

A Data warehouse-oriented Methodology for Qualitative Semi-Structured Web Information and Social Networking Sites' User Status Search

Md. Shahriar Kabir
Lakehead University

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science



Lakehead
UNIVERSITY

All rights reserved. This thesis may not be reproduced in whole or in part, by
photocopying or other means, without the permission of the author

Supervisory Committee

Dr. Jinan Fiaidhi, Supervisor and Chair
(Department of Computer Science, Lakehead University, Canada)

Dr. Sabah Mohammed, Internal Examiner
(Department of Computer Science, Lakehead University, Canada)

Dr. Shan Du, Internal Examiner
(Department of Computer Science, Lakehead University, Canada)

Dr. Abdulsalam Yassine, External Examiner
(Department of Software Engineering, Lakehead University, Canada)

ABSTRACT

Finding most desired and useful information from the diverse information and content embedded on webpages has become more challenging due to the rapid growth of websites and webpages, dynamic changes and updates of information and content on webpages, the lack of well-formed structure of webpage content and so on. Information search seems a trivial task when plain text, hyperlink texts, embedded images, videos that all make up webpage content remain in semi-structured form. Semi-structured webpage content do not have predefined structure and remains in hierarchically nested HTML tags of a webpage body. Unlike structured webpage content, heterogeneous semi-structured webpage content can't be neatly formatted, organized and modeled directly into relational database. One of the most important information types on the web is web user's emotion expressed in user-posted status on Social Networking Sites like Facebook, Twitter. Publicly posted user status is informative enough to know user's daily thoughts, feelings, emotions through textual self-description. The data warehouse-oriented methodology of semi-structured webpage content extraction and modeling into database introduces a simplified and less labor intensive XML-based semi-structured webpage content extraction technique that overcomes the limitations of existing pre-defined specification file and Wrapper-based techniques to adapt rapid changes of webpage content and to extract same piece of information on different webpages having differentiated nested HTML structure. This methodology also introduces Multidimensional Fact data modeling technique for semi-structured webpage content storage into relational database. Our implemented methodology ensures qualitative search result in terms of hyperlinks to most desired webpages appearing first with a relatively very low fractional amount of minute.

Our implemented social networking sites' user status updates extraction and modeling into database approach introduces a lexical approach for analyzing emotion-related features in status updates to map emotion related linguistic features with psychological traits. It ultimately ensures constant higher percentage (ranging between 81%-86%) of correct user status updates appearing in search results that accurately reveal psychological traits as exhibited by user.

Contents

Abstract	iii
Contents	iv
List of Tables	vii
List of Figures	viii
List of Abbreviations	xi
Acknowledgments	xii
Dedication	xiii
1 Introduction	1
1.1 Background	1
1.2 Web Mining	2
1.3 Semi-structured webpage content.....	3
1.4 Semi-structured webpage content extraction and mining methodologies.....	3
1.5 User posted status on social networking sites and Psycholinguistic meaning of words	4
1.6 Research questions	5
1.7 Objectives.....	5
1.8 Thesis outline	7
2 Related Work	8
2.1 Semi-structured webpage content extraction for relational database schema mapping.....	8
2.2 Identification of emotion in Social Networking Sites’ user status to reveal psychological traits	10
2.3 Conclusion.....	12

3 Semi-structured Webpage Extraction and Modeling into Relational database	14
3.1 Introduction	14
3.2 A simplified technique of Semi-structured webpage content extraction and modeling into Database	17
3.2.1 Extracting Webpage Content out of HTML DOM Tree.....	18
3.2.2 HTML Data to XML Conversion and XML File Generation.....	20
3.2.3 XML data file integration, Mapping and Data loading.....	23
3.2.4 Data Modeling	24
3.3 Implementation of Web Application	29
3.3.1 Tool selection	30
3.3.2 Web API implementation	31
3.4 Experimental Execution and Results	33
3.4.1 Experimental output	33
3.4.2 Performance result and Analysis of result	36
3.5 Conclusion.....	38
4 Processing Social Networking Sites' User Status to Reveal User's Psychological Trait.....	40
4.1 Introduction	40
4.2 Implemented Lexical approach for Identification of Emotional and Psychological Traits in Facebook user status updates and Twitter tweets.....	41
4.2.1 Data extraction.....	41
4.2.2 Data preprocessing.....	44
4.2.2.1 Data cleaning.....	44
4.2.2.2 Tokenization	47
4.2.3 Lexicon detection.....	49
4.2.4 Data loading according to Multidimensional Fact Data model.....	51
4.3 Web API implementation.....	52
4.4 Experiment and Evaluation of Experimental Results.....	55
4.4.1 Experimental output.....	55
4.4.2 Evaluation Metrics	58
4.4.3 Performance result and Analysis of result.....	59
4.5 Conclusion.....	67

5 Discussion, Future Work and Conclusion	68
5.1 Overview	68
5.2 Main Contributions	68
5.3 Current Exceptions and Scope for Improvement	69
5.4 Conclusion.....	70
Bibliography.....	71

List of Tables

<u>Table</u>	<u>Page</u>
2.1 Advantages of our approach over major existing semi-structured webpage content extraction and database schema mapping techniques	12
3.1 List of web API development tool used	30
3.2 List of web API development technology and libraries used.....	30
3.3 Time taken in each experimental test.....	36
3.4 Performance of search query execution compared to other web search query applications of Semi-structured webpage content.....	38
4.1 Four subcategories of LIWC dictionary for human psychological states.....	42
4.2 Data fields in extracted Facebook dataset and their descriptions.....	43
4.3 Data fields in extracted Twitter dataset and their descriptions.....	43
4.4 Four subcategories of LIWC dictionary for human psychological states.....	49
4.5 Lexicon detected based on Tweet (T)	50
4.6 Mapping of opinion word, amplifier word and emoji wirh psychological trait.....	51
4.7 Initial data records in datasets and processed data records in database	60
4.8 Experimental test result on dataset of January 7 to 8, 2019	61
4.9 Experimental test result on dataset of January 14 to 15, 2019.....	61
4.10 Accuracy of experimental test.....	66
4.11 Precision and Recall measures in experimental test of emotion classification in tweets by Srinivasu Badugu and Matla Suhasini.....	67

List of Figures

<u>Figure</u>	<u>Page</u>
3.1 Same type of information on two different webpages having different HTML structure	15
3.2 UML object diagram showing a complete webpage as a complex object.....	17
3.3 Extraction pattern specification of a webpage in web extractor’s extraction pattern.....	19
3.4 Part of the HTML structure of specified webpage region.	20
3.5 Part of HTML DOM tree built by web extractor.....	20
3.6 Webpage content or data elements within HTML DOM structure are grouped into tabular HTML data elements.	21
3.7 Partial tabular HTML data structure.....	22
3.8 Part of XML data after conversion of tabular HTML data.	22
3.9 View of a data record in intermediate transactional database table	24
3.10 Multidimensional fact data model for semi-structured webpage content database storage.	25
3.11 Partial view of data records in THESISDBDW.DWB_TBL_WP_CONTENT base fact table.....	26
3.12 Data structures of dimensional tables	26
3.13 Partial view of data records in Ontology dimensional table THESISDBDIM.DWL_TBL_ONTLGY.....	27
3.14 Formation of dimensional table THESISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP.	27

3.15	View of data records of keywords in table THEISISDBDIM.DWL_TBL_KEYWORDS_INFO.....	28
3.16	Partial view of data records in THEISISDBDW.DWB_TBL_SHOPPING_WP_CONTENT Level-1 derived table.....	29
3.17	GUI of web API.....	31
3.18	Search results for search query ‘google pixel phone’	33
3.19	Search result for input search query ‘google pixel’.....	34
3.20	Search result for input search query ‘google pixel phone’	34
3.21	Search result for input search query ‘google pixel phone 32GB’	35
3.22	Image search result for input search query ‘google pixel phone 32GB’	36
3.23	Graphical representation.....	37
4.1	Data preprocessing technique used on both Facebook and Twitter datasets	44
4.2	Example Tweet (T)	45
4.3	Partial data view of our opinion dictionary	49
4.4	Partial data view of our emoticon dictionary.....	50
4.5	Multidimensional fact data model for Social Networking Sites’ user status database storage	51
4.6	Input search query for Facebook status updates and tweets that contain keywords “what an amazing year”.....	53
4.7	Search result of Facebook user status and Twitter tweets that contain keywords “what an amazing year”.....	55
4.8	Partial view of search result for input search query “what an amazing year”	56
4.9	Partial view of search result for input search query “very much tensed”	57
4.10	Partial view of search result for input search query “I really hate”	57

4.11	Partial view of search result for input search query “feeling very much sad”	58
4.12	Graphical representation of experimental test result on dataset of January 7 to 8, 2019	63
4.13	Graphical representation of experimental test result on dataset of January 14 to 15, 2019	65

List of Abbreviations

LIWC	Linguistic Inquiry and Word Count
HTML	Hypertext Markup Language
XML	Extensible Markup Language
NLTK	Natural Language ToolKit
JSON	JavaScript Object Notation
CSV	Comma-Separated Values
ASCII	American Standard for Information Interchange
PTBTokenizer	Penn Tree Bank Tokenizer

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Dr. Jinan Fiaidhi for all her invaluable guidance, support and encouragement as my supervisor. Her guidance, constructive suggestions, and encouragement are the reason I was able to learn and grow as a researcher. It was an absolute privilege to work with her on this thesis work. She was always accessible, motivated and willing to help me in my research even in her busy schedule.

I would also like to thank the examination committee members, Dr. Sabah Mohammed, Dr. Shan Du and Dr. Abdulsalam Yassine for reviewing this thesis and providing their helpful comments, constructive suggestions thereby.

I would like to thank my family and friends for their continuous encouragement and help, especially my parents for giving necessary advice and guidance and arranging all facilities to make learning easier. I choose this moment to acknowledge their contribution gratefully.

DEDICATION

I would like to dedicate this thesis to:

My parents, who gave me the foundation of something they always enjoyed, education. For inspiring me and always believing in my ability to accomplish anything I set my mind to. For making sure I had access to all the resources I need to succeed no matter the circumstances.

My wife, who always took care of my chores when I was busy studying and made me laugh whenever I would feel low.

Chapter 1

Introduction

1.1 Background

With the phenomenal growth of the World Wide Web, today's websites and webpages are experiencing an ever-increasing volume of information and content published for open access. It is evident that finding most desired and useful information from the diverse information and content embedded on websites has become more challenging task due to the rapid growth of websites day-by-day, dynamic changes or updates of information and content on webpages, the lack of structure of webpage content and so on. Moreover, one of the most important information types on the web is web user's emotion or expression in user generated content, for example, user status updates on Social Networking Sites, user reviews of products, forum posts, posts on microblogs, remarks and comments on online news portals etc. User's status updates on Social Networking Sites like Facebook, Twitter help us to identify discovery and recognition of positive or negative emotion as people expressed on diverse subject matters of interest. User posted comments are often convincing and their indicators can be used as motivation when making choices and decisions on the patronage of certain products and services, endorsement of a political event or even reactions to social issues. User status updates extraction and processing from ever increasing and massive volume of Social Networking Sites, online blogs, product review websites, online news websites has become a difficult task in Information Retrieval (IR) domain that has been eased by Natural Language Processing (NLP), Machine Learning, Text Mining techniques over the past years.

1.2 Web Mining

Web Mining brings all to address challenges, problems of information and content extraction from ever growing websites and mining useful data. The emerging field of web mining aims at finding and extracting relevant information that is hidden in web-related data, particularly in hypertext documents published on the web. This new area of research has been defined as an interdisciplinary field (or multidisciplinary) that uses techniques borrowed from data mining, text mining, databases, statistics, machine learning, multimedia, etc. Web Mining is classified into three basic categories as: (i) Web Content Mining, (ii) Web Usage Mining and (iii) Web Structure Mining.

Web Content Mining refers to the discovery of the most desired and useful information from webpage or website content [3]. Text, images, embedded videos, hyperlink texts, data blocks all make up webpage content. Research activities in [1],[2],[3] recognized web content mining problems to improve the way that pages are presented to end users, improving the quality of search results and extract interesting content pages.

Web content mining could be differentiated from two point of views [5]:

- (i) Agent-based approach: The agent-based approach uses so called Web agents to collect relevant information from the World Wide Web. A web agent is a program that visits a website and filters the information the user is interested in. There are three subtypes for the agent-based approach: Intelligent Search Agents, Information Filtering or Categorization and the Personalized Web Agents.
- (ii) Database approach: The database approach of Web Content Mining tries to develop techniques for organizing semi-structured data on the web into more structured collections of information resources. Standard database querying mechanisms and data mining techniques can be used to analyze those collections.

As the Agent-based approach aims on improving the information finding and filtering on a given website only, for example, information finding and filtering on a website that has custom search option to search the entire website, the Database approach of Web Content Mining aims on

modeling the data on the web into more structured form in order to apply standard database querying mechanism that eventually aids to build a keyword-based web search application.

1.3 Semi-structured webpage content

Information search becomes a trivial task when text, embedded images, videos, hyperlink texts, data blocks that all make up webpage content remain in semi-structured form. Semi-structured webpage content do not have a predefined structure and remains in hierarchically nested HTML tags of a webpage body. No semantic is applied to semi-structured web data and analysis of words or sentences is required for extracting relevant information. For example, webpages of many Consumer-to-Consumer (C2C) e-commerce websites having product's features specification, image, price etc., online financial websites having stock market data and information, weather forecasting websites attribute semi-structured data [42]. Unlike structured data that can be neatly modeled, organized, and formatted into ways that are easy for us to manipulate and manage, semi-structured webpage content does not attribute to have enough regular structure to qualify for the kinds of management and automation usually applied to structured data. Structural irregularity in semi-structured data can serve the following purposes: obtaining general information contents, facilitating the integration of data from several information sources, improving storage, optimizing query evaluation. The use of semi-structured data can be felt in the areas involving raw data which does not have any fixed format. Semi-structured data is convenient for data integration. Most data mining algorithms are not designed for semi-structured data and should at least be adapted in order to deal with such data [43].

1.4 Semi-structured webpage content extraction and mining methodologies

Existing traditional techniques and methodologies of extracting and mining semi-structured webpage content are:

- (i) Object Exchange Model (OEM) with Schema Knowledge Mapping
- (ii) Top-down Extraction
- (iii) Natural Language Processing (NLP)

(iv) Web Data Extraction Language (WDEL)

In OEM technique the extraction process is based on pre-defined specification file and the outcome is OEM object that contains individual piece of embedded web content together with information about the structure. The content and structure of the resulting OEM object are defined by the specification file. On top of OEM object, Schema Knowledge is built to query against semi-structured data [1]. Top-down approach extracts complex object from data rich websites. To extract information from set of data rich websites, description of what to extract is needed [4]. NLP requires preprocessed structured format of the semi-structured data and application of Supervised Wrapper Induction and Automatic Information Extraction in its Information Integration step [1]. Information Synthesis stage of NLP requires linguistic knowledge that attributes use of user-defined grammar rules. Systems based on WDEL converts web data to relational data model and store it in relational database. WDEL is based on Tree language tree automata. Web document is treated as the form of directed graph upon which explicit Tree language grammar is applied to map between hierarchical structure to relational data model [1][6].

1.5 User posted status on social networking sites and Psycholinguistic meaning of words

One of the most important information types on the web is web user's expression in user-posted status on Social Networking Sites like Facebook, Twitter. Personal emotions, feelings, thoughts as expressed in our daily status, posts on social networking sites are informative enough to reveal short-term and long-term affective states of mind e.g. happy, surprised, annoyed, angry, sad etc. Psycholinguistic meaning of many different words used by users to express emotions, feelings, thoughts in status updates are strong indicators of psychological traits of users. Most of the time people post status updates by writing sentences without following grammatical rule. Psycholinguistic meaning of many implicit words used in user status helps to understand the underlying causes and motivations of our reactions to different subject matter of interests, to each other within group(s) or even to surrounding situations.

1.6 Research questions

To evaluate capability and performance of existing methodologies used for extracting and mining semi-structured webpage content as well as performance of methodologies used for processing Social Networking Sites' user status updates from a researcher's point of view, a set of questions may come into mind that are but not limited to:

1. As the information and content on the websites and webpages are being updated or changed rapidly, how efficient are the existing methodologies used for extracting and mining semi-structured webpage content to process such changes during extraction and integration phase?

2. As the same piece of information may appear on different websites and webpages having differentiated hierarchical and nested HTML structure, to what extent existing techniques and methodologies are able to extract and store such kind of web data and information efficiently?

3. Existing XML based techniques have different methodologies for HTML to XML conversion or transformation in order to extract semi-structured data from websites and to map XML data with relational database but are they simple and less labor-intensive methodologies?

4. NLP techniques and classification-based text mining algorithms are generally used to identify the context of user comment by analyzing associated words used in that textual comment but do the existing NLP techniques can address and eliminate NLP problems like contextual information, word sense disambiguation, meaning extraction?

1.7 Objectives

The goal of the thesis is to implement a data warehouse-oriented semi-structured webpage content, Social Networking Sites' user status updates extraction and modeling into relational database methodology which will help us to obtain qualitative information search results in terms of producing a list of hyperlinks to most desired webpages and user posts on Social Networking Sites according to given search keywords. Following key objectives have been set up for intended thesis:

1. To apply a simplified XML-based technique of extracting webpage content that remains in semi-structured form on webpages in order to store into relational database within a data warehouse. Irrespective of differentiated nested HTML structure of webpages and rapid changes of information and content on webpages, a simplified XML-based technique is a convenient way of transforming semi-structured webpage content into exchangeable data format for direct integration with relational database.

2. To use XML to relational database schema mapping and to apply Ontology on extracted and stored web data. Ontology information of a webpage is a vocabulary representation for referring words and terms used on that webpage to indicate specific subject area or domain to which the webpage belongs.

3. To use data modeling concept and technique for efficient data management and better information retrieval process for the most desired information according to user's query. Data modeling concept provides a best way of data management and data retrieval technique for large volume of data records across database tables and datamarts within a data warehouse.

4. To use a lexical approach for identification of psycholinguistic features in Facebook user status updates and Twitter tweets for analyzing psychological traits as exhibited by users for reactions to different subject matter of interests and to surrounding situations. Lexical approach for analyzing psycholinguistic features in status updates helps us identify short-term and long-term affective states of human mind having potential connections with psychological traits.

5. To evaluate performance of the implemented data warehouse-oriented methodology and to analyze experimental test results. Evaluation of experimental test results drives us to measure quality of search results.

1.8 Thesis outline

Five chapters are included in this thesis.

Chapter 1 introduces and explains the background, research motivation and objectives of the research.

Chapter 2 briefly discusses existing accomplished research work related to semi-structured webpage content extraction and modeling into database. This chapter also discusses existing accomplished research work related to identification of emotion, trait emotionality (psychological trait) and context in user posted status on Social Networking Sites based on linguistic features of emotional words.

Chapter 3 introduces a simplified XML-based methodology for extracting webpage content in semi-structured form and presents data modeling technique for XML to relational database schema mapping.

Chapter 4 presents a lexical approach for identification of psycholinguistic features in Social Networking Sites' user status updates for analyzing emotional and psychological traits which helps us to understand the underlying causes and motivations of reactions exhibited by other users.

Chapter 5 discusses the main contribution, current exceptions and the scope for improvement as well as overall conclusion of this thesis.

Chapter 2

Related Work

2.1 Semi-structured webpage content extraction for relational database schema mapping

ANDES, an XML-based framework features a multistep extraction and transformation process of extracting semi-structured data from given websites and transforming the data into a well-structured form i.e. relational database tuples [9]. Target HTML pages are retrieved through a configuration file-based crawler and passed into extractor that attributes extensive use of Extensible Stylesheet Language Transformations (XSLT) rules on pre-processed XHTML in order to produce XML output file. XML data is then mapped to relational database schema and inserted into relational database, thereby making the extracted data available to database applications and tools for data mining. The main disadvantage of this XML-based framework is that it requires HTML to XHTML conversion and use of XSLT file to produce XML format to map with relational database schema but as websites change, an XSLT file may fail to correctly extract data from the pages that changed.

Another XML-based technique for extracting semi-structured content from webpages and storing webpage content into relational database tables is to derive object-oriented database schema for storing XML data [6][7][8]. The idea is to use the document type descriptor (DTD) to derive an object-oriented schema. DTDs describe the structure of XML documents and are considered as the schemas for XML documents. Several inlining techniques like basic, shared and hybrid solve complex DTD associated with the XML document by simplification rules. After simplification of DTD, inlined DTD graph is drawn and then generate database schema according to the inlined

DTD graph. Methodologies that use XML DTD use DTD to derive an object-oriented schema but the technique is rigid because the object-oriented schema cannot accommodate XML data that does not conform to the given DTD.

Zhoa Li and Wee Keong Ng [56] described a Web data extraction system, WICCAP and its internal Web Data Extraction Language (WDEL) to describe the web documents to be extracted, output data model and the mapping relation between them. They introduced the grammar of WDEL and the grammar of WDEL can be viewed as the schema of web documents and output data. WDEL is based on a formal tree language to provide a unified view of representing original web documents to be extract, output of extraction and transforming between them. WDEL grammars can be used as basic criteria to cluster extracted data and web documents but authors did not consider to combine the semantic information (e.g., the words extracted in output) and the schema to find relationship among web data resources.

Yuanying Mo and Tok Wang Ling [57] utilized ORA-SS (Object-Relationship-Attribute model for Semi-Structured data) as data model and used object-relational database management systems to store and manage semi-structured data. They outlined an algorithm that maps ORA-SS schema diagrams to the object-relational model. Such an algorithm demonstrates how semi-structured data can efficiently and consistently be stored in a nested relational database management system like Oracle 8i or its newer version Oracle 9i.

The methodology for extracting heterogeneous semi-structured web data from webpages and mapping into relational database implemented by Hicham Snoussi *et al.* [44] uses W4F (World Wide Web Wrapper Factory) and JEDI (Java Extraction and Dissemination of Information) wrappers that contain a set of extraction and transformation instructions, including rules and codes of control. A rule contains a syntactic constraint describing the data (character strings) to be extracted. The goal of a wrapper is to indicate how to generate a representation in XML for the extracted data. Authors used SOX (Schema for Oriented Object XML), a language of definition of schemas for XML files. SOX has been developed to overcome the insufficiencies of

DTD. The most useful case of such an extraction process is on webpages that present dynamic contents, but with fixed structures. The main disadvantage of this methodology is that users have to learn specific languages to write Wrappers additionally and the syntax of the language makes it difficult to use wrappers.

Multilevel database approach to organize web-based information focused on the idea that the lowest level of the database contains primitive semi-structured information stored in various Web repositories, such as hypertext documents [5]. At the higher level(s), metadata or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented databases.

2.2 Identification of emotion in Social Networking Sites' user status to reveal psychological traits

The emotion words people generally use provide important psychological cues to their thought processes, emotional states, intentions, and motivations. Tausczik and Pennebaker discussed that Linguistic Inquiry and Word Count (LIWC), a transparent text analysis program counts words in psychologically meaningful categories [19]. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including emotionality, social relationships, thinking styles, and individual differences. Authors have summarized some of the LIWC dimensions that reflect language correlates of attentional focus, emotional state, social relationships, thinking styles, and individual differences. This review is very much brief and selective. Word use is highly contextual and many of the findings may not hold with different groups of people or across a wide range of settings.

Srinivasu Badugu and Matla Suhasini [45] proposed a knowledge-based approach for detecting the emotion in tweets. The proposed approach enables us to classify large amounts of short texts messages into four classes of emotion: (i) Happy-Active Class, (ii) Happy-Inactive Class, (iii) Unhappy-Active Class and (iii) Unhappy-Inactive Class. Classifying short texts according to finer-grained classes of emotions provides rich and informative data about the emotional states of

individuals. The accuracy with the system is 85%. With the proposed system it is possible to understand the deeper levels of emotions. In this system authors considered only the text part and they ignored emojis, but emojis also play major role in the detection of emotion.

Some researchers applied lexical approach to identify emotions in text. For example, Strapparava and Mihalcea [46] constructed a large lexicon annotated for six basic emotions: anger, disgust, fear, joy, sadness and surprise. Through comparative evaluations of several knowledge-based and corpus-based methods, authors tried to identify the methods that work best for the annotation of emotions. As a scope of future work, this work should be extended to explore the lexical structure of emotions and integrate deeper semantic processing of the text into the knowledge-based and corpus-based classification methods. In another work, Choudhury *et al.* [47] identified a lexicon of more than 200 moods frequent on Twitter. Inspired by the circumplex model, they measured the valence and arousal of each mood through mechanical turk and psychology literature sources. Then, they collected posts which have one of the moods in their mood lexicon in the form of a hashtag at the end of a post.

Rajwa *et al.* [21] introduced basic psychological needs corpus consisting of 6,334 tweets. The corpus was annotated with emotion categories, psychological needs, levels of satisfaction of the needs, types of social contexts and life domains. As a scope of future work, the dataset authors used can be extended to cover more types of needs based on other theories and can be enriched to cover more contexts and life domains.

Wang showed how various linguistic features correlate with each personality trait and to what extent can we predict personality traits from language [22].

Results from studies by Wang *et al.* [48], Schwartz *et al.*[49] exploring the relationship between linguistic indicators extracted from Facebook and emotional well-being are mixed and inconclusive. These studies reported only weak-to-moderate relationship between psychological and linguistic indicators.

Bharat Gaiind *et al.* [60] addressed the problem of detection, classification and quantification of emotions of text collected from Twitter. Authors proposed a method to classify text into six different Emotion-Categories: Happiness, Sadness, Fear, Anger, Surprise and Disgust. Authors managed to create a large bag of emotional words, along with their emotion intensities. On testing, their model provides significant accuracy in classifying tweets taken from Twitter. Their work should be extended to establish a system for automatically updating the bag-of-words.

2.3 Conclusion

Usually the most useful and important webpage content e.g. plain texts, images, video, hyperlink texts remain inside the hierarchically nested HTML tags of webpage body. Compared to existing work done by Jussi Myllymaki, 2001 [9], Kshitija Pol *et al.*, 2008 [6], Niki Kapadia *et al.*, 2012 [7], Mustafa Atay *et al.*, 2005 [8], Hicham Snoussi *et al.*, 2001 [44] our data warehouse-oriented approach of semi-structured webpage content extraction and modeling into database introduces a simplified and less labor intensive XML-based semi-structured webpage content extraction technique that has following advantages:

Table 2.1: Advantages of our approach over major existing semi-structured webpage content extraction and database schema mapping techniques

Authors [existing research work]	Webpage content extraction and database schema mapping technique	Implemented work in this research	Advantages
Jussi Myllymaki [9]	XML-based, HTML to XHTML conversion and use of XSLT file	A data warehouse-oriented approach of semi-structured webpage content extraction and	Simplified and less labor intensive XML-based technique that does not require HTML-XHTML conversion and use of XSLT as XSLT file may fail to correctly extract content from webpages that change rapidly
Kshitija Pol <i>et al</i> [6], Niki Kapadia <i>et al.</i> [7], Mustafa Atay <i>et al.</i> [8]	XML-based, Object-oriented database schema generation using DTD for storing XML data		Does not require DTD. Instead, HTML-XML conversion for direct mapping with database table and modeling in database

Hicham Snoussi <i>et al.</i> [44]	XML-based, application of Wrapper having extraction rules and use of SOX for definition of schemas for XML files	modeling into database	Independent of Wrapper hence does not require writing extraction rules in Wrapper
-----------------------------------	--	------------------------	---

Compared to existing work done by Srinivasu Badugu and Matla Suhasini [45], Strapparava and Mihalcea [46], Wang *et al.*, 2012 [48], Schwartz *et al.*, 2014 [49] our implemented lexical approach for identification of psychological traits as exhibited by users in Social Networking Sites' user status updates has following advantages:

- As emojis play major role in the detection of emotion, we have a set of emojis majorly used in social networking sites and build a emoticon dictionary. In this dictionary we associated each emoticon to a sentiment annotation.
- To identify emotions in user status updates on Social Networking Sites, our lexical approach integrates deeper semantic processing of the emotion related words into knowledge base.
- To show relation between linguistic features and psychological traits, our lexical approach has a psycholinguistic repository that contains mapping of positive and negative opinion words, emojis, amplifier words.

Chapter 3

Semi-structured Webpage Content Extraction and Modeling into Relational database

3.1 Introduction

Webpage content in semi-structured form usually remains in hierarchically nested HTML tags of a webpage body. Unlike structured data that can be neatly modeled, organized, and formatted into relational database, semi-structured webpage content or webpage data does not attribute to have enough regular structure to qualify for the kinds of management and automation usually applied to structured data. Webpages of many Consumer-to-Consumer (C2C), Business-to-Consumer (B2C) e-commerce websites having product's feature specification, image, price etc., online financial websites having stock market data and information, online news websites attribute semi-structured content.

As the same piece of information may appear on different webpages having differentiated hierarchical and nested HTML structure, it is evident that extracting the most desired, useful webpage content from the diverse websites and modeling into relational database has become challenging. Figure 3.1 shows two differentiated webpages of two different job searching websites that contain almost the same type of information of a job advertisement but with different hierarchical and nested HTML structure. XML has become a standard for the exchange of semi-structured data. Integrating XML data into data warehouses is a hot topic. Storing and managing webpage content as XML data using relational databases is an attractive area of research for the researchers since relational databases are mature.

ca.indeed.com/jobs?q=Flexible%20Part%20Time&l=Thunder%20Bay%2C%20ON&jt=parttime&sort=date&vjk=ab44cef5e6c7136

Sort by: relevance - date

Part-time (undo)

Distance: within 50 kilometers

Salary Estimate: \$30,700+ (32), \$51,100+ (12)

Location: Thunder Bay, ON (50)

Company: Shoppers Drug Mart / Phar... (8), McDonald's (5)

Title: Part-time Crew Member (5)

Part Time Cook
Beendigen
Thunder Bay, ON
Part Time (4 on 4 off rotation). Flexible (20 hours per week). The Cook is based at Ontario Street, however all employees may be required to work across other...
Easily apply
Just posted save job more...

Part Time Cook
Beendigen - Thunder Bay, ON
Apply Now

Position Title
Cook - Part Time

Hours of work
Flexible (20 hours per week)

Location
Ontario Street

Reports to
Crisis Home Operations Administrator

Review
The Cook is reviewed during the probationary period by the

(a)

workopolis.com/jobsearch/find-jobs?ak=part+time&l=thunder+bay+ontario&lg=en&st=true&job=ZIZyb6ISDW-qsZ_S9CoBYOdeOd-ScI4pUDM_e9F1Z7s67cNbWcDZ_A

Part Time Cook
Beendigen — Thunder Bay, ON
Position Title Cook - Part Time Hours of work Flexible (20 hours per week) Location Ontario Street Reports to Crisis Home Operations Administrator...
Estimated: \$35,000 - \$46,000 a year Today

Veterinary Technician
Crossroads Veterinary Clinic — Thunder Bay, ON
Crossroads Veterinary Clinic is seeking a friendly, team orientated and enthusiastic "Veterinary Technician" to join our expanding rural practice in Thunder...
Estimated: \$42,000 - \$52,000 a year Today

Line Cook/Prep Cook
The Eagle's Landing — Thunder Bay, ON
Part-Time ** The Eagle's Landing Restaurant @ The Landmark Hotel* ** Hiring ASAP* ** Under new management and new menu to be launched mid-September*** ...
Estimated: \$25,000 - \$34,000 a year Today

Seasonal Cashier
Sephora — Thunder Bay, ON
Job ID: 173099 Store Name/Number: ON-Thunder Bay (0870)

Part Time Cook
Beendigen
Thunder Bay, ON Today Estimated: \$35,000 - \$46,000 a year Apply Now

Position Title
Cook - Part Time

Hours of work
Flexible (20 hours per week)

Location
Ontario Street

Reports to
Crisis Home Operations Administrator

Review
The Cook is reviewed during the probationary period by the
with annual evaluations conducted by the Crisis Home

(b)

Figure 3.1: Same type of information on two different webpages having different HTML structure:
 (a) “Part Time Cook” job advertisement on indeed.ca (b) “Part Time Cook” job advertisement on workopolis.com

Two types of strategies arise when we want to store XML data into relational databases and query [8]:

- (i) When the XML file is associated and validated with a Document Type Descriptors (DTD) file and the DTD or XML schema definition file is used to translate XML to database;
- (ii) When no DTD or no XML schema is associated with the XML file i.e. model mapping approach.

While the DTD-based mapping or XML schema mapping schemes [10],[11],[12],[13],[14] have presented different strategies to generate a good database schema from an XML schema, there has been no published work presenting algorithms for mapping XML files to relational data that will fit into the generated database schema and preserve the XML data order [8]. Several model mapping schemes [10],[11],[12],[15],[16],[17] are used to store and query XML documents. These approaches are not dependent on the complexity of the XML schemas or its structure.

In this chapter, section 3.2 describes a step-by-step our simplified XML-based technique for extracting webpage content in semi-structured form on websites to store into relational database. An efficient DOM-based linear data mapping approach has been followed which handles the order-preserving requirement of XML data in files to map with relational database table. It also describes use of Multidimensional Fact data model concept for efficient data management and better information retrieval technique for the most desired information according to user's query. Section 3.3 contains a brief discussion about the implementation of a web application programming interface (API) which we developed as part of the total implementation. This web API helps us to obtain qualitative information search results in terms of list of hyperlinks to most relevant webpages according to given search keywords. Section 3.4 contains experimental outputs and performance result of the web application based on our given input search keywords. In this section we also analyzed the experimental results. Section 3.5 concludes the chapter by discussing what we achieved.

3.2 A simplified technique of Semi-structured webpage content extraction and modeling into database

If we view a complete webpage as a real-world complex object, then a complex object is characterized by its attributes. The UML object diagram shown in Figure 3.2 represents a webpage as a complex object generalizing all its attributes or embedded content.

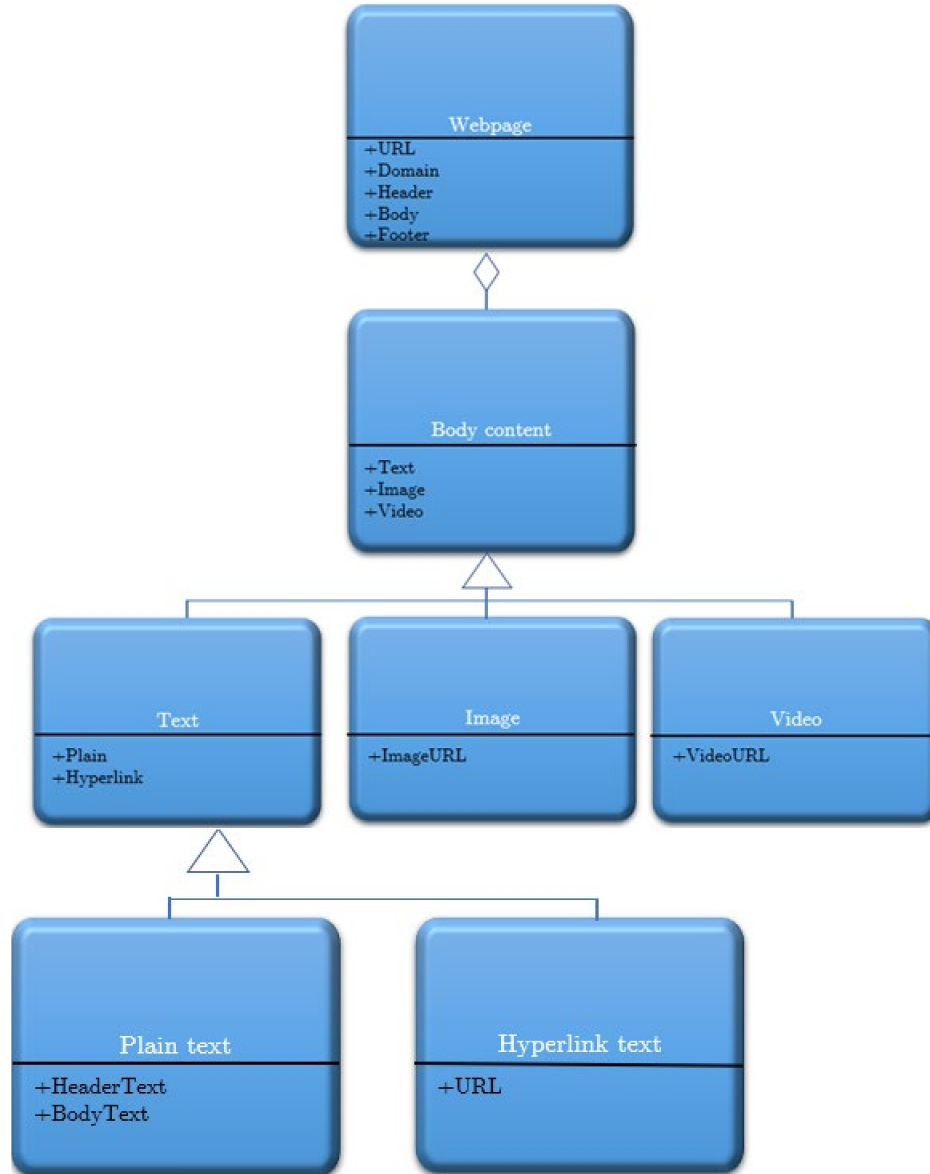


Figure 3.2: UML object diagram showing a complete webpage as a complex object

Generally, a webpage (complex object) is composed of URL address, header, body, footer sub-contents. Text, images, video contents all make up body of a webpage. Typically, two types of text remain:

- (i) Plain text
- (ii) Hyperlink text.

Types of content considered e.g. text, image, video for integration in a data warehouse all bear characteristics that can be used for indexing. We choose Oracle database Release 12.2.0.1.0 (Oracle 12c) as database for storing extracted semi-structured webpage content into database tables as per our developed data model in the database.

The overall XML-based technique for extracting semi-structured webpage content and storing webpage content into relational database is divided into four steps as described below:

3.2.1 Extracting Webpage Content out of HTML DOM Tree

To extract the most useful plain texts, image URL, video URL, hyperlink text out of the hierarchical and nested HTML, we used web2mine [50], an easy-to-use webpage extractor that extracts the content (text, hyperlink text, image URL, video URL) from webpages and transforms results into multiple data file formats e.g. XML, CSV etc. The extraction process of web2mine is divided into three main steps:

- (i) Extraction pattern specification – configuration of extraction pattern by specifying webpage regions that contain HTML structures with embedded webpage content or data
- (ii) Scheduling an automatic extraction task – scheduling an automatic extraction task at a specified time that automatically finds webpage URLs and extracts content based on extraction pattern
- (iii) Extracted webpage content export – automatic generation and exporting of data files which contain extracted content.

Figure 3.3 shows a typical configuration of an extraction pattern of an e-commerce website's webpage in web2mine. In an extraction pattern, we need to specify webpage regions which we want to extract as in figure 3.3, we marked by “Yellow” color (The high-definition figure can be zoomed in for details).

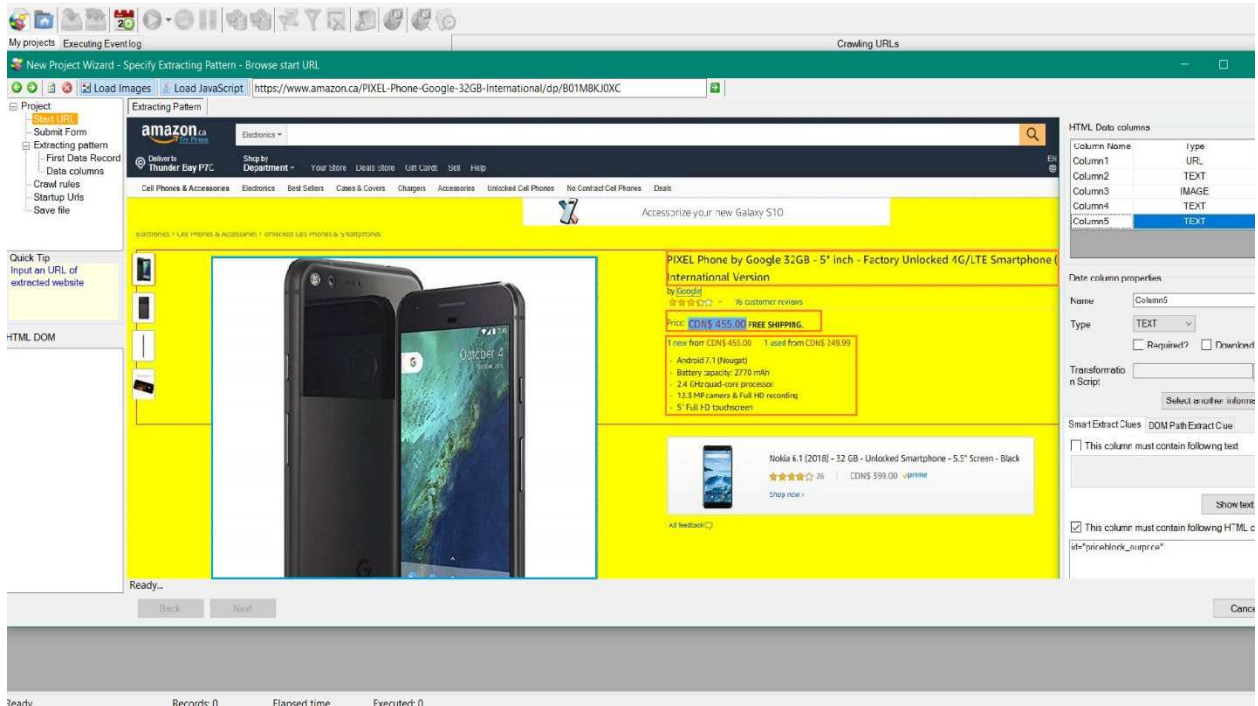


Figure 3.3: Extraction pattern specification of a webpage in web extractor's extraction pattern

(The high-definition figure can be zoomed in for details)

Webpage content of specified webpage region adheres to HTML DOM tree structure. Figure 3.4 shows part of the nested HTML structure of the webpage content for which we specified webpage regions in extraction pattern specification in Figure 3.3 and Figure 3.5 shows part of the corresponding HTML DOM structure built by the web extractor.


```

<div id="centerCol" class="centerColAlign">
  <div id="title_feature_div" class="feature" data-feature-name="title" data-cel-widget="title_feature_div"> </div>
  <div id="qpeTitleTag_feature_div" class="feature" data-feature-name="qpeTitleTag" data-cel-widget="qpeTitleTag_feature_div">
  </div>
  <div id="bylineInfo_feature_div" class="feature" data-feature-name="bylineInfo" data-cel-widget="bylineInfo feature div">
  <div id="bylineInfo_feature_div" class="feature" data-feature-name="bylineInfo" data-cel-widget="bylineInfo_feature_div"> </div>
  <div id="cmrsSummary_feature_div" class="feature" data-feature-name="cmrsSummary" data-cel-widget="cmrsSummary_feature_div">
  <ul class="a-unordered-list a-vertical a-spacing-none">
    <li>
      ::marker
      <span class="a-list-item">Android 7.1 (Nougat)</span>
    </li>
  </ul>

```

Figure 3.4: Part of the HTML structure of specified webpage region

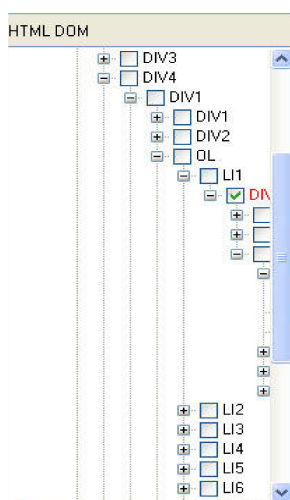


Figure 3.5: Part of HTML DOM tree built by web extractor

The web extractor has the ability to analyze the page structure and generating an output with elements grouped based on the hierarchy. It uses JAXP (from Sun Microsystems) as a DOM parser. Access to a DOM structure is done by following a path from the root in a hierarchy.

3.2.2 HTML Data to XML Conversion and XML File Generation

Web extractor is responsible for grouping data elements (text, image URL, video URL, hyperlink text) within DOM structure of specified webpage regions into tabular data format (i.e. as HTML data columns with associated values) to generate XML data file with necessary XML tags. Generally, all the data elements within a HTML “<div>” tag having ‘role’ attribute value ‘main’

are grouped together by extraction task. Based on the specified webpage regions that contain HTML “div” tag having ‘role’ attribute value ‘main’, the extraction task identifies each ‘’ tag having text and ‘’ tag having image and video URL and transforms into individual HTML data column.

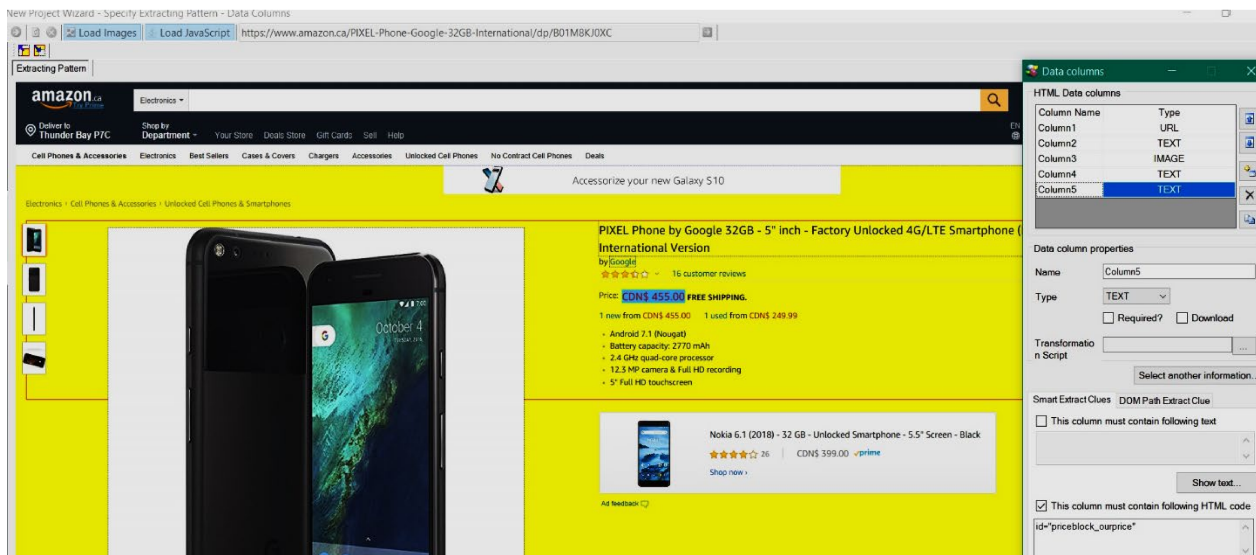

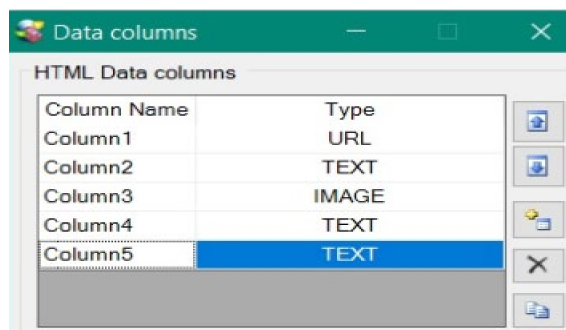


Figure 3.6: Webpage content or data elements within HTML DOM structure are grouped into tabular HTML data elements

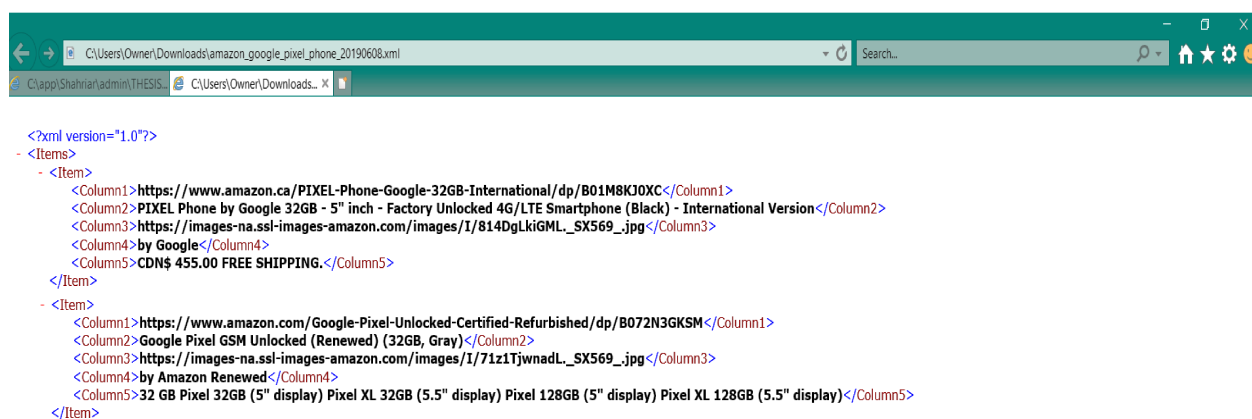
Figure 3.6 shows that webpage content or data elements within DOM structure of specified webpage regions of the webpage for which we specified extraction pattern in figure 3.3 grouped into tabular data format i.e. HTML data columns with associated values. Generally, the extraction task builds a long list of HTML data elements in tabular format. Figure 3.7 shows the partial tabular structure of HTML data elements with associated values which represents corresponding data elements in each of the smaller regions of the webpage as marked by “Orange” colored box (“”) in figure 3.3.



Column Name	Type
Column1	URL
Column2	TEXT
Column3	IMAGE
Column4	TEXT
Column5	TEXT

Figure 3.7: Partial tabular HTML data structure

Initially we scheduled extraction task to trigger at our specified time every day. It automatically starts to extract webpage content or data elements out of the HTML DOM tree, binds data elements into tabular HTML data columns with associated values to generate XML data with corresponding XML tags and finally dumps XML files into database file directory. Files usually have naming convention which is the website's organizational name in the domain name plus a keyword that indicates the type of webpage and formatted date (for example, amazon_google_pixel_phone_20190606.xml). Figure 3.8 shows part of XML data after the conversion of tabular HTML data elements.



```

<?xml version="1.0"?>
<Items>
  <Item>
    <Column1>https://www.amazon.ca/PIXEL-Phone-Google-32GB-International/dp/B01M8KJ0XC</Column1>
    <Column2>PIXEL Phone by Google 32GB - 5" inch - Factory Unlocked 4G/LTE Smartphone (Black) - International Version</Column2>
    <Column3>https://images-na.ssl-images-amazon.com/images/I/814DgLiGML_SX569_.jpg</Column3>
    <Column4>by Google</Column4>
    <Column5>CDN$ 455.00 FREE SHIPPING.</Column5>
  </Item>
  <Item>
    <Column1>https://www.amazon.com/Google-Pixel-Unlocked-Certified-Refurbished/dp/B072N3GKSM</Column1>
    <Column2>Google Pixel GSM Unlocked (Renewed) (32GB, Gray)</Column2>
    <Column3>https://images-na.ssl-images-amazon.com/images/I/71z1TjwnadL_SX569_.jpg</Column3>
    <Column4>by Amazon Renewed</Column4>
    <Column5>32 GB Pixel 32GB (5" display) Pixel XL 32GB (5.5" display) Pixel 128GB (5" display) Pixel XL 128GB (5.5" display)</Column5>
  </Item>

```

Figure 3.8: Part of XML data after conversion of tabular HTML data

3.2.3 XML data file integration, Mapping and Data loading

The scheduled extraction task of the web extractor dumps all XML files into a database file directory. We utilized one of the default data file directories of Oracle 12c, named 'DATA_PUMP_DIR'. Box 3.1 shows a database procedure insXMLWebPageContent() who is responsible for mapping XML data values of XML file with database table.

```

01 Procedure insXMLWebPageContent()
02 Begin
03   V_FILE_NAME := NULL;
04   L_NULL := NULL;
05   V_DIRECTORY := 'C:\app\Shahriar\admin\THEISISDB\dpdump\*.xml'; //Physical path of
                                                                    data file directory
06   L_DIRECTORY := V_DIRECTORY;
07   SYS.DBMS_BACKUP_RESTORE.SEARCHFILES(L_DIRECTORY, L_NULL); //Searches
                                                                    a XML file in a given directory and get full physical
                                                                    path of XML file
08   For each x IN (select fname_krbmsft as DIRECTORY_FILE_NAME from x$krbmsft)
09     do DBMS_OUTPUT.PUT_LINE(x.DIRECTORY_FILE_NAME)
10   End For
11   Insert into Table1 Select fname_krbmsft as DIRECTORY_FILE_NAME From x$krbmsft
12     Select SUBSTR(DIRECTORY_FILE_NAME,(INSTR(DIRECTORY_FILE_NAME,'\',-
1,1)+1),length(DIRECTORY_FILE_NAME)) into V_FILE_NAME from Table1
13   Insert into Table2
with tt as
(Select(Select xmltype(bfilename('DATA_PUMP_DIR',V_FILE_NAME),
nls_charset_id('UTF8')) xmlcol From Dual)
From tt, TABLE(XMLSequence(extract(xmlcol,'/Items/Item'))))
14   UTL_FILE.FREMOVE('DATA_PUMP_DIR', V_FILE_NAME) //Removes XML file after
                                                                    data loading
15 End

```

Box 3.1: Database procedure for mapping XML data values of XML file with database table

Box 3.2 shows a database procedure insTransactTblWebContent() which is responsible for loading mapped XML data into an intermediate transactional database table named as THEISISBDW.THEISISBDW.DWR_TRANSACT_TBL_WP_CONTENT.

```

01 Procedure insTransactTblWebContent()
02 Begin
03   generateKeywordFromURL(); //Generates keywords from full webpage URL address value
04   Insert into table THESISDBDW.DWR_TRANSACT_TBL_WP_CONTENT values
      ((Select Table3.* from Table3), (Select Table2.* from Table2))
05 End

01 Procedure generateKeywordFromURL()
02 Begin
03   For each x IN (Select KEYWORDS from (select
      regexp_replace(regexp_replace(regexp_substr(column1,'[/|]+' , 1, level),
      '[^0-9A-Za-z]',' ','') {2,}' , ' ') as KEYWORDS from Table2
      connect by regexp_substr(column1,'[/|]+' , 1, level) is not NULL) From Dual)
04   do
05     Create a tuple of table Table3 as
      Select LIST_ELEMENT(x. KEYWORDS, 1, ' '), LIST_ELEMENT(x. KEYWORDS, 2, ''),...,
      LIST_ELEMENT(x. KEYWORDS, length(KEYWORDS), ' ') From Dual//LIST_ELEMENT
      is a data structure for storing ordered sets of
      items
06   End For
07 End

```

Box 3.2: Database procedure for loading mapped XML data into an intermediate transactional database table

Figure 3.9 shows partial view of a data record in the intermediate transactional database table THESISDBDW.DWR_TRANSACT_TBL_WP_CONTENT.

URL_DOMAIN_NAME	URL_KEYWORD_1	URL_KEYWORD_2	URL_KEYWORD_3	URL_KEYWORD_4	URL_KEYWORD_5	WP_CONTENT_DATA_VAL_1	WP_CONTENT_DATA_VAL_2
1 www.amazon.com	Google	Pixel	Phone	128	GB	https://www.amazon.com/Google-Pixel-Phone-128-GB/dp/B01M1Y19IH	Google Pixel Phone 128 GB - 5 inch Display (Factory Unlocked US W

Figure 3.9: View of a data record in intermediate transactional database table

3.2.4 Data Modeling

For efficient data management, indexing of data records of heterogeneous semi-structured webpage content (text, image URL, video URL) and better data retrieval process, we utilized

multidimensional fact data modeling technique [52]. Multidimensional fact data model provides a logical model of organizing data records in database tables that eventually improves database query processing time on stored data records for better data retrieval process. In multidimensional fact data model, generally there are two types of database tables: (i) Fact table and (ii) Dimensional tables. Each data record in fact table generally contains foreign keys which are eventually primary keys of two or more individual dimensional tables.

Figure 3.10 shows partial view of our multidimensional fact data model. We choose Oracle SQL Developer version 18.4.0 (64-bit) to build our multidimensional fact data model.

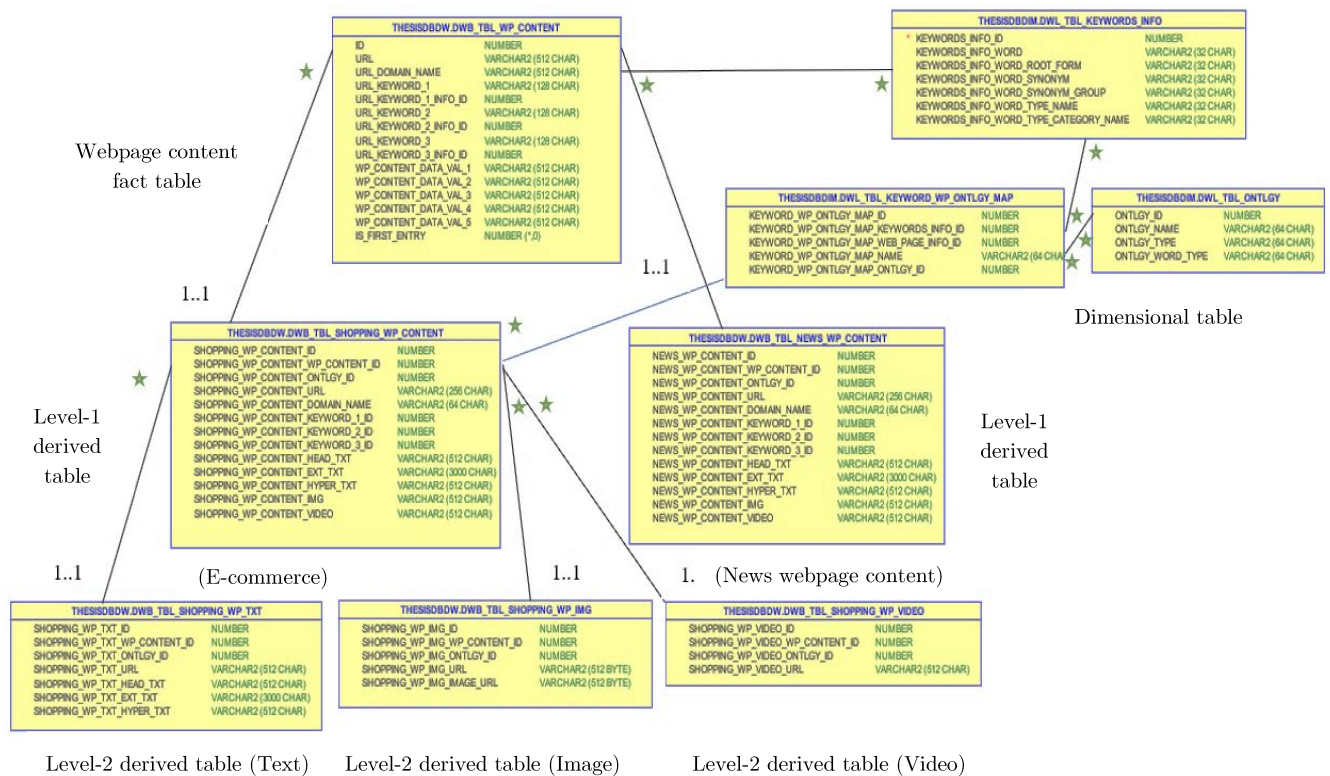


Figure 3.10: Multidimensional fact data model for semi-structured webpage content database storage (*The high-definition figure can be zoomed in for details*)

We maintained a base fact table named as **THESSISDBD.DWB_TBL_WP_CONTENT** to which we fed data records of intermediate transactional database table **THESSISDBD.DWR_TRANSACTION_TBL_WP_CONTENT** as mentioned in section 3.2.3.

Figure 3.11 shows partial view of data records in the base fact table THESISDBDW.DWB_TBL_WP_CONTENT.

	URL	URL_DOMAIN_NAME	URL_KEYWORD_1	URL_KEYWORD_1_INFO_ID	URL_KEYWORD_2	URL_KEYWORD_2_INFO_ID	URL_KEYWORD_3	URL_KEYWORD_3_INFO_ID	WP_CONTENT
1	https://www.amazon.ca/PIXEL-Phone-Google-32GB-International/dp/B01M8KJ0XC	www.amazon.ca	PIXEL	1	Phone	3	Google	5	https://www.
2	https://www.ebay.ca/itm/Google-Pixel-32GB-Factory-Unlocked-4G-LTE-Android-N...	www.ebay.ca	Google	5	Pixel	1	32GB	8	https://www.
3	https://www.amazon.com/Google-Pixel-Unlocked-Certified-Refurbished/dp/B072N...	www.amazon.com	Google	5	Pixel	1	Unlocked	9	https://www.
4	https://www.walmart.com/ip/Google-Pixel-Phone-128-GB-5-inch-Display-Factory...	www.walmart.com	Google	5	Pixel	1	Phone	3	https://www.
5	https://www.newegg.com/p/238-001E-00024	www.newegg.com	Google	5	Pixel	1	32 GB	8	https://www.
6	https://slickdeals.net/f/11853511-google-pixel-32gb-128gb-brand-new-factory...	www.slickdeals.net	unknown	99999	google	5	pixel	1	https://slic
7	https://www.ctvnews.ca/canada/thunder-bay-ignored-community-wish-for-outsid...	www.ctvnews.ca	thunder bay	11	ignored	(null)	(null)	(null)	https://www.
8	https://www.cbc.ca/news/canada/thunder-bay/bortuzzo-stanley-cup-1.5210317	www.cbc.ca	bortuzzo	16	stanley-cup	17	thunder bay	11	https://www.
9	https://www.thnewswatch.com/local-news/firefighters-respond-to-man-submerge...	www.thnewswatch...	firefighters	18	respond	19	man	20	https://www.
10	https://globalnews.ca/news/5476939/bombardier-lay-off-thunder-bay/	globalnews.ca	bombardier	21	(null)	(null)	thunder bay	11	https://glob

Figure 3.11: Partial view of data records in THESISDBDW.DWB_TBL_WP_CONTENT base fact table (*This figure can be zoomed in for details*)

In our multidimensional fact data model, there are several dimensional tables. Figure 3.12 shows structure of three important dimensional tables in our data model.

THEISISBDDIM.DWL_TBL_KEYWORDS_INFO	
* KEYWORDS_INFO_ID	NUMBER
KEYWORDS_INFO_WORD	VARCHAR2 (32 CHAR)
KEYWORDS_INFO_WORD_ROOT_FORM	VARCHAR2 (32 CHAR)
KEYWORDS_INFO_WORD_SYNONYM	VARCHAR2 (32 CHAR)
KEYWORDS_INFO_WORD_SYNONYM_GROUP	VARCHAR2 (32 CHAR)
KEYWORDS_INFO_WORD_TYPE_NAME	VARCHAR2 (32 CHAR)
KEYWORDS_INFO_WORD_TYPE_CATEGORY_NAME	VARCHAR2 (32 CHAR)

THEISISBDDIM.DWL_TBL_WEB_PAGE_INFO	
* WEB_PAGE_INFO_ID	NUMBER
WEB_PAGE_INFO_NAME	VARCHAR2 (32 CHAR)
WEB_PAGE_INFO_DOMAIN_NAME	VARCHAR2 (64 CHAR)
WEB_PAGE_INFO_TYPE	VARCHAR2 (32 CHAR)

THEISISBDDIM.DWL_TBL_ONTLGY	
ONTLGY_ID	NUMBER
ONTLGY_NAME	VARCHAR2 (64 CHAR)
ONTLGY_TYPE	VARCHAR2 (64 CHAR)
ONTLGY_WORD_TYPE	VARCHAR2 (64 CHAR)

Figure 3.12: Data structures of dimensional tables

Each data record in base fact table THESISDBDW.DWB_TBL_WP_CONTENT contains foreign key to dimensional table THESISBDDIM.DWL_TBL_KEYWORDS_INFO. Dimensional table THESISBDDIM.DWL_TBL_KEYWORDS_INFO is a keyword dictionary

table that contains information of keywords which usually appear in webpage URL. Dimensional table THESISDBDIM.DWL_TBL_ONTLGY is a formal vocabulary representation of Ontology for referring to the keywords related to webpages of a wide range of subject areas or domains. Figure 3.13 shows partial view of data records in THESISDBDIM.DWL_TBL_ONTLGY table.

	ONTLGY_ID	ONTLGY_NAME	ONTLGY_TYPE	ONTLGY_WORD_TYPE
1	1	shopping	business	product
2	2	news	information	(null)
3	3	finance	trade	stock
4	4	scientific	science	(null)
5	5	generalize	general	(null)
6	6	shopping	business	brand

Figure 3.13: Partial view of data records in Ontology dimensional table

THESISDBDIM.DWL_TBL_ONTLGY

For mapping between keyword and Ontology, we built dimensional table THESISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP. Figure 3.14 shows formation of dimensional table THESISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP. This table is a combination of all data records in keywords and Ontology dimensional tables.

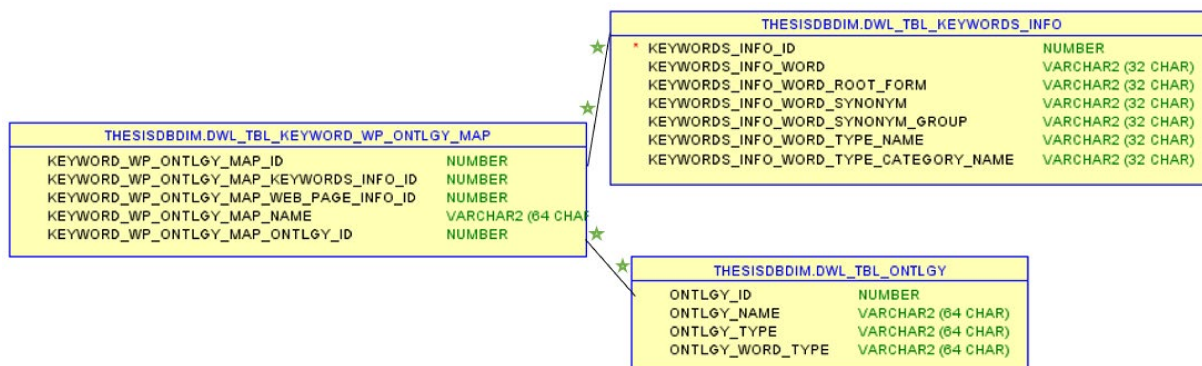
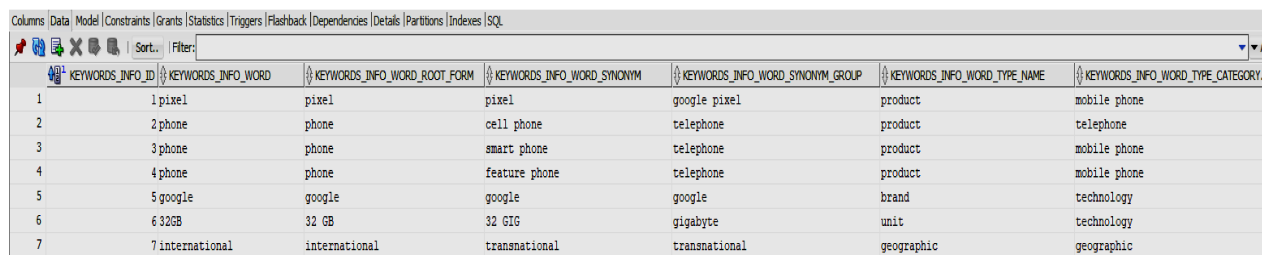


Figure 3.14: Formation of dimensional table

THESISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP

The base fact table THESISDBDW.DWB_TBL_WP_CONTENT contains all the data records of webpage content of many different websites. Based on meaningful keywords found in a webpage's URL and Ontology information of a website i.e. specific subject area to which a website belongs (for example, e-commerce, news, finance), we derived Level-1 derived tables from fact table THESISDBDW.DWB_TBL_WP_CONTENT. In figure 3.3 we specified extraction pattern task in web extractor to extract content of a webpage having URL <https://www.amazon.ca/PIXEL-Phone-Google-32GB-International/dp/B01M8KJ0XC>.

Meaningful keywords in URL after the domain name are 'PIXEL', 'Phone', 'Google', '32GB', 'International'. Figure 3.15 shows how we kept these keywords as data records in table THESISDBDIM.DWL_TBL_KEYWORDS_INFO.



KEYWORDS_INFO_ID	KEYWORDS_INFO_WORD	KEYWORDS_INFO_WORD_ROOT_FORM	KEYWORDS_INFO_WORD_SYNONYM	KEYWORDS_INFO_WORD_SYNONYM_GROUP	KEYWORDS_INFO_WORD_TYPE_NAME	KEYWORDS_INFO_WORD_TYPE_CATEGORY...
1	1 pixel	pixel	pixel	google pixel	product	mobile phone
2	2 phone	phone	cell phone	telephone	product	telephone
3	3 phone	phone	smart phone	telephone	product	mobile phone
4	4 phone	phone	feature phone	telephone	product	mobile phone
5	5 google	google	google	google	brand	technology
6	6 32GB	32 GB	32 GIG	gigabyte	unit	technology
7	7 international	international	transnational	transnational	geographic	geographic

Figure 3.15: View of data records of keywords in table THESISDBDIM.DWL_TBL_KEYWORDS_INFO

In figure 3.15 we can see the unique column value of 'KEYWORDS_INFO_WORD_TYPE_NAME' column in table THESISDBDIM.DWL_TBL_KEYWORDS_INFO is 'product' and the unique column value of 'KEYWORDS_INFO_WORD_TYPE_CATEGORY' is 'mobile phone'. In figure 3.13 we can see the ontological meaning associated with the word type 'product' is 'shopping'. Then we mapped all the keywords having word type category 'mobile phone' with 'shopping' in THESISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP table. So, as we stored all the content of the webpage having URL <https://www.amazon.ca/PIXEL-Phone-Google-32GB-International/dp/B01M8KJ0XC> in table THESISDBDW.DWB_TBL_WP_CONTENT as a single data record, subsequently we stored a single data record in Level-1 derived table

THEISISDBDW.DWB_TBL_SHOPPING_WP_CONTENT. Figure 3.15 shows how we stored these keywords as these data records in table THEISISDBDW.DWB_TBL_SHOPPING_WP_CONTENT. We maintained indexing of data records of all Level-1 derived tables in fact table THEISISDBDW.DWB_TBL_WP_CONTENT. Indexing improves the speed of data retrieval operations on a database table at the cost of additional writes and storage space.

SHOPPING_WP_CONTENT_ONTLGY_ID	SHOPPING_WP_CONTENT_URL	SHOPPING_WP_CONTE...	SHOPPING_WP_CONTENT_KEYWORD_I_ID	SHOPPING_WP_CONTENT_KEYWORD...	SHOPPING_WP_CONTENT_KEYW...	SHOPPING_WP_CONTENT_HEAD_TXT
1	1 https://www.amazon.com/Google-Pixel-Unl...	www.amazon.com	5	1	9	Google Pixel GSM Unlocked (Renewed) (32G
2	1 https://www.walmart.com/ip/Google-Pixel...	www.walmart.com	5	1	9	Google Pixel Phone 128 GB - 5 inch Displ
3	1 https://www.newegg.com/p/238-001E-00024	www.newegg.com	5	1	8	Google Pixel 32GB (Factory Unlocked) 5-i
4	1 https://slickdeals.net/f/11853511-googl...	www.slickdeals.net	99999	5	1	Google Pixel 32GB/128GB Brand New Factor
5	1 https://www.amazon.ca/PIXEL-Phone-Googl...	www.amazon.ca	1	3	5	PIXEL Phone by Google 32GB - 5" inch - F
6	1 https://www.ebay.ca/itm/Google-Pixel-32...	www.ebay.ca	5	1	8	Details about Google Pixel 32GB Factory

Figure 3.16: Partial view of data records in THEISISDBDW.DWB_TBL_SHOPPING_WP_CONTENT Level-1 derived table

Each data record of a Level-1 derived table contains foreign key to dimensional table THEISISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP. Each data record of a Level-1 derived table can be retrieved using keywords-Ontology mapping information in dimensional table THEISISDBDIM.DWL_TBL_KEYWORD_WP_ONTLGY_MAP as it is in Figure 3.10. Furthermore, text, image URL, video URL all make up each data record in Level-1 derived table, so we derived individual Level-2 derived tables from Level-1 derived tables DWB_TBL_SHOPPING_WP_TXT, DWB_TBL_SHOPPING_WP_IMG, DWB_TBL_SHOPPING_WP_VIDEO as we showed in figure 3.10 for storing text, image, video content in each of those Level-2 derived table.

3.3 Implementation of Web Application

In order to obtain qualitative information search results in terms of list of hyperlinks to most relevant webpages according to given search keywords, we developed a web application as part of the total implementation. Our web application has a web API which is responsible for transferring

user given input search keywords to a web server that ultimately transfers forward to the database to retrieve data records. The web API is also responsible for displaying list of hyperlinks to most relevant webpages as search results that the database sends back to the application interface through the web server.

3.3.1 Tool selection

Table 3.1: List of web API development tool used

Name	Versions
Visual Studio Community 2019	16.1.29001.49

Table 3.2: List of web API development technology and libraries used

Technology	
Name	Versions
ASP.Net (C#) Core Application for Windows Web API	4.0.30319
Libraries and assemblies	
Name	Versions
Microsoft .Net Framework	4.0.30319
Oracle Data Provider for .Net, 64-bit (Oracle.ManagedDataAccess)	4.122.1.0

Table 3.1 shows the tool we used to develop the web API. Microsoft's Visual Studio Community 2019 is an integrated development environment (IDE) framework for developing ASP.NET Core application that features communication between client-side (Browser) to database and vice-versa through a web server. Table 3.2 shows technology and libraries we used to develop the web API. To develop a database driven web application, we choose ASP.Net with C# code-behind that provides a wide range of client-side tools to design browser-based user interface as well as supports built-in and user-defined functions and methods written in C#. We utilized class libraries provided by .Net framework 4.0 to build C# code and to run the web application. We also utilized Oracle.ManagedDataAccess assembly provided by Oracle Data Provider for Microsoft's .NET

development platform. Oracle.ManagedDataAccess assembly provides a large collection of class libraries used to access data stored in Oracle database.

3.3.2 Web API implementation

As part of the web API, the graphical user interface (GUI) that we designed to run on a web browser has specific input area (Textbox) to allow users to enter input search keywords to find a specific information contained on a webpage. Figure 3.16 shows the GUI of the web API where a user enters input search keywords to find his/her desired information.

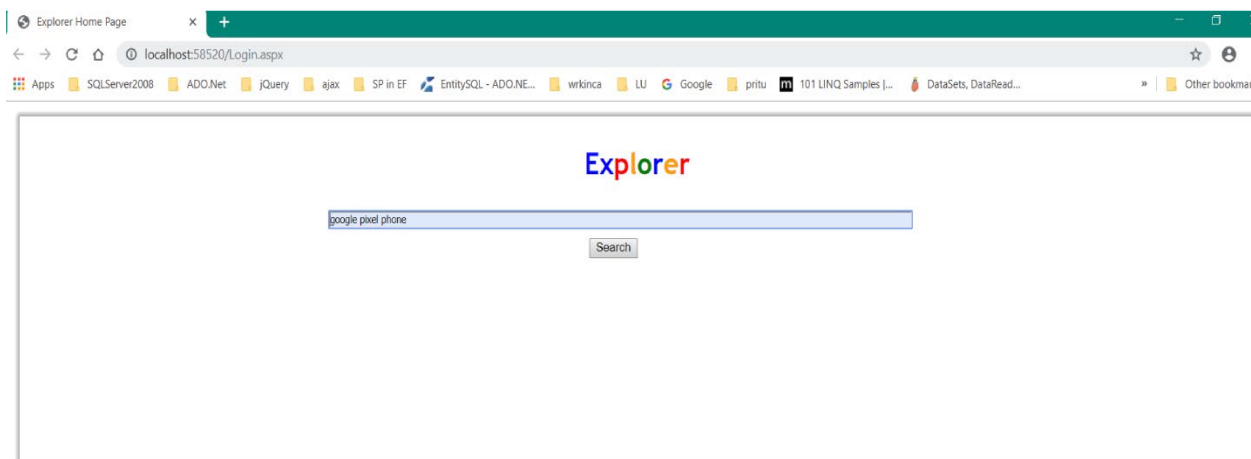


Figure 3.17: GUI of web API

The input search keywords form a string that goes through sequential multiple user-defined functions written in C# as described below:

- **GetSearchKeywordType** (string[] keywords): This method takes input search keywords as array of string once the input search keywords are splitted and white spaces among the splitted keywords are removed by built-in Split() function.

GetSearchKeywordType (string[] keywords) function passes array of string to data access layer function **SearchKeywordType**(string[] keywords) that finds unique KEYWORDS_INFO_WORD_TYPE_NAME column value of dimensional table THESISDBDIM.DWL_TBL_KEYWORDS_INFO in the database.

- **GetSearchKeywordInfoAndOntology** (string keywords_info_word_type): This method takes keyword word type (single word) as input identified in first step and passes to data access layer function **SearchKeywordInfoAndOntology** (string keywords_info_word_type) that finds all the ids of corresponding each search keywords, unique KEYWORD_WP_ONTLGY_MAP_NAME column value and corresponding keyword-ontology mapping id in keyword-ontology mapping dimensional table.
- **GetWebPageSearchResult**(int[] keywords_id, int keyword_ontology_map_id, string ontology_map_name): Once keyword-ontology mapping id is identified in step two, this method passes necessary keyword id values stored in an integer array, keyword-ontology mapping id and ontology mapping name to data access layer function **SearchWebPage** (int[] keywords_id, int keyword_ontology_map_id, string ontology_map_name) that finds corresponding Level-1 derived table name in the database which contains ontology mapping name, retrieves all the data records stored in the Level-1 derived table and returns search result to user interface to show as list of hyperlinks to webpages that contain texts, images URL, video URL closely related to search keywords.

Figure 3.18 shows search result obtained for given input search string ‘google pixel phone’. Unique word type value and word type category value for ‘google’ ‘pixel’ ‘phone’ are ‘product’ and ‘mobile phone’. The search result below shows all the hyperlinks to webpages of shopping domain (ontology) that contain search query ‘google pixel phone’.

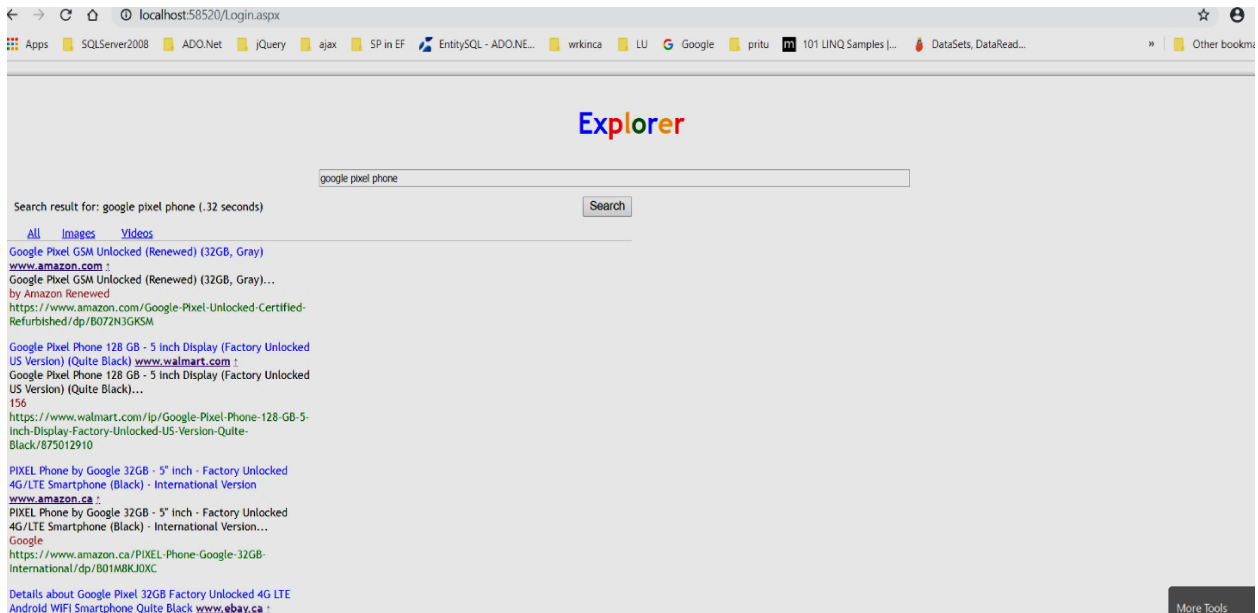


Figure 3.18: Search results for search query ‘google pixel phone’

3.4 Experimental Execution and Results

In order to obtain experimental output as well as performance result of our experiment, we executed the web application several times on a local web server ‘IIS 7’ running on a local computer. To obtain desired search results according to our given search keywords, we provided three combinations of search keywords as inputs to the web application for three experimental attempts.

3.4.1 Experimental output

We fixed three combinations of search keywords as inputs:

‘google pixel’

‘google pixel phone’

‘google pixel phone 32 GB’

Figure 3.19 shows as list of hyperlinks to all relevant webpages that contain text, image, video content having words ‘google’, ‘pixel’ embedded with content.

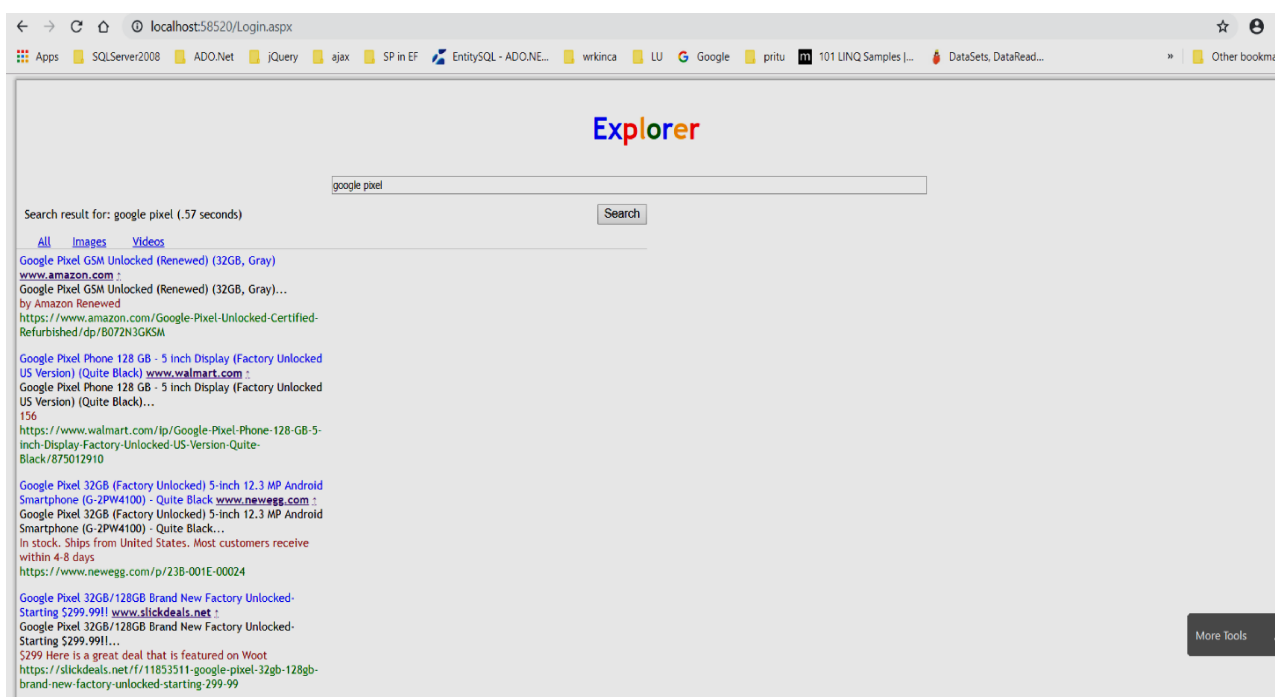


Figure 3.19: Search result for input search query 'google pixel'

Figure 3.20 shows search result as list of hyperlinks to most relevant webpages appear first that contain text, image, video content having words 'google', 'pixel', 'phone' embedded with content.

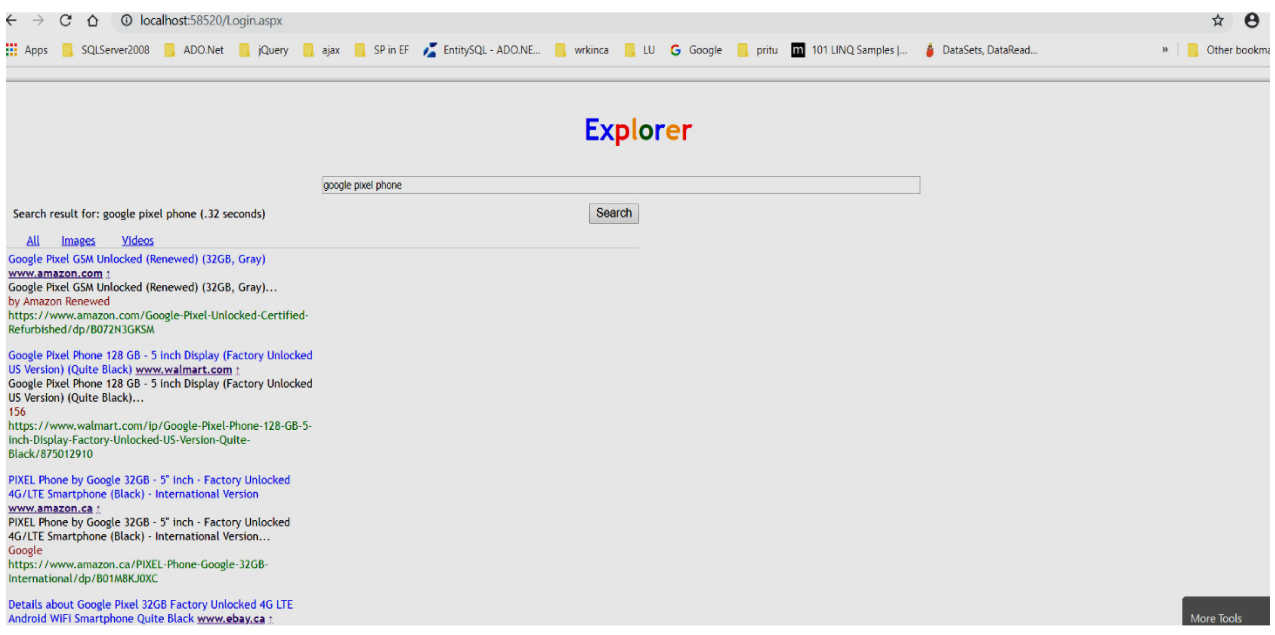


Figure 3.20: Search result for input search query 'google pixel phone'

Figure 3.21 shows search result as list of hyperlinks to most relevant webpages appear first that contain text, image, video content having words ‘google’, ‘pixel’, ‘phone’, ’32GB’ embedded with content.

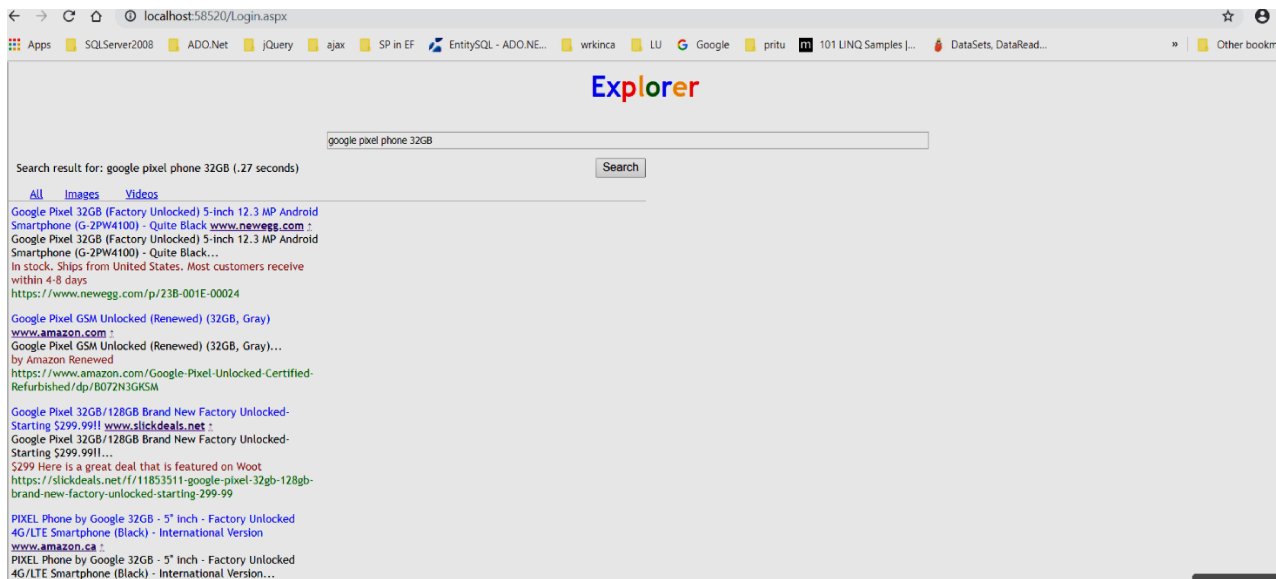


Figure 3.21: Search result for input search query ‘google pixel phone 32GB’

Figure 3.22 shows search result as list of hyperlinks to most relevant webpages appear first that contain image content having words ‘google’, ‘pixel’, ‘phone’, ’32 GB’ embedded with content.

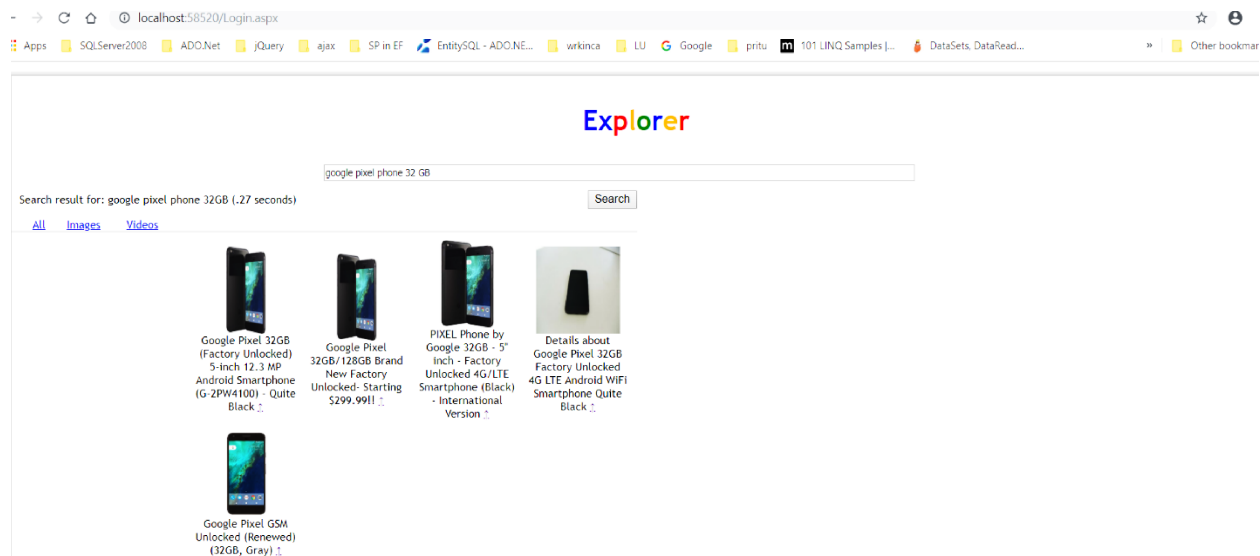


Figure 3.22: Image search result for input search query 'google pixel phone 32GB'

3.4.2 Performance Result and Analysis of Result

For each of the three experimental outputs in section 3.4.1, we recorded the time taken for the web application that passes input search keywords as search query to execute on stored data records in database for retrieval of data records closely related to search keywords and displaying the data records as search result (list of hyperlinks to relevant webpages). We used utility function **TimeSpan()** that calculates the time interval (in milliseconds) between system time and the start time when the input search keywords as a single string value go through sequential multiple user-defined functions written in C# to retrieve data records stored in database tables. Table 3.3 shows recorded time taken in each of the three experimental outputs.

Table 3.3. Time taken in each experimental test

TIME (Milliseconds)	TIME (in Second)	No. of search keywords
568	0.57	2
320	0.32	3
270	0.27	4

Figure 3.23 shows graphical representation of time taken in each of the three experimental outputs.

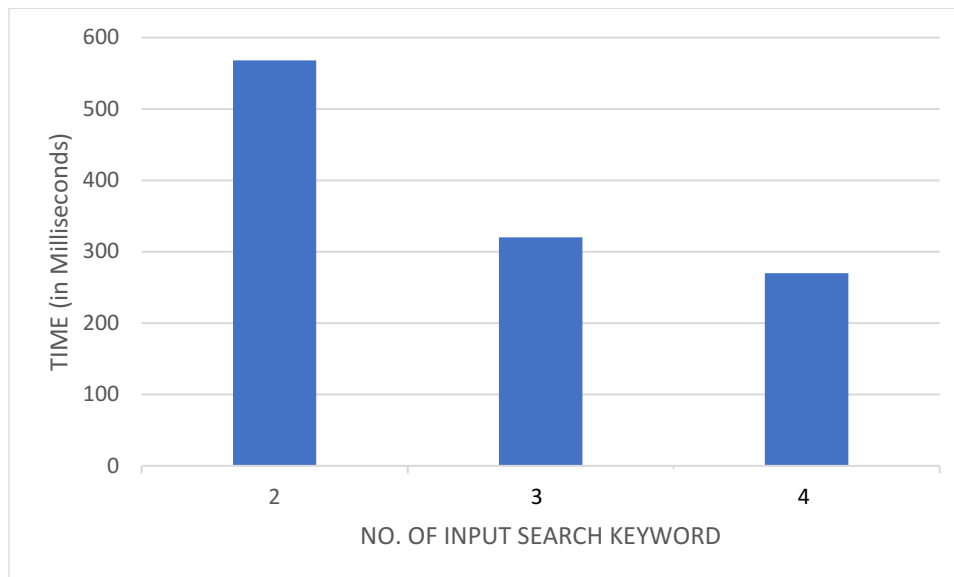


Figure 3.23: Graphical representation

In figure 3.23 of the graphical representation, we see that the time taken for the web application that passes input search string 'google pixel' (no. of search keywords=2) through sequential multiple user-defined functions to retrieve data records as search result (list of hyperlinks to relevant webpages) from database table is 0.57 seconds and the search result (figure 3.19) contain all the hyperlinks to relevant webpages that contain text, image, video content having words 'google', 'pixel' embedded with content whereas the time taken for the application that passes input search string 'google pixel phone' (no. of search keywords=3) through sequential multiple user-defined functions to retrieve data records as search result is 0.32 seconds and the search result (figure 3.20) contain hyperlinks to most relevant webpages appeared first that contain text, image, video content having words 'google', 'pixel', 'phone' embedded with content. Furthermore, when the web application passes input search string 'google pixel phone 32GB' (no. of search keywords=4) through sequential multiple user-defined functions to retrieve data records as search result, it displays hyperlinks to most closely and desired webpages appeared first (in figure 3.21) that contain text, image, video content having words 'google', 'pixel', 'phone', '32GB' embedded with content and takes relatively lowest time 0.27 seconds. Using the unique word type value

‘mobile phone’ of keywords ‘google’, ‘pixel’, ‘phone’, ‘32GB’ we mapped these search keywords with ontology name ‘shopping’. So instead of searching all the data records in THESISDBDW.DWB_TBL_SHOPPING_WP_CONTENT Level-1 derived table, the query passed to the database searches only the data records that contain particular keyword-ontology mapping id of ‘shopping’ and return the list of hyperlinks of most closely relevant webpages appearing at first in search result.

Generally, an important metric of search result performance for any keyword-based web search application is the time taken for the web application that passes input search keywords or search query to execute on stored data records in a database for retrieval of data records of semi-structured webpage content. Compared to other web search query systems that utilizes different SQL-like languages to execute on database stored records of semi-structured webpage content, our web application attributes to take an overall good less amount of average time to execute on stored data records in a database for retrieval of data records of semi-structured webpage content.

Table 3.4: Performance of search query execution compared to other web search query applications of Semi-structured webpage content

Web search query applications of Semi-structure webpage content search	Execution time range of search query	Implemented web application in this research work	Execution time range of search query
WebSQL [58]	from 10 seconds to above 1 minute	Explorer	from 10 seconds to 60 seconds
WebDB [59]	from 10 seconds to 70 seconds		

3.5 Conclusion

In our multidimensional fact data model, we maintained a base fact table that contains data records of webpage content in semi-structured form of many different websites. Based on meaningful keywords found in a webpage’s URL, Ontology information of a webpage i.e. specific subject area to which that webpage belongs (for example, e-commerce, news, finance), we derived Level-1 derived tables from fact table. We maintained indexing of data records of all Level-1

derived tables in base fact table. Indexing improves the speed of data retrieval operations. So, the more efficiently we do mapping of more and more keyword and ontology as well as indexing of data records in database tables, the better search result with high speed of data retrieval time we can expect.

Chapter 4

Processing Social Networking Sites' User Status to Reveal User's Psychological Trait

4.1 Introduction

Facebook and Twitter are the two most popular Social Networking Sites as Facebook, the biggest Social Network Site worldwide, has 2.41 billion monthly active users as of the second quarter of 2019¹ and Twitter has over 321 million active monthly users². People frequently use Facebook and Twitter to broadcast their activities and share their opinions about a wide variety of topics and events. Social Networking Site's user status posted as public is informative enough to know user's daily thoughts, feelings, emotions through textual self-description. User status on Social Networking Sites like Facebook and Twitter represent a vast and relatively new source of ecological data with potential connections with individual behavioral and psychological characteristics or traits. Traits reflect people's characteristic patterns of thoughts, feelings, and behaviors [53]. Traits are based on Trait theory in personality psychology. Trait theory is an approach to the study of human personality [54]. Trait theorists are primarily interested in the measurement of traits, which can be defined as habitual patterns of behavior, thought, and emotion. Traits are something a person either has or does not have, but in many others traits have dimensions such as positive (good behaviors) and negative (bad behaviors) [55].

User status on Social Networking Sites like Facebook and Twitter is a rich source of information to help understand the behaviors and affective states of individuals. Some of the existing research

¹ <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/> Date accessed: August 7, 2019.

² <https://en.wikipedia.org/wiki/Twitter>. Date accessed: August 7, 2019.

focuses on general dimensions sentiment analysis and opinion mining, using three polarity categories: positive, negative and neutral [31], [32]. Balabantaray *et al.*[33], Yan *et al.* [34] went beyond the general sentiment analysis and deeply recognize distinct and dimensional emotional categories. Golbeck *et al.* [35], Bollen *et al.*[36] have also explored more distinguishable long-term affective states such as personality and mood.

In this chapter, section 4.2 describes a lexical approach for identification of emotional and psychological cues expressed in extracted Facebook user status updates and Twitter tweets. This section describes use of Multidimensional Fact data model concept [52] for efficient data management and better information retrieval technique for the most desired search result according to user's query. Section 4.3 contains a brief discussion about the implementation of web API module that retrieves people's tweets and status updates from our datamart of processed Facebook user status and Twitter tweet data records based on mapping of opinion words, emojis with psychological traits. Section 4.4 contains experimental outputs and evaluation of experimental test results. Section 4.5 concludes the chapter by discussing what we achieved.

4.2 Implemented Lexical approach for Identification of Emotional and Psychological Traits in Facebook user status updates and Twitter tweets

The overall Facebook user status updates, Twitter tweets dataset extraction as well as emotional, psychological cues identification in extracted Facebook user status updates and Twitter tweets process is divided into four steps as described below:

4.2.1 Data extraction

People frequently use the two top-most popular Social Networking Sites Facebook and Twitter to broadcast their activities and share their opinions about a wide variety of topics and events. Facebook user status updates and Twitter tweets posted as public are informative enough to know user's daily thoughts, feelings, emotions through textual self-description. We collected Facebook dataset of 9,917 status updates of Facebook's English speaking active users through Facebook

Graph API¹ and this dataset contains 9,917 publicly posted user status between January 7, 2019 to January 8, 2019 having words used to express emotions that represent below LIWC's² (Linguistic Inquiry and Word Count) [19] 4 subcategories of human psychological states (out of 64 different subcategories in the LIWC dictionary):

Table 4.1: Four subcategories of LIWC dictionary for human psychological states

LIWC Category	LIWC Subcategory	Example word
Affective processes - Positive emotion	Positive emotion	Love, Amazing
Affective processes - Negative emotion	Anxiety	Tensed
	Anger	Hate
	Sadness	Sad

LIWC includes a semantic dictionary that maps words and word stems to a set of 64 different categories of emotional and cognitive aspects of human behavioral traits [51]. To obtain publicly posted user status updates or page posts, we followed API calling method for Facebook object type 'post'³. We specified Facebook object type 'post' having 'message' field containing example words as listed in Table 4.1. During extraction of user status updates through Graph API, we specified our desired fields to remain in the Facebook dataset. The Facebook dataset we collected was in JSON format. We converted JSON datafile into CSV file format with data fields as listed in Table 4.2 below.

¹ <https://developers.facebook.com/docs/graph-api/overview>

² https://www.researchgate.net/publication/282124505_The_Development_and_Psychometric_Properties_of_LIWC2015

³ <https://developers.facebook.com/docs/graph-api/reference/v4.0/post>

Table 4.2: Data fields in extracted Facebook dataset and their descriptions

Data column name	Description
from	Official Facebook author id
id	Unique Facebook post id
target	Facebook display name on the profile
message	Status
link	Link attached to user's public post

Using **nlk.twitter** v.3.4 python library [40], we collected Twitter dataset of recent 3,240 tweets between January 7, 2019 to January 8, 2019 (Twitter live public stream) which is a sample of all Tweets published by active users during January 7, 2019 to January 8, 2019 having examples words as listed in Table 4.1. We utilized **Twitter** class¹ of **nlk.twitter** package and we used **tweets()** function of **Twitter** class to specify example words (listed in Table 4.1) as keyword to pass through **tweets()**. Tweets retrieved from the Twitter Search API were in JSON which a simple structured text format and we used **nlk.twitter.common.json2csv**² utility function of **nlk.twitter** package to extract selected fields from a file of line-separated JSON tweets and to write in a file of CSV format. Data fields in the Twitter dataset were as listed in Table 4.3.

Table 4.3: Data fields in extracted Twitter dataset and their descriptions

Data column name	Description
user_id	Official Twitter's author id
id	Unique Tweet id
screen_name	Display name on the profile
text	Twitter textual data
source	URL attached to a tweet posted as public

¹ <http://www.nltk.org/howto/twitter.html#simple>

² https://www.nltk.org/_modules/nltk/twitter/common.html#extract_fields

4.2.2 Data preprocessing

Our collected datasets of Facebook status updates and Twitter tweets went through extensive pre-processing steps. Figure 4.1 shows the overall process of data preprocessing we used for our Facebook and Twitter datasets:

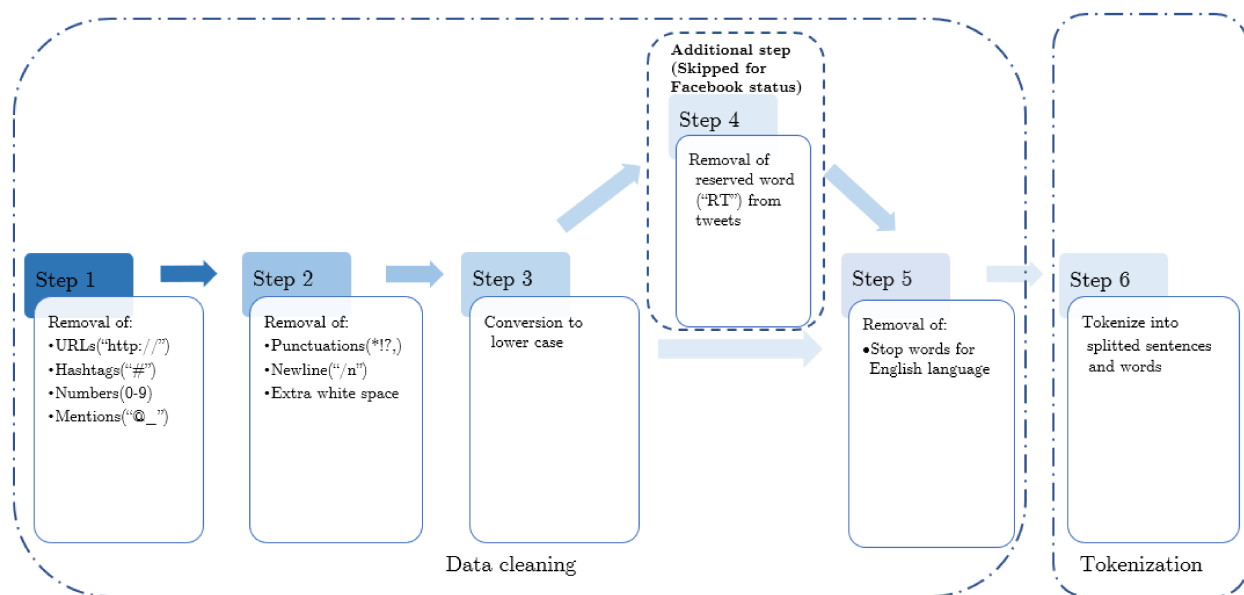


Figure 4.1: Data preprocessing technique used on both Facebook and Twitter datasets

4.2.2.1 Data cleaning

The data cleaning stage starts with removing content that usually makes text data (Facebook status updates and Twitter tweets) noisy, for example, URL("http://"), mentions (starting with "@"), hash tags ("#"), numbers ("0-9"). Then in step 2 we removed punctuations (*!?, except "."), newline returns("/n"), and extra spaces furthermore because they also make text data noisy. At first we used regular expressions in Python source code to eliminate them. Emojis (😊, 😬, 😞, 🙏), punctuation mark "." (full stop) were retained. In step 3 we converted all the words within text into lower case. Step 4 is an additional step for removing reserved words ("RT") from tweets only. We skipped this step during cleaning Facebook dataset. We also used the Natural Language Toolkit (NLTK) v3.4.4, an open source Python library in steps 3-5. NLTK v3.4.4 contains an extensive collection of modules for text (encoded in both ASCII code and Unicode) processing.

We used `nltk.corpus` module of NLTK v3.4.4 for removing stop words in step 5. Box 4.1 is the result of data cleaning task applied on a tweet (T), illustrated in figure 4.2.

Check out these awesome bulletin boards [@Jessica_Beth230](#) has ready for the students at WMS! We are almost ready to start **what** will be **an AMAZING** school **year!** [#InspireExcellence](#) [#WMS](#)

Figure 4.2: Example Tweet (T)

“check out these awesome bulletin boards has ready for the students at wms we are almost ready to start what will be an amazing school year inspireexcellence wms”

Box 4.1: Tweet (T) after preprocessing

Box 4.2 shows Python code snippet of different functions used in cleaning the example tweet in figure 4.2.

```

import re, string
import nltk

from nltk import word_tokenize
from nltk.corpus import stopwords

def remove_url(text): #Removes URL
    text = re.sub(r'\b(?:https?|ftp)://)?\w[\w-]*(?:\.[\w-]+)\S*(?![,.])', ' ', text.lower())
    words1 = re.findall(r'[a-z.]+', text)
    return ' '.join(words)

string_punctuation1 = "#,@"

def remove_punctuation1(text): #Removes hashtag(#), Mentions(@)
    no_punct = ""
    for letter in text:
        if letter not in string_punctuation1:
            no_punct += letter
    return no_punct

def remove_punctuation2(text): #Removes punctuations except full stop(".")
    text = re.sub("[^a-zA-Z' ]+", ' ', s)
    words2 = re.findall("[^\n\t\d:.,]+", text)
    return ' '.join(words2)

def remove_numbers(text): #Removes numbers
    output = ''.join(c for c in text if not c.isdigit())
    return output

def to_lower(text): #Conversion to lower case
    return ' '.join([w.lower() for w in word_tokenize(text)])

stop_words = set(stopwords.words('english'))

def remove_stopwords(sentence): #Removes stop words
    clean_sent = []
    for w in word_tokenize(sentence):
        if not w in stop_words:
            clean_sent.append(w)
    return " ".join(clean_sent)

```

Box 4.2: Python code snippet of different functions used in data cleaning stage


```

from nltk.internals import find_jar, config_java, java, _java_options
from nltk.tokenize.api import TokenizerI
from nltk.parse.corenlp import CoreNLPParser

class StanfordTokenizer(TokenizerI):
    def __init__(
        self,
        path_to_jar = None,
        encoding = 'utf8',
        options = None,
        verbose = False,
        java_options = '-mx1000m',
    ):
        self._stanford_jar = find_jar(
            self._JAR,
            path_to_jar,
            env_vars = ('STANFORD_POSTAGGER',),
            searchpath = (),
            url = _stanford_url,
            verbose=verbose,
        )
        self._encoding = encoding
        self.java_options = java_options

        options = {} if options is None else options
        self._options_cmd = ','.join(
            '{0}={1}'.format(key, val) for key, val in options.items()
        )

    def _parse_tokenized_output(text):    #Output after tokenization
        return text.splitlines()

    def tokenize(self, text):            #Calls Stanford tokenizer's PTBTokenizer and pass tweet text
        cmd = ['edu.stanford.nlp.process.PTBTokenizer']
        return self._parse_tokenized_output(self._execute(cmd, text))

```

Box 4.3: Python code snippet of function used in tokenization of tweet text

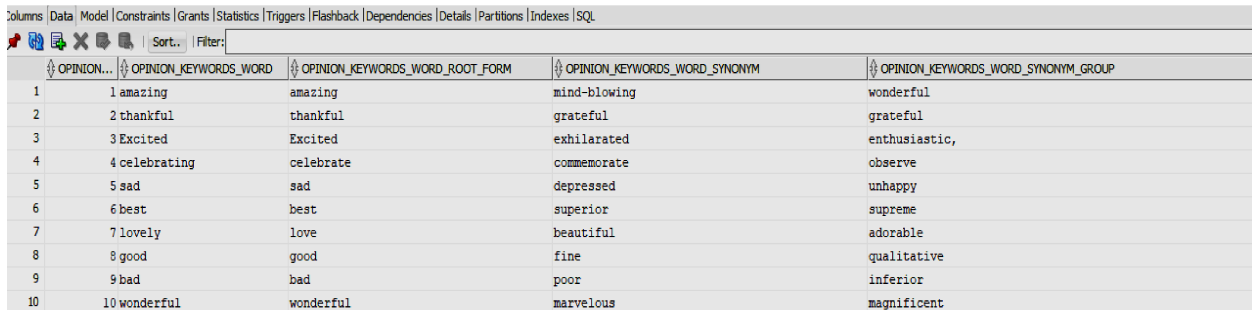
4.2.3 Lexicon detection

In our lexicon approach for identification of psycholinguistic meaning of words used in Facebook user status updates and Twitter tweets for analyzing psychological traits, we at first build two database dictionaries: (i) Opinion words dictionary containing opinion words that express desirable (e.g. amazing, happy etc.) or undesirable (e.g. sad, pissed etc.) states and (ii) Emojis dictionary containing emojis majorly used. To build our opinion dictionary, we relied on LIWC's¹ semantic dictionary. Table 4.4 shows four subcategories of LIWC dictionary.

Table 4.4: Four subcategories of LIWC dictionary for human psychological states

LIWC Category	LIWC Subcategory	Example word	No. of words
Affective processes - Positive emotion	Positive emotion	Happy, Amazing, Love, Sweet	406
Affective processes - Negative emotion	Anxiety	Tensed, nervous	91
	Anger	Hate, Pissed	184
	Sadness	Sad, Cry	101

Figure 4.3 shows partial view of our opinion dictionary.



The screenshot shows a database data view with the following columns: OPINION_KEYWORDS_WORD, OPINION_KEYWORDS_WORD_ROOT_FORM, OPINION_KEYWORDS_WORD_SYNONYM, and OPINION_KEYWORDS_WORD_SYNONYM_GROUP. The data is as follows:

OPINION_KEYWORDS_WORD	OPINION_KEYWORDS_WORD_ROOT_FORM	OPINION_KEYWORDS_WORD_SYNONYM	OPINION_KEYWORDS_WORD_SYNONYM_GROUP
1 amazing	amazing	mind-blowing	wonderful
2 thankful	thankful	grateful	grateful
3 Excited	Excited	exhilarated	enthusiastic,
4 celebrating	celebrate	commemorate	observe
5 sad	sad	depressed	unhappy
6 best	best	superior	supreme
7 lovely	love	beautiful	adorable
8 good	good	fine	qualitative
9 bad	bad	poor	inferior
10 wonderful	wonderful	marvelous	magnificent

Figure 4.3: Partial data view of our opinion dictionary

¹ http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

In order to include emojis in the text analysis performed with LIWC, we collected a set of emojis majorly used in social networking sites and build a emoticon dictionary. In this dictionary we associated each emoticon to a sentiment annotation presented in [37]. Figure 4.4 shows partial view of our emoticon dictionary.

EMOTICON_ID	EMOTICON_SIGN	EMOTICON_MEANING	EMOTICON_SENTIMENT_TYPE
1	1 :))	big smiley	Positive
2	2 :-)	smiley	Positive
3	3 :	pleasant	Positive
4	4 :- (extremely sad	Negative
5	5 :(sad	Negative
6	6 :D	grin	(null)
7	7 :O	gasp	(null)
8	8 ;)	wink	(null)
9	9 :/	unsure	(null)
10	10 :' (cry	Negative
11	11 O:)	blessed	Positive
12	12 <3	heart	Negative
13	13 3:)	crazy	(null)
14	14 >:(grumpy	(null)

Figure 4.4: Partial data view of our emoticon dictionary

Considering opinion words to determine the user expression or opinion, amplifier terms can increase or decrease the intensity of the affected opinion word (for example, the word “so” in the sentence “it’s so excellent” increases the intensity of the opinion word “excellent”). So, we build another dictionary for amplifier words.

The output of the tokenization step becomes input for lexicon detection. Table 4.5 shows output of lexicon detection after matching the tokenized words of Tweet (T) in section 4.2.2.1 with our opinion, emoticon, amplifier lexicons.

Table 4.5: Lexicon detected based on Tweet (T)

Opinion words	awesome	ready	amazing
Amplifier	-	almost	-
Emoticon	-	-	-

4.2.4 Data loading according to Multidimensional Fact Data model

We utilized multidimensional fact data modeling technique [52] to store processed data records of Facebook user status and Twitter tweet into datamart designed in our database. Figure 4.5 shows our multidimensional fact data model used to store processed data records of Facebook user status and Twitter tweet into database table.

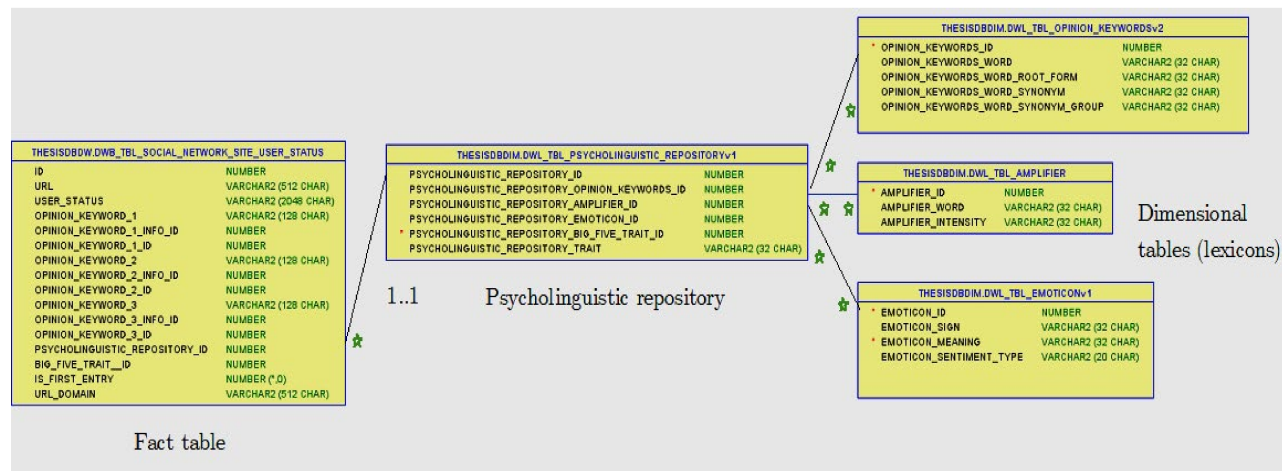


Figure 4.5: Multidimensional fact data model for Social Networking Sites' user status database storage (*The high-definition figure can be zoomed in for details*)

In our multidimensional fact data model we have a psycholinguistic repository named `THEISISDBDIM.DWL_TBL_PSYCHOLINGUISTIC_REPOSITORY` as database table which is a combination of opinion, emoticon and amplifier dictionary tables. The psycholinguistic repository table contains mapping of positive and negative opinion words, emojis, amplifier words with LIWC's 4 subcategories of human psychological traits. For example, we mapped opinion word "happy", amplifier word "so" and emoticon ":-)" to LIWC subcategory "joyful". Table 4.6 shows the mapping of opinion word "happy", amplifier word "so" and emoticon ":-)" to LIWC subcategory "Joy".

Table 4.6: Mapping of opinion word, amplifier word and emoji with psychological traits

Opinion word	Amplifier	Emoticon	Psycholinguistic trait
happy	so	:-)	joy

We stored all processed data records of Facebook user status and Twitter tweet into a fact table `THEISISDBDW.DWB_TBL_SOCIAL_NETWORK_SITE_USER_STATUS` which serves as datamart of Facebook user status and Twitter tweets. Each data record in this table contains associated opinion keywords and foreign key to `THEISISDBDIM.DWL_TBL_PSYCHOLINGUISTIC_REPOSITORY` table.

4.3 Web API implementation

To retrieve people's Facebook status updates and Twitter tweets from our datamart of processed Facebook user status and Twitter tweet data records based on mapping of positive and negative opinion words, emojis, amplifier words with LIWC's different categories of psychological states in our psycholinguistic repository, we utilized the same web API we developed for searching most relevant webpages in terms of list of hyperlinks according to given search keywords. But to find out Facebook status updates and Twitter tweets from our datamart of processed Facebook user status and Twitter tweet data records, we implemented separate user-defined functions in C# through which a given formatted input search query (string) passes sequentially. Figure 4.6 shows the GUI of the web API that transfers input search query of desired Facebook status updates and Twitter tweets down to database through a Http web request and displays Facebook user status and Twitter tweets with URL attached to each user status and tweet post as search result returned by database.

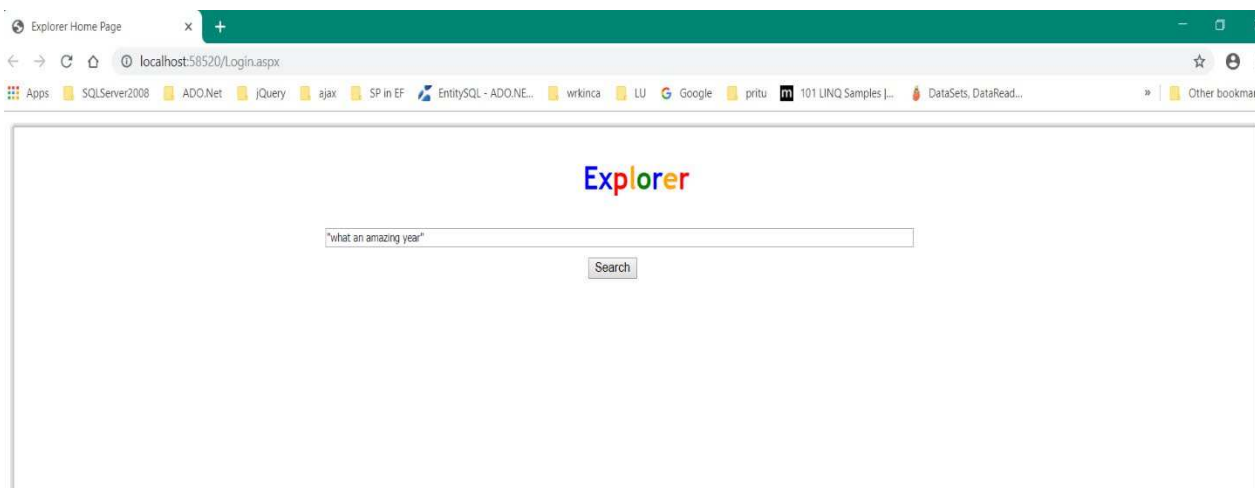


Figure 4.6: Input search query for Facebook status updates and tweets that contain keywords “what an amazing year”

As shown in figure 4.6, we need to enter search keywords enclosed in quotation (“”) and the search keywords form a string which goes through sequential user-defined functions as described below:

- **GetSearchKeywordType** (string[] keywords): This method takes input search keywords as array of string once the input search keywords are splitted as single keyword enclosed in quotation (“”) and white spaces among the splitted keywords are removed by built-in Split() function. GetSearchKeywordType (string[] keywords) function passes array of string to function **SearchKeywordTypeCategory**(string[] keywords). As the splitted keywords are enclosed in quotation, at this stage they are identified as words used in narrative sentence for tweets and user status on Social Networking Sites like Twitter, Facebook.
- **GetOpinionAmplifierWord**(string[] keywords): Once keywords are identified as words used in narrative sentence for tweets and user status, function GetOpinionAmplifierWord(string[] keywords) takes each keyword out of quotation (“”) and forms an array of string which goes through data access layer function **SearchOpinionAmplifierWord**(string[] keywords). Function

SearchOpinionAmplifierWord(string[] keywords) identifies opinion words, amplifier words by matching opinion and amplifier lexicons.

- **GetOpinionAmplifierWordId**(string[] keywords): Once opinion and amplifier words are identified, this function forms an array of string comprising opinion and amplifier words. This function passes the array of string to data access layer function `SearchOpinionAmplifierWordId(string[] keywords)`. Function `SearchOpinionAmplifierWordId(string[] keywords)` finds associated id of each opinion word and amplifier word.
- **GetPsycholinguisticMapping**(int[] keywords_id): This function takes integer id of each opinion word and amplifier word as an array of integer which the function passes to `SearchPsycholinguisticMapping(int[] keywords_id)`. This function finds the mapping of opinion, amplifier word to associated psycholinguistic trait in psycholinguistic repository of our database.
- **GetSocialNetworkingSiteUserStatus**(int[] opinion_words_id, int[] psycholinguistic_trait_map_id): Once psycholinguistic word mapping id is identified, this method passes necessary opinion word id values stored in an integer array, psycholinguistic trait mapping id to data access layer function `SearchSocialNetworkingSiteUserStatus (int[] opinion_words_id, int[] psycholinguistic_trait_map_id)` that retrieves all the data records stored in the datamart of tweets and user status updates and returns closely related Facebook user status and Twitter tweet data records that match certain psycholinguistic trait mapping and search keywords to show as search result of Facebook user status and Twitter tweets with URL attached to each user status and tweet post on the GUI.

Figure 4.7 shows the search result Facebook user status and Twitter tweets with URL attached to each user status and tweet post as displayed by the web API.

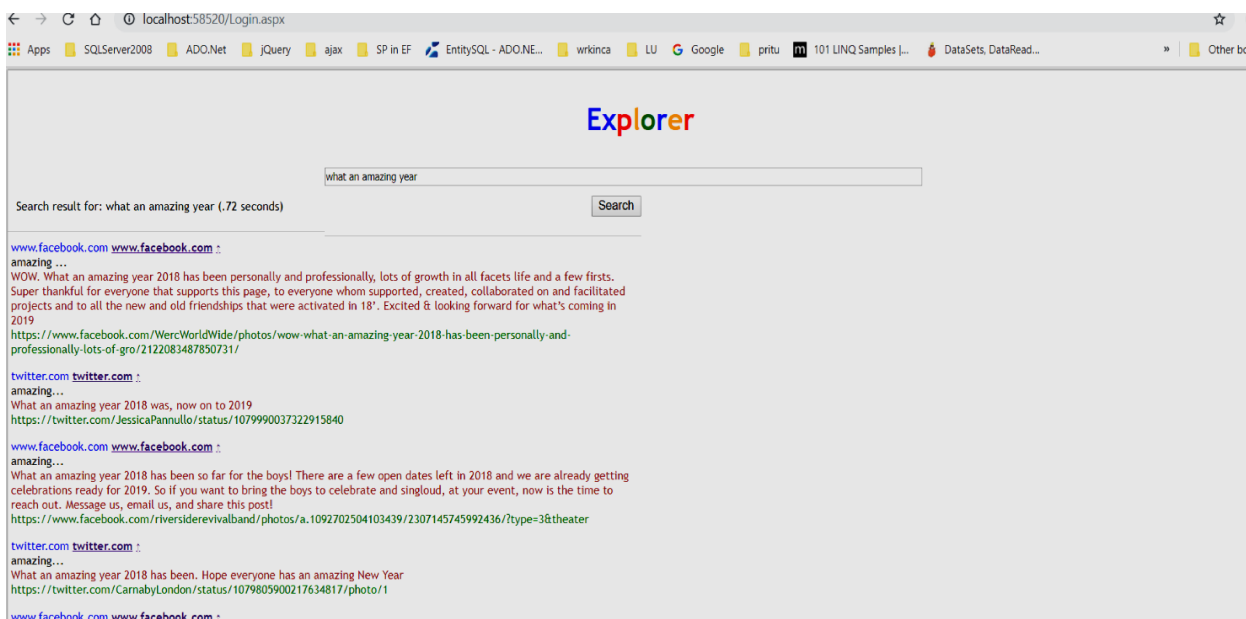


Figure 4.7: Search result of Facebook user status and Twitter tweets that contain keywords “what an amazing year”

4.4 Experiment and Evaluation of Experimental Results

As part of experiment, to search desired Facebook user status updates and Twitter tweets from our datamart of processed Facebook user status and Twitter tweet data records based on mapping of positive and negative opinion words, emojis, amplifier words with LIWC’s different categories of psychological states in our psycholinguistic repository we executed the web API with four different types of fragmented sentences which may remain in Facebook user status updates and Twitter tweets. These four different types of fragmented sentences contain four different opinion words ‘Amazing’, ‘Tensed’, ‘Hate’, ‘Sad’ that represent LIWC’s four different subcategories of human emotional and psychological states ‘Positive emotion’, ‘Anxiety’, ‘Anger’ and ‘Sadness’ respectively.

4.4.1 Experimental output

Through the GUI of the web API we searched most relevant tweets and status updates that exhibit user reactions to different subject matter of interests and contain below four fragmented sentences:

“what an amazing year”

“very much tensed”

“I really hate”

“feeling very much sad”

Figure 4.8 shows the search result as Facebook user status and Twitter tweets with URL attached to each user status and tweet post for search query “what an amazing year”.

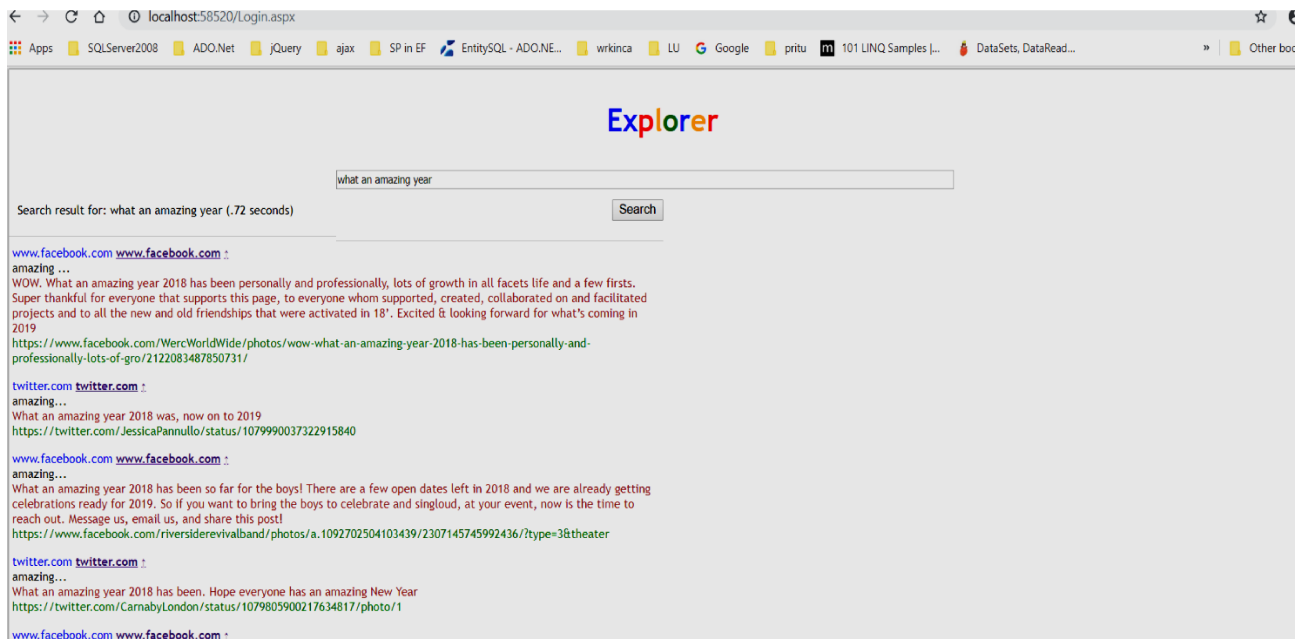


Figure 4.8: Partial view of search result for input search query “what an amazing year”

Figure 4.9 shows the search result as Facebook user status and Twitter tweets with URL attached to each user status and tweet post for search query “very much tensed”.

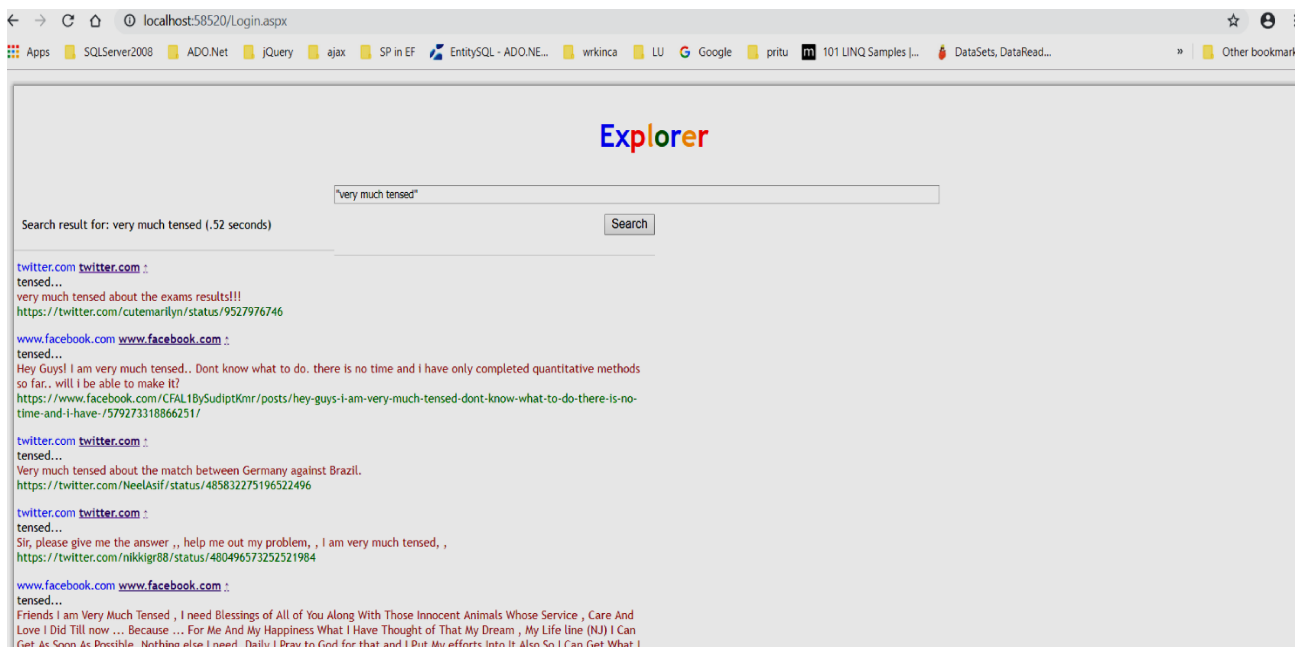


Figure 4.9: Partial view of search result for input search query “very much tensed”

Figure 4.10 shows the search result as Facebook user status and Twitter tweets with URL attached to each user status and tweet post for search query “I really hate”.

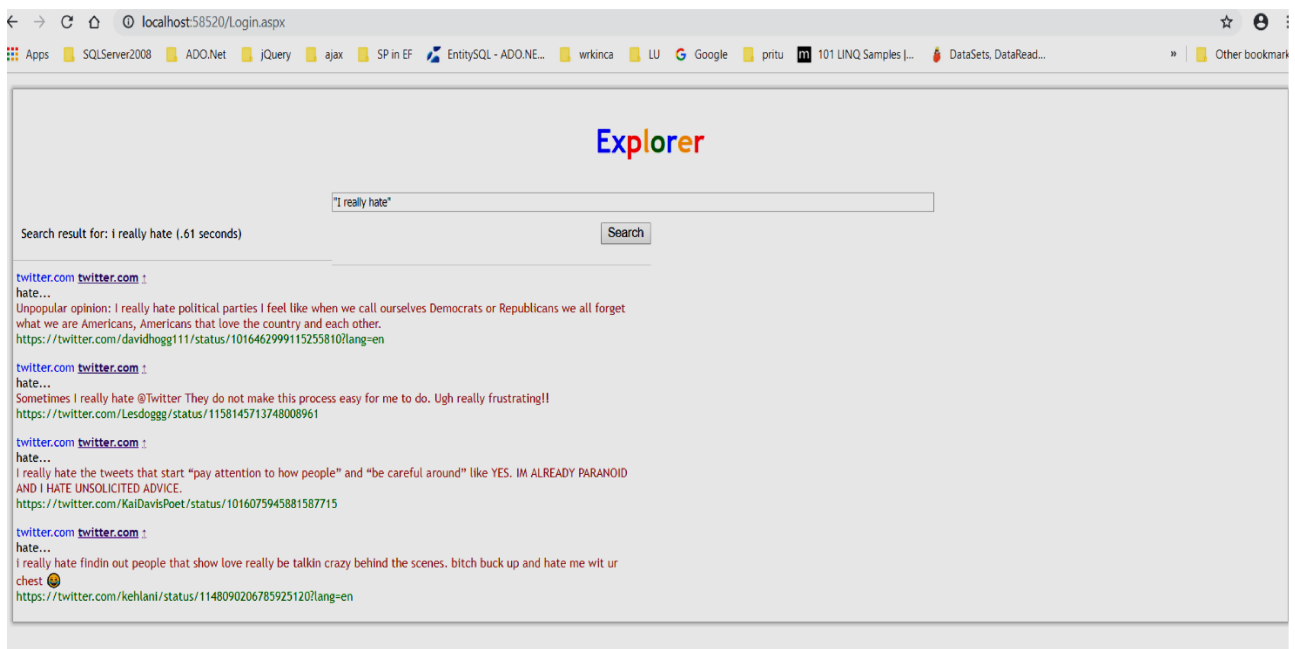


Figure 4.10: Partial view of search result for input search query “I really hate”

Figure 4.11 shows the search result as Facebook user status and Twitter tweets with URL attached to each user status and tweet post for search query “feeling very much sad”.

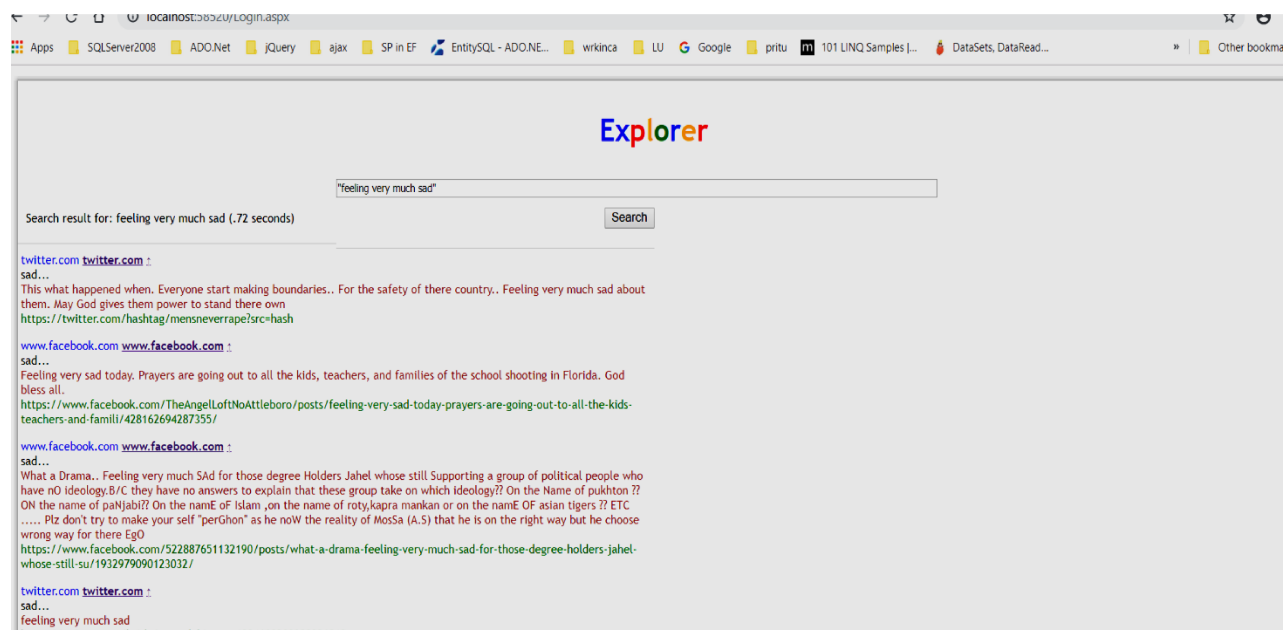


Figure 4.11: Partial view of search result for input search query “feeling very much sad”

4.4.2 Evaluation Metrics

To see how accurately the Facebook user status updates and Twitter tweets having opinion words in search result reveal corresponding true mapping with psychological traits, we executed the four experiments as described in section 4.4.1 to execute on pair of Facebook and Twitter datasets as collected through Facebook Graph API and Twitter Search API for two time durations:

- (i) January 7, 2019 to January 8, 2019
- (ii) January 14, 2019 to January 15, 2019

For each pair of datasets we were interested to see out of total processed and stored user status update and tweet data records that contain opinion words as listed in Table 4.1, how many user status updates and tweets appeared in search result contain certain sentence fragments that accurately reveal true mapping of opinion words, emojis and amplifier words with corresponding psychological traits.

To measure how user status updates and tweets that appear in search result are truly relevant to accurate psychological trait and to measure how much user status updates and tweets that appear in search result are correct out of certain number of user status updates and tweets that should have been returned we used ‘Precision’¹ and ‘Recall’¹ statistical metrics respectively.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

where TP and FP are the number of true positive and false positive results.

Recall is commonly referred to as sensitivity or true positive rate and is the odds of getting a positive test outcome given a positive case.

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$$

4.4.3 Performance Result and Analysis of Result

For January 7, 2019 to January 8, 2019, we collected 9,917 publicly posted status updates of many English-speaking active users of Facebook and 3,240 tweets of many English-speaking active users of Twitter. For January 14, 2019 to January 15, 2019, we collected 8,653 publicly posted status updates of many English-speaking active users of Facebook and 3,198 tweets of many English-speaking active users of Twitter. All these user status updates and tweets contained example opinion words used to express emotions that represent LIWC’s 4 subcategories of human psychological states (Table 4.1). we executed the four experiments as described in section 4.4.1 to execute on pair of combined Facebook and Twitter datasets stored in the database. Table 4.7 shows the no. of total data records in our collected two datasets and the no. of processed data records that we stored in database after data cleaning.

¹ https://en.wikipedia.org/wiki/Precision_and_recall

Table 4.7: Initial data records in datasets and processed data records in database

Data set	Total no. of data records (Facebook user status + Twitter tweets) in Dataset	No. of processed data records after Data Cleaning
January 7 to 8, 2019	13,157	11,619
January 14 to 15, 2019	11,851	9,435

For each pair of datasets we were interested to see out of total processed and stored user status update and tweet data records in database table that contain opinion words as listed in Table 4.1, how many user status updates and tweets appeared in search result contain certain sentence fragments that accurately reveal true mapping of opinion words, emojis and amplifier words with corresponding psychological traits. We were specifically interested to see % of correct user status updates and tweets in search result that contain certain sentence fragment and reveal true mapping of opinion word with corresponding psychological trait.

The percentage (ranging between 81%-86%) of correct user status updates and tweets appeared in search results we obtained is uniform to the percentage of accuracy (80%-85) as obtained by Srinivasu Badugu and Matla Suhasini [45] in their experimental result of emotion classification in tweets methodology. The higher percentage indicates better identification of emotion that accurately reveal true psychological traits as exhibited by Facebook and Twitter user.

Table 4.8 shows the result that we obtained after executing the four experimental tests (section 4.4.1) on our first dataset (January 7 to 8, 2019).

Table 4.8: Experimental test result on dataset of January 7 to 8, 2019

	Facebook user status and Twitter tweets that contain opinion word:			
	"Amazing"	"Tensed"	"Hate"	"Sad"
No. of user status and tweet data records in database table	3,904	2,859	3,014	1,842
No. of user status and tweets appeared in search result that contain certain sentence fragments and accurately reveal true psychological traits	3,154	2,363	2,570	1,497
Out of total no. of user status and tweet data records in database table having certain opinion word, % of correct user status updates and tweets in search result that contain certain sentence fragment and reveal true mapping of opinion word with corresponding psychological trait	80.79%	82.65%	85.27%	81.27%

Table 4.9 shows the result that we obtained after executing the same four experimental tests on our second dataset (January 14 to 15, 2019).

Table 4.9: Experimental test result on dataset of January 14 to 15, 2019

	Facebook user status and Twitter tweets that contain opinion word:			
	"Amazing"	"Tensed"	"Hate"	"Sad"
No. of user status and tweet data records in database	3,526	2,307	3,180	422
No. of user status and tweets appeared in search result that contain certain sentence fragments and accurately reveal true psychological traits	2,829	1,870	2,603	306
Out of total no. of user status and tweet data records in database table having certain opinion word, % of correct user status updates and tweets in search result that contain certain sentence fragment and reveal true mapping of opinion word with corresponding psychological trait	80.23%	81.06%	81.86%	81.99%

From table 4.8 we can see that out of 11,619 processed and stored Facebook user status updates and Twitter tweets, about 3,904 user status updates and tweet data records in database contain opinion word “Amazing”. Out of 3,904 stored user status updates and tweet data records that contain opinion word “Amazing”, we found 3,154 user status updates and tweets in search result that contain “what an amazing year” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “what an amazing year” sentence fragment and reveal true mapping of opinion word “Amazing” with corresponding positive psychological trait is 80.79%. Out of 11,619 processed and stored Facebook user status updates and Twitter tweets, about 2,859 user status updates and tweet data records in database contain opinion word “Tensed”. Out of 2,859 stored user status updates and tweet data records that contain opinion word “Tensed”, we found 2,363 user status updates and tweets in search result that contain “very much tensed” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “very much tensed” sentence fragment and true mapping of opinion word “Tensed” with corresponding negative trait “Anxiety” is 82.65%. Out of 11,619 processed and stored Facebook user status updates and Twitter tweets, about 3,014 user status updates and tweet data records in database contain opinion word “Hate”. Out of 3,014 stored user status updates and tweet data records that contain opinion word “Tensed”, we found 2,570 user status updates and tweets in search result that contain “I really hate” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “I really hate” sentence fragment and true mapping of opinion word “Hate” with corresponding negative trait “Anger” is 85.27%.

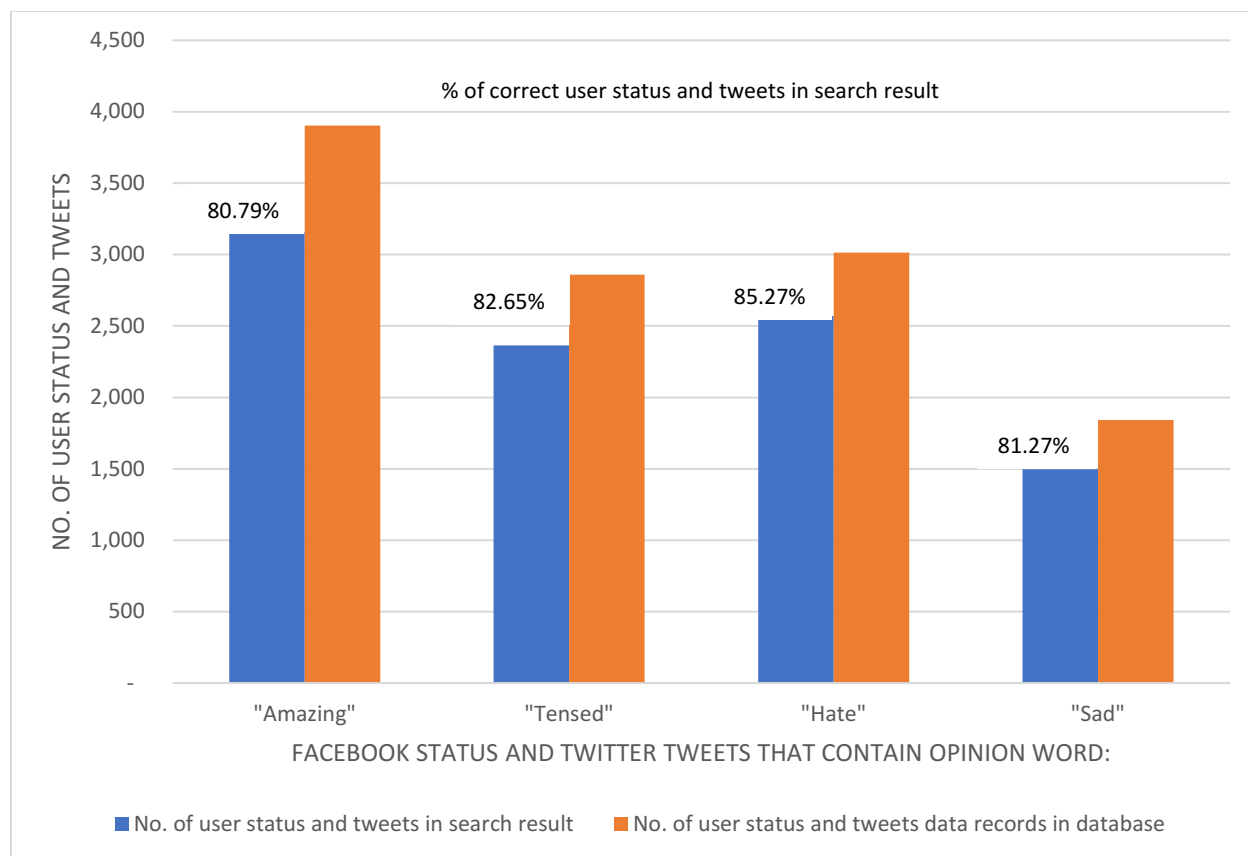


Figure 4.12: Graphical representation of experimental test result on dataset of January 7 to 8, 2019

Out of 11,619 processed and stored Facebook user status updates and Twitter tweets, about 1,842 user status updates and tweet data records in database contain opinion word “Sad”. Out of 1,842 stored user status updates and tweet data records that contain opinion word “Sad”, we found 1,497 user status updates and tweets in search result that contain “feeling very much sad” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “feeling very much sad” sentence fragment and true mapping of opinion word “Sad” with corresponding negative trait “Sadness” is 81.27%.

Figure 4.12 shows the graphical representation of experimental results in table 4.8. The graphical representation shows % of correctly identified user status and tweets in search result that contain specific sentence fragment and true mapping of opinion word with corresponding positive and negative psychological trait.

From table 4.9 we can see that out of 9,435 processed and stored Facebook user status updates and Twitter tweets, about 3,526 user status updates and tweet data records in database contain opinion word “Amazing”. Out of 3,526 stored user status updates and tweet data records that contain opinion word “Amazing”, we found 2,829 user status updates and tweets in search result that contain “what an amazing year” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “what an amazing year” sentence fragment and reveal true mapping of opinion word “Amazing” with corresponding positive psychological trait is 80.23%. Out of 9,435 processed and stored Facebook user status updates and Twitter tweets, about 2,307 user status updates and tweet data records in database contain opinion word “Tensed”. Out of 2,307 stored user status updates and tweet data records that contain opinion word “Tensed”, we found 1,870 user status updates and tweets in search result that contain “very much tensed” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “very much tensed” sentence fragment and true mapping of opinion word “Tensed” with corresponding negative trait “Anxiety” is 81.06%. Out of 9,435 processed and stored Facebook user status updates and Twitter tweets, about 3,180 user status updates and tweet data records in database contain opinion word “Hate”. Out of 3,180 stored user status updates and tweet data records that contain opinion word “Tensed”, we found 2,603 user status updates and tweets in search result that contain “I really hate” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “I really hate” sentence fragment and true mapping of opinion word “Hate” with corresponding negative trait “Anger” is 81.86%. Out of 9,435 processed and stored Facebook user status updates and Twitter tweets, about 422 user status updates and tweet data records in database contain opinion word “Sad”. Out of 422 stored user status updates and tweet data records that contain opinion word “Sad”, we found 346 user status updates and tweets in search result that contain “feeling very much sad” sentence fragment and the percentage (%) of correct Facebook status updates and Twitter tweets that contain “feeling very much sad” sentence fragment and true mapping of opinion word “Sad” with corresponding negative trait “Sadness” is 81.99%. Figure 4.11 shows the graphical representation of individual percentage (%) i.e. % of correctly identified user

status and tweets in search result that contain specific sentence fragment and true mapping of opinion word with corresponding positive and negative emotional traits. Figure 4.13 shows the graphical representation of experimental results in table 4.9. The graphical representation shows % of correctly identified user status and tweets in search result that contain specific sentence fragment and true mapping of opinion word with corresponding positive and negative psychological trait.

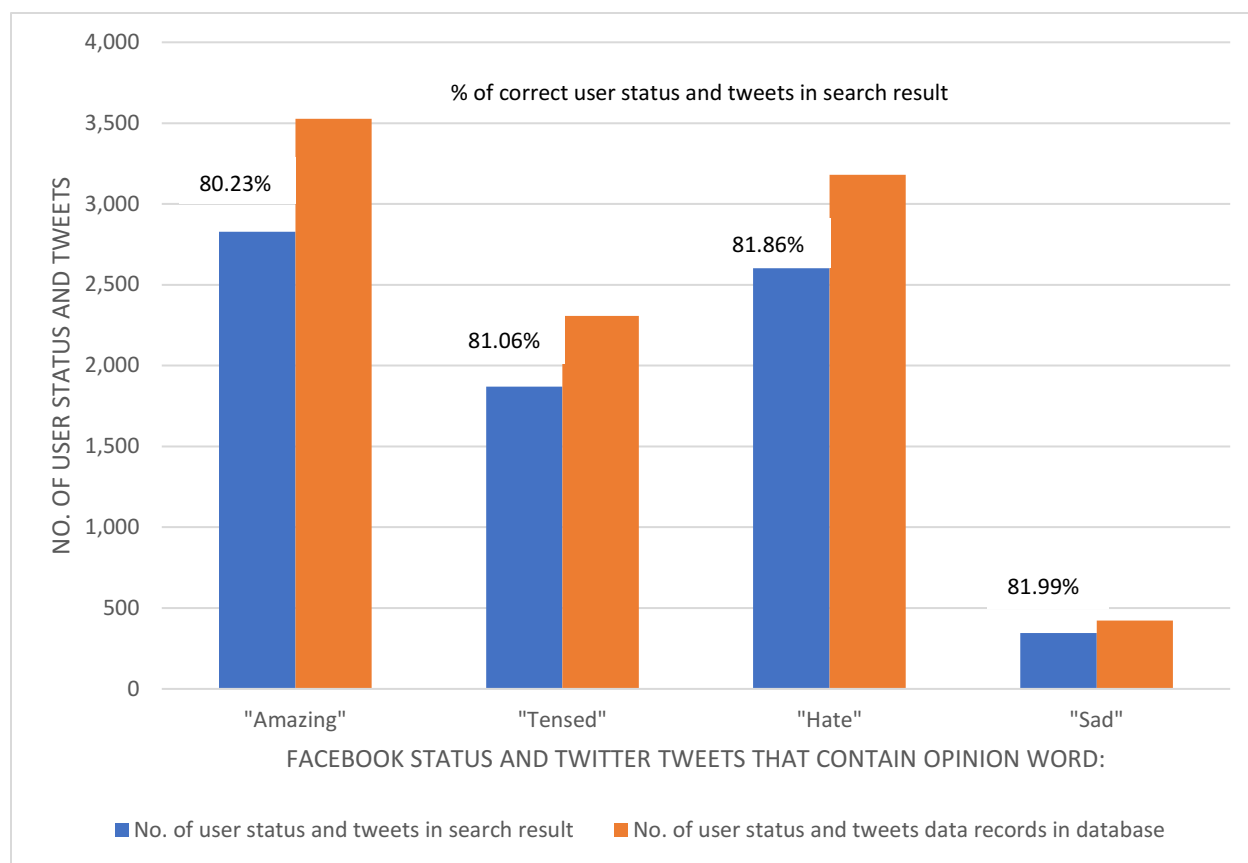


Figure 4.13: Graphical representation of experimental test result on dataset of January 14 to 15, 2019

From the combined search result obtained after combining the four experimental test results for dataset of January 7 to 8, 2019, we calculated precision to see "how useful the search results are" i.e. how user status updates and tweets that appear in search result are truly relevant to accurate psychological trait. We calculated recall to see "how complete the search results are" i.e. how

much user status updates and tweets that appear in search result are correct out of certain number of user status updates and tweets that should have been returned we used.

From the combined search result obtained after combining the four experimental test results for dataset of January 14 to 15, 2019, we also calculated precision, recall to measure quality of search result. We also calculated F-measure (range between 0 to 1) and F-measure (F-score) is a measure of test accuracy.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Table 4.10 shows the calculated results of precision, recall and F-measure.

Table 4.10: Accuracy of experimental test

User status and tweets search result (based on psychological trait mapping)		
Experimental test on:		
	January 7 to 8, 2019 dataset	January 14 to 15, 2019 dataset
Precision	0.901	0.881
Recall	0.907	0.91
F-measure	0.904	0.895

The higher F-measure is for an experimental test, the better accuracy the experimental test has.

Table 4.11 shows precision, recall as obtained by Srinivasu Badugu and Matla Suhasini [45] in their experimental result of emotion classification in tweets for four category of emotions: (i) Happy-Active Class, (ii) Happy-Inactive Class, (iii) Unhappy-Active Class and (iv) Unhappy-Inactive Class.

Table 4.11: Precision and Recall measures in experimental test of emotion classification in tweets by Srinivasu Badugu and Matla Suhasini

Emotion Category	Precision	Recall
Happy-Active	0.66	0.80
Happy-Inactive	0.79	0.92
Unhappy-Active	0.90	0.80
Unhappy-Inactive	0.94	0.90

Higher precision, recall indicates usefulness and completeness of the search results¹.

4.5 Conclusion

Mapping of positive and negative opinion words, emojis, amplifier words with LIWC's 4 subcategories of human psychological traits in our psycholinguistic repository yields significant good search results of appropriate Facebook user status updates and Twitter tweets. In each of the experimental tests executed

on two datasets of Facebook user status updates and Twitter tweets, higher percentage (ranging between 81%-86%) of correct user status and tweets in search result indicate that mapping of opinion words, emojis, amplifier words with human psychological traits in our psycholinguistic repository yields good search result of user status updates and Twitter tweets that accurately reveal true emotional and psychological traits as exhibited by Facebook and Twitter user. High F-scores of experimental tests on two datasets indicate good accuracy of the tests.

¹ https://en.wikipedia.org/wiki/Precision_and_recall

Chapter 5

Discussion, Future Work and Conclusion

5.1 Overview

Our approach of semi-structured webpage content extraction is a simple and less labor intensive methodology in terms of eliminating extra labor intensive HTML preprocessing work required during HTML to XML conversion i.e. HTML to structured XHTML and XHTML to XML conversion. The implemented multidimensional fact data modeling methodology that we applied for semi-structured webpage content database storage ensures a good speed of data retrieval operations on database table. Our implemented lexicon approach that we applied on sample datasets of Facebook user status updates and Twitter tweets to detect certain emotion-related linguistic features in user status updates and tweets and to incorporate them into psycholinguistic repository yields significant good search results in terms of correct Facebook user status updates and Twitter tweets that truly represent appropriate emotional and psychological traits. But there are some exceptions, limitation and identified areas which need to be considered for future scope of work.

5.2 Main Contributions

Our multidimensional fact data modeling technique for semi-structured webpage content database storage and indexing of data records in database tables yield qualitative information search results in terms of most relevant webpages appearing first with a high speed of data retrieval time. Our implemented lexicon approach that we applied on a sample dataset of Facebook user status updates and Twitter tweets to detect emotion-related linguistic features in user status updates and Twitter tweets and to incorporate them into psycholinguistic repository provide an efficient

way to identify psycholinguistic meaning of words used in user status and tweets which helps us to understand many different emotional and psychological traits exhibited by users on Social Networking Sites.

5.3 Current Exceptions and Scope for Improvement

In our data warehouse-oriented approach of semi-structured webpage content extraction and modeling into relational database, extraction fails when the complete HTML structure, not the content of a webpage changes rapidly. Our work is limited to extract webpage content of many C2C, B2C e-commerce websites, online news websites, job advertisement websites. As a scope of future work, the implemented extraction and modeling approach should be extended to adapt webpage content of other websites like online financial websites having stock market data and information, weather forecasting websites etc. that also attribute to have semi-structured content.

Regarding processing Social Networking Sites' user status for analyzing psychological traits to understand the underlying causes and motivations of user reactions to different subject matter of interests and to surrounding situations, the main limitation of our research work regards the sample dataset size which is limited in number. This is due to the exploratory nature of this research. More data should be collected to consolidate its findings. There are many instances, when some words are used in an entirely opposite sense, because of negations like “not”, “never”, “don't”, etc. For example, the tweet “Not at all feeling excited for school!” may result in getting “Joy” as its emotion category (because of the opinion word “excited”), if negations are not accounted for. In our lexical approach for identification of psycholinguistic features in Facebook user status updates and Twitter tweets we did not consider the negation effect of words. As part of future work:

- our lexical approach for identification of psycholinguistic features in Facebook user status updates and Twitter tweets can be extended to apply on a large sample of Facebook user status and Twitter tweet dataset in a cost-effective way with the given relative ease and low cost of data collection procedure.

- our research work can be extended to consider negation effect of words by maintaining additional lexicon of ‘Negation’ words.
- we need to perform more experimental executions of our implemented web API with other example opinion words of LIWC’s four subcategories of psychological states to measure the persistent quality of search results of Facebook user status updates and Twitter tweets.
- our work can be extended to cover user status updates of Non-English Speaking users on Facebook and Twitter as emotion-related many linguistic features in user status updates posted using other international languages are strong indicators of psychological behavior as exhibited by Non-English Speaking users.

5.4 Conclusion

A simplified and less labor intensive XML-based methodology for extracting webpage content in semi-structured form has been discussed. We discussed our multidimensional fact data modeling technique for semi-structured webpage content database storage and indexing of data records in database tables. Regarding Social Networking Sites’ user status for analyzing psychological traits, we discussed the lexicon approach to identify certain emotion-related linguistic features in user status updates and tweets that eventually helped us to understand the psycholinguistic meaning of words used in user status and tweets for realizing underlying causes and motivations of user reactions to different subject matter of interests and to surrounding situations. This data warehouse-oriented methodology for extracting and relational database schema mapping of semi-structured webpage content and Social Networking Sites’ user status provides qualitative information search results in terms of list of hyperlinks to most relevant webpages and user posts on Facebook and Twitter.

Bibliography

- [1] A. Kumar, R. K. Singh. A Study on Web Content Mining. *International Journal of Engineering and Computer Science*, ISSN: 2319-7242, vol. 6 issue 1, pages 20003-20006, Jan 2017.

- [2] N. Parmer, V. Ricchariya, J. Maurya. An Exploratory Review of Web Content Mining Techniques and Methods. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5 issue 5, May 2016.

- [3] T. S. Anushya devi, M. Selvanayaki. An Overview of Web Content Mining and its Techniques. *International Journal of Advanced Science Engineering and Technology*, vol. 3 issue 2, pages 48-53, Apr 2014.

- [4] F. Johnson, S. K. Gupta. A Survey on Web Content Mining Techniques: A Survey. *International Journal of Computer Science and Technology*, vol. 47 no. 11, Jun 2012.

- [5] K. Patidar, P. Purohit, K. Sharma. Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library, *International Journal of Computer Applications*, vol. 2 issue 1, Mar 2011.

- [6] K. Pol, N. Patil, S. Patankar, C. Das. A Survey on Web Content Mining and Extraction of Structured and Semi structured Data. *IEEE First International Conference on Emerging Trends in Engineering and Technology*, pages 543-546, 2008.

- [7] N. R. Kapadia, K. Patel, M. C. Parikh. Partitioning Based Web Content Mining. *International Journal of Engineering Research and Technology*, vol. 1 issue 3, 2012.

- [8] M. Atay, A Chebotko, D. Liu, S. Lu, F. Fotoubi. Efficient schema-based XML to relational data mapping. *Information systems*, Elsevier, 2005.

- [9] J. Myllymaki. Effective Web data extraction with standard XML technologies. In *Proceedings of the 10th international conference on World Wide Web*, Hong Kong, pages 689-696, May 01 - 05, 2001.
- [10] D. Lee, W. Chu. Constraints-preserving transformation from XML document type definition to relational schema. Elsevier, 2000.
- [11] M. Mani, D. Lee. XML to relational conversion using theory of regular tree grammars. *VLDB Workshop on EEXTT*, 2002.
- [12] P. Bohannon, J. Freire, P. Roy, J. Simeon. From XML schema to relations: a cost-based approach to XML storage. ICDE, 2002.
- [13] S. Lu, Y. Sun, M. Atay, F. Fotouhi. A New inlining algorithm for mapping XML DTDS to relational schema. In *Proceedings of the First International Workshop on XML Schema and Data Management, in conjunction with the 22nd ACM International Conference on Conceptual Modeling*, Chicago, IL, October 2003.
- [14] J. Shanmugasundaram, K. Tufte, C. Zhang, G. He, D. Dewitt, J. Naughton. Relational Databases for Querying XML Documents: Limitations and opportunities. VLDB 1999, pages 302-314.
- [15] H. Zafari, K. Hasami, M. Ebrahim Shiri. Xlight, an Efficient relational schema to store and query XML data. In *Proceedings of the IEEE International conference in Data Store and Data Engineering*, pp: 254-257, 2011.
- [16] M. Ibrahim Fakharaldien, J. Mohamed Zain, N. Sulaiman. XRecursive: An efficient method to store and query XML documents. *Australian Journal of basic and Applied Sciences*, vol. 5 issue 12, pp: 2910-2916, 2011.
- [17] M. Sharkawi, N. Tazi. LNV : Relational database Storage structure for XML documents. The *3rd ACS/IEEE International Conference On Computer Systems And Applications*, pp:49-56, 2005.
- [18] E. Triantaphyllou, G. Felici. Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques. Springer Science and Business Media, Sep 2006.

- [19] Yla R. Tausczik, James W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, vol. 29, Issue 1, pages 24-54, 2010.
- [20] M. Kosinski, D. Stillwell, T. Graepel. Private traits and attributes are predictable from digital records of human behavior. In *Proceedings of the National Academy of Sciences* 110(15), Mar 2013.
- [21] Rajwa Alharthi, Benjamin Guthier, Camille Guertin, Abdulmotaleb El Saddik. A Dataset for Psychological Human Needs Detection From Social Networks. *IEEE Access*, ISSN: 2169-3536, vol. 5, pages 9109 - 9117, May 2017.
- [22] Y. Wang, Understanding Personality through Social Media. Stanford University Publications, 2015.
- [23] A. Khan, B. Baharudin, K. Khan. Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews. *Trends in Applied Sciences Research*, 6(10) pages 1141-1157, 2011.
- [24] Attardi, G. and M. Simi, Blog mining through opinionated words. In *Proceedings of the 15th Text Retrieval Conference*, pages 2-7, Nov 2006.
- [25] Hu, M. and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 755-760, 2004.
- [26] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the Conference on Web Search and Web Data Mining customer reviews*’, ACM, 2010.
- [27] R. B. W. N. Jeonghee Yi, T Nasukawa. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *ICDM*, IEEE, 2003.
- [28] B. B. K. Khan, A. Khan. Sentence based sentiment classification from online customer reviews. ACM, 2010.
- [29] C. P. W. C.C. Yang, Y.C. Wong. Classifying web review opinions for consumer product analysis. ACM, 2009.

- [30] P. H. Theresa Wilson, Janyce Wiebe Proceedings of human language technology conference and conference on empirical methods in natural language processing. Association for Computational Linguistics, page 347354, 2005.
- [31] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Language Resources and Evaluation Conference (LREC)*, vol. 10, 2010.
- [32] E. Kouloumpis, T. Wilson, J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! In *the International AAAI Conference on Web and Social Media (ICWSM)*, pages 538-541, 2011.
- [33] R. C. Balabantaray, M. Mohammad, N. Sharma. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, vol. 4, no. 1, pages 48–53, 2012.
- [34] J. L. S. Yan, H. R. Turtle. Exploring fine-grained emotion detection in tweets. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 73–80, 2016.
- [35] J. Golbeck, C. Robles, K. Turner. Predicting personality with social media. CHI, ACM, pages 253–262, 2011.
- [36] J. Bollen, H. Mao, A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. The *International AAAI Conference on Web and Social Media (ICWSM)*, vol. 11, pages 450– 453, 2011.
- [37] G. Vashisht, S. Thakur. Facebook as a corpus for emojis based sentiment analysis. *International Journal of Emerging Technologies and Advanced Engineering*, vol. 4, pages 904-908, May 2014.
- [38] <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>
Date accessed: August 7, 2019.
- [39] <https://en.wikipedia.org/wiki/Twitter>. Date accessed: August 7, 2019.
- [40] <https://www.nltk.org/api/nltk.twitter.html>.

- [41] Stanford Core NLP,
<https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/process/PTBTokenizer.html>.
- [42] S. Mahmood, I. Hamidah, A. Mustapha, L. Abdullah. A Framework for Extracting Information from Semi-Structured Web Data Sources. In *Proceedings of Third International Conference on Convergence and Hybrid Information Technology*, vol. 2, pages 27-31, 2008.
- [43] S. Abiteboul, P. Buneman, D. Suciu. *Data on the Web : From Relations to Semi-structured Data and XML*. Morgan Kaufmann, 2011.
- [44] H. Snoussi, L. Magnin, Jian-yun Nie. Heterogeneous Web Data Extraction using Ontology. *Third International Bi-Conference Workshop on Agent-oriented Information Systems*, 2001.
- [45] S. Badugu, M. Suhasini. Two Step Approach for Emotion Detection on Twitter Data. *International Journal of Computer Applications*. vol. 179, no. 53 pages 12-19, 2018.
- [46] C. Strapparava, R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, ACM, pages 1556-1560, 2008.
- [47] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz. Predicting depression via social media. In *International AAAI Conference on Weblogs and Social Media*, 2013 The AAAI Press.
- [48] N. Wang, M. Kosinski, D. J. Stillwell, J. Rust. Can well-being be measured using Facebook status updates? Validation of Facebook's gross national happiness index. *Soc.Indic. Res.* vol. 115, pages 483–491, 2012.
- [49] H.A. Schwartz, G. Park, M. Sap, E. Weingarten, J. Eichstaedt, M.L. Kern et al. Towards assessing changes in degree of depression through Facebook. In *Poster Workshop on Computational Linguistics and Clinical Psychology : From Linguistic Signal to Clinical Reality*, pages 118–127, 2014.
- [50] <http://webextract.net/index.aspx>. Date accessed: August 7, 2019
- [51] J. Pennebaker, R. Boyd, K. Jordan, K. Blackburn. The Development and Psychometric Properties of LIWC2015. DOI: 10.15781/T29G6Z.

- [52] https://en.wikipedia.org/wiki/Dimensional_fact_model Date accessed: August 7, 2019
- [53] E. Diener R. E. Lucas. Introduction to Psychology: The Full Noba Collection R. Biswas-Diener and E. Diener (Eds), Noba Textbook Series: Psychology. Champaign, IL, 2019
- [54] https://en.wikipedia.org/wiki/Trait_theory#cite_note-1 Date accessed: August 7, 2019
- [55] S. Kassin. *Psychology*. 4th edition, Prentice-Hall, Inc, 2003.
- [56] Z. Li, W. K. Ng. WICCAPP: from semi-structured data to structured data. In *Proceedings. 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems*, pages. 86-93, 2004.
- [57] Y. Mo, T. W. Ling. Storing and Maintaining Semi structured Data Efficiently in an Object-Relational Database. In *Proceedings of the 3rd International Conference on Web Information Systems Engineering*, 2002.
- [58] A.O. Mendelzon, G. Mihaila, T. Milo. Querying the World Wide Web. *International Journal on Digital Libraries*. vol. 1, pages. 80-91 1997.
- [59] W. S. Li, J. Shim, K. Candan. WebDB: A system for querying semi-structured data on the web. *Journal of Visual Languages and Computing*, vol. 13, no. 1, pages. 3-33, 2002.
- [60] B. Gaiind, V. Syal, S. Padgalwar. Emotion Detection and Analysis on Social Media. The Global Journal of Engineering Science and Researches. ISSN 2348 – 8034, pages. 78-89, 2019.