own group, such as *NN* as noun, *JJ* as adjective and *RB* as adverb. In searching for a pair of noun and adjective or adverb, this research model detected the noun first and put it in a temporary variable file until it found the nearest adjective or adverb based on *POS* tag symbol. This process was repeated until no noun was detected in the comment.

**Figure 5.6:** Process of searching noun, adjective and adverb

### 5.3.2 Transformation Method

Figure 5.7 details the transform process in this research method. The pre-processed data from the extraction part was the main input for the transformation part. In the transformation part, the extracted nouns were matched with the feature ontology (*FO*) to understand the features that
Figure 5.7: Detail of Transform Process
customers intended to review. Consequently, the polarity of extracted sentiment words would be calculated. A pair of feature and polarity was inserted into the fact comment table with the related dimension key. The matching process of the output of this process was then loaded to the data warehouse. The key points for this process are feature matching and opinion matching.

A record of comment, which includes product, customer information, date, pairs of entity descriptor (entity is a noun, and a descriptor is an adjective or adverb) was captured by the system and transferred into the database. Besides that, feature ontology and its synonym, and sentiment word and its synonym were also part of the system’s database.

Opinion matching data maintenance includes sentiment data maintenance and sentiment word list maintenance. The SentimentData.csv data was inserted in the sentiment data table. The adjective and adverb detected in the previous process were stored in this table. As mentioned by Liu et al. [2005], to obtain a collection of sentiment words, there is a need for continuous works in capturing new sentiment word, either positive or negative word. Besides storing sentiment word from the extraction process, this research project also stored the sentiment word list from Liu et al. [2005] because they had a large data of sentiment word collected from 2005 until now. One idea to enrich the sentiment data is using the synonyms for every word in the sentiment data table. This sentiment data’s synonym is currently maintained manually by finding huge linguistic datasets (synonyms) or using sentimentdata.csv generated by Java based on huge linguistic dataset in the future. This sentiment word’s synonym process is important to make sure that any new word used by a customer can be matched with the current sentiment word in the database if it has the same meaning. This combination of sentiment words then gave the polarity in the table based on sentiwordnet’s website.

**Feature Ontology**

In more specific terms, we assumed that we have feature ontology for all products. In this task, we denoted $A$ as a set of attributes of all products for an enterprise, and $g$ as a group of customers. Based on these two definitions, we formally defined the meaning of feature ontology and opinion sentences (Yaakub et al. [2011]) as follows:

**Definition 1** Feature Ontology ($FO$): A Feature Ontology $FO = \langle A, R \rangle$, is the combination of a set of attributes $A$ and their relations, where $(a_1, a_2) \in R$, if attribute $a_1$ is a more general concept than $a_2$, or say $a_2$ is more specific than $a_1$. This concept can be described
clearly by example of location’s field in real life. The entity of a location is too general, so we can divide it into more detailed using more than one attributes such as through the attributes of number, street, city, province or state, zip code, and country.

The typical rule generation algorithm is the Apriori rule generation algorithm. It generates association rules for every frequent itemset (also called pattern) $X$ discovered in the frequent pattern mining process. At first, it partitions a frequent itemset $X$ into two parts $A$ and $B$, where $A \cup B = X$ and $A \cap B = \emptyset$, obtaining an association rule if $A$ then $B$. Then it calculates the confidence value of this rule and keeps it if the confidence value is larger than a subjectively specified minimum confidence threshold. This process iterates until all possible partitions for all frequent itemsets are produced.

Association rules were used to produce the level of feature ontology based on the concept of if and then in order to find the relationship between a group of products. if was used to show some states of items based on some conditions and then was used to refer to another item, which has different term from the previous item, but has some similarities in condition. These association rules are important to predict another feature in our feature ontology. An association rule has two parts, which are if and then. In our research work, if is a feature found in the dataset, and then is a feature that is found in combination with the if. For example, the rule with if phone and nokia then picture appeared more than 3 times in our system, then system detected picture as feature for the group of product for phone and Nokia.

Based on feature ontology, we defined the attributes in four levels: street $<$ city $<$ province or state $<$ country. A lattice was formed because the attributes of a dimension were organized hierarchically, adopting the time dimension: day $<$ month $<$ year.

**Definition 2 Opinion Sentence (OS):** Based on sentiment-feature pairs, opinion sentence is defined as:

$$OS = \{(f_1, s_1), \ldots, (f_m, s_m)\}$$

where, $f$ and $s$ represent a set of feature-sentiment pairs.
5.3.3 Opinion Extraction

To understand the features that customers reviewed, each frequent noun was mapped with feature ontology. Normally, there were four cases to map a feature $f$ to a product attribute $a$. Those four cases are: exact match attribute $a$, partial match attribute (synonym) $p$, association rules $b_1$ and no match $b_2$, in which the system ignores it for feature $f$. These definitions are as follows:

\begin{align*}
    a, & \quad \text{if } \exists a \in A, a = f \\
    p, & \quad \text{if } \forall a \in A, f \in \text{synonym}(a) \\
    b_1, & \quad \text{if } \exists a \in A, f \rightarrow a \text{ is an association rule} \\
    b_2, & \quad \text{otherwise}
\end{align*}

Therefore, the function to map a feature $f$ to a product attribute $a$ is presented as follows:

\[
f_a(f) = \begin{cases} 
    f; & \text{for case } a \text{ and } p \\
    b_1; & \text{if } f \rightarrow a \text{ is an association rule} \\
    b_2; & \text{otherwise}
\end{cases} \tag{5.3}
\]

To explain the equation, in the first case the attribute $a$ appears in feature ontology $FO$. In this case we easily mapped the attribute to the feature and understood the feature that the customer wants to review. In the second case, where an attribute $a$ was not found in feature ontology, then we obtained all possible synonyms for attribute $a$ and mapped the synonym to feature ontology. If none of the first two cases worked then an association mining between attribute $a$ or the attribute synonym with other attributes in the review would be used. Based on the relation between attribute $a$ and other attributes in the review, attribute $a$ could be mapped to feature ontology. If there was no useful relation found, then attribute $a$ was ignored.

The algorithm $GetFeatures()$ as shown in Algorithm 1, describes how the process of extracting comments in the model captured the frequent nouns to represent the features and attributes of a product. The process of entity matching was done based on feature ontology.
Algorithm 1: GetFeatures()

Input : $EC$ - a list of extracted nouns from comments, $FO$ - Feature Ontology; $FS$ - Feature Synonyms.

Output: A list of Frequent Nouns $FN$.

1. Let $FN=\emptyset$;
2. for each $e \in EC$ do
3. \hspace{1em} $fk = NULL$;
4. \hspace{2em} if $e \in FO$ then
5. \hspace{3em} Let $fk$ be $e$'s feature key in the product dimension table;
6. \hspace{2em} else
7. \hspace{3em} if $e \in FS$ then
8. \hspace{4em} Let $e_1 \in FO$ be the synonym feature of $e$;
9. \hspace{4em} Let $fk$ be $e_1$'s feature key in the product dimension table;
10. \hspace{3em} else
11. \hspace{4em} end if
12. \hspace{3em} end if
13. \hspace{2em} end if
14. \hspace{2em} if all associated noun $e_1$ of $e$ do
15. \hspace{3em} if $e_1 \in FO$ then
16. \hspace{4em} $fk$ = Feature key in the product dimension table;
17. \hspace{4em} return;
18. \hspace{3em} else
19. \hspace{4em} if $e_1 \in FS$ then
20. \hspace{5em} $fk$ = Feature key by $fs$ relationship with $FO$ dimension table;
21. \hspace{5em} else
22. \hspace{6em} Let $fk = NULL$;
23. \hspace{4em} end if
24. \hspace{3em} end if
25. \hspace{2em} end
16. end if
17. end for
18. If $fk \neq NULL$ then
19. $FN = FN \cup \{fk\}$;
Table 5.1: Seven Level Polarity System

<table>
<thead>
<tr>
<th>Polarity</th>
<th>Terms</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Distinguish</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Accept</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>Neutral</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>Reject</td>
<td>1</td>
</tr>
<tr>
<td>-2</td>
<td>Poor</td>
<td>0</td>
</tr>
<tr>
<td>-3</td>
<td>Worst</td>
<td>0</td>
</tr>
</tbody>
</table>

*FO*, feature synonyms *fs*, noun mining, and associated nouns. Feature ontology is based on our four levels ontology as shown in Figure F.1, where the levels are divided into product, features and attributes. Feature synonyms are a selection of different words with the same meanings as the main features based on the dictionary provided by thesaurus.com. The last technique that we used was noun mining, in which we remove all terms from reviews except the nouns. Each potential association rule, which includes the support, confidence, input noun support and also predicted noun were acquired to mine in the mining model table.

The output of mapping attributes for feature ontology is a list of features. To get the opinion of the customer about a specific feature, each feature was associated with the closest sentiment words. Then the polarity for each sentiment word was calculated.

5.3.4 Attribute Polarity

One of the model tasks is to capture the attributes, as well to evaluate the polarity of the customers’ comments. All attributes were captured using the feature ontology and its synonyms. This mapping is essential because most customers always give their comments for specific features. It is difficult to implement in the real world without putting features with the same characteristics in one group. Attribute Polarity is a list of product’s attributes with customers’ comments including their strength of polarity.

Table 5.1 shows the seven level polarity system. The polarity for our system ranged between 3 and -3 (3 as the strongest positive and -3 as the worst negative comment). These words were used by customers to express their opinions in the comment. Type 0 means that the word is already in the database repository in our model and type 1 means that the model detects the new
5.4. OPINION INTEGRATION

<table>
<thead>
<tr>
<th>Comment</th>
<th>Attribute Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(general,3)</td>
</tr>
<tr>
<td>2</td>
<td>(general,1),(general,-3),(application,-3)</td>
</tr>
<tr>
<td>3</td>
<td>(general,2), (general,2)</td>
</tr>
<tr>
<td>4</td>
<td>(general,-1),(accessories,1)</td>
</tr>
</tbody>
</table>

Table 5.2: Attribute Polarity

Word and saves it in the database. The output of the attribute polarity is shown in Table 5.2. Based on customer comments, the output of attribute polarity is a pair of product feature term from product ontology and the sentiment word’s polarity. Most of polarity system used by other researchers used positive scales such as *fare*, *good*, *excellent*, *verygood*, and *outstanding* in showing the sentiment orientation. This is different to our research, as we used new score of polarity based on Table 5.1. This calculation is important to use in our OGC model. Details about polarity calculation and OGC is explained in subsection 5.4 Opinion Integration.

The algorithm `AttributePolarity()` as shown in Algorithm 2 describes the process of transferring an opinion sentence into a list of attribute polarity pair. In this algorithm, we used sentiwordnet calculation as our reference to scale any sentiment $s$ by a number between -3 to 3 based on the seven level polarity system.

### Algorithm 2: AttributePolarity()

**Input**: OS - an opinion sentence  
**Output**: A list of Attribute Polarity Pairs $AP$.

1. Let $AP = \emptyset$;
2. **for each** $(f, s) \in OS$ **do**
3.   Let $m$ be the scale of sentiment $s$;
4.   $a = f_a(f)$;
5.   **if** $\exists x$ such that $(a, x) \in AP$ **then**
6.     $AP = (AP - \{(a, x)\}) \cup \{(a, \max(x, m))\}$;
7. **else**
8.     $AP = AP \cup \{(a, m)\}$;

5.4 Opinion Integration

In this section, we discuss how to integrate customers opinions with an existing CRM database. We first propose a data warehouse based design to make linkage between product, customers
and their opinions. We also present equations for calculate orientations of a group customers to a certain product or its category. At last, we discuss how to report what users want.

5.4.1 Data Warehouse Based Design

Figure 5.8 shows the main dimension used in this research development. FactComment is the centre for other dimensions. It contains the ID of all dimension such as CustomerID and ProductID. In FactComment we developed the data cubes for OGC report. All of these dimensions were from the CRM system, except the comment dimension.

Data cube is one of the main parts in data warehouse as it is considered as a generalization of three-dimensional spreadsheet. In our research, data cube was the cornerstone of our main report as we used it to produce a lot of reports from the dataset. Some companies used data cube to summarize their businesses based on financial data either by product, geography or location, and time-period. In our case, we used product, time, location and comments as the main dimensions of our data. Figure 4.2 shows the data cube for comments from our research work. Numerous dimensions were used to calculate the polarity of product based on the fact dimension of factcomment, which has the formula to calculate OGC. Based on OGC’s formula, we produced a report that shows the orientation of opinionated comments for every single feature and attributes of a particular product.

5.4.2 Orientation

Based on the data cubes, we worked out the orientation based on product features for a given group of customers, \( g \) to a category of product, \( c \) as follows (Yaakub et al. [2011]). The data for group of customer and product are selected from microsoft access’ repository to represent the customer and product of CRM.

\[
o(g, c) = \begin{cases} 
  \text{positive, if } ogc(g, c) > 0 \\
  \text{negative, if } ogc(g, c) < 0 \\
  \text{neutral, otherwise}
\end{cases}
\] (5.4)
5.4. OPINION INTEGRATION

Figure 5.8: The Main Dimension in Data Warehouse
This equation was used to find the orientation of customers comments in general, without using ontology level. It only calculates the polarity of the whole sentence. The \( ogc \) is a formula to calculate product features orientation based on set of customer, product and users opinion on a particular product feature(s). The \( ogc \) is referred to calculation of product feature based on product’s ontology level and sentiment word, meanwhile \( o(g, c) \) is orientation of product sentiment based on customer comments. The orientation is positive when \( ogc(g, c) \) for the group of customer and product is greater than 0, or negative when \( ogc(g, c) \) is less than 0. Neutral will be set if \( ogc(g, c) \) is equal to 0. The detailed calculation of \( ogc(g, c) \) is as follows.

\[
ogc(g, c) = \sum_{l=-3}^{3} (l \ast \text{polarity}(c, g, z))
\]

(5.5)

where, polarity is the total support of the comments for the sentiment value \( z \), based on the group of customers \( g \), and the feature or category of product \( c \) from the CRM database from our system. While \( l \) refers to level of polarity from ontology, in this equation it only refers to level one of ontology.

This equation only refers to general comments by customer about a product without using detailed product feature. As mentioned earlier, this equation works for level one ontology based on \( l \). For example, this calculation only refers to a general product such as camera and mobile phone, not the details about the product feature. For example, if user gave a comment regarding the phone(category of product) without any feature with sentiment word excellent, the calculation is 1 from level of ontology and times with 3 from polarity system, then the OGC for this calculation is 3 (positive).

This formula will produce inaccurate result of \( ogc \) because \( c \) only represents the general ontology, which is level one. Most customers’ comments include the general aspect of a product and also the specific features of the product in detail. This means that the reviews do not just mention the general features such as mobile phone. For example, a user likes the WiFi of a mobile phone very much, but we cannot simply say that the user likes the communications very much. We solved this problem by enhancing \( c \) in the formula. Every level of ontology was calculated differently with the review count depending on the level of the features in the feature ontology.
The polarity of product $c$ for a group of users $g$ at $z$ scale can be calculated as follows:

$$polarity(c, g, z) = \frac{\text{total support to } c \text{ at } z}{2^{p-n}}$$

where this equation refers to the polarity of a product in specific either positive, negative or neutral. In this equation, $n$ is the level number of $c$ in the feature ontology. For example, if the feature is in level two of ontology, so $n$ is 2 and $p$ is the length of the feature ontology, which, in this research the length of feature ontology is four.

5.4.3 Report

Based on the data cubes we produced numerous reports depending on the company’s need in SQL Server Reporting Service (SSRS). Currently, there are more than 20 reports in this research project, based on the data cubes creation. This research model has the ability to produce a product report based on customers opinionated comments in more than one level such as polarity in general, and also detail the polarity for every feature and attribute of product.

Figure 6.7 shows a report of polarity and orientation of product from opinionated comments on a particular product in general. The polarity of sentiment for this particular mobile phone was positive in general, but this figure also shows different orientations for different product attributes. Product attributes such as Accessories, Others and General were positive, but connectivity was negative. Meanwhile, three others attributes were neutral in term of customer’s orientation based on the OGC formula.

Meanwhile, Figure 5.10 shows the detail on connectivity attributes. Even though, the polarity of orientation for connectivity (a product attribute) was negative, not all of the comments received from customers were negative comments in all connectivity’s features. The feature of WIFI, for example, had positive comments and other features such as GPRS and Infrared were negative, while Bluetooth was neutral. Only two features commented by customers were negative, which were EDGE and 3G. This result shows that the polarity of sentiment words influenced the orientation of comments. Every single sentiment word in this process was counted, which means that higher polarity level used by customer in their comments will influence the result.

These reports showed that this research project has produced a new dimension in the CRM
## CHAPTER 5. IMPLEMENTATION

**Figure 5.9**: Summary of Polarity and Orientation based on Product Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Index</th>
<th>Total</th>
<th>Other</th>
<th>General</th>
<th>Communication</th>
<th>Accessibility</th>
<th>Total OCG Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>222 Polarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>222 Polarity</td>
<td>600</td>
<td>62</td>
<td>22</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>31 Polarity</td>
<td>0</td>
<td>22</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>33 Polarity</td>
<td>100</td>
<td>6</td>
<td>22</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>53 Polarity</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>64 Polarity</td>
<td>32</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>65 Polarity</td>
<td>75</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table represents the summary of polarity and orientation based on product features with specific data entries for each entry.*
## 5.4 OPINION INTEGRATION

### Figure 5.10: Details of Polarity and Orientation based on Connectivity Attributes

<table>
<thead>
<tr>
<th>Connectivity</th>
<th>Poor (-3)</th>
<th>Reject (-2)</th>
<th>Weak Reject (-1)</th>
<th>Neutral (0)</th>
<th>Accept (1)</th>
<th>Distinguish (2)</th>
<th>Excellent (3)</th>
<th>Total</th>
<th>OGC</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3G</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>-7</td>
<td>Negative</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>Neutral</td>
</tr>
<tr>
<td>EDGE</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>Positive</td>
</tr>
<tr>
<td>GPRS</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>-14</td>
<td>-14 Negative</td>
<td></td>
</tr>
<tr>
<td>Infrared</td>
<td>0</td>
<td>14</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>33</td>
<td>-37</td>
<td>-37 Negative</td>
<td></td>
</tr>
<tr>
<td>WIFI</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Positive</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>15</td>
<td>46</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>4</td>
<td>80</td>
<td>-52 Negative</td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>7</td>
<td>30</td>
<td>3</td>
<td>25</td>
<td>32</td>
<td>9</td>
<td>109</td>
<td>59</td>
<td>59 Positive</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>10</td>
<td>13</td>
<td>40</td>
<td>0</td>
<td>32</td>
<td>14</td>
<td>23</td>
<td>132</td>
<td>33 Positive</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>22</td>
<td>6</td>
<td>62</td>
<td>31 Positive</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>64</td>
<td>144</td>
<td>10</td>
<td>144</td>
<td>120</td>
<td>88</td>
<td>604</td>
<td>222 Positive</td>
<td></td>
</tr>
</tbody>
</table>

Mobile Phone

Accessories
Application
Communication
Connectivity
Entertainment
General
Other
Total

---
system with the ability to present the polarity and orientation of each product attribute and feature. This new model in CRM system gives more choices for a company to perform customer analysis. The current CRM system only uses structured data from their data warehouse to forecast future business strategy. But with this new model, companies have two types of data to analyse as they can also use the opinions of customers about a particular product. As mentioned earlier in previous chapters, data mining result is more relevant if more data used for its analysis process. Meanwhile, customers also benefit from this model as they can obtain new information about a particular product from other customers who have experienced the particular product.

5.5 Summary

In this chapter, we divided our work into two parts, which the first part extracted and integrated structured (CRM system) and unstructured data from customers’ comments; and second part produced report on the polarity of a product based on customers’ opinionated comments. In developing our model, we used Microsoft SQL Server 2008, Microsoft Visual Studio 2008 and Microsoft SQL Server 2008 Report Builder 2.0 to perform the integration process. Besides that, JAVA programming and Part-of Speech (POS) tagging were also applied in the data mining process.

This research used multidimensional data warehouse and some methods in ETL process for customer information and comments extraction, transformation and loading. This model transfers comments from the traditional text file, and then load them into a multidimensional data warehouse, which involves fact table and other dimensions (customer, products, time and location). The scope of this research involved designing of multidimensional models, data transformation and loading, and data report building, which were executed in data warehouse, as well as performed customer comments extraction (which is also recognized as opinion mining) using Java programming. In other word, this project involved data collection, data warehouse design, data transfer and implementation, data analysis, and evaluation.

This research project expressed the importance of ontology and synonym in the extraction process as these two techniques can remove duplicated data and detect product features with good result. The report on the orientation of every single feature of a product is a new contribution in the CRM system. Based on the author’s knowledge, there is no previous research with
similar research finding as this research project.
Chapter 6

Evaluation and Discussion

6.1 Introduction

This research model focused on finding the sentiment from customers comments based on product’s features and summarizing product features based on customers’ opinion orientation. This research used ontology and synonym to define the features (product’s attributes) of products. This system was built in JAVA and used Microsoft SQL Server 2008 R2 for developing data warehouse.

For the evaluation, all reviews were read and evaluated by human beings. A comprehensive list of features in the reviews was extracted. Similarly, customers opinion were extracted and identified as positive or negative based on the provided scales. In this paper, we were only interested in sentences with opinions. The main objectives for our evaluation were divided into two perspectives:

- Identification of Opinion Orientation
  This evaluation was calculated to find Recall, Precision and also F Score

- Identification of Product’s Orientation based on Customer Comments
  The ability of the model to give a conclusion from customers’ comments about the opinion orientation of every product features and also a summary of the product itself. We performed manual evaluation because it was hard to find any baseline model for this part.
We divided the evaluation into two parts based on the research objectives to find the recall and precision, F Score, and the report on customer’s opinion orientation for the particular product. The extraction of product feature was the first phase of this research work. First, we developed the ontology of products based on four layers ontology from the general definition to the product’s attribute instance based on the theory mentioned in previous chapter. This ontology for all products used, which frequent words are selected as features for all 4 tiers of ontology. Meanwhile, infrequent words that appeared in system more than five times and matched with the features synonym also included as product feature. Then, we used the synonyms for product features to get a more accurate result because comments from users vary and unpredictable in term of the terms, which they used. We used Recall and Precision as measurements for this phase. Five models from Minqin Hu and Bing Liu work were used as the baseline models (Hu and Liu [2004]) and also work from Su Su Htay and Khin Thidar Lynn (Htay and Lynn [2013]).

For the extraction of product feature pairs and opinion’s orientation, we used more than 1000 random data items from five groups of dataset from Minqin Hu and Bing Liu’s work as training data to mine the patterns. We used ontology and synonym for product features. For opinion orientation, we used the lexicon of orientation from Minqin Hu and Bing Liu’s project (Hu and Liu [2004]). We improved the lexicon by using synonyms for every term in the database. We also used the seven polarity system for every word in the lexicon based on the wordnet system. We did this to distinguish the difference between positive and negative words. Accuracy was used to compare our findings in this phase with the baseline models. The reason for this task was to find any pattern between product features and its sentiment word.

The last task in this research was to find the conclusion in report form from customers’ comments about the opinion of every product feature and also the summary of the product in general. This part was developed using the results in the extraction of pairs of product feature and opinion’s orientation. Even though the results were good, it was not 100 percent accurate. However, the originality of the result was not affected because there was only less than one percent error in every dataset. The report on product’s orientation from customers was based on the equation of OGC discussed in the previous chapter.
6.1.1 Testing Environment

Figure 6.1 shows the entire environment for our testing development.

![Diagram of testing environment]

**Figure 6.1**: Our Testing Environment

The testing environment was divided into three layers, which are:

- **System Layer**: Experimental Platforms /Infrastructure component
  
  This basic experimental system was developed in *MS Windows* 7 and *JAVA* 1.6. *JAVA* was the main programming language used in the entire system development, and without this programming language, the extraction of product’s comments or reviews cannot be done properly, especially in testing the model with multiple techniques such as *POS*, association mining, and detection of distance (noun-adverb or noun-adjective).

- **Application Layer**: Application System /Technical components/Data
  
  In the application layer, this research model used *Microsoft SQL Server* 2008 as the main tool for developing data warehouse. All essential tables from the *CRM* system and customer’s review were stored in data warehouse, meanwhile *Microsoft Visual Studio* 2008 was the tool for coding *SQL* and the main tool for *ETL* processes development.
Microsoft SQL Server 2008 Report Builder 2.0 was the tool used for presenting all reports in this research based on OGC formula. For JAVA development tool, this research development used Eclipse 3.6.

- Process Layer: Datawarehouse/ETL/SSRS Report

This layer involves three main processes, which are design of multidimensional model, implementation of ETL processes and development of research reports.

### 6.1.2 Dataset

The data is collected from the Minqin Hu and Bing Liu’s project (Hu and Liu [2004]). This dataset is chosen because it is popularly used among sentiment analysis’ researchers. Based on literature review, we chose this database because it provides a judgment for relevant features. Besides that, Minqin Hu and Bing Liu always update their lexicon for positive and negative sentiment from 2004 until now. This dataset is also currently used by other researchers such as Htay and Lynn (Htay and Lynn [2013]), Khan, Baharudin and Khan (Khan et al. [2014]) and Bagheri, Saraee and Jong (Bagheri et al. [2013]). So far, they have provided up to five different datasets of product reviews from the customers, which are the set of mobile phone, the set of DVD player, the set of MP3 Player, and two set of camera digital from website of amazon.com. We chose these data because it has a large data pool regarding the customer comments on the particular product with more than three different techniques. This data comes with the tagging based on Minqin Hu and Bing Liu’s work. As mentioned earlier, we deleted the tagging first to get the raw data from this data collection. Then we processed again the needed raw data to produce the new tagging with the manually correction for the small amount of wrong tagging by the system. Bing Liu, Minqin Hu, and Junsheng Cheng (Liu et al. [2005]) also mentioned, the automatic tagging is not 100 percent correct. An ontology based on the features and attributes of the products was used before we used our own tagging process (Please see all ontologies in the appendix C-F). The main objective of using ontology is to make sure that all the product features can be easily detected by the research model. The feature ontology’s synonym is also used to make sure that simple word such as msg can be detected because it also means message (Please see the synonym list in Appendix). The pruning process, verb matching and opinion lexicon are the main techniques used in this research work. Opinion lexicon used in our model are based on worked by Bing Liu and Minqin Hu (Hu and Liu [2004]), which already divided
sentiment words into two groups as positive words and negative words.

6.1.3 Baseline Models

We use two baseline systems. They are the system developed by Minqin Hu and Bing Liu (Hu and Liu [2004]) and the system developed by Su Su Htay and Khin Thidar Lynn (Htay and Lynn [2013]). Minqin and Bing’s technique started with crawling the reviews, and then put it in the review database. By using the NLProcessor, the Part of Speech (POS) tags are produced to isolate the reviews into the categories. They identified the features by finding the frequent nouns or noun phrases using the association mining. This is the major difference between their method and our method because we use the ontology to define the features of product(s). After that, Minqin Hu and Bing Liu use adjective words to find the infrequent features. Then, they identified the opinion sentence based on the adjective word close to the feature word, before producing the summary of review. The following are the five different approaches tested by Hu and Liu:

- Opinion Sentence Extraction
  Minqin Hu and Bing Liu (Hu and Liu [2004]) defined opinion sentence as a sentence that has at least one product features and at same time has one or more opinion words. From the customer’s product review, a frequent feature the nearby adjective recorded as opinion words.

- Frequent Feature
  The frequent feature identification used association mining technique for that particular purpose. Frequent feature in this part is based on itemsets, where itemsets is referred as product feature. In this case, a set of words or phrase that appear many time together known as frequent itemset, which appear more than one percent (minimum support) of the total review sentence. The reason why association mining was used because the frequent itemset is most probably the product feature.

- Compactness Pruning
  This method main function is to check the useful of feature phrases among features that
appear more than one time. Those features that appear meaningless is removed from the
system.

- **P-support Pruning**
  
  The P-support is used to remove the duplicate features. In p-support, all of the sentences
cannot have any feature phrase if the noun or noun phrase already has the same feature.

- **Infrequent Feature**
  
  The infrequent feature defined as a noun that has opinion word near to it but this noun
itself is not a frequent noun. Hu and Liu [2004] acknowledged that this infrequent noun
is always misunderstood as irrelevant to the product’s items and they ignored most of
infrequent feature as they mentioned that only around 15-20 percent of the total number
of features in experimental results.

Meanwhile, Su Su Htay and Khin Thidar Lynn (Htay and Lynn [2013]) used opinion words
or opinion phrases as their patterns to detect product feature from product reviewed based on
*adjective, adverb, verb, and noun*, which they used *POS* tag to identify these phrases. After
*POS* tagging process, features are extracted from a review based on opinion word or opinion
phrases from sentences. This process is known as patterns of words, where the domain of this
noun was identified based on manually tagged training corpus. Pattern knowledge was used to
pair the feature with the nearest opinion word in the sentence, which was identified based on
*adjective, adverb, verb and noun*. They used these four types of words to produce patterns
based on formula such as (*adjective, noun*), (*adjective, noun, noun*), (*adverb, adjective*),
(*adverb, adjective, noun*), (*verb, noun*), and so forth. The pattern was used in pattern knowl-
edge only for first three words after the system identified a feature word. If an opinion word is
located in this area, the system will pair it with a feature. This technique only can be used when
we already knew the orientation of sentiment analysis, either positive or negative.

### 6.1.4 Evaluation Measures

Table 6.1 shows the confusion table to calculate precision and recall. True Positive (*TP*) represents
the number of correct classification of the positive examples. False Negative (*FN*) represents
the number of incorrect classification of the positive examples. False Positive (*FP*) represents
6.1. INTRODUCTION

Table 6.1: The Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Classified Positive</th>
<th>Classified Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

the number of incorrect classification of the negative examples. True Negative \((TN)\) represents the number of correct classification of the negative examples. The formula is shown as

\[
p = \frac{TP}{(TP + FP)} \quad r = \frac{TP}{(TP + FN)}
\]

Precision and recall are the selected measurements for evaluation because they measure how precise and how complete the classification is on the positive class (Reichheld and Teal [2001]). The confusion matrix table shown in table 6.1 was used to calculate precision \((p)\) and recall \((r)\).

\[
F - Score = \frac{2(p)(r)}{(p + r)}
\]

In addition, the \(F\) Score formula above was used to calculate the average result of all products to compare the research model with Bing and Minqin’s technique (Hu and Liu [2004]) and that of Su Su Htay and Khin Thidar Lynn (Htay and Lynn [2013]).

6.1.5 The Proposed Model

Customers’ comments from the amazon.com’s website based on five products’ dataset were used as the main data source. In this model, data source (comments) is extracted into nouns and sentiment words using the ETL (Extraction, Transformation, and Loading) process. This research used part of speech \((POS)\) tagging to pre-process the data in the source file. In this process, all kinds of words were deleted except for noun, adverb, and adjective with the assistance of the wordnet system. The main part of this phase was to capture frequent nouns because most of them were products features. The ontology-based product features, which was developed early in Microsoft SQL Server was used to make sure that the frequent nouns selected
were the right product features. The synonyms in the ontology based on thesaurus.com, which are updated manually were also used to make sure that the noun with the same meaning can be captured by the system, such as \textit{msg} for \textit{message}. Only the synonyms process was done manually because this step indirectly captures the infrequent nouns.

The main data used by the current CRM system as discussed before is \textit{Customer}. Meanwhile, the research model enhanced the current design of CRM’s data warehouse based on four main tables, named as product, customer, time and opinion. Data in the opinion table is a combination of product dimension, customer dimension and customers’ comments from the extracted of unstructured data. The difference between this research model and the current CRM system were data used in Customer Analysis. As mentioned, the current CRM system used structured data from customer and product, however the research model used both of these structured data and unstructured data from customer’s comments.

Data cubes are created to analyze the orientation of some groups of customers for products at certain levels. Based on the \textit{ogc} formula mentioned in the previous section, the system produced a report regarding the orientation comments from customers about product’s features.

### 6.2 Results and Discussion

In order to mine opinions and product features, we selected all data items from five products datasets from Minqin Hu and Bing Liu (Hu and Liu [2004]). We divided the dataset into two groups with equal size as training and testing data. Training dataset is used to train our model with expected output, especially product feature. That’s the limitation of the current opinion mining. People do not know how to train the opinion mining system. Instead, to ensure finding accurate features and sentiment words. For testing data we randomly divided the dataset into five groups where each containing more than 1000 data.

First, we executed our proposed feature ontology technique to find the precision and recall. Figure 6.2 shows that the results for Precision, which in average it increased slightly from the first testing until the third testing. In the fourth testing, almost all results showed a slightly decrease in performance, before it increased again in the fifth and final testings. The average of Precision for all products are promising as the lowest performance is 0.78 for \textit{MP3Player} and the highest performance is 0.944 for \textit{DVDPPlayer}. Analyses from this trend of results,
where the declining in the fourth testing because the research model cannot recognised some of the nouns representing the product’s features, which are not in the product ontology because of unfamiliar term such text that represent the features of message in product ontology. This problem occurred because some feature’s ontology did not include the training process. The last testing was improved tremendously compared with the fourth testing.

![Graph](image.png)

**Figure 6.2:** Results for Precision using Our Ontology for All Products

Meanwhile, performance for Recall by using feature ontology also shows a good result. Figure 6.3 shows the results for Recall are high as the average for these performances are around 0.78 and 0.941 for MP3Player and DVDPlayer respectively.

Our second technique is feature ontology and its synonym. We tested this technique by running same approach as our first technique (feature ontology only). Figure 6.4 shows the precision result for all products using feature ontology and synonym. This result shows that our second technique slightly improves when compared to the first technique except for camera2 in average. Range of precision using this technique is between 0.786 and 0.956, which is higher than our first technique.

For Recall, feature ontology and synonym technique also shows some improvement compared to our first technique. Figure 6.5 shows the Recall result for all products using our feature ontology and synonym.

Based on this result, the proposed model with ontology and synonym outperformed the same approach used only for ontology in all testing sets. This shows that by using synonyms, product features in ontology has the ability to detect similar meaning of different word that referred to
CHAPTER 6. EVALUATION AND DISCUSSION

Figure 6.3: Results for Recall using Our Ontology for All Products

Figure 6.4: Results for Precision using Our Feature Ontology and Synonym for All Products
6.2. RESULTS AND DISCUSSION

We compared our research model with five techniques developed by Bing and Minqin (Hu and Liu [2004]) and one technique developed by Su Su Htay and Khin Thidar Lynn (Htay and Lynn [2013]). First, the average results for recall and precision of the proposed techniques for all baseline models produced by Minqin Hu and Bing Liu are shown in Table 6.2 and Table 6.3. In Table 6.2 and Table 6.3, the proposed technique showed significant improvement compared to the best baseline model in both precision and recall by the model used by Minqin and Bing (Hu and Liu [2004]). Compared to the best result in recall (Infrequent Feature), the proposed model showed about $9\% = \frac{0.872 - 0.80}{0.80}$ improvement. Moreover, compared with the best result in precision (P-support pruning), the proposed model showed about $12\% = \frac{0.890 - 0.790}{0.790}$ improvement. A comparison of the method used by Su Su Htay and Khin Thidar Lynn is shown in Table 6.4 and also in Table 6.5.

In Table 6.4 and Table 6.5, the proposed technique shows slight improvement compared to the baseline model in precision and significantly improved in recall compared to the model used by Su Su Htay and Khin Thidar Lynn. When compared with the best result in recall (Ontology and Synonym), the proposed model showed about $1\% = \frac{0.872 - 0.857}{0.857}$ improvement. Moreover, when compared to the best result in precision (P-support pruning), the proposed model showed about

**Figure 6.5:** Results for Recall using Our Feature Ontology and Synonym for All Products
### Table 6.2: Result for Recall

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Frequent</th>
<th>Infrequent</th>
<th>Support</th>
<th>Compaction</th>
<th>p - Support</th>
<th>OSE</th>
<th>Frequent</th>
<th>Infrequent</th>
<th>Support</th>
<th>Compaction</th>
<th>p - Support</th>
<th>OSE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera1</td>
<td>0.872</td>
<td>0.855</td>
<td>0.80</td>
<td>0.67</td>
<td>0.657</td>
<td>0</td>
<td>0.872</td>
<td>0.855</td>
<td>0.80</td>
<td>0.67</td>
<td>0.657</td>
<td>0</td>
<td>0.872</td>
</tr>
<tr>
<td>Camera2</td>
<td>0.774</td>
<td>0.765</td>
<td>0.652</td>
<td>0.594</td>
<td>0.594</td>
<td>0</td>
<td>0.774</td>
<td>0.765</td>
<td>0.652</td>
<td>0.594</td>
<td>0.594</td>
<td>0</td>
<td>0.774</td>
</tr>
<tr>
<td>Mobile Phone</td>
<td>0.832</td>
<td>0.816</td>
<td>0.731</td>
<td>0.592</td>
<td>0.592</td>
<td>0</td>
<td>0.832</td>
<td>0.816</td>
<td>0.731</td>
<td>0.592</td>
<td>0.592</td>
<td>0</td>
<td>0.832</td>
</tr>
<tr>
<td>MP3 Player</td>
<td>0.774</td>
<td>0.765</td>
<td>0.652</td>
<td>0.594</td>
<td>0.594</td>
<td>0</td>
<td>0.774</td>
<td>0.765</td>
<td>0.652</td>
<td>0.594</td>
<td>0.594</td>
<td>0</td>
<td>0.774</td>
</tr>
<tr>
<td>DVD Player</td>
<td>0.950</td>
<td>0.932</td>
<td>0.857</td>
<td>0.774</td>
<td>0.774</td>
<td>0.754</td>
<td>0.950</td>
<td>0.932</td>
<td>0.857</td>
<td>0.774</td>
<td>0.774</td>
<td>0.754</td>
<td>0.950</td>
</tr>
</tbody>
</table>
Table 6.3: Result for Precision.

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Frequent</th>
<th>Compactness</th>
<th>$P - Support$</th>
<th>Infrequent</th>
<th>OSE</th>
<th>Ontology</th>
<th>Ontology and Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera1</td>
<td>0.552</td>
<td>0.634</td>
<td>0.825</td>
<td>0.747</td>
<td>0.643</td>
<td>0.928</td>
<td>0.952</td>
</tr>
<tr>
<td>Camera2</td>
<td>0.594</td>
<td>0.679</td>
<td>0.781</td>
<td>0.710</td>
<td>0.554</td>
<td>0.852</td>
<td>0.836</td>
</tr>
<tr>
<td>Mobile Phone</td>
<td>0.563</td>
<td>0.676</td>
<td>0.828</td>
<td>0.718</td>
<td>0.815</td>
<td>0.918</td>
<td>0.924</td>
</tr>
<tr>
<td>MP3 Player</td>
<td>0.573</td>
<td>0.683</td>
<td>0.754</td>
<td>0.692</td>
<td>0.589</td>
<td>0.780</td>
<td>0.786</td>
</tr>
<tr>
<td>DVD Player</td>
<td>0.531</td>
<td>0.634</td>
<td>0.765</td>
<td>0.743</td>
<td>0.607</td>
<td>0.944</td>
<td>0.956</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.56</strong></td>
<td><strong>0.66</strong></td>
<td><strong>0.79</strong></td>
<td><strong>0.72</strong></td>
<td><strong>0.642</strong></td>
<td><strong>0.884</strong></td>
<td><strong>0.890</strong></td>
</tr>
</tbody>
</table>
Table 6.4: Result for Precision between Pattern Knowledge and Our Techniques

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Pattern Knowledge</th>
<th>Ontology</th>
<th>Ontology and Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera1</td>
<td>0.739</td>
<td>0.928</td>
<td>0.952</td>
</tr>
<tr>
<td>Camera2</td>
<td>0.712</td>
<td>0.852</td>
<td>0.836</td>
</tr>
<tr>
<td>Mobile Phone</td>
<td>0.736</td>
<td>0.918</td>
<td>0.924</td>
</tr>
<tr>
<td>MP3 Player</td>
<td>0.696</td>
<td>0.780</td>
<td>0.786</td>
</tr>
<tr>
<td>DVD Player</td>
<td>0.782</td>
<td>0.944</td>
<td>0.956</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.733</strong></td>
<td><strong>0.8844</strong></td>
<td><strong>0.8908</strong></td>
</tr>
</tbody>
</table>

Table 6.5: Result for Recall between Pattern Knowledge and Our Techniques

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Pattern Knowledge</th>
<th>Ontology</th>
<th>Ontology and Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera1</td>
<td>0.921</td>
<td>0.924</td>
<td>0.950</td>
</tr>
<tr>
<td>Camera2</td>
<td>0.812</td>
<td>0.824</td>
<td>0.832</td>
</tr>
<tr>
<td>Mobile Phone</td>
<td>0.821</td>
<td>0.819</td>
<td>0.855</td>
</tr>
<tr>
<td>MP3 Player</td>
<td>0.762</td>
<td>0.768</td>
<td>0.774</td>
</tr>
<tr>
<td>DVD Player</td>
<td>0.970</td>
<td>0.941</td>
<td>0.952</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.8572</strong></td>
<td><strong>0.8552</strong></td>
<td><strong>0.8726</strong></td>
</tr>
</tbody>
</table>

21\% = \frac{0.890 - 0.733}{0.733} \text{improvement.}

The results also proved that our model with ontology and synonym technique not only detected frequent features, it also detected infrequent features such as \textit{msg}. Even though, Hu and Liu [2004] mentioned that infrequent feature is only involved in less than five percent in customers' comments, the result showed that even in small percentage, it had better result compared to only detecting frequent features.

Based on table 6.6, the proposed technique also showed significant improvement compared to the best baseline model. Compared to the best result in F Score, by the Infrequent Feature technique and Pattern Knowledge, the proposed model showed more than 16\% = \frac{0.880 - 0.757}{0.757} and 11\% = \frac{0.880 - 0.790}{0.790} improvement respectively. Our earlier technique without synonym had already outperformed all of the baseline models. This result proved the importance of ontology usage in detecting product features.

In summary, all results proved the significant improvement of the proposed techniques compared to the baseline models. Precision, Recall and even \textit{F Score} showed that the research project was better than all of the baseline models. The research model’s capability to detect
Table 6.6: Result for F Score.

<table>
<thead>
<tr>
<th>Model name</th>
<th>FScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT Ontology and Synonym</td>
<td>0.880</td>
</tr>
<tr>
<td>PT Ontology</td>
<td>0.867</td>
</tr>
<tr>
<td>Pattern Knowledge</td>
<td>0.790</td>
</tr>
<tr>
<td>Opinion Sentence Extraction</td>
<td>0.661</td>
</tr>
<tr>
<td>Frequent Feature</td>
<td>0.6134</td>
</tr>
<tr>
<td>Compactness Pruning</td>
<td>0.665</td>
</tr>
<tr>
<td>P-support Pruning</td>
<td>0.711</td>
</tr>
<tr>
<td>Infrequent Feature</td>
<td>0.757</td>
</tr>
</tbody>
</table>

product, product attributes and product features was the main reason for better result than its baseline models.

6.3 Case Study for Orientation

The proposed model not only extracts product features and orientation words, it also caters the need of most customers when choosing the right product by showing the orientation of customers’ comments for all product features. As mentioned earlier, the results of pairs of product features and sentiment words orientation were used to implement the report for a product orientation by customers. The SQL Server was used to produce the report for this purpose. The equation of \( ogc \) was used to develop the report, which included the group of customers, and the category of product’s features. Meanwhile, seven levels polarity system was used for every sentiment word to calculate an accurate \( ogc \)’s result.

This is another contribution to sentiment analysis research area. As far as we know, at this time no other researchers do (conducted any) the calculation of sentiment polarity based on sentiment and feature word as what we did. Most of the research in this area only used sentiment word to calculate the polarity to rate sentiment orientation either positive or negative. The common techniques used are star rating ([Dragut and Fellbaum 2014], [Guzman and Maalej 2014], [Koukourikos et al. 2012]), reader rating ([Adeborna and Siau 2014]), ([Pontiki et al. 2014]), ([Pappas and Popescu-Belis 2013]), ([Maks and Vossen 2013])) and emotion signal ([Hu et al. 2013]). The closest research to our technique used is by Pontiki et al. [2014], but their research only emphasized on four aspects only (features) and their polarity calculation also
only use the sentiment word.

The proposed model not only emphasized customers’ information as the current practice of CRM, but it also summarized the orientation of customers’ comment on a product in four layers, which were the product in general, product’s attributes, product’s features, and feature’s instance. The main advantage of this model is that it can show potential customers about the orientation of every feature in a particular product, for example, the mobile phone. Figure 6.6 shows the orientation for a mobile phone in general. A total of 604 features were detected with OGC 222, which means that this product had positive comment from customers.

![Table: Mobile Phone General Polarity](image)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Poor (-3)</th>
<th>Reject (-2)</th>
<th>Weak (-1)</th>
<th>Neutral (0)</th>
<th>Accept (1)</th>
<th>Distinguish (2)</th>
<th>Excellent (3)</th>
<th>Total</th>
<th>OGC</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect</td>
<td>34</td>
<td>64</td>
<td>144</td>
<td>10</td>
<td>144</td>
<td>120</td>
<td>88</td>
<td>604</td>
<td>222</td>
<td>Positive</td>
</tr>
</tbody>
</table>

**Figure 6.6:** General Polarity for Mobile Phone

Normally, a customer gives his/her opinion about a product based on product feature, but not the product in general. As shown in Figure 6.7, our model shows every product attributes in detail. Even though, the product received positive comments in general, not every product attribute receives a positive review. In this case, connectivity had a negative result. Others product attributes had positive orientations.

Customer reviews are different between one customer to other. For example, a customer may not be happy with the Connectivity reception on the mobile phone, but more specifically, they may be happy with the "WIFI" feature. Figure 6.8 shows this situation, which shows that customers opinion on a product differs based on product features, not on the product in general.

This actual result for every feature of product was calculated using the ogc equation. From the customer’s point of view, this model generates report, which gives them more options to help them to make a better decision before they decide to buy a product. In this report, the customer will have a clearer picture about the polarity of every product feature commented by other customers. The advantage of this model report is its capability to present the orientation of a product in three parts; general (product), product’s attributes and product features. Customers can compare any product feature or attribute of their interest. The company can promote the product based on the most popular features or attributes discussed by other customers in the comments.
<table>
<thead>
<tr>
<th>Mobile Phone</th>
<th>Poor (-3)</th>
<th>Reject (-2)</th>
<th>Weak Reject (-1)</th>
<th>Neutral (0)</th>
<th>Accept (1)</th>
<th>Distinguish (2)</th>
<th>Excellent (3)</th>
<th>Total</th>
<th>OGC</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessories</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>29</td>
<td>9</td>
<td>75</td>
<td>44</td>
<td>Positive</td>
</tr>
<tr>
<td>Application</td>
<td>3</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>24</td>
<td>1</td>
<td>24</td>
<td>70</td>
<td>55</td>
<td>Positive</td>
</tr>
<tr>
<td>Communication</td>
<td>3</td>
<td>3</td>
<td>13</td>
<td>0</td>
<td>25</td>
<td>21</td>
<td>13</td>
<td>76</td>
<td>84</td>
<td>Positive</td>
</tr>
<tr>
<td>Connectivity</td>
<td>3</td>
<td>13</td>
<td>40</td>
<td>0</td>
<td>15</td>
<td>1</td>
<td>4</td>
<td>80</td>
<td>-32</td>
<td>Negative</td>
</tr>
<tr>
<td>Entertainment</td>
<td>7</td>
<td>3</td>
<td>30</td>
<td>3</td>
<td>25</td>
<td>32</td>
<td>9</td>
<td>109</td>
<td>59</td>
<td>Positive</td>
</tr>
<tr>
<td>General</td>
<td>10</td>
<td>13</td>
<td>40</td>
<td>0</td>
<td>32</td>
<td>14</td>
<td>23</td>
<td>132</td>
<td>33</td>
<td>Positive</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>22</td>
<td>6</td>
<td>62</td>
<td>51</td>
<td>Positive</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>64</td>
<td>144</td>
<td>10</td>
<td>144</td>
<td>120</td>
<td>88</td>
<td>604</td>
<td>222</td>
<td>Positive</td>
</tr>
</tbody>
</table>

**Figure 6.7:** Detail of Customer’s Comments based on Product Attributes
CHAPTER 6. EVALUATION AND DISCUSSION

**Figure 6.8**: Detail of Customer's Comments based on Product Attributes' Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Total</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>100</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>Feature 2</td>
<td>80</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td>Feature 3</td>
<td>60</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Feature 4</td>
<td>40</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Feature 5</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

**Diagram Description**: The diagram illustrates the distribution of comments across different features, highlighting the positive and negative feedback for each feature.
6.4 Summary

The experimental results and the case study show that the proposed approach is superior than the baseline models. The reasons for the improvement are as follows:

- Development of Product Ontology
  Ontology is a tree-like structure to represent the relationship between categories or features in a domain. This research project used product ontology to detect nouns that represent a product feature in a four tier ontology. These tiers represent product general name, product attributes, product attributes’ features and instances of the attribute. The ontology matches product reviews (comments) to find the features, especially the frequent features. This ontology can also detect infrequent feature if the current ontology is already designed with particular infrequent features in its database. Detail about product ontology especially mobile phone already mentioned in chapter 5 and the Appendix.

- Synonym on Product Ontology
  Synonyms of product features were considered in the proposed research. For every product feature, at least five synonyms word based on a thesaurus website and popular short term words used by current customer for example ”text” and ”msg”, which refer to message. This synonym is important especially when customers refer to the same item with other terms. Even though some time the synonyms give wrong definition for the same item, our model removes the synonym that does not represent the product features. Figure 6.9 shows this situation, in this case synonym words for message. We only selected the top five synonyms for every ontology used.

Although this research produced good results, it can still be improved, especially when this product ontology only represents four products. A new ontology must be built to handle comments for multiple products and also multiple domain in the future.

The product orientation report currently depends on the OGC formula. This formula is quite accurate for every feature that it represents, but it can lead to misunderstanding especially for those, unclear with the OGC formula. For example, if the system have 3 products polarity which the first two are -1 and -1, which accumulate as -2 and another one is +3. Even though, the number of product polarity for -1 is more than +3 in term of it appearance in system,
but the result of orientation is positive because the amount of positive which is +3 is bigger than negative (-2), even though the feature that represents the negative orientation is bigger in number. The numbering system for polarity needs to be addressed in a more technical manner in order to come out with a better solution.

Our new architecture for CRM combined customers’ personal record, products record, and also feedback regarding a particular product that they have used. This new architecture is very significant for companies and manufacturers to gather the best opinion from customers regarding the particular product, especially for future product improvement. Our testing and evaluation also showed better results compared to our baselines models with a precision of 0.890 and a recall of 0.872. Meanwhile, our result for F Score was also better than the baseline models with 0.880.

In addition, the report on ogc can cater customers’ need when they are still undecided about a particular product because the report generated by SQL Server shows the polarity of every product features commented by others customers. Currently, our model has been developed to cater five products with ontology. In the business world, more products are produced to meet customers demand. Our next target in the future is to develop a model with ontology that can be used on multiple products. The final report for ogc can also be upgraded for multiple products’

Figure 6.9: Synonym of Product Ontology
reports.
Chapter 7

Conclusions

The research strengths and limitations are discussed in this chapter. This chapter concludes with few recommendations for future works or researches in this particular area.

7.1 The Challenges

The emergence of the Internet in the corporate world and its high usage penetration among the people today has changed the business world in general, and business strategy in specific, in attracting new customers and retaining current loyal customers. Companies must take full advantages of this new environment either to improve their business strategy or to attract customers using online business strategy, or the best choice would be to take both. Mastering these two strategies is one step ahead for a company compared to their competitors. Changes in marketing paradigms are compulsory in competing with other competitors. Before the emergence of the Internet, most businesses used broad marketing target, but with the Internet at the finger tip of customers, companies must target more customers, which means one to one marketing. As mentioned by Campbell and Cunnigham [1983], the greatest asset that every company has is the customers. Typically, customers in this generation have a good Internet access, which brings the business to a new dimension in marketing their products or services. Demographical issue is no longer a barrier for companies to market their products because customer can purchase them everywhere at anytime. Thus, customers need accurate information about a particular product to get informative description about it, compare the price and do the transaction if they are interested in any product or service.
Companies must gather a lot of information to provide informative description about their products or services. This is the toughest task as many information spread on the Internet from a lot of resources without knowing which is an accurate fact or information. The main issue in customer’s side that company need to provide is accurate information. From the company’s side, they also need to change their business culture by setting up new business dimensions to cope with the fast growing technology and the Internet in their industries.

CRM is one of the tools available in market to help the company to manage their business transactions from the front end (customer service department) to the back end (HR, Marketing and other department). Customer analysis is the most important application in CRM that analyses a lot of information about customer’s transaction with a company based on past records. The current CRM only analyses structured data from their databases about the customers and the company’s own records such as product, sales and marketing. This shows that the current CRM only analyses data for future marketing plan based on customer only. It does not have any ability to give any analyses on the future of a product.

Our research has shown the importance of opinion mining in CRM. The current CRM model emphasises only on customer details such as personal information, and characteristics of customers to predict the future pattern of purchasing the products. Besides that, the current CRM model only captures product details from manufacturers or suppliers without information from the customers’ feedback regarding a particular product used by them.

This research has produced a new architecture for the CRM model by combining customers’ personal record, product record and feedback regarding a particular product that they have already used. One big problem to combine these two information between the current data in CRM database and comments from customers is the different format used by CRM (structured data) and comments (unstructured data). We extracted the comment using some techniques as mentioned in chapters four and five to convert unstructured to structured data. This research enhanced the current CRM model by adding unstructured data based on customer’s comment into CRM’s database or data warehouse.

The final outcome of this research produces a comprehensive report based on the opinions of customers about a particular product. The usage of SQLServer has help the presentation of the report as simple and easy to use by staffs of any company. This report shows a summary of
customer orientation on a particular product based on either the product in general or product features, in specific, product attributes and product instances.

This new architecture is very significant for companies and manufacturers to obtain the best comprehensive summary of opinions from customers regarding a particular product, especially for product improvement in the future. Our testing and evaluation also showed better results compared to our baseline models, with a precision of 0.890 and recall of 0.872.

7.2 Contribution (Findings)

This research has few contributions toward data mining research, especially in opinion mining. In this section, we discuss the research contribution based on the four research objectives mentioned in chapter one, Introduction.

- Acquire Users’ Comments (Comments from Users)

This research used two types of data; one is the structured data, which represents the CRM data, such as customer, transaction history, and financial. This data were collected from Microsoft’s sample data from SQLServer software, which contains a large amount of samples. Another data is acquired from Hu and Liu [2004], which contains more than a thousand data consisting of comments for five products from various customers, and has unstructured data format. 500 comments from this data were used as training data and all data including first 500 data were used as testing data. Both portion of these data were merged later in CRM’s environment.

Normally, in current CRM environment, the data used in processing only involves structured data from company records such as records from different departments, such as marketing, sales and service, and customer service. Only data from current customers are used to analyse the customer patterns, even though there are data about products. But, the product data are more on its description such as the top product sold by the company, the stock of the product and the supplier of that particular product.

This research produces a new way to collect data for CRM, which is by gathering structured data used by the current CRM model and also data from customer opinionated
comments. These comments have a lot of information on customer’s thoughts about a particular product. This research shows the importance of these data combination to produce a new analysis for CRM model. As the current CRM only emphasizes on customer data to forecast the future business for a company, this new idea of combining customers comments allows company more substantial analysis than before. Furthermore, the new analysis produces customer purchasing patterns, and also pattern of the product features of customers’ interest, which adds value to the current CRM model.

- Development of Product Ontology

It was not easy for a model to capture and understand about product and its features. In this research, the first technique to capture the product and its features used frequent noun as all products and their features are noun. This technique was problematic as the model detected too many nouns and most of them were not of the product or its features. Then, we used the pruning technique to remove some unnecessary words such as the and a. As these two techniques cannot solve the problem, the model later used ontology followed by synonyms for product features.

The creation of product ontology is innovative to detect noun in the model. The Concept of Hierarchy was used as the main idea to build this ontology. The basic idea such postal format for address used to describe every tier of ontology. For example, this address structure: House No > Street Name > City > State > Country, shows clear information about address. This tree system starts from specific to more general information. In this ontology, reverse structure was used compared to the postal address system, as it started from general to specific information. The idea was to use this ontology in the future such that it can comply with multiple products.

The product ontology has four tiers of definition as shown in Figure F.1. The first tier is general definition of the product, followed by the features of the product in the second tier. Attributes of product features in third tier and last but not least is instances of the attributes in the fourth tier. It was shown that this product ontology can recognize almost the product and its features from customers’ comments. It can also cater terms referred by customer in their comments, either in general terms or more specific (technical) terms. Comments from customers are not only properly spelled words, but also in text short
messageform. For examples, for the word message, some users prefer to write it as msg, which is the most popular way to write it this technology era. The model cannot recognise this new features name. One solution for this problem was by enhancing the ontology with synonyms. All of the products and their features were equipped with the five most popular synonyms based on thesaurus.com website. Moreover, common words not available in thesaurus.com were added manually into database such as msg for message. This enhancement to the ontology slightly increased the results for recall and precision of the research model.

• Summarize Opinions

This research project had two main sections; First was to integrate unstructured data (comments) and structured data (CRM), and second was to calculate the polarity of customer orientation on a particular product. The first work was covered by the first two objectives, while the second work was based on the third objective. The introduction of OGC formula (Yaakub et al. [2011]) in this research opened new dimensions for companies to calculate the product orientation based on customer comments. The OGC formula was based on three groups of data, namely comments with opinionated word, group of customer and the product itself.

Cubes development was based on the needs of the company and its customers. This research model has the ability to produce many reports in SQLServer based on the product in the general, or product features in the more specific report. The idea of creating multi levels reports was to give accurate result about the product. For example, if one customer wants to know about the product in general, but does not care about the technical part of the product, then this model will produce a report about product orientation in general. Otherwise, it will produce more detailed and specific report if the customer or the company wants it. The cube’s creation is based on four ontology levels, which means that this model can produce four levels of report.

• Model Evaluation

This is a common part in research development, which is to evaluate the research result. The evaluation was done in two parts. The fiirst evaluated the significance of combining
structured and unstructured data. Five baseline models from Hu and Liu [2004] and baseline model from Htay and Lynn [2013] were used to evaluate this model. The results showed that this model is very significant as this model performed better than other baseline models. Our model had better results compared to its baseline models in all three evaluation models named recall, precision and F-score. This research technique outperformed the best result of the baseline models in recall measurement by almost 9 percent, which was 0.872 compared to 0.80 for Bing Liu and Minqin’s model and slightly improved by 1 percent, which was 0.872 compared to 0.857 for baseline model, Pattern Knowledge (Htay and Lynn [2013]). Furthermore, the result for precision showed that the research technique had 0.890 compared to 0.790 for the best result for the baseline model by Bing Liu and Minqin Hu, which was almost 12 percent improvement and almost 21 percent improvement from the Pattern Knowledge technique. For F Score, the best performance by the baseline models was 0.790, while the research technique showed more impressive result with 0.880. The result of F Score was more than 11 percent better than the baseline models. The performance of this research technique was better than five baseline models by more than 11 percent for every evaluation conducted in this research, which showed the significance of this technique.

7.3 Strength and Limitation

To the author’s knowledge, this study is the first, to model an enhanced CRM model by integrating unstructured data from customer comments with structured data. This integration model is important in business today, especially for companies using CRM as a tool to help them analyse customer’s purchasing pattern for future business strategy. As mentioned earlier, current practices in CRM only uses structured data from customer records to analyse the next purchase by the customer (s). This structured data only covers 20 percent of the information in the company’s website, while another 80 percent is unstructured data, which means that the company is not using all of its resources for customer analysis. This research project has solved this issue using customer and product data from CRM with customers comments data from the company’s website.

The creation of OGC formula has produced a new report for customers and companies. This formula is used in building data cubes to calculate the polarity of customer orientation on
a particular product. This research model has been built to cope with customer’s demand in getting information about product features in detail. Customers who use this model can easily get a summary of the features of every product based on their needs. For example, a customer who wants to analyse customers’ orientation about certain features of a product can get the report using this research model.

The development of product ontology has proven that this model has the ability to capture and recognise the features of a product better than the frequent nouns technique. The combination of product ontology and synonyms have produced good result for the model to recognise the product and its features in customer comments.

The limitation of this project is that the product ontology can handle five products at a time. Time constraint is the single factor that prevented the researcher from developing the ontology for multiple products.

7.4 Future Works

Currently, this research model has been developed to cater only for five products with ontology. In the business world, more products are produced to meet customers demand. The next target in the future is to develop a model with ontology that can be used for multiple products. The development of multiple product ontologies need a thorough review, especially for product features and attributes with the same term, but different functions or definitions. This problem is quite complex and a good technique is needed to solve this problem.

Changes in product ontology means some of the model’s algorithms need to change as well. OGC formula and algorithm must follow the direction of the model. OGC formula needs to be upgraded, especially in order to follow the changes in ontology. Changes in any ontology tier level will change the formula of OGC as well to make sure that the final result in the report is accurate.

7.5 Summary

CRM is a system used by companies to improve income and profit based on a customer analysis model by analysing customer behaviours, especially when doing transaction with the company.
Existing CRM system stresses on customer as their main input when doing customer analysis. We enhanced this current CRM by combining product and customer data from a CRM system, which are in structured format. Furthermore, we add a new entity to CRM system from the company’s product website and reviews or comments from customers regarding the product. The problem is, however, reviews from customer have unstructured data.

Opinion mining was used to solve this problem in this research. Unstructured data (review) was extracted using some opinion techniques such as association rules, frequent features and measurement of distance between noun and adjective or noun and adverb. ETL processes from a data warehouse was used to extract, transform and load the data to the CRM system. Product ontology was built with synonym to capture the product and product features from a review. A pair of product features and sentiment words was the output for the extraction process. The performance of this research project was far better than the six baseline models with precision, and F-Score of more than 11 percent. These promising results proved the importance of product ontology and synonym in the extraction process.

The final result of this research is a report of product’s orientation based on customers’ product review. The ability of our research model to produce the orientation for every product features is a new finding in this area and it adds value to the CRM system.
## Appendix A

### Product Ontology and Synonym

<table>
<thead>
<tr>
<th>Product Ontology</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>mobile, telephone, call, handphone</td>
</tr>
<tr>
<td>Product</td>
<td>motorola, panasonic, nokia, sanyo, sony, microsoft, nikon, canon</td>
</tr>
<tr>
<td>Service</td>
<td>facility, resource, solution, system</td>
</tr>
<tr>
<td>Camera</td>
<td>video camera, camcoder, video</td>
</tr>
<tr>
<td>Price</td>
<td>cost, payment, charge, fee, fare</td>
</tr>
<tr>
<td>Picture</td>
<td>image, painting, print, sketch, photo</td>
</tr>
<tr>
<td>Message</td>
<td>text, sms, email, mms, chat</td>
</tr>
<tr>
<td>Signal</td>
<td>edge, gsm, triband, communication</td>
</tr>
<tr>
<td>Connectivity</td>
<td>infrared, bluetooth, wireless, 3g, gprs</td>
</tr>
<tr>
<td>Storage</td>
<td>memory, memory card</td>
</tr>
<tr>
<td>Sound</td>
<td>voice, audio, tone, ringtone, ring</td>
</tr>
<tr>
<td><strong>Product Ontology</strong></td>
<td><strong>Synonym</strong></td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Cover</td>
<td>sleeve, casing, wrapping</td>
</tr>
<tr>
<td>Screen</td>
<td>shutter, display, monitor, background, screensaver</td>
</tr>
<tr>
<td>Size</td>
<td>area, bulk, compact</td>
</tr>
<tr>
<td>Capacity</td>
<td>volume</td>
</tr>
<tr>
<td>Accessory</td>
<td>extra, wallpaper, adapter, body, earphone</td>
</tr>
<tr>
<td>Application</td>
<td>software, program, calendar, game, organizer</td>
</tr>
<tr>
<td>Design</td>
<td>model, interface, color</td>
</tr>
<tr>
<td>Performance</td>
<td>production, speed, weight, quality</td>
</tr>
<tr>
<td>Zoom</td>
<td>lens, focus, closeup</td>
</tr>
<tr>
<td>Flash</td>
<td>light, backlight</td>
</tr>
<tr>
<td>Button</td>
<td>click, key, keypad, panel</td>
</tr>
<tr>
<td>Radio</td>
<td>receiver, transistor, hifi</td>
</tr>
<tr>
<td>Headset</td>
<td>earpiece</td>
</tr>
<tr>
<td>Macro</td>
<td>command, function, shortcut</td>
</tr>
<tr>
<td>Scene</td>
<td>part, section, extract</td>
</tr>
<tr>
<td>Indoor</td>
<td>interior, enclosed, covered</td>
</tr>
<tr>
<td>Movie</td>
<td>moving picture, film</td>
</tr>
<tr>
<td>Adapter</td>
<td>electric plug, connector, converter</td>
</tr>
<tr>
<td>Control</td>
<td>switch, controller, regulator</td>
</tr>
<tr>
<td>Resolution</td>
<td>tenacity, firmness</td>
</tr>
<tr>
<td>Product Ontology</td>
<td>Synonym</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Construction</td>
<td>building, creation, structure</td>
</tr>
<tr>
<td>Exposure</td>
<td>contact, experience</td>
</tr>
<tr>
<td>Body</td>
<td>form, figure, frame</td>
</tr>
<tr>
<td>Strap</td>
<td>band, fast, fastening, leash, strip</td>
</tr>
<tr>
<td>Navigational</td>
<td>directional, route, navigation</td>
</tr>
<tr>
<td>Folder</td>
<td>binder, file</td>
</tr>
<tr>
<td>Play</td>
<td>show</td>
</tr>
<tr>
<td>Warranty</td>
<td>guarantee, contract, pledge</td>
</tr>
<tr>
<td>Style</td>
<td>class, smartness, charm</td>
</tr>
<tr>
<td>Sync</td>
<td>synchronize, coordinate, match, harmonize</td>
</tr>
<tr>
<td>Forward</td>
<td>onward, advancing</td>
</tr>
<tr>
<td>Rewind</td>
<td>reverse</td>
</tr>
<tr>
<td>Machine</td>
<td>mechanism, device</td>
</tr>
</tbody>
</table>
### Appendix B

#### Sentiment Word

<table>
<thead>
<tr>
<th>Sentiment Word</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>excellent, great</td>
<td>+3</td>
</tr>
<tr>
<td>perfect, best</td>
<td>+3</td>
</tr>
<tr>
<td>amazing, exceptional</td>
<td>+3</td>
</tr>
<tr>
<td>strength, superb</td>
<td>+3</td>
</tr>
<tr>
<td>outstanding, incredible</td>
<td>+3</td>
</tr>
<tr>
<td>fantastic, enthusiastic</td>
<td>+3</td>
</tr>
<tr>
<td>marvellous, terrific</td>
<td>+3</td>
</tr>
<tr>
<td>superior, favourite</td>
<td>+2</td>
</tr>
<tr>
<td>sturdy, better</td>
<td>+2</td>
</tr>
<tr>
<td>well, easier</td>
<td>+2</td>
</tr>
<tr>
<td>quieter, awesome</td>
<td>+2</td>
</tr>
<tr>
<td>Sentiment Word</td>
<td>Polarity</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>useful, solid</td>
<td>+2</td>
</tr>
<tr>
<td>attractive, wise</td>
<td>+2</td>
</tr>
<tr>
<td>impressive, powerful</td>
<td>+2</td>
</tr>
<tr>
<td>unbelievable</td>
<td>+2</td>
</tr>
<tr>
<td>bless, recommended</td>
<td>+1</td>
</tr>
<tr>
<td>good, decent</td>
<td>+1</td>
</tr>
<tr>
<td>cute, pretty</td>
<td>+1</td>
</tr>
<tr>
<td>like, okay</td>
<td>+1</td>
</tr>
<tr>
<td>quick, creative</td>
<td>+1</td>
</tr>
<tr>
<td>worth, nifty</td>
<td>+1</td>
</tr>
<tr>
<td>glad, cool</td>
<td>+1</td>
</tr>
<tr>
<td>fair, pleased</td>
<td>+1</td>
</tr>
<tr>
<td>pro, beautiful</td>
<td>+1</td>
</tr>
<tr>
<td>comfy, excited</td>
<td>+1</td>
</tr>
<tr>
<td>interesting, enjoy</td>
<td>+1</td>
</tr>
<tr>
<td>satisfy, adequate</td>
<td>+1</td>
</tr>
<tr>
<td>acceptable</td>
<td>+1</td>
</tr>
<tr>
<td>average</td>
<td>0</td>
</tr>
<tr>
<td>Sentiment Word</td>
<td>Polarity</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>horrendous, horrible</td>
<td>-3</td>
</tr>
<tr>
<td>torture, disgust</td>
<td>-2</td>
</tr>
<tr>
<td>unbearable, crazy</td>
<td>-2</td>
</tr>
<tr>
<td>weird, terrible</td>
<td>-2</td>
</tr>
<tr>
<td>useless, disappointed</td>
<td>-2</td>
</tr>
<tr>
<td>suck</td>
<td>-2</td>
</tr>
<tr>
<td>expensive, unpredictable</td>
<td>-1</td>
</tr>
<tr>
<td>tedious, danger</td>
<td>-1</td>
</tr>
<tr>
<td>destructible, teeny</td>
<td>-1</td>
</tr>
<tr>
<td>loss, bored</td>
<td>-1</td>
</tr>
<tr>
<td>minus, negative</td>
<td>-1</td>
</tr>
<tr>
<td>poor, disturbance</td>
<td>-1</td>
</tr>
<tr>
<td>inconvenient, lack</td>
<td>-1</td>
</tr>
<tr>
<td>confusing, paranoid</td>
<td>-1</td>
</tr>
<tr>
<td>dissatisfy, problem</td>
<td>-1</td>
</tr>
<tr>
<td>complicated, tricky</td>
<td>-1</td>
</tr>
<tr>
<td>distort, unfortunate</td>
<td>-1</td>
</tr>
</tbody>
</table>
Appendix C

MP3 Ontology
Figure C.1: MP3 Ontology
Appendix D

DVD Player Ontology
Figure D.1: DVD Player Ontology
Appendix E

Camera Ontology
Figure E.1: Camera Digital Ontology
Appendix F

Mobile Phone Ontology


Raphaël Féraud, Marc Boullé, Fabrice Clérot, Françoise Fessant, and Vincent Lemaire. The orange customer analysis platform. In *Proceedings of the 10th industrial conference on*


