Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

People's Democratic Republic of Algeria
University Abou Bakr Belkaid - Tlemcen
Faculty of Science
Computer Science Department

Master Thesis

For obtaining the Master's degree in Computer Science

Option: Système d'information et de connaissance (S.I.C)

Theme

Quality of experience (QoE) estimation of Web services via text mining tools

Realized by:

- **Yasmina DALI YOUCEF**

Presented on July 04, 2022 in front of jury composed of:

- Mr. Mohammed Amine BENTALLAH (President)
- Mr. Mohammed Ismail SMAHI (Supervisor)
- Mme. Amal HALFAOUI (Examiner)

Année Universitaire 2021/2022

I dedicate this modest work to

My dearest parents, Tarik and Malika, who instilled in me all the right values, who offered me the best education and who gave me all the necessary courage and willpower to reach my goals. My big sister Manel who has always supported me. And to the friends that became family who believed in me and contributed in one way or another to my success.

Acknowledgement

First of all, I deeply thank **ALLAH** who has always helped me with his great generosity by giving me courage, will and determination to carry out all my projects of studies.

I also thank **Mr. M. I. SMAHI** who has been an exemplary memoir director. I learned a lot from him, both humanly and professionally. I thank him for his listening as well as the rigor of his follow-up and his availability.

I express my sincere gratitude to Professors **A. HALFAOUI** and **M. A. BENTALLAH** for all the knowledge that they transmitted to us respectively during the third year of undergraduate studies and these two years of Master. I particularly thank them for honoring me by accepting to participate and to chair my jury.

Furthermore, my thanks go to all the rest of my dear professors who have contributed to my training and learning throughout my five years of study, by sharing their knowledge and their scientific experience. My special thanks go to **Mr. H. MATALLAH**, Head of the Department of Informatics for his assistance, help and his availability during my university training.

Finally, I am very grateful to my parents, my sister and my close friends for all their sacrifices, support, patience and trust. Their presence was indispensable.

Abstract

Web services may be seamlessly and dynamically integrated thanks to web service composition. The performance of a composition as a whole is determined by the actions of participating Web services. Therefore, while selecting services for service composition, good quality is crucial. Current methods for choosing and discovering Web services focus on functional factors(availability and response time), or quality of service. Despite the fact that these factors are essential for choosing Web services, they could not accurately reflect user perceptions of quality. We begin this master's thesis by looking at Web Services from a broad perspective. Second, we investigate what Text Mining is and the fundamental components and techniques that make it what it is. Last but not least, we parameter a python algorithms that use Text Mining method to mine user reviews and measure their polarity and the outcome give us a sense of how the technique performs.

Keywords: Web Services, Text Mining, Quality of service, Quality of experience.

ملخص

يمكن دمج خدمات الويب بسلاسة وديناميكية من خلال تكوين خدمة الويب. يتم تحديد أُداءْ التكوين ككل من خلال إجراءات خدمات الويب المشاركة. لذلك، عند اختيار الخدمات لتكوين الخدمة، فإن الجودة أُمر بالغ الأهمية. تركز الأساليب الحالية لاختيار واكتشاف خدمات الويب على عوامل الجودة أُو غير الوظيفية للخدمة، مثل التوافر ووقت الاستجابة. على الرغم من أُن هذه العوامل ضرورية لاختيار خدمات الويب، إلا أُنها قد لا تعكس بدقة تصور المستخدمين للجودة. نبدأُ أُطروحة الماستر هذه من خلال النظر إلى خدمات الويب من منظور عام. بعد ذلك، ننظر إلى ما هو التنقيب عن النص، والمكونات والتقنيات الأُساسية التي تجعله على ما هو عليه. أُخيرًا، نقدم طريقة لاستخراج النص تستغل ملاحظات المستخدم لقياس قطبتيها والنتيجة تعطينا فكرة عن أُداءْ التقنية المختارة.

الكلمات الدالة : خدمات الويب، التنقيب عن النص، جودة .الخدمة، جودة الخبرة

Résumé

Les services web peuvent être intégrés de manière transparente et dynamique grâce à la composition des services web. Les performances d'une composition dans son ensemble sont déterminées par les actions des services Web participants. Par conséquent, lors de la sélection des services pour la composition de services, la qualité est cruciale. Les méthodes actuelles de choix et de découverte des services Web se concentrent sur des facteurs fonctionnels(disponibilité et le temps de réponse) ou sur la qualité de service. Malgré le fait que ces facteurs sont essentiels pour le choix des services Web, ils peuvent ne pas refléter avec précision la perception de la qualité par les utilisateurs. Nous commençons ce mémoire de maîtrise en examinant les services Web d'un point de vue général. Ensuite, nous étudions ce qu'est lextraction de texte, ainsi que les composants et les techniques fondamentaux qui en font ce qu'elle est. Enfin, nous présentons un code python que nous avons choisis de parametre car il utilise des methodes qui exploite les commentaires des utilisateurs afin de mesurer leur polarité et les résultats nous donne une idée de la performance des technique utilise dans le code.

Mots-clés: Services Web, Extraction de texte, Qualité du service, Qualité de l'expérience.

Contents

List of Figures

List of Tables

1

General Introduction

Sommaire

1.1 Context & Problematic

Web services are a logical outcome of how the Web has developed. They are rising in popularity and starting to influence many facets of life. The Web has been expanding its reach to make it possible for increasingly complex types of interaction since its inception as a tool to share and disseminate information on a worldwide scale, thereby becoming a huge dispersed content library, [1], [2]. Within the next five years, services are predicted to rule the software sector.Because of our growing reliance on these services, serious issues with service reliability, security, and timeliness are emerging. In the literature, a number of strategies have been put out for providing services, [3], [4].

Because there are more and more services available on the Internet and because there are many Web services that are functionally identical, choosing a Web service has lately been a challenging task. We need new concept technologies to participate in.

QoS is a useful criterion for distinguishing functionally comparable Web services. The original thought was to provide Quality of Service (QoS) information via static release and runtime monitoring.The limitation of this strategy is that quality values are acquired only in certain regulated conditions and platforms. The results would most likely alter if the Web service was invoked from a different platform or from a different geographical location.

Then come the complementary if not main help Quality of Experience (QoE). This exists thanks to another source of quality information easily accessible, which can provide a more subjective view of the quality of service. User reviews are a valuable source of knowledge, where users can freely publish their opinions and experiences to the public. Frequently this still also a tedious and time consuming, because users can face a large number of reviews to analyze, and a long list of services to compare, [5]. For this method to function, we need a mechanism to extract the users' perspectives using Text Mining methods.

Text mining has therefore become an interesting research subject in an effort to extract usable information from unstructured texts. Extremely information-dense unstructured texts cannot easily be used by computers for additional processing. Text mining is a practice that covers several academic areas since it involves extracting useful information from text. It is similar to text data mining, which is applied to text-based data. It is used to analyze and read textual data using a vast and expanding range of approaches, [6]–[9].

1.2 Manuscript plan

We gathered all of the relevant information about all of these topics, which in many circumstances blend nicely together for a greater good. We made certain to expose and explain all of the key fundamentals that make Web Services what they are in Chapter 1. From the foundation of their framework, architecture, work-flow, reputation and many complimentary concepts.

Following that (Chapter 2), we made an attempt to explain what Text Mining is all about, including its method, numerous approaches, and applications. We have selected a unique approach to revealing the enormous amount of information, by providing simple definitions and even examples in the majority of cases to facilitate and enhance the comprehension. Last but not least, In order

to conduct this research, we added a few parameter to a Python program that encompasses all the areas that we have decided to further the knowledge in. Using it we have exposed the main step followed in it and explained briefly the result obtained in it. The overall conclusion will be a simple recap of the main point that you will read in this thesis.

2

Introduction to The Web Services

Sommaire

2.1 Introduction

An introduction to the fundamentals of Web Services is provided in this first chapter. To properly understand the revolutionary influence it has on the world up to this age, it is essential to disclose the key element of it, such as its framework, architecture, workflow, and even its strength and weaknesses. In this section, the closer look that we will take on those few components will be helpful for the results from Chapter 3 that shows how the forms of Web Services method may be used to solve actual problems.

2.2 Web Services

In these recent years a light has been shined on Web Services and everyone is looking more and deeper into this subject. Web services developed themselves into this giant information distributor all over the world and refined the communication between client and servers: single form-based interactions, retail e-commerce applications, and more complex business-to-business interactions. Without changing the core of the Webs fundamentals based on human-

to-applications interaction, the focus today is on applications-to-applications interactions due to the large resource sharing.

If the question what is Web Services? Get to your head its actually self-explanatory, it refers to access services over the Web with a little development on it to make it possible like architecture and business models. There are many views on the full definition of Web Services but they differ on small mater for example:

- The IBM Web service define it as: "Web services are a new breed of Web application. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Web services perform functions, which can be anything from simple requests to complicated business processes" [2].

- Now according to the W3C: "Web service is defined as a software system designed to support interoperable machine to machine interaction over a network. Web services are frequently nothing more than Web APIs that may be accessed through a network, such as the Internet, and executed on a remote system that provides the required services" [10].

Web Services definitions incorporates a wide range of systems, the phrase or term most commonly used to refer to clients and servers that communicate using SOAP(Simple Object Access Protocol) standard (XML messages). The assumption that there is also a machine-readable representation of the operations handled by the server, a definition in the WSDL (Web Services Description Language), is common in both the area and the terminology.

The latter isn't necessary for a SOAP endpoint, but it is in the common Java and.NET SOAP frameworks for automatic client-side code generation. Some industry associations, such as the WS-I (Web Services Interoperability Organization), demand both SOAP and WSDL as component of their Web service definition.
Web services are Internet-based business functions that are self-contained. They are written to rigorous open specifications so that they can function together and with other components of a similar nature.

Web services are beneficial to businesses because they allow systems from various firms to communicate with one another more effortlessly than before.

Organizations need the ability to link up their established systems quickly and efficiently with other companies as they necessitate closer operations between suppliers and customers, engage in more joint ventures, and face the prospect of more mergers and acquisitions. As a result, Web services enable businesses to conduct more e-business with more prospective business partners, in more and various ways than before, and at a lower cost.

The rapid expansion in Internet use of the World Wide Web has resulted in a huge boost in demand for Web services. In the development of distributed application systems, web services have received a lot of momentum. Because of the benefits of interoperability, reusability, and adaptability, certain vital applications contemplate implementing the Web services model. Existing Web service models must be modified to enable important applications in order to ensure their survival.

2.3 Web Services Key words

The World Wide Web Consortium (W3C) Founded in 1994 by Tim Berners-Lee who previously developed World Wide Web (WWW). It is an international community dedicated to enhancing the internet. It consists of several hundred member companies from numerous connected IT industries. To encourage collaboration and interoperability among all web stakeholders.

International Business Machines (IBM) t is an American-based manufacturer and service provider of IT gear on a global scale. With the rise of early computers in the 1950s and 1960s, IBM became a leader in computing hardware. However, the business suffered major losses as a result of its inability to keep up with the transition to personal computers in the 1980s. The majority of IBM's offerings today are hosting, consulting, and enterprise-level technology hardware.

The Web Services Interoperability Organization (WS-I) The organization was founded through a proposal by Microsoft and IBM. It is a multi-industry initiative that aims to hasten the development and distribution of interoperable Web services that would run on a variety of platforms, programs, languages, and computer systems throughout the Internet. Companies interested in establishing best practices for Web services are welcome to join the association.

However, the group doesn't design or develop Web service standards; instead, it develops guidelines, evaluates the interoperability of existing standards, and then offers suggestions in light of the evaluations.

Organization for the Advancement of Structured Information Standards (OASIS) In the computer science world, OASIS establishes open standards designed to promote innovation and lower costs on a global scale. Security, cloud computing, content technology, Web services, and e-government are just a few of the important technology sectors in which the organization has created open standards. Each OASIS standard is developed with the assistance of its members, and it is only made public after receiving majority approval.

Extensible Markup Language (XML) XML is a common format that the W3C maintains for the encoding and transmission of structured data on the web or between apps. By enabling users to generate custom defined tags in accordance with XML Document Type Definition (DTD) standards, the language employs a structured representation. A Document Object Model (DOM) is a tree that can be used to represent the structure of an XML document (DOM).

You can find the definitions in [11].

2.4 Web Services Framework

New Web-based frameworks (Figure 2.1) must be based on standards to ensure broad adoption. The markup format of HTML is extended and formalized by XML, which also offers a way to describe structured data. As a result, XML has replaced HTML as the industry standard for information representation on the Internet and provides the foundation for new Web technologies.

XML has been criticized for being verbose and requiring extensive parsing. However, when it facilitates communication between several sets of systems, verbosity actually becomes one of its greatest virtues. The Web services architecture is best built on XML because of its explicit representation of structured data. For more details see [1].

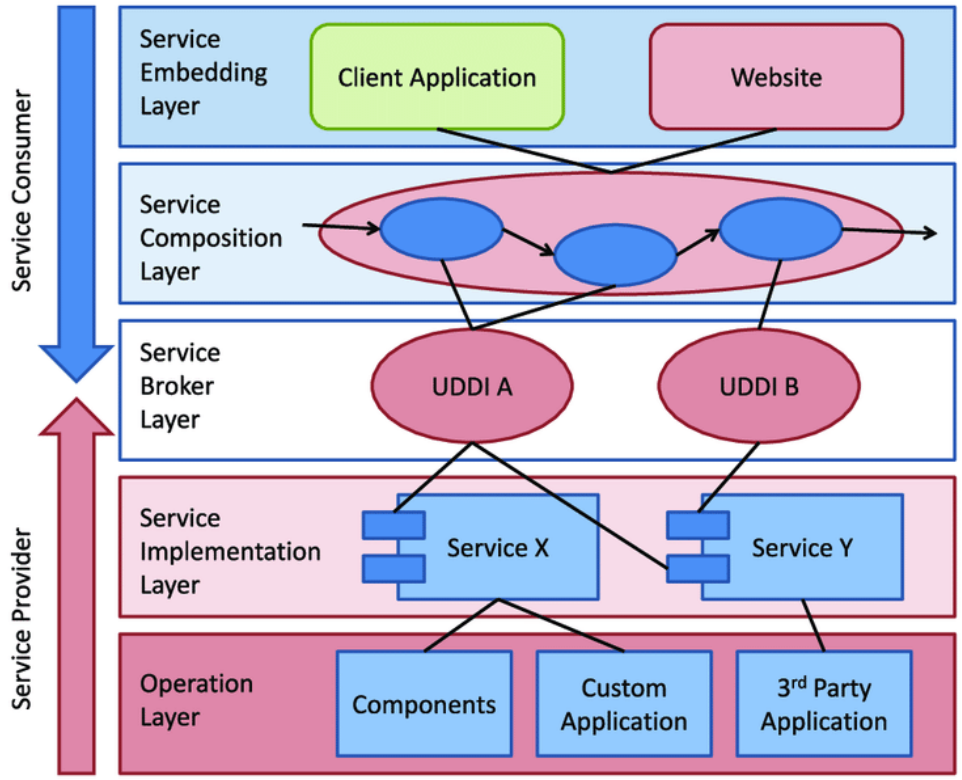

Figure 2.1: Example of a Web Service Framework, [12].

2.5 Web services Architecture

A digital library's design seems to be very similar to the architecture needed to support web services. The digital library community is used to verbs like publish and locate are often utilized to describe documents. Such verbs are used in the web services architecture to define applications. In reality, instead of being a distributed digital library for content/data, the web services architecture may be thought of as a distributed digital library for services. This indicates that several of the challenges addressed by the digital library community, such as metadata for discovery, authentication and authorization, and business models for accessing intellectual property, are indeed appropriate to online services and must be handled in that environment. The added complication with the web services design is interacting with services after obtaining them. For electronic information resources, this is a challenge since you must understand how to provide the resource to a user. Fortunately, the infrastructure for this is well-established; you can send a MIME type with a document and then use an

application that can handle that document type. The issue with services is that they may come in far more varieties than document varieties, and that their interactions with online services are much more intricate and sophisticated. It is now unfeasible to envision that an application could come upon a brand-new category of service and begin engaging in intricate interactions with it. As a result, customers using the web services architecture must know ahead of time what kind of service they will be communicating with, although they can dynamically learn about a specific implementation of that service. This is similar to a cross-searching application, which mixes the results from several search engines that accept a specific search protocol. The protocol is known in advance, but the specific search engines may be chosen dynamically from a search engine registry.

IBM has released its web services architecture, which describes the technology needed to enable web services in terms of three roles: service provider, service requestor, and service registry. The verbs "publish," "find," and "bind" are used to describe interactions between these roles (Figure 2.2). The process known as "bind" enables an application to establish a connection with and begin communicating with a web service at a certain web address.

A service is an implementation of a service description above this architecture, and a service description is the metadata describing the service. The interface and location of the service must be sufficiently described in this metadata for a service requestor to be able to access it; resource discovery metadata like classification may also be included:

- A service registry receives a service description published by a service provider.

- The service registry is then used by a service requester to locate the service description.

- The service requestor can connect with the service provider and use the service by using the information supplied in the service description.

Note: The architecture does permit the constrained situation where a service requester has direct access to the service description through another channel, such as hard-coding. To learn more, go to [6].

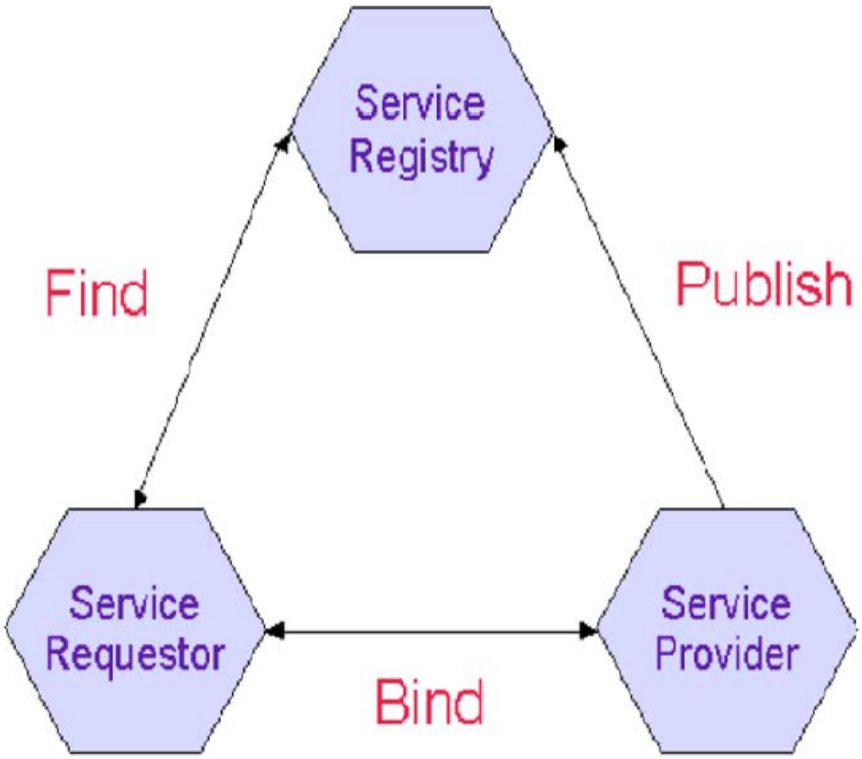

Figure 2.2: BM web services architecture, [13].

2.6 Work-Flow Web services

There is no single document that covers all of the requirements used to create Web services because they are modular in design. Furthermore, there isn't a single, reliable set of requirements. There are a few "key" requirements that may be reinforced by other ones depending on the situation and the technology chosen, such as: SOAP, WSDL and UDDI, for more details about the blow definition go look at [11].

- Simple Object Access Standard (SOAP): An XML-based protocol, used so an application can share information over HTTP and HTTPS. There are some that have been written for bindings for SMTP and XMPP. It was developed to allow for application-to-application communication. The ability for programs to communicate over the Internet is crucial for application development.

- Web Services Description Language (WSDL): is the Extensible Markup Language's format (XML). By conveying information about each other's

functionality and features, it facilitates communication across web services. Without your prior knowledge of the web service, WSDL uses a straightforward framework to define what each one offers. It makes available communication endpoints, which are points where clients can access the service and where web applications can communicate with one another.

- Universal Description, Discovery, and Integration (UDDI): is an XML-based registry that allows companies from all around the world to list themselves online, basically a Web services metadata protocol that allow publishing and discovering. Its ultimate objective is to speed up online transactions (design time or runtime) by making it possible for businesses to locate one another online and interact directly their systems for e-commerce.

With the explanation of the new technology chosen and the architecture of Web service that we say in (Figure 2.2), we now can visualize a little bit more in (Figure 2.3) and explain roughly the basic steps of a Web services work-flow.

- Establishing a Web service and its service specification is done by a service provider.
- A service provider makes the service available to the general public using a service registry that corresponds to the Universal Description, Discovery, and Integration (UDDI) specification standard.
- Now a service requester can locate a Web service via the UDDI interface since it has been published.
- A WSDL service description and a URL (uniform resource locator) directing to the requested service are both issued by the UDDI registry to the service requester.
- The service requester can then immediately bind to the service and activate it using this information.

2.7 Problems of Current Web Services

Transaction Atomicity is not provided. The single instant between its invocation and its response is actually stateless.

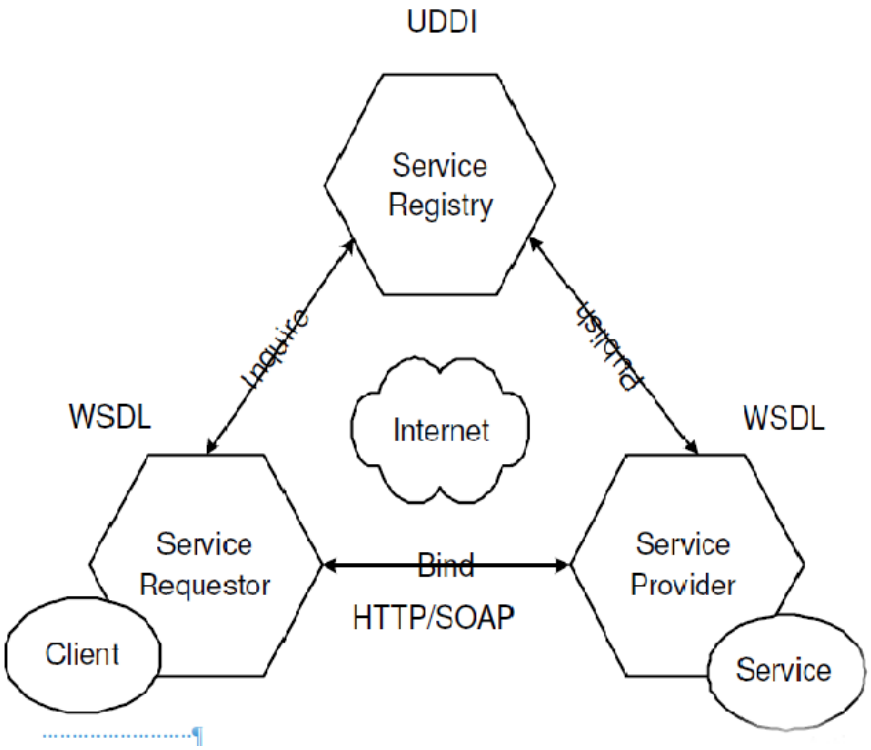

Figure 2.3: Web services work-flow, [10].

Security Unsafe transportation. To address the insecure Internet flow, further measures like encryption are required.

Interoperability Not establish new norms. To compile "profiles" of guidelines from the OASIS, W3C, and other organizations is what Web services Interoperability Organization (WS-I) aimed at. The profiles are collections of linked standards that can be used to generate conformance tests and certifications.

Reliability The Online world is by its very nature unpredictable, unstable and unreliable. The critical issue lies in the message reliability because there are no transport protocols that can address it all for the teamwork between e-business and e-transaction.

Composition From an industrial standpoint, a clear and unambiguous goals are typically unavailable. The proper management or production of data objects that appear in permanent documents is the implicit purpose of a business process. In spite of it, in these models the true purpose of a firm is frequently left unstated and is instead conveyed at a superior stage.

To go deeper into the web services work-flow visit for further information [10].

2.8 Strength

The strength of the method can be summarized as follows:

- Web services enable service providers and merchants to market their products and services by announcing their availability here on Internet.
- The separation of service interfaces from implementations and platform considerations, the ability to execute dynamic service binding, and a step closer to cross-language, cross-platform interoperability are all advantages of web services.
- These advantages emerge from the WSDL's standard XML interface and access descriptions (Web Services Description Language). Enterprise application integration, B2B integration, and grid computing all these benefit and advantages emerges from the WSDL description.
- Web services can connect programs running on different platforms, communicate database information, and make applications designated for internal usage available over the Internet.
- When implemented as utilities or as a pay-per-use application, web services have also established a proper and stronger market.

2.9 Weakness

The Weakness of the method can be summarized as follows:

- The discovery agencies are heavily used by the web services framework to showcase their presence. A poor implementation of the discovery mechanism would cause the web service's reach to its intended users and markets to suffer a significant setback.
- The Interoperability requirements are still being established; they are far from being finished. The Web Services Interoperability Organization must promptly publish the planned Web Services interoperability standards

- These still-evolving infrastructure standards must be completed and deployed before other layers and components can start constructing their frameworks on top of them.

You can have a better view about the strength and weakness of web services in [3], [4].

2.10 Web Services Reputation

In the last ten years, web service selection has grown in importance as a subject of study for web information systems. There are now several studies on this subject in the literature. Due to the popularity, several Web services that are comparable and remarkably similar have emerged, making it practically impossible to select the most appropriate one for a given functionality. Bring these two innovative concepts to life (see Figure 2.4):

2.10.1 Quality of service (QoS)

The ability to provide various applications, users, or data flows varying priority or to ensure a specific degree of performance for a data flow is known as quality of service. All parts of a connection are subject to regulations for quality of service, including response time, loss, signal-to-noise ratio, crosstalk, echo, interruptions, frequency response, volume levels, and others. In application layer services like telephone and streaming video, QoS is occasionally used to refer to a statistic that represents or forecasts the subjectively experienced quality. QoS is more than simply a guideline for a set of performance measures related to customers. It also gives a name to the idea that service providers actively control quality by prioritizing and filtering overall network traffic.

2.10.2 Quality of Experience (QoE)

In the past, Quality of Service (QoS), which seeks to quantify service characteristics objectively, gave rise to QoE. The Quality of Experience definition is that it refers to how happy or annoyed a person is with a product or service (e.g., web browsing, phone call, TV broadcast). It comes from the program or service meeting the user's expectations for usefulness and/or satisfaction in light of the user's personality and current, [5]. Overall, we may define QoE as a blueprint for

all human subjective and objective quality demands and perspectives that result from a person's interactions with technology and business organizations in a specific setting. This growing multidisciplinary discipline called QoE, which focuses on comprehending the general criteria for human quality, is based on social psychology, cognitive science, economics, and engineering science. QoE is used to enhance the likelihood that ALL network users will be able to access applications and services in a productive manner. Quality of Experience is a significant metric that matters to service users. They may learn what could be wrong with their services and how to fix them by being able to measure it in a controlled manner.

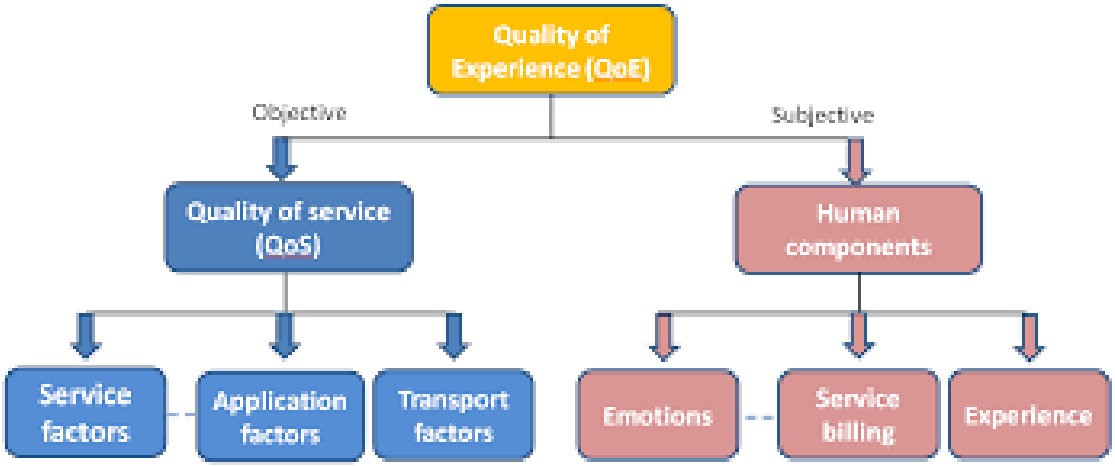

Figure 2.4: QoE vs QoS, [14].

2.11 Conclusion

Web services improved client-server communication and transformed themselves into a massive global information distributor. It is a piece of software that is identifiable by a URI and whose interfaces and binding can be specified, explained, and found using XML artifacts. An overview of Web Services is provided in this chapter. It clarifies basic and important concepts like what Web Services are, a few helpful Key Words, the Framework, Architecture, Work-Flow, even its strengths and weaknesses, and reputation. Next, we will go into a brand-new, crucial idea called "Text mining and its component" that must be thoroughly discussed for the benefit of the thesis.

3

Introduction to Text Mining

Sommaire

3.1 Introduction

The second chapter of this thesis will be devoted to a brand-new idea that is still being studied by several research institutions throughout the globe due to the wide range of opportunities it opens. Text mining has been the subject of several publications that have been and are still being released; we will highlight its main ideas and some essential components. We will cover a wide range of topics in the subject of text mining, including its process, techniques and their

comparison, applications, and main benefits and disadvantages. This new part will also demonstrate in chapter 3 how essential it is and how seamlessly it integrates with the Web Service.

3.2 Text Mining

Unstructured text makes up a significant portion of the digital data produced as a result of technological innovation and widespread use. This free-form text offers important knowledge and information. It takes applying mining algorithms to the textual data to extract knowledge from unstructured text. The non-trivial extraction of hidden, unknown, and possibly important information from textual material is referred to as text mining. This is an important area of research, it is very often referred to as text analysis, text data mining, and knowledge discovery in text (KDT).

There is a huge amount of data that is unstructured, structured, or semi-structured that is saved in various locations. Examples include text files, pdf files, emails, online chats, SMS messages, product reviews, html files, and xml files. Unstructured text is difficult for computers to process further. Therefore, a method is required that may be used to extract valuable data from unstructured text. Following that, these details are kept in text database format, which includes both structured and a few unstructured fields. By collecting it using various ways, text mining finds new pieces of information from textual material that were previously unrecognized or secret. The multidisciplinary topic of text mining includes information retrieval, information extraction, categorization, clustering, visualization, summarization. You can see in more detail at [6], [9].

There are five fundamental text mining steps (see Figure 3.1):

- Information gathering from unstructured data.
- Transform the supplied information into structured data.
- Analyze structured data to find the pattern.
- Analyze the pattern
- Retrieve the important data, then register it in the database.

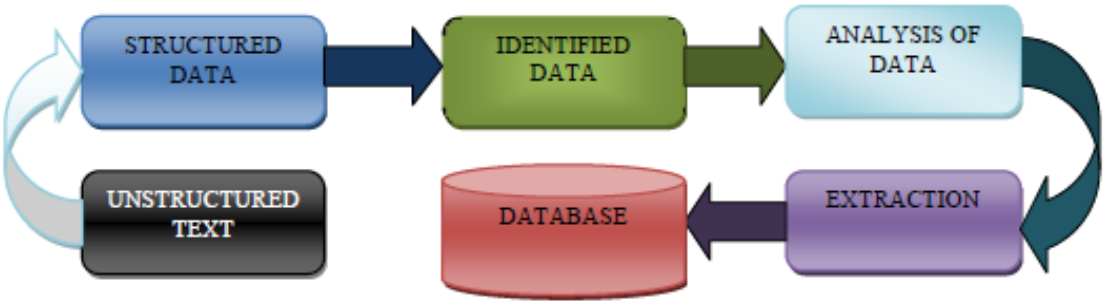

Figure 3.1: Text Mining Steps, [15].

3.3 Text Mining Process

3.3.1 Text Preprocessing

On a collection of documents containing unstructured or semi-structured data, text pre-processing is used. A raw text file is transformed into a clearly defined sequence of linguistically significant units by a text pre-processing operation. It involves the subsequent type of processing.

Text Cleanup In order for machines to understand human language, raw text must be cleaned before it can be used for natural language processing (NLP) It carries out activities like removing advertisements from websites and removing tables, figures, and other content.

Tokenization Tokenization is the process of dividing a text into token-sized pieces. Individual words, phrases, or even complete sentences can be used as tokens. Some characters, such as punctuation marks may be dropped and spaces deleted during the tokenization process.

Filtering (Stop word Removal) It removes words with little or no content information, such as articles, conjunctions, and prepositions. Words that appear a lot are also removed. Filtering constructs basic word forms in order to identify words based on their roots. (See Figure 3.2).

Lemmatization It reduces the word to its proper linguistic root, which is the verb's base form. Understanding the context comes first, followed by determining the position of a word in a sentence, and then finding the "lemma." Because

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Figure 3.2: Example of Stop Word Removal, [16].

it has to do with the semantics and POS of a sentence, its implementation is challenging, see Figure 3.3.

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Figure 3.3: Example of lemmatization, [17].

Linguistic processing Word Sense Disambiguation (WSD), Semantic Structure, and Part-of-Speech Tagging (POS) are all involved.

Part-of- speech tagging It establishes the linguistic category of a word. Each token is assigned a word class. There are eight lexeme classes in the English language: noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection, see Figure 3.4.

World Sense Disambiguous (WSD) It is the operation of determining whether a given word in a text is ambiguous. It is also the task of assigning the most

Figure 3.4: Example of Stop Word Removal, [18].

appropriate meaning to a polysemous (multiple meanings) word in a given context automatically.

Semantic structure We have two methods for constructing semantic structure:

- *Full parsing*: For a sentence, this yields a full parse tree. It frequently fails due to poor tokenization, POS tagging errors, novel words, incorrect sentence splitting, grammatical inaccuracy, and so on.

- *Partial parsing*: Also known as word chunking, generates syntactic constructs such as Noun Phrases and Verb Groups.

3.3.2 Text Transformation

It generates features before moving on to feature selection. Feature generation represents documents by the words they hold and the number of times they appear, where the order of the words is unimportant. The process of selecting a subset of important features for use in model creation is known as feature selection. It reduces dimensionality by removing features that are redundant or irrelevant.

3.3.3 Text Mining Methods

Text mining methods include classification, clustering and summarizationetc

If you want more pointer related to this section you can go and see [9], [19], [20].

3.4 Text Mining Techniques

3.4.1 Information Retrieval

A method for locating significant phrases and relationships within text to draw out important information from enormous amounts of text. Information Retrieval (IR) is the technique of identifying pertinent and related patterns from a corpus of words or phrases. Text mining and information retrieval for textual data are closely related. The most well-known (IR) are Google search engines, which identify Web pages that are related to a given set of keywords. Even though various algorithms are employed in IR systems to monitor user behavior and find pertinent info in accordance. This search engine uses algorithms that are query-based to monitor trends and produce more useful results. It is regarded as a development of document retrieval in which the returned documents are processed to extract the relevant information for the user. This offers users more pertinent and useful information that meets their needs, see Figure 3.5.

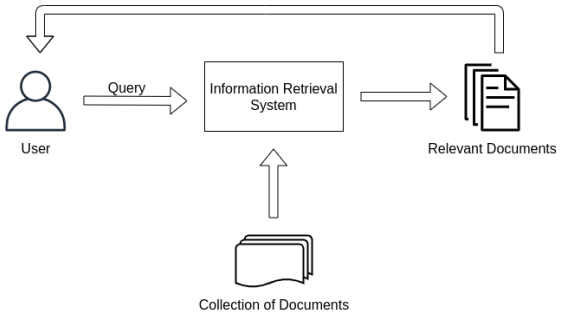

Figure 3.5: Information Retrieval, [21].

3.4.2 Information Extraction

Researchers in the relevant domain define the attributes and connections. To complete this task, a pattern matching algorithm is employed to search text for redetermined sequences. Specific characteristics and entities are extracted from the document using IE systems, and their connections are then established. Tokenization, named entity identification, sentence segmentation, and part-of-speech assignment are all tasks involved in the information extraction process. A database is used to store the extracted corpus and process it later. To verify and

assess the applicability of results on the retrieved data, the precision and recall procedure is employed. To accomplish information extraction and produce more relevant results, comprehensive and in-depth knowledge of the pertinent field is required. General information extraction process is as shown in Figure 3.6

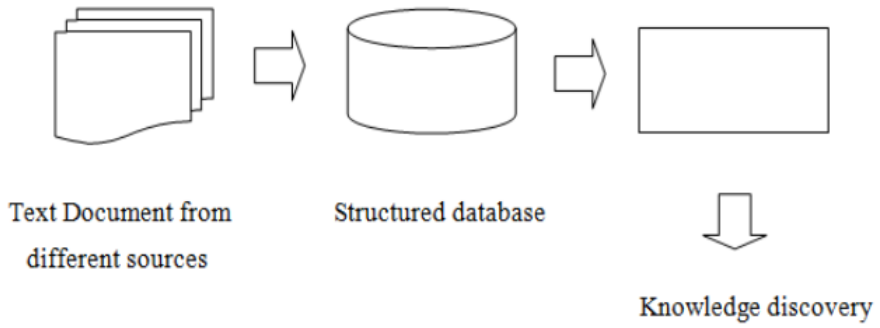

Figure 3.6: Information Extraction, [7].

3.4.3 Categorization

Text categorization is a type of "supervised" learning where the groupings for each training document are pre-determined. So, a free text document is automatically categorized into one or more categories. When a document in a normal language is categorized, its content is determined by a preset set of categories. The process of determining the appropriate themes or topics for each text document is a collection of text documents. To put it another way, the aim of categorization is to train a classifier using documented examples, after which unknown examples are automatically classified.

3.4.4 Clustering

One of the most fascinating and significant areas in text mining is clustering. You can use the clustering method to identify collections of papers having similar content. It makes use of similarity metrics to group various items into classes based on how similar they are to one another and how dissimilar they are to one another. In contrast to categorization, it clusters things without considering their class identities first (In a cluster, terms or patterns that are similar across several papers are gathered.). So, the challenge is to use no prior knowledge to appropriately cluster the supplied unidentifiable collection. For instance, document clustering helps with retrieval by establishing connections between similar

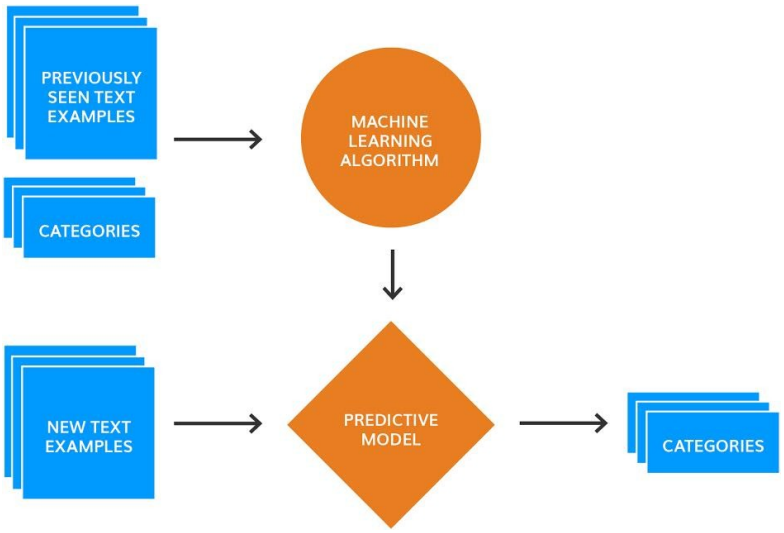

Figure 3.7: Categorization, [7].

documents, which, in turn, facilitates the retrieval of related documents whenever one of the documents is determined to be pertinent to a search. Therefore, clustering assures that a helpful item will not be missed from the search results because materials can appear in several subtopics, see Figure 3.8.

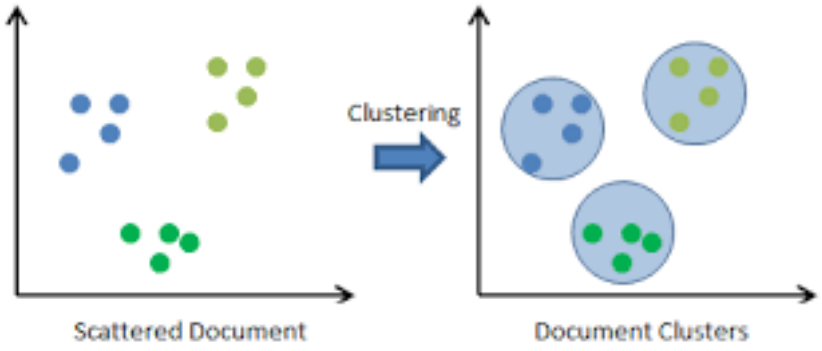

Figure 3.8: General idea about Clustering, [22].

3.4.5 Visualization

In text mining, visualization techniques can help to speed up and improve the discovery of relevant information. Text flags are used to indicate document category and density, as well as to represent specific documents or groups of documents. Visual text mining organizes large textual sources visually. The user can make changes to the document by zooming and resizing it.

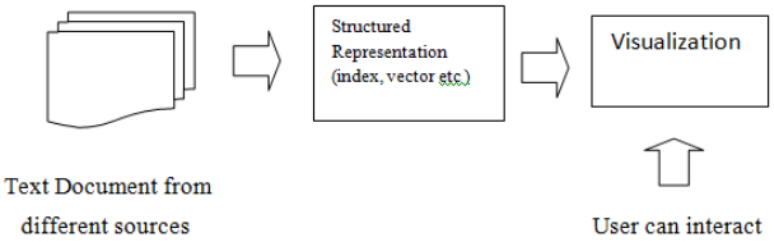

Figure 3.9: Visualization, [7].

3.4.6 Summarization

Text summarization is an old problem in text mining that requires urgent attention from researchers in the fields of computational intelligence, machine learning, and natural language processing. It is the process of condensing a text into a shorter form while retaining the information and overall meaning. The process of automatically creating a compressed version of a given text that provides useful information to the user is known as text summarization. Because researchers in large organizations or companies do not have time to read all documents, they summarize documents and highlight summary with main points. Previously, automatic text summarization was performed based on the presence of a specific word or phrase in a document. Later, additional text mining methods were introduced alongside the standard text mining process to improve the relevance and accuracy of results, see Figure 3.1.

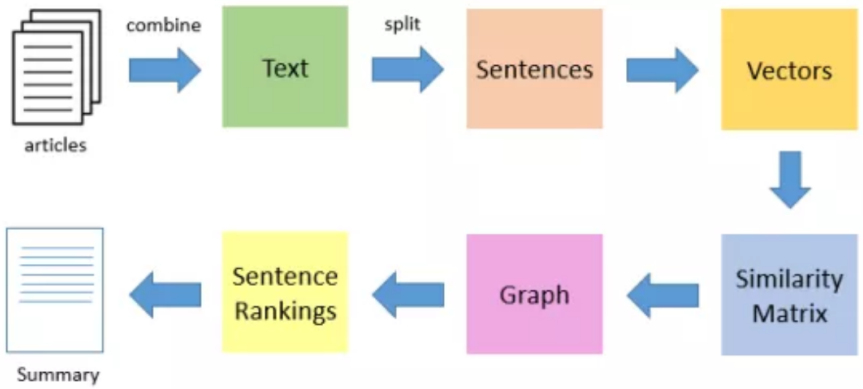

Figure 3.10: Automatic text summarization, [23].

The majority of text mining techniques are the same, however some are preferred by researchers over others. You may visit [6]–[9] to see more of these techniques or even more alternatives.

3.5 Comparison Text Mining Technique

The primary models, algorithms, and tools are displayed in this section. Many different strategies are used in text mining, and they all have a significant impact, [6], [19].

- The information retrieval method utilized unstructured text, from which it could extract useful information.
- Information extraction takes data from databases with structured data and extracts it.
- The material is summarized using the summarization approach, which shortens the text while maintaining its significance.
- The categorization is a procedure that is under supervision and makes use of preset set documents based on their material.
- Clustering is a technique for recognizing underlying data structures and arranging them into coherent groupings for further research and analysis. Furthermore, it works with highly dimensional data, revealing surprising data patterns. Another feature is the grouping of data of a similar type and the links that exist between them.

3.6 Application of Text Mining

3.6.1 Classification of Scientific Documents

Text mining techniques like as classification, clustering, and regression analysis are used by scientists to find articles that are relevant to their research. The subject score engine replaced keyword searches with a structured automated method, which accelerated research and improved the accuracy of the results.

Bioinformatics Biomedical text mining is text mining applied to biomedical and molecular biology texts and publications. The inspiration for this study comes mostly from the tremendous growth in the quantity of publications in this area, making it practically impossible for many biologists to keep up with the associated literature. The term "biomedical text mining" refers to text mining techniques used on texts and publications in the field of molecular biology

and biomedicine. This effort mostly comes from biologists who are confronted with significant publications increase in their area, making it practically impossible for many scientists to keep up with the associated literature. Bio-entity research is still in its early stages. Its purpose is to detect and categorize scientific terms used in molecular biology that are connected to particular instances of notions useful to biologists. With the massive increase of reported data caused by high throughput experimental procedures, entity recognition is becoming increasingly important.

Life Science The life science and healthcare industries generate massive volumes of textual and numerical data concerning, among other things, patient records, sicknesses, drugs, disease symptoms, and therapy. It is tough to find a text that is appropriate and relevant for making a decision. The vocabulary used in medical records is diverse, convoluted, extensive, and specialized, making knowledge finding a particularly difficult task. The application of text mining methods in the biomedical field allows for the extraction of valuable data as well as the association and inference of correlations between various diseases, species, and genes. By comparing various diseases, symptoms, and their courses of therapy, the use of suitable text mining technologies in the medical area aids in evaluating the efficacy of medical therapies. Text mining is employed in many sectors, including biomarker discovery, the pharmaceutical industry, clinical trade analysis, preclinical safe toxicity research, competitive intelligence and landscape design for patents, disease mapping using genes, and studying focused identifications.

Academic and Research Field The educational field uses a variety of text mining methods and methodologies to examine regional educational patterns, students learning motivation in particular fields, and employment rates. Text mining in the research industry makes it simpler to identify and arrange research papers and relevant data from several domains in one spot. It is possible to access student achievement in a variety of categories and discover how various factors impact study choice.

3.6.2 Security

Text mining methods are increasingly being used in the field of security. Many text mining software tools are offered for security purposes, particularly for monitoring and analyzing online plain text sources such as Internet news, blogs, and mail. Because of the urgency of the situation and the enormity of the problem, text mining approaches have enormous promise in this subject.

Junk Emails Junk emails are unwanted or unsolicited email messages sent by a firm for marketing or promotional objectives. Email is used in many legal processes, including the exchange of information and documents. Unfortunately, it may also be misused. Spam, or unsolicited email, has grown at an exponential rate in recent years, undermining email's value. Anti-spam filters are one solution. Hand-written rules and blacklists are used in the bulk of commercially available filters.

Social Media Text mining software packages may be used to monitor and analyze online plain text from internet news, blogs, and other sources. Text mining algorithms may be used to find and evaluate the number of posts, likes, and followers on social media networks. This form of research illustrates how people react to various articles and news items, as well as how they spread. It shows how individuals of a certain age group or community act when expressing similar and differing ideas about the same article.

3.6.3 Business Intelligence

Text mining is significantly used in business intelligence to analyze consumers and competitors so that organizations and enterprises may make better decisions. Text mining technologies are useful for identifying a certain topic. It provides a better grasp of business as well as expertise on how to boost customer satisfaction and get a competitive advantage. It also helps with business and commerce applications, telecommunications, and customer chain management systems.

Human Resource Management The management of human resources can also benefit from text mining. HRM refers to formal procedures designed for the administration of personnel inside a business. Human resource managers

duties mostly come under the headings of staffing, employee remuneration and benefits, and defining/designing work. The ideal use of office mining, for instance, is the analysis of employee perspectives. Mining technology may be used to store fresh CVs as well as to evaluate business success and worker overall satisfaction, both of which are significant tasks.

Customer Relationship Management CRM is responsible for responding to each client correspondence or query as soon as possible. Through textual analysis, these communications or enquiries are sent to the appropriate party or service for further action.

3.6.4 Other point:

- Banks, insurance and financial markets.
- Classification of NEW as Text.
- Multilingual App of Natural Language Processing.
- Telecommunications, energy and other services industries.
- Web search Enhancement.

3.7 Advantage and Disadvantage in Text Mining

3.7.1 Advantages of Text Mining

- Since databases can only hold so much data, text mining has been used to tackle this issue.
- Managing such a large volume of unstructured data for the purpose of quickly finding patterns has been a challenging task that has been resolved by text mining.
- Using a method like information extraction, it is simple to find the names of various entities and the relationships between them in a corpus of texts.
- The most remarkable aspect of text mining is our capacity to interpret data statistically on a scale that was previously unimaginable.

3.7.2 Disadvantages of Text Mining

- Programs cannot be used to directly evaluate unstructured text in order to mine it for knowledge or information.
- The text papers do not provide the beginning information required.
- Reliance on multilingual text refining causes issues.
- The majority of texts are written in natural language.
- The ambiguity/vagueness issue is present in natural language as well. You can read [6]–[8], [19] to get some further complementary viewpoints on the pros and cons.

3.8 Conclusion

Today a huge volume of digital data is available in computer world and most of them in textual form. To extract information from this unstructured document, text mining techniques are applied. These techniques are used to analyze the interesting and relevant information effectively and efficiently from large amount of unstructured data. In this chapter about text mining, several text mining process, techniques, a comparison of different text mining technique, applications in various fields have been discussed and the advantage/ disadvantage of it. We have noticed that the selection and use of right techniques and tools according to the domain, help to make the text mining process easy and efficient.

4

Reputation Assessment

Sommaire

4.1 Introduction

Finally, the chapter that will englobe and give meaning to the two previous chapters. For this part of the thesis, we will see all the previous knowledge that we explained being used to it fullness. We took some codes written in python and implemented them with the right parameter. We made it into the perfect example for our prospect. We will explain the work that was made in it and add some clarification to show the ingenuity of it.

4.2 Used directory and evaluation metrics

4.2.1 ProgrammableWeb

An information and news source regarding the Web as a programmable platform is called ProgrammableWeb, [24]. Over 19,000 open online APIs and hundreds of apps have been documented by the website, which also maintains a library of web APIs, mashups, and applications. TechCrunch referred to it as

the "journal of the API economy." ProgrammableWeb is divided into two main sections:

1. Technology and API Directories.

2. News, commentary, and reviews.

4.2.2 G2Crowd

Previously known as G2, it is a portal for peer reviews. The business specializes in compiling user reviews of business software. Users of the platform may log in using their LinkedIn/Facebook/Email accounts and provide reviews of the things they use. Some users are rewarded by G2.com with gift cards, contest prizes, and reputation points on the website in order to entice reviewers to participate. G2 uses an algorithm to detect workers of companies rating their own goods and employees of firms reviewing their competitors' products in an effort to spot fake user evaluations. In order to deter spam and useless remarks, reviews are further manually reviewed and decided upon by the community. For reasons of certification, G2 also asks for screenshots of the reviewer using the product, [25].

4.2.3 Evaluation metrics

To evaluate the performance of the method, we use two metrics *Precision* and *Recall*. The first one is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p):

$$Precision = \frac{T_p}{T_p + F_p} \tag{4.1}$$

The second one is defined as the number of true positives over the number of true positives plus the number of false negatives (F_n):

$$Recall = \frac{T_p}{T_p + F_n} \tag{4.2}$$

4.3 Modelization

We used "Modelio" to draft some simple diagrams.They show us the essential steps of the algorithm and how it works. In this Class diagram we have shown the fundamental action of the followed algorithm and how they relate to each other, see 4.1.

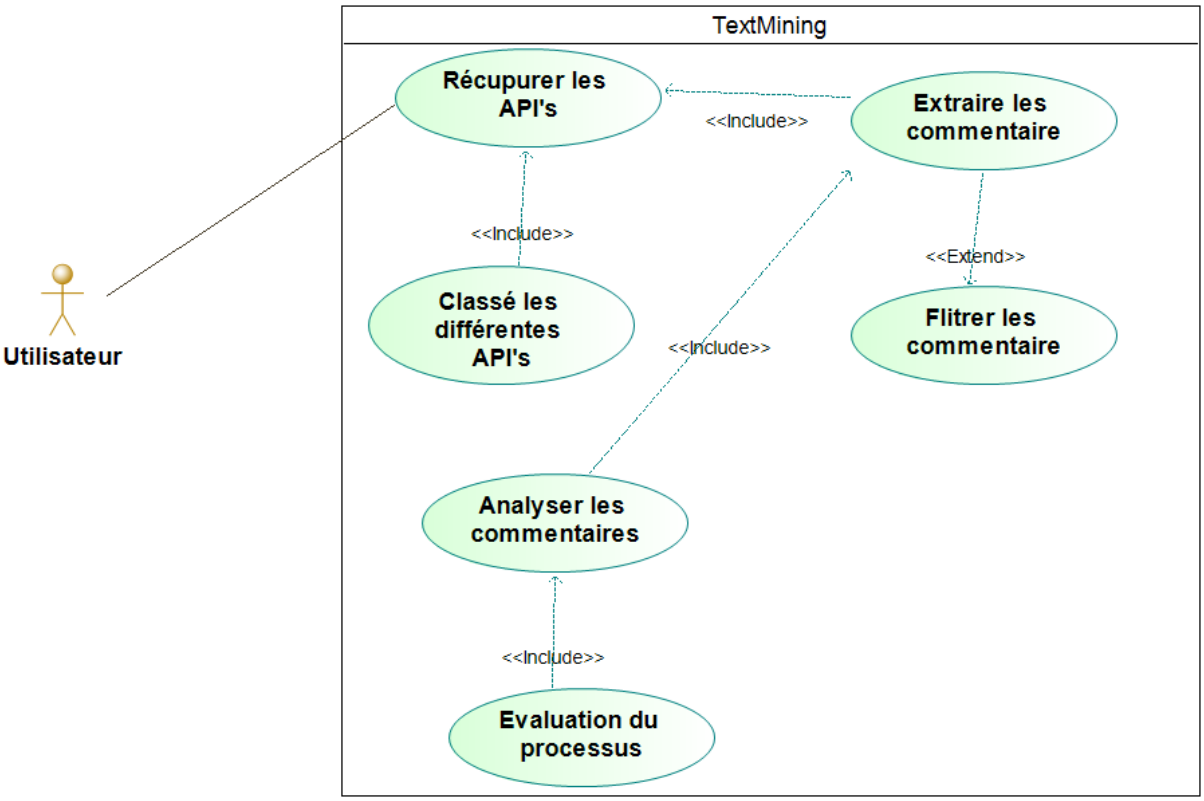

Figure 4.1: Class diagram.

This sequence diagram demonstrate that we take the data stocked in our dataset and execute algorithmes/methodes on them, see 4.2

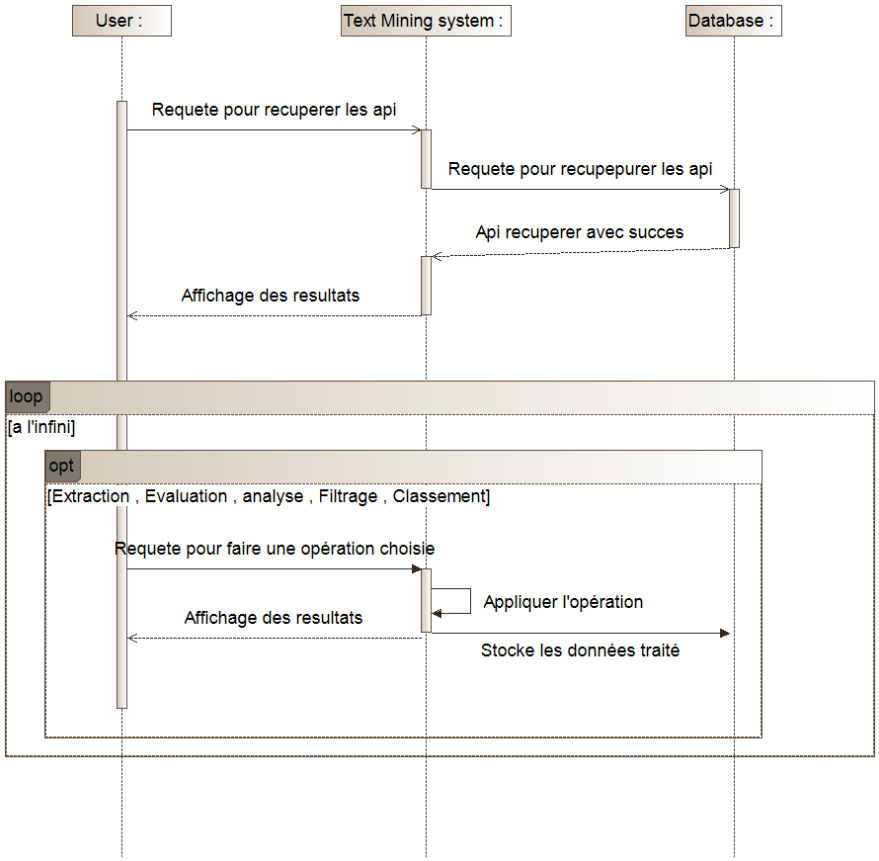

Figure 4.2: Sequence diagram.

The activity diagram present in our views is the best fit diagram for a work like this. We show in it how the choices are made in the algorithm (work-flow) from a simple API data to the rank of service reputation, everything can be seen at 4.3.

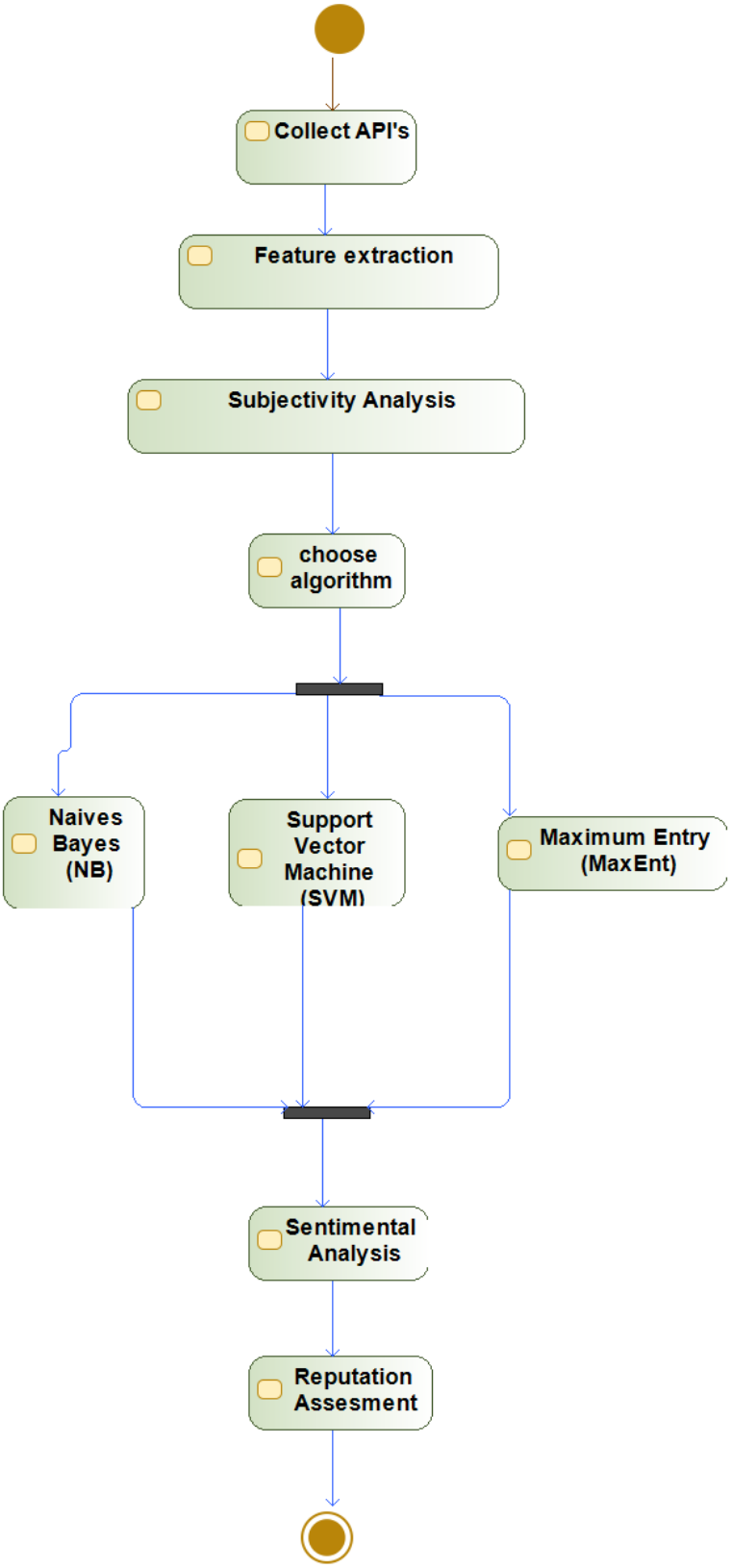

Figure 4.3: Activity diagram.

4.4 Assessment Process / Experimental Evaluation

The many phases that will be followed to mine service user evaluations are described in this section. The phases are as follows:

1. Prepossessing of data
2. Feature Extraction.
3. Subjectivity Analysis.
4. Sentiment Analysis.
5. Reputation Assessment.

4.4.1 Prepossessing of Data

The process of getting the most user reviews of Web services is known as data collection (two Web service directories, ProgrammableWeb and G2Crowd, were crawled to collect users reviews which express their experience when using the Web service). To document the many methods that consumers express their ideas, the reviews got gathered from various sources. Furthermore, we anticipate that reviews frequently include a range of perspectives. The review that provides two opposing viewpoints, get divided into sentences to conduct a sentence-level analysis. A scraper automatically browses through each Web service to collect all of the evaluations and the Web service description in order to get the data. Web scraping produced 12 962 and 2 718 Web services, respectively, with 715 and 16 132 reviews from ProgrammableWeb and G2Crowd. These services fall under the following categories: social, sales, marketing, development, messaging, images, video, and music. Because the G2Crowd repository already separates user ratings into positive and negative reviews. The supervised machine learning algorithms can be trained more easily thanks to this prelabeled data. Figure 4.4 depicts an illustration of a user review.

4.4.2 Feature Extraction

In chapter 2, feature extraction was described. It is carried out utilizing several methodologies that might result in many models of the text. The following lists them in order of simplicity to degree of complexity:

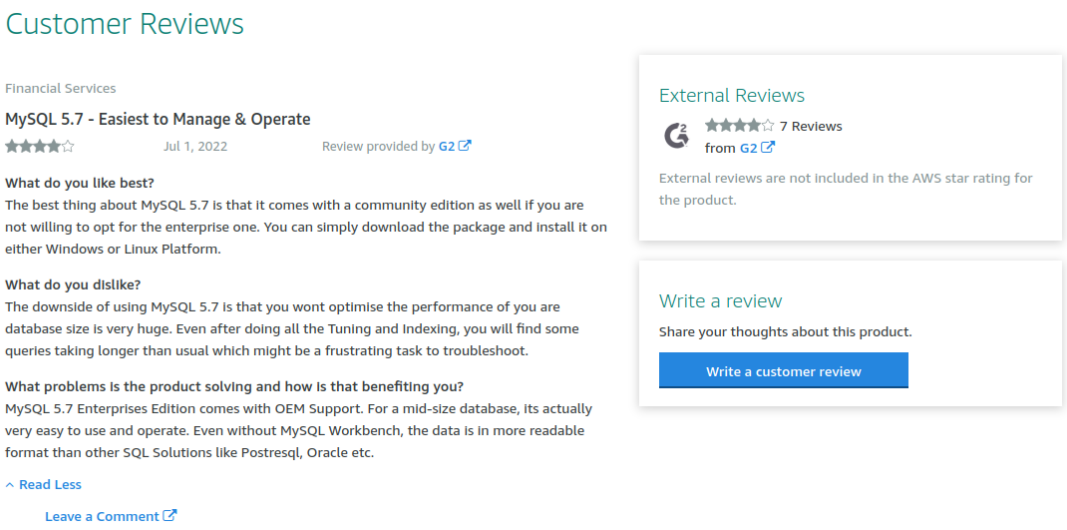

Figure 4.4: User Review Example.

1. Bag of Words (BoW).
2. N-grams.
3. Part of speech (PoS).
4. Term Frequency-Inverse Document Frequency (TF-IDF).

Bag of Words (BoW) This is a straightforward and adaptable methodology that builds a vocabulary of terms based on the word occurrences in a text. In order to locate comparable sentences based on their content and understand the meaning of the phrase, this is called a "bag" since information about the structure or order is disregarded.

N-grams By constructing a vocabulary made up of word groupings, n-grams enhance the BoW technique by helping to better understand the text. Every collection of words is referred to as a "gram," therefore an n-gram is a run of n words. Two-word (2-gram) and three-word (3-gram) sequences are referred to as bigrams and trigrams, respectively.

Part of speech (PoS) This approach labels words according to the grammatical function they serve in user evaluations (nouns, adjectives, adverbs, and so on). The goal of this strategy is to give each word its own treatment by differentiating them based on their function in the sentences.

Term Frequency-Inverse Document Frequency (TF-IDF) The TF-IDF information retrieval approach gives each word in a document a weight, where the weight's value indicates how significant the word is to the text. The term frequency and the inverse document frequency, two statistical measures of word frequency, are combined to create the TF-IDF.

1. Term Frequency (TF): TF measures how frequently a word appears in a text.

2. Inverse Document Frequency (IDF): IDF measures a word's importance across all texts.

The product of multiplying each word's TF and IDF is known as TF-IDF. The term is odd if TF-IDF has a high value, and more frequent if TF-IDF has a low value. Finally, a characteristic vector is constructed using the TF-IDF of each word in a phrase.

Each user review got broke up into sentences because a single review might have several viewpoints but a sentence often only has one perspective. Following that, features from sentences are extracted in order to create a simplified representation of each sentence that may be used by machine learning algorithms. Before extracting features from the user review, punctuation and stopwords were deleted.

For instance, "The durability of the storage is quoted as being able to survive concurrent faults, which is absolutely awesome." and "at times the retrieval at data can incur some delay." Look at the table below.

Note: Bag of Words created vector representations for each phrase based on word frequency. Words were stemmed and lemmatized for Bigrams in order to treat words with varied conjugations, verbal time, or number (plural or singular) as the same word. Part of Speech and TF-IDF, on the other hand, standardize the words to determine the right grammatical function of the words and weights for each word.

4.4.3 Subjectivity Analysis

The next stage is to determine whether a sentence reflects an opinion or not after the pertinent components have been removed from sentences. This procedure divides each sentence into subjective and objective sentences using the qualities that were derived from the previous step.

Bag of Words
[the, durability, of, the, storage, is, quoted, as, being, able, to, survive, concurrent, faults, which, is, absolutely, awesome]
[at, times, the, retrieval, at, data, can, incur, some, delay]
Bigrams
[(The, durability), (durability, of), (of, the), (the, storage), (storage, is), (is, quoted), (quoted, as), (as, being), (being, able), (able, to), (to, survive), (survive, concurrent), (concurrent, faults), (faults, which), (which, is), (is, absolutely), (absolutely, awesome)]
[(at, times), (times, the), (the, retrieval), (retrieval, at), (at, data), (data, can), (can, incur), (incur, some), (some, delay)]
Part of Speech (durability, NN), (storage, NN), (quote, VVN), (able, JJ), (survive, VVP), (concurrent, JJ), (fault, NNS), (absolutely, RB), (awesome, JJ)
[(times, CC), (retrieval, NN), (datum, NNS), (incur, VVP), (delay, NN)]
TF-IDF
durabl: 0.3482, storag: 0.2751, quot: 0.3482, abl: 0.2751, surviv: 0.3482, concurr: 0.3482, fault: 0.3482, absolut: 0.3482, awesom: 0.3482
time: 0.4185, retriev: 0.3769, data: 0.4771, incur: 0.4771, delay: 0.4771

Table 4.1: Sentences Example.

- *Subjective sentences*: These are statements that indicate a user viewpoint, whether it be favorable or bad, and originate from a person.
- *Objective sentences*: these statements represent a fact, a description, or factual information rather than a user's opinion. These statements lack polarity and cannot be categorized as either positive or negative.

Multiple supervised machine learning approaches been used to conduct this categorization, including Naive Bayes, Support Vector Machines, and Maximum Entropy.

Naives Bayes (NB) This straightforward probabilistic approach involves: The assumption is that features-occurrences of a particular instance of a text feature-are independent variables in the calculation of the likelihood that a user review would fall into a certain class (for instance, an "objective" or "subjective" sentence in a review). A pre-processed dataset of user reviews and their preset class labels are used to train the constructed classifier (objective and subjective, for instance). Then, it makes it possible to forecast a class with a particular probability using conditional probabilities for a given fresh review that is characterized by its text

attributes.

Support Vector Machine (SVM) It is a non-probabilistic supervised learning technique in this instance, where user evaluations are represented as points in a multi-dimensional space based on their textual properties. The algorithm then determines a (maximum-margin) hyperplane that allows the reviews in various classes to be separated (objective reviews and subjective ones). The same dataset used in the prior method's training is used to train the classifier created using this technique. Then, for a fresh review, it may forecast class membership based on its location in the area.

Maximum Entry (MaxEnt) It is a non-probabilistic supervised learning technique in this instance, where user evaluations are represented as points in a multi-dimensional space based on their textual properties. The algorithm then determines a (maximum-margin) hyperplane that allows the reviews in various classes to be separated (objective reviews and subjective ones). The same dataset used in the prior method's training is used to train the classifier created using this technique. Then, for a fresh review, it may forecast class membership based on its location in the area. Table 4.2 displays the findings of the study.

	Subjective Sent			Objective Sent.		
	NB	SVM	MaxEnt	NB	SVM	MaxEnt
	Bag of Words					
Precision	0.9417	0.9347	0.9407	0.9192	0.9220	0.9319
Recall	0.9038	0.9109	0.9134	0.9404	0.9276	0.9332
	Bigrams					
Precision	0.9415	0.9398	0.9217	0.9212	0.8926	0.9039
Recall	0.9148	0.8711	0.8832	0.9533	0.9449	0.9325
	Part of Speech					
Precision	0.8785	0.9125	0.9200	0.9322	0.9047	0.9115
Recall	0.9300	0.7974	0.9106	0.8833	0.9069	0.9183
	TF-IDF					
Precision	0.9767	0.9013	0.8816	0.5237	0.9077	0.8723
Recall	0.9103	0.9453	0.9342	0.9997	0.8967	0.0214

Table 4.2: Subjectivity Classification Results.

Note: In contrast to many others, these three approaches were chosen for their simplicity (just a few parameters to modify the method) in addressing this two-class (objective or subjective) classification issue. The first and final

approaches are complimentary probabilistic simple procedures, meanwhile the third is a well-known efficient text-mining method. Furthermore, as compared to most other machine learning algorithms, they all deliver in average rapid training times and minimal overfitting.

4.5 Sentimental Analysis

In this section, the opinion polarity of each subjective statement is determined. It is a field that investigates people's feelings, appreciations, attitudes, and emotions toward a thing or any aspect of an entity. Products, services, organizations, events, people, or concepts are examples of these entities. Attitude analysis is concerned with opinions that indicate or imply a positive or negative sentiment. Subjective sentences are those that communicate a feeling or an opinion, as opposed to objective statements that declare or describe a fact. Because sentiment analysis detects opinions in text, it is necessary to specify what constitutes an opinion.

An opinion is a wide notion that denotes a person's mood, assessment, admiration, or attitude toward a certain item (the opinion target) (the opinion holder). On the other side, the opinion target is the person whose viewpoint is being supported by the opinion holder, who displays a certain emotion during the contact.

The sentiment, on the other hand, is a feeling, attitude, judgment, or emotions connected to the viewpoint that has a polarity. Through this approach, the sentiment connected to the subjective statements is identified, along with its polarity. Similar to the previous procedure, this procedure uses several machine learning algorithms for supervised classification on the characteristics retrieved from the phrase.

Several tests have been included in the sentiment classification literature to assess the performance of various feature types and classification methods. 56 386 positive sentences and 56 386 negative sentences have been chosen in this instance. Eighty percent of these phrases were used for training and twenty percent were used for testing. Similar to before, k-fold cross validation with k=5 and precision, recall, and f-measure were used to compare and validate classification algorithms.

Table 3.3.4 displays the findings of the study.

Note: Sentiment classification behaved similarly to subjective classification.

	Subjective Sent			Objective Sent.		
	NB	SVM	MaxEnt	NB	SVM	MaxEnt
	Bag of Words					
Precision	0.7981	0.8001	0.8210	0.7749	0.7549	0.7952
Recall	0.7387	0.7033	0.8974	0.6957	0.7413	0.8072
	Bigrams					
Precision	0.7322	0.7521	0.7953	0.8117	0.7919	0.7743
Recall	0.8250	0.8017	0.7982	0.7787	0.7746	0.8167
	Part of Speech					
Precision	0.7268	0.7102	0.7602	0.7103	0.7186	0.7558
Recall	0.7030	0.7262	0.7543	0.7294	0.7005	0.7591
	TF-IDF					
Precision	0.8186	0.7831	0.5918	0.5049	0.6942	0.7089
Recall	0.0397	0.7581	0.9164	0.9909	0.6728	0.0186

Table 4.3: Sentiment Classification Results.

Bigrams outperformed the other features, and Naive Bayes outperformed SVM and Maximum Entropy. Once again, the likelihood of a given bigram belonging to a class is greater than that of a bag of words, part of speech, or TF-IDF.

Multiple characteristics provided no additional advantage since features with a low chance of belonging solely to one class were included. Similarly, SVM produced superior results when bigrams were used to create more distinct feature vectors between classes than other features. The addition of more features just enhanced the similarity between feature vectors in both groups.

The limitations in the most frequent features worked effectively in bag of words, bigrams, and even part of speech, but not in TF-IDF since the inverse frequency in each word for positive and negative phrases was identical.

On the other hand, because the analysis was done at the sentence level, the possibility of a bigram being present in separate sentences is quite low (When a phrase with bigrams is not included in the training set, a priori information can be computed from each word individually in order to calculate the polarity of the bigram.). Because each word is independent of the others, Naive Bayes produced superior results.

4.6 Reputation Assessment

Reputation Evaluation: In this stage, sentence polarities are combined to provide a single reputation score for each Web service. Based on the quantity of both good and/or negative user evaluations, a single score is computed for each Web service in this procedure.

Note: Using a subtraction of positive reviews from negative reviews or measuring the average of positive reviews to get the Web service score is not an acceptable approach since it ignores the ratio of positive to negative ratings. For instance, suppose we had a service with:

- A) 5 positive reviews and 0 negative.
- B) 13 positive reviews and 8 negatives.
- C) 11 positive reviews and 0 negative.

A subtraction gives us:

- A) 5
- B) 5
- C) 11

This signifies that services A and B are both excellent. This is not true because service B has received unfavorable feedback. The average for favorable ratings, on the other hand, is 1, 0.619, and 1 for services A, B, and C, respectively. This is also incorrect since it implies that services A and C are equally good. To provide a balanced score for each Web service, the Bayesian Average algorithm is used.

The application procedure was carried out on an unlabeled dataset of Web service in order to make a fair and objective evaluation of the quality assessment phase. Bigrams and Naive Bayes were the features and classifiers they chose since they produced the greatest outcomes in the earlier stages of the process. Their application in emotion classification and subjective categorization. Then performed a brief text pre-processing on our ProgrammableWeb dataset prior to running the procedure to remove distracting material from the reviews, such as misspellings and strange characters, and we trained our classifiers using the entire dataset gathered from G2Crowd.

Table 4.4 displays a representative sample of the Web service ranking result list based on its evaluation score. Score values were trimmed to four decimals since such sparse data need a high level of accuracy.

Name	Positive	Negative	Score
Wordstream Keyword Niche Finder	8	0	0.5000
PeerIndex	6	0	0.4960
Basecamp	6	2	0.4920
FotoFlexer	3	1	0.4879
Shopzilla	3	2	0.4859
FedEx	1	0	0.4857
EarthTools	0	0	0.4836
Salesforce.com	1	2	0.4817
Yelp	0	1	0.4816
Shutterfly	1	3	0.4798
Active	0	2	0.4796
Trulia	1	4	0.4779
Shopping.com	2	11	0.4669

Table 4.4: Representative sample of the Web service ranking.

Notes:

- To begin, bigram characteristics were retrieved for each review.
- Second, subjectivity classification distinguished between subjective and objective statements.
- Finally, sentiment classification classified subjective statements as favorable or negative. A score of 20 is assigned to each Web service depending on the quantity of good and negative sentences.

4.7 Results

Following the collection of all user ratings for the reviews, we identified three different sorts of results that fall into the Included/Excluded category:

- *Included*:
  - Reviews with a unified opinion: Reviews with converged ratings, where each Web service's three user ratings are higher than zero.

- *Excluded*:
  - Reviews with inconsistent user ratings: These are reviews in which the user ratings vary widely or in which at least one rating deviates significantly from the others.
  - Reviews that are neutral: These reviews have a total user score of zero.

The divergence amongst users is attributable to the review's obscurity/ambiguity or the users' perception of it, according to a more in-depth observation. Each user's perspective affects it in its own unique way. As a consequence, 174 reviews for 70 Web services were compiled, with 121 of them being good and 53 of them being negative.

Table 4.5 demonstrates that using user feedback to gauge Web service quality produced positive outcomes. However, we found that recall in positive statements and relatively poor precision in negative ones. These ideals were brought about by:

- The inclusion of positive lines in evaluations that the students had rated as unfavorable.
- The existence of negative sentences inside reviews that were given favorable ratings by the students.

In fact, the procedure analyzes sentences at the sentence level, and the student ratings were for the entire evaluation. Since many evaluations begin with positive lines and conclude with negative ones or begin with negative sentences and end with positive ones, students admitted that they often struggled to decide between negative and positive throughout an entire review. Additionally, the same review may be categorized differently by various users due to the ambiguity of certain reviews, making it impossible for the classifier to guarantee accuracy of 100 percent. A significant portion of the dataset and the manual analysis of "abnormal" manual scoring situations were done in order to reduce this danger.

4.8 Conclusion

In this chapter, we have presented a piece of work that demonstrates the integration of the previous two chapters' applications, with further clarification

	Precision	Recall
Positive Sentences	0.91	0.72
Negative Sentences	0.64	0.88

Table 4.5: Comparison of text mining techniques.

of key algorithms, a broad conceptual context, and primarily the outline of the algorithm that we have set to the ideal parameter. By identifying them and linking their outcomes, we have demonstrated the five primary steps that were followed in the process. This chapter's sole goal was to present a piece of work demonstrating the coherence of Web Services and Text Mining.

5

General Conclusion

These three themes Web Services, Web Service Reputation, and Text Mininghave been the focus of our search and information gathering efforts for this thesis. To make it easy for the reader to follow the work, we made an effort to shrewdly select the most important material and deliver it to them in a coherent manner. We selected a certain element for each chapter that, in our viewpoint, serves as the foundation for all of them.

The information you should possess after reading the chapter on Web Services, in our opinion, must be connected to the Framework, Architecture, strengths and weaknesses, work-flow, and reputation. With the this research objective in mind, each one of them needed to be explained in order to move further. The next step is text mining, which presents the most challenge due to the variety of its content. We have demonstrated the procedure, methods, and Text Mining application. Each of these parts is made up of a number of separate components, each of which has had its procedures, approaches, or algorithms explained in the most straightforward manner possible. Some of these components have also been given examples to help you understand them properly. Finally, we have demonstrated how Service Web and Text Mining are compatible. You can see how the two of them work together to improve user experience, as well as all the Text Mining techniques and stages utilized on user reviews obtained from the ProgrammableWeb and G2Crowd directories. The end outcome serves as education on application and differences of the applied substance. Our goal was to educate the audience on all of these concepts, and we did so by using a

specific practical python algorithm parametered by us as a means of doing so.

This effort will be a little service to all students and researchers who want clear, concise explanations of the fundamentals.

Bibliography

[1] F. Curbera, W. Nagy, and S. Weerawarana, "Web services: Why and how," in *Workshop on Object-Oriented Web Services-OOPSLA*, sn, vol. 2001, 2001.

[2] T. Gardner, "An introduction to web services," *Ariadne*, vol. 29, 2001.

[3] C. Ferris and J. Farrell, "What are web services?" *Communications of the ACM*, vol. 46, no. 6, p. 31, 2003.

[4] H. Kreger, "Fulfilling the web services promise," *Communications of the ACM*, vol. 46, no. 6, 29–ff, 2003.

[5] *Qualinet, European Network on Quality of Experience in Multimedia Systems and Services*. [Online]. Available: `http://www.qualinet.eu/`.

[6] S. Dang and P. H. Ahmad, "Text mining: Techniques and its application," *International Journal of Engineering & Technology Innovations*, vol. 1, no. 4, pp. 22–25, 2014.

[7] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," *International Journal of Computer Applications*, vol. 85, no. 17, 2014.

[8] N. Goel, "A study of text mining techniques: Applications and issues," *Pramana Research Journal*, vol. 8, pp. 2249–2976, 2018.

[9] L. Gohil, "Text mining: Process and techniques," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 3, no. 3, pp. 2347–5552, 2015.

[10] C. P. Wah, "Building reliable web services: Methodology, composition, modeling and experiment," PHD, 2008.

[11] *Techopedia*. [Online]. Available: `https://www.techopedia.com/`.

[12] J. S. de Bruin, "Service-oriented discovery of knowledge: Foundations, implementations and applications," Ph.D. dissertation, Leiden University, 2010.

[13] S. Baloch, "Temporal reasoning for web services composition for personal assistants," M.S. thesis, 2020.

[14] *Framework, for Assessment of Quality of Service of Telecommunications Systems and Services*. [Online]. Available: https://www.ca.go.ke/wp-content/uploads/2018/02/Framework-for-Assesment-of-Quality-of-Service-of-Telecommunications-Systems-and-Services-1.pdf.

[15] *upgrad, Digital Journal, What is Text Mining: Techniques and Applications*. [Online]. Available: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/.

[16] *Geeksforgeeks, Digital Journal, Removing stop words with NLTK in Python*. [Online]. Available: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/.

[17] *freecodecamp, Forum, All you need to know about text preprocessing for NLP and Machine Learning*. [Online]. Available: https://www.freecodecamp.org/news/all-you-need-to-know-about-text-preprocessing-for-nlp-and-machine-learning-bc1c5765ff67/.

[18] *Mosaix, Artificial Intelligence to Understand Human Language, Forum, Applying Unsupervised Machine Learning to Sequence Labeling*. [Online]. Available: https://medium.com/mosaix/deep-text-representation-for-sequence-labeling-2f2e605ed9d.

[19] S. Sheela and T. Bharathi, "Analysing different approaches of text mining techniques and applications," *International Journal of Computer Science Trends and Technology*, vol. 6, no. 4, pp. 23–29, 2018.

[20] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text mining: Techniques, applications and issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, 2016.

[21] *Analytics Vidhya, Forum, Information Retrieval using word2vec based Vector Space Model*. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/08/information-retrieval-using-word2vec-based-vector-space-model/.

[22] *Medium, Stay Currious, Discover stories, thinking and expertise from writers from any topics, Forum, K-means Clustering and its use-case in the Security Domain*. [Online]. Available: https://ajaymorya538.medium.com/k-means-clustering-and-its-use-case-in-the-security-domain-f9328cb2ead8.

[23] *Analytics Vidhya, Forum, An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation)*. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/.

[24] *Programmableweb*. [Online]. Available: https://www.programmableweb.com/.

[25] *Where you go for software?* [Online]. Available: https://www.g2.com/.

Abstract

Web services may be seamlessly and dynamically integrated thanks to web service composition. The performance of a composition as a whole is determined by the actions of participating Web services. Therefore, while selecting services for service composition, good quality is crucial. Current methods for choosing and discovering Web services focus on functional factors(availability and response time), or quality of service. Despite the fact that these factors are essential for choosing Web services, they could not accurately reflect user perceptions of quality. We begin this master's thesis by looking at Web Services from a broad perspective. Second, we investigate what Text Mining is and the fundamental components and techniques that make it what it is. Last but not least, we parameter a python algorithms that use Text Mining method to mine user reviews and measure their polarity and the outcome give us a sense of how the technique performs.

Keywords: Web Services, Text Mining, Quality of service, Quality of experience.

ملخص

يمكن دمج خدمات الويب بسلاسة وديناميكية من خلال تكوين خدمة الويب. يتم تحديد أداء التكوين ككل من خلال إجراءات خدمات الويب المشاركة. لذلك، عند اختيار الخدمات لتكوين الخدمة، فإن الجودة أمر بالغ الأهمية. تركز الأساليب الحالية لاختيار واكتشاف خدمات الويب على عوامل الجودة أو غير الوظيفية للخدمة، مثل التوافر ووقت الاستجابة. على الرغم من أن هذه العوامل ضرورية لاختيار خدمات الويب، إلا أنها قد لا تعكس بدقة تصور المستخدمين للجودة. نبدأ أطروحة الماستر هذه من خلال النظر إلى خدمات الويب من منظور عام. بعد ذلك، ننظر إلى ما هو التنقيب عن النص، والمكونات والتقنيات الأساسية التي تجعله على ما هو عليه. أخيرًا، نقدم طريقة لاستخراج النص تستغل ملاحظات المستخدم لقياس قطبيتها والنتيجة تعطينا فكرة عن أداء التقنية المختارة.

الكلمات الدالة : خدمات الويب، التنقيب عن النص، جودة .الخدمة، جودة الخبرة

Résumé

Les services web peuvent être intégrés de manière transparente et dynamique grâce à la composition des services web. Les performances d'une composition dans son ensemble sont déterminées par les actions des services Web participants. Par conséquent, lors de la sélection des services pour la composition de services, la qualité est cruciale. Les méthodes actuelles de choix et de découverte des services Web se concentrent sur des facteurs fonctionnels(disponibilité et le temps de réponse) ou sur la qualité de service. Malgré le fait que ces facteurs sont essentiels pour le choix des services Web, ils peuvent ne pas refléter avec précision la perception de la qualité par les utilisateurs. Nous commençons ce mémoire de maîtrise en examinant les services Web d'un point de vue général. Ensuite, nous étudions ce qu'est lextraction de texte, ainsi que les composants et les techniques fondamentaux qui en font ce qu'elle est. Enfin, nous présentons un code python que nous avons choisis de parametre car il utilise des methodes qui exploite les commentaires des utilisateurs afin de mesurer leur polarité et les résultats nous donne une idée de la performance des technique utilise dans le code.

Mots-clés: Services Web, Extraction de texte, Qualité du service, Qualité de l'expérience.