Logic Based Models for Information Retrieval in Social Media Contexts

Von der Fakultät für Ingenieurwissenschaften
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation

von

Firas Sabbah M.Sc.
aus Nablus, Palästina

1. Gutachter: Prof. Dr.-Ing. Norbert Fuhr
2. Gutachter: Prof. Dr. Ingo Frommholz

Tag der mündlichen Prüfung: 31 Juli 2023

Abstract

The massive volume of information available in social media nowadays has created an urgent need for information retrieval models to assist users in various retrieval tasks in this domain. One of the most prominent fields where users gain important information from social media is the process of evaluating products or services based on customers' reviews. However, the usage of such data in information retrieval tasks raises many major issues. Firstly, social contributions contain information from a wide variety of users, therefore the credibility of the information becomes questionable. Secondly, the absence of information in social contributions does not necessarily mean "nothing", as in many cases, missing information would be implicitly meaningful. Finally, social contributions may contain contradictions that can confuse the users and limit the usefulness of the information. Approaches for modelling information retrieval based on social media contributions have been widely discussed, including probabilistic multi-valued logic-based models. The main strength of these models is their ability to address states like “unknown” and “inconsistent” for information matching besides the states of the traditional binary logic (true and false). Therefore, probabilistic multi-valued logic is well suited to model information retrieval in social contributions-based contexts. However, so far, they have not been utilised to model information retrieval in such environments. In this thesis, we investigate the utilisation of two types of multi-valued probabilistic logics for information retrieval tasks in a social contributions based environment. In the first part of this thesis, we investigated the utilisation of four-valued and subjective logics in the domain of hotel reviews in a system-oriented study. In the second part, we have conducted user studies, to test the effectiveness of a logical model as an algorithm to rank items in a laptop store. Our results have shown powerful abilities of the multi-valued logical models in ranking tasks.

Zusammenfassung

Die riesige Menge an Informationen, die heutzutage in sozialen Medien verfügbar ist, hat einen dringenden Bedarf an Information Retrieval-Modellen geschaffen, die den Nutzern bei verschiedenen Retrieval-Aufgaben in diesem Bereich helfen. Einer der prominentesten Bereiche, in dem Nutzer wichtige Informationen aus sozialen Medien gewinnen, ist der Prozess der Bewertung von Produkten oder Dienstleistungen auf der Grundlage von Kundenrezensionen. Allerdings wirft die Verwendung solcher Daten bei der Informationssuche viele wichtige Fragen auf. Erstens enthalten soziale Beiträge Informationen von einer Vielzahl von Nutzern, so dass die Glaubwürdigkeit der Informationen fraglich wird. Zweitens bedeutet das Fehlen von Informationen in sozialen Beiträgen nicht notwendigerweise "nichts", da in vielen Fällen fehlende Informationen implizit sinnvoll sind. Schließlich können soziale Beiträge Widersprüche enthalten, die die Nutzer verwirren und den Nutzen der Informationen einschränken können. Ansätze zur Modellierung der Informationsbeschaffung auf der Grundlage von Beiträgen in sozialen Medien wurden vielfach diskutiert, darunter auch probabilistische, auf mehrwertiger Logik basierende Modelle. Die Hauptstärke dieser Modelle liegt darin, dass sie neben den Zuständen der traditionellen binären Logik (wahr und falsch) auch Zustände wie "unbekannt" und "inkonsistent" beim Informations abgleich berücksichtigen können. Daher ist probabilistische mehrwertige Logik gut geeignet, um die Informationsbeschaffung in Kontexten zu modellieren, die auf sozialen Beiträgen basieren. Bislang wurde sie jedoch noch nicht zur Modellierung der Informationsbeschaffung in solchen Umgebungen eingesetzt. In dieser Arbeit untersuchen wir die Verwendung von zwei Arten von mehrwertigen probabilistischen Logiken für Information Retrieval Aufgaben in einer auf sozialen Beiträgen basierenden Umgebung. Im ersten Teil dieser Arbeit haben wir in einer systemorientierten Studie die Verwendung von vierwertigen und subjektiven Logiken im Bereich der Hotelbewertungen untersucht. Im zweiten Teil haben wir Nutzerstudien durchgeführt, um die Effektivität eines logischen Modells als Algorithmus für das Ranking von Artikeln in einem Laptop-Shop zu testen. Unsere Ergebnisse haben gezeigt, dass die mehrwertigen logischen Modelle bei Ranking-Aufgaben sehr leistungsfähig sind.

Acknowledgments

I would like first to thank my supervisor, Prof. Norbert Fuhr, for his unwavering support and guidance throughout my studies. Despite giving me the freedom to pursue my own ideas, he has always provided valuable feedback and suggestions, helping me to refine and enhance the quality of my work. His dedication to working on our publications has always inspired me to work harder and deliver high quality work. I am also grateful to the members of the examination committee for taking the time to read and evaluate my thesis. Furthermore, I would like to acknowledge the support and encouragement of my colleagues and friends at Duisburg-Essen University. Also, I would like to thank Dagmar Kern and Andrea Papenmeier from GESIS Institute for their help in designing the user studies of my research. Lastly, I would like to express my heartfelt appreciation to my family, whose unbounded support has been a constant source of strength during the challenging years of my studies.

This work was supported by the German Research Foundation (DFG) under Grant No. GRK 2167, Research Training Group "User Centered Social Media".

Contents

d

Chapter 1

Introduction

Social media has become a fundamental part of our daily life and has also become an important source of information. We interact with social networks in various ways, for example by sharing posts with others, commenting on news or reviewing products and services. We generate therefore large amounts of data on a daily basis which can be useful in a wide variety of applications including business, entertainment, education, healthcare and many others.

Retrieving information from social media is of great importance in many practical areas in the IT field such as search engines and recommender systems. In recent years, the field of information retrieval has made significant progress in the area of content understanding with the goal of creating more accurate and effective models and systems for information retrieval from traditional web sources like web pages. However, building information retrieval models based on social content is challenging due to its unique characteristics such as the unstructured nature of social data and the large amount of the data which makes it much harder to identify and extract relevant information. Therefore, research on efficient approaches to modelling information in social media is a challenging and important research problem.

Social information retrieval (SIR) is a sub-field of information retrieval that focuses primarily on studying information retrieval in the context of social content. The goal of SIR is to provide effective search capabilities for locating relevant information from online social networks that can be used to satisfy the user's needs of information. Within this area of research, there has been a significant amount of works focused on the modelling of social content and tackling many of the challenges associated with it. However, the majority of the works have treated this task as purely syntactical in nature i.e. they focus on extracting features from the social contents and use them as inputs for other tasks or applications. While this is a good starting point to achieve

the task of SIR, these approaches ignored mostly some of the most important issues of the social data that cannot be identified or extracted as syntactic features, such as the credibility of the data, the inconsistency of the data among users as well as the fact that some information may not be present explicitly in the texts but only can be indirectly inferred.

The credibility issue in social content originates primarily from the fact that it is written often by anonymous users who may be biased, incorrect or misinformed, and therefore the information they provide may not be reliable. In addition, due to the highly dynamic nature of the social content, it often contains outdated information and so there is a need to detect and remove such information when processing the social content in order to provide users with accurate information. The credibility issue has been studied extensively in the mainstream NLP literature (Wawer et al. [2014]; Aladhadh et al. [2018]; Zhou et al. [2017]). However, while most of that work is focusing on the assessment process of the credibility, only a few concentrated on the process of involving it in IR tasks (Weerkamp and de Rijke [2012]; Zhang et al. [2012]; Poongodi et al. [2019]).

One more critical issue about the social content is the contradictions among the sources. As social content is generated by users based on their personal perceptions and opinions, it is possible that the information from different sources is conflicting with each other. In such cases, it is difficult to obtain an accurate representation of the actual information available about a topic in social content. It is therefore very important to identify the conflicting views in the data and consider these while processing the content. Similar to the credibility issue, most of the work (Dori-Hacohen and Allan [2015]; Badache et al. [2018]; Garimella et al. [2018]) in the area has focused on the detection of the contradicting information rather than the integration and handling of the conflicting information in the IR tasks (Könsgen et al. [2018]; Ali et al. [2021]).

Another important issue in social data is the presence of missing or incomplete information. In many cases, the users only provide partial descriptions of a specific topic. Therefore, it is very difficult for IR systems to obtain complete information when a user asks a question related to a particular topic. This problem arises due to the fact that there are many sources, especially user reviews, that contain only a few sentences describing opinions on a topic while other topics may not have reviews at all. It also happens in some cases that the relevant information about the topic is not extracted by the IR system due to the absence of proper keywords in the query submitted by the user. Thus, it becomes essential for the IR system to handle the missing in-

formation in order to provide more accurate search results to the users. This issue was discussed in previous works. One common approach to deal with it is to fill in the missing information using different techniques or sources (Asadi and Regan [2019]; Pujianto et al. [2019]). However, the completion techniques are inappropriate in the field of social content. Such data often contain subjective assessments and emotions rather than facts that can be automatically inferred from existing data.

1.1 Contribution

Although some research has been done in the area of addressing these issues, these works have been more focused on addressing the issues individually rather than considering them together. Our novel contribution in this thesis addresses these three important topics related to social data in an integrated manner, and provides a unified framework for offering search capabilities to end users when carrying out IR tasks using social data. The proposed framework consists of two main components: (1) a system for indexing a special type of web documents containing user opinions toward different topics, and (2) a system for integrating information from these resources and evaluating the relevance score of the search results.

(1) Indexing

In traditional IR systems, the indexing process relies on keywords found in documents. However, keyword-based approaches suffer from several limitations, particularly the lack of context consideration. This often leads to irrelevant results or missed relevant documents because the system fails to comprehend the intent or meaning behind the query or document, particularly in social contexts where informal language, shorthand, and sarcasm are prevalent. Consequently, keyword-based approaches encounter various retrieval problems, including the vocabulary mismatch problem, ambiguity, difficulties in handling misspellings and synonyms.

For the issues of social contexts we are discussing, keyword-based indexing is also considered inappropriate in addressing these challenges. Let us start with the issue of credibility. Although some studies have attempted to evaluate credibility based on keyword analysis, such as the works of Wawer et al. [2014] and Aladhadh et al. [2018], relying solely on keywords is not enough

to accurately determine the credibility of the content. The problem of credibility is inherently related to the semantic understanding of the user's intent, which is a complex task that cannot be fully captured by keyword analysis alone. The credibility of information is influenced by several factors beyond the words present in a post or comment. These include factors such as the user's profile, activities, and other contextual information.

Social contexts are vulnerable to contradictions because people tend to express their opinions and experiences differently on social media platforms, leading to contradictory statements or ideas. For example, one person's contribution may express a belief that the heavy earthquake caused the collapse of houses, while another person's contribution may contradict this belief by attributing the collapse to the poor quality of construction. Keyword-based approaches, such as the one proposed by Badache et al. [2018], are insufficient in addressing contradictions because they rely solely on the explicit presence of terms and antonyms without considering the broader context or meaning behind the opinion.

When addressing the issue of missing information in Information Retrieval (IR), it is necessary to handle it under the concept of the Open-World Assumption, which treats any missing information as "unknown" due to the absence of evidence to determine its truth or falsehood. However, most traditional keyword-based approaches lack a systematic approach to dealing with such information. Consequently, in these cases, IR systems typically respond to user queries with a simple message of "No matching results" because no explicit keyword matches were found in the searched documents.

To overcome the limitations of keyword-based approaches, we propose a new method for representing and indexing social content using logical statements rather than simple keywords. In our approach, each logical statement represents a user's opinion toward aspects or topics of some domain. We restrict the logical statements to a specific domain in order to be able to recognise arbitrary statements that may contradict each other. A logical statement comprises several components, including the level of credibility of the opinion, the polarity of the opinion toward the aspect, and a weight indicating the sentiment toward the aspect. To represent cases of missing information, we consider a special case that points to an unknown or uncertain stance toward the queried aspect by replacing the polarity and sentiment weight of the opinion.

To illustrate how this could be done, consider the following simple example where we need to provide information regarding the topic “cleanliness” of a

hotel based on users' opinions. In this case, the level of credibility for each user's opinion would be expressed using a numeric scale. The polarity of the opinions is expressed by using a binary value indicating whether the opinion is positive or negative. In order to distinguish between opinions like "The hotel was not clean" and "The hotel could have been cleaner", weight is also assigned to the opinion to represent the degree of sentiment that the user expresses toward this topic. Finally, in the case where some user opinions are not relevant to the required information ("cleanliness"), e.g. "Great location", we indicate such opinions as being unknown to the topic of interest.

(2) Retrieval

In traditional IR approaches, the matching process between queries and indexed documents is handled under the general formalisation of the boolean logic. That is, a logical value 'true' is assigned if a keyword appears within a given document while the logical value 'false' is assigned if the term does not appear in the document. However, this formalisation is considered insufficient to handle our proposed representation of social data. As described above, an opinion extracted from social data has three cases of relevance to the query: positive, negative, and unknown relevance. In addition, these opinions can have varying weights to express the strength of the sentiment being expressed. To deal appropriately with such situations, we propose the use of probabilistic multi-valued logic as a formalism to represent the matching process.

As a result of using the multi-valued logic representation instead of the boolean logic, it is now problematic to perform the ranking operation straightforwardly. In the ranking operation, the goal is to compute the value of the maximum relevance score of a document with respect to the query. Under the use of the multi-valued logic, it is not sufficient to consider the documents with only the positive relevance score (as is the case with boolean logic), while ignoring the other logical values (i.e negative, unknown). As a result, there is a need for a method to estimate the overall relevance of each document to the query based on more than a single value. We tackle this problem by computing a combined relevance score for the document based on weighted importance measures of the values of the three relevance cases (positive, negative, and unknown).

To evaluate our methodology, we concentrated on one case of social content which is online reviews. We applied the framework on two example appli-

cations. For the first application, we followed a system-oriented approach to evaluate the applicability of the approach in the domain of hotel reviews. For the second application, we focused on a user-oriented approach for evaluating the method using reviews from the laptop domain as inputs. In both cases, we used multi-valued logical models to extract relevancy information between documents (i.e. represented by reviews) and queries (i.e. domain aspects like e.g cleanliness from the hotel domain, battery from the laptop domain).

Experiments showed that our system demonstrated an improved retrieval performance compared to the baseline systems that only considered relevance based on a single binary relevance case. Moreover, the experimental results particularly from the first application showed that users are more influenced by negative reviews than positive reviews when evaluating hotel aspects. It also exposed that users have a positive bias toward the aspects that have not been reviewed. In the second application, results demonstrated that our approach is able to support the users in selecting products based on the experience or opinions of previous reviewers.

While it might be possible to achieve better one-dimensional rankings via tuning deep learning methods (in case there is enough training data available), the advantage of the logic-based approaches is that they handle contradictions, omission and credibility in a transparent way, which also can be made visible for the end-user. For researchers, this transparency allows for a better understanding of the problems, identifying the major influencing factors and spotting possible improvements. As IR research is paying more attention to the transparency of the methods employed [Culpepper et al. 2018], our work is along these lines.

To the end of this thesis, our contribution consists of the development of approaches and representation methods in different phases of the IR process. We present our work based on Goker and Davies [2009] representation of an IR process - as shown in figure 1.1 - by focusing on five of its components. We focused on a specific type of documents in the social content domain for the input documents; these are the items containing contributions such as products with reviews. For the input queries, we formulated a type of aspect-oriented search queries that suited the task we are tackling. In the indexing process, we indexed the items as logical statements; tuples of aspects, polarities, and weights defined based on the domain of interest. In the matching phase, we matched the queries with the indexed items, aggregated the matching results using multi-valued logical models, and ranked the results accordingly. Finally, in the feedback process, we evaluated the quality

of the rankings using both system- and user-oriented approaches.

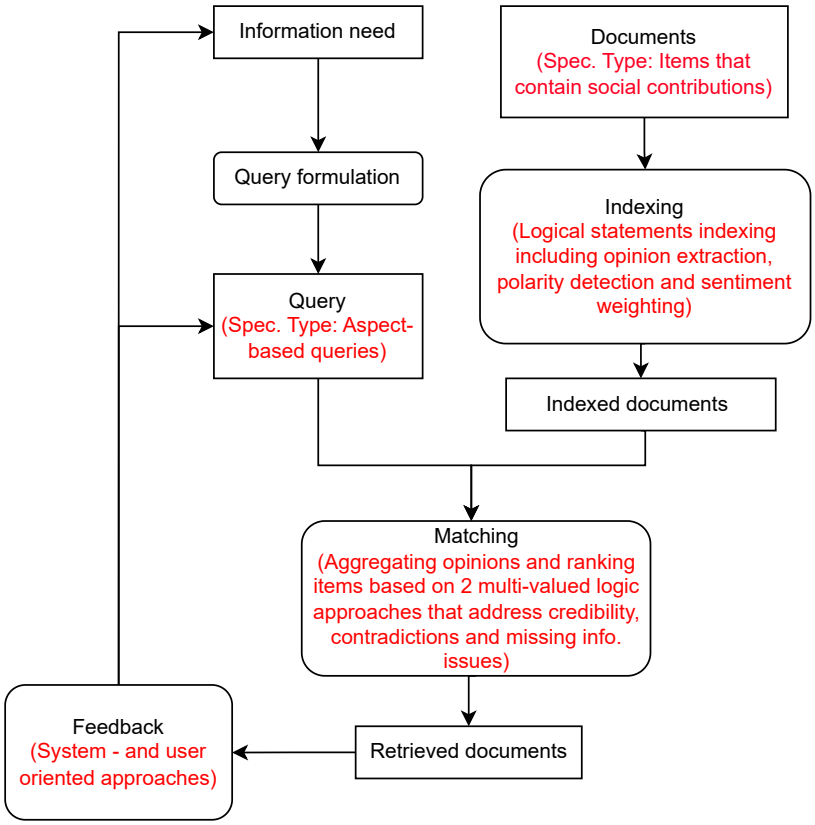


Figure 1.1: Goker's information retrieval process. Our contribution is shown in red.

1.2 Structure of the thesis

The remainder of this thesis is organised in two parts: the first part is focused on the theoretical background and related work. The second part presents two experimental studies carried out to test our proposed framework in both systems-oriented and user-oriented approaches.

Chapter-oriented coverage is as follows: Chapter 2 provides a literature review of IR in social media. It also provides an overview of related research that discusses the issues of social media retrieval we are investigating in our thesis work. Moreover, it discusses the popular IR approaches used in the field of online reviews and highlights their limitations. Finally, it shows some relevant work on multi-valued logic. Chapter 3 describes the multi-valued logical framework and its application to model IR in social content. It includes a demonstration of the theoretical foundations of the proposed models. Chapter 4 demonstrates the first application of the multi-valued models. It represents the system-oriented approach for evaluating the proposed models. Chapter 5 shows the user-oriented evaluation approach. It includes details about two user studies performed for the second application. Finally, chapter 6, presents the conclusion and future work related to this thesis work.

Part I

Technical Foundations

Chapter 2

IR in Social Media Contexts

In this chapter, we present an overview on how information retrieval in social media contexts was handled. Starting from section 2.1 which defines and presents basic knowledge about information retrieval in general and how the change in web structure represented by the existence of social media contexts has changed the way that IR models use to deal with such content. In section 2.2, we define some of the major issues that face retrieval in social media and how the literature addressed them. In sections 2.3 and 2.4, we focus on the reviews and the review-based retrieval as the case study of retrieval in social media, which this thesis is mainly considering. Finally, section 2.5 gives an overview of the logical models and their use for modelling the IR.

2.1 IR in Social Contexts

Information retrieval (IR) is the process of maintaining and providing access to relevant information based on the users' needs [Baeza-Yates et al. 1999]. An IR system's primary function is to provide users with timely access to the needed information based on their queries or requests. It is also necessary that the users are able to navigate through the information easily and efficiently, as information is retrieved from a huge collection of data and resources. In order to achieve this, the IR system should store documents and queries as textual objects, often represented as a collection of weighted terms, and match them to provide a ranked list of documents identified to be relevant to the user's query.

Modelling In IR is the task of representing a given document collection in terms that allow the efficient retrieval of documents of interest to a user from that collection. The boolean model, the vector space model [Salton 1971],

and the probabilistic model [Robertson and Jones 1976] are only the most prominent categories of these. Research in IR has developed over the past years a lot of IR model extensions based on each one of these categories. As of the most prominent models of the category vector space models, we have developed our work in this thesis based on models term frequency (TF), term frequency-inverse document frequency (TFIDF) and Okapi BM25.

Retrieval evaluation is a process of associating a quantitative metric to the results produced by an IR system in response to a set of user queries. Many different metrics for evaluating the retrieval quality of IR systems and algorithms have been proposed, i.e. the quality of the results. Such metrics are Precision and Recall, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Discounted Cumulative Gain (DCG) and many others.

With the emergence of the social web, a large range of applications and services have emerged to provide users with an easier and more effective way to interact with social networks and other online repositories of information. This information can be very useful in information retrieval tasks for both user and resource modelling. Therefore, the need for information retrieval support in the social context has resulted in new types of IR models. Such models aim at providing users with relevant information based on their profile and their behaviour within the social space or based on the information that other participants contributed in the same social space. Examples of such models include collaborative filtering and recommendation systems, Q&A systems and others. Given the potential of such models for building customised and personal views of information for users, there have been growing efforts to incorporate social network information in IR systems in order to obtain improved search quality in IR tasks.

Many works in the literature discuss IR in social contexts. Works in this field can be categorised into two categories: 1) IR tasks supported with the assistance of social data. And 2) recommender systems that use social data for improving the recommendations.

For the first category, social information was integrated in all IR task phases which include indexing, querying, ranking and retrieving. In indexing, researchers generally focused on developing new document representations by involving additional metadata extracted from social contents [Carmel et al. 2010; Lee et al. 2012]. For querying, works such as [Zhou et al. 2012; Do et al. 2016; Massoudi et al. 2011] used social data in different techniques for forming or expanding queries in order to improve the search results. Finally, for the ranking and retrieving, there is a very wide range of works that

discuss ranking based on social content. However, we mention here some of the most prominent works that focus on the ranking procedure from the point of view of social popularity and importance. An example of this is the SocialSimRank and SocialPageRank [Bao et al. 2007] which extended the well-known PageRank based on social annotations for calculating similarity and popularity. In the same direction, [Takahashi and Kitagawa 2009] proposes an extension of the well-known HITS [Kleinberg 1999] approach.

Research on social data has also gained a lot of attention in the field of recommender systems. It is very often that recommender systems in this area choose collaborative filtering based methods as their basic models. The common idea behind the usage of these algorithms in this area is that users who are socially connected, are very likely to share common interests. The domain of recommender systems in this field is very wide. One example on the common issues addressed with the help of social data is the cold-start problem. Works like Sedhain et al. [2014; 2017]; Li et al. [2019] have proposed different approaches to tackle this issue.

2.2 Common Issues in Social Media Retrieval

Clearly, social data is a valuable source for enhancing and developing effective and efficient IR models. However, such usage of data in IR models is raising questions about the quality of such models as this kind of data often suffers from several limitations, which consequently affect the models they are dependent on. In this thesis, we take a closer look at some of the most common issues that can arise when employing social data for IR purposes and discuss some relevant works dealing with these issues. Specifically, we focus on the following three issues: (1) the credibility/trustworthiness of user-generated content, (2) the contradictions that can sometimes exist between users' contributions, and (3) the incomplete or missing information that makes analysing the data collected in social media an inherently challenging task. To address these issues, we first give a general overview of the problems that social data face and provide a brief overview of the existing approaches for dealing with them.

2.2.1 Credibility

One of the primary concerns of IR in social content is credibility. The challenges arise due to the fact that the majority of social content contributors are anonymous or unaccountable for the veracity of their comments. Social platforms have very small control over the quality of social content. On such platforms, users are not bound by editorial guidelines. As a result, the quality of the generated content is quite variable, which may mislead users who rely on such content to make decisions, such as purchasing a product based on user reviews.

Indeed, credibility is seen from different perspectives. At the source level, credibility is determined according to whether users are trustworthy and honest in their contributions. At the content level, credibility is determined by the quality and the truthfulness of the input text. At the system level, credibility refers to the trustworthiness of a system which processes and manages the social content. Credibility can be also viewed in a broader sense to include other aspects such as reliability, validity, objectivity, transparency, freshness, accuracy, and completeness. All these different aspects contribute to the overall quality of a social platform and the overall quality of the data that can be obtained from it. Credibility is therefore a very broad and complex concept that is dependent on the context in which it is being applied.

There exists a wide variety of works that discuss the subject from various points of view by scanning the literature on credibility and other relevant concepts, such as trust and reliability in social content. In our literature survey, we concentrate on two key aspects of social content credibility. First, we look at the studies that utilised various methods for assessing the credibility of social content. Second, we observe the inclusion of the social contributions' credibility in enhancing the quality of the retrieved results in the IR domain.

Credibility Assessment

Credibility assessment measures the degree of trust that users have in a piece of information. There are different approaches that can be used to measure the credibility of a piece of information on a social platform. Some of the approaches include text-based approaches like linguistic feature analysis and sentiment analysis. Another popular approach is user-based methods where contents are rated according to the user's level of trust in the sources of the

information they post. There are some studies that have combined both text and user-based approaches in their evaluations of social content credibility.

In the text-based approaches, the extracted features are used to identify the characteristics of the content. These features can then be compared with the characteristics of known reliable and unreliable sources of information to identify the features that distinguish reliable sources from unreliable ones. Linguistic features are one of the common approaches in this field, they can be used to assess the trustworthiness of a text including grammatical and lexical errors, the number of references used in the text and many other features. An example of such an approach is described in [Wawer et al. 2014; Aladhadh et al. 2018]. For sentiment analysis approaches, the texts are analysed to detect the presence of positive or negative emotions in the texts. These are then compared with the extracted emotions of reliable sources to assess the credibility. An example of such approaches is discussed more in [Alonso et al. 2021]

For user-based credibility, it was common that works focused on a specific platform like Twitter to assess the credibility of the users. The reason for this is that such a platform provides a content-rich informative network of real-world users. The approaches for user credibility include using of explicit information from users such as their biography information (e.g. name, age, etc.) and tweeting patterns; or implicit information such as their connections within the network. Examples on this category of credibility are [Poongodi et al. 2019; Zhou et al. 2017]

While in this section we provided a quick overview of the various methods for assessing the credibility of social content, we emphasise that this thesis's goal does not include assessing credibility. The existing literature in this field provides a wide range of methods and frameworks that can be easily integrated into our research. The purpose of this research is to introduce an IR model that incorporates credibility as well as other aspects of social content into the ranking of search results. The next section focuses specifically on works that consider credibility in IR tasks.

Credibility in IR

Credibility involvement in IR tasks was not discussed much in the literature. However, this area seems to be an emerging research field that is gaining more and more attention in recent years. The reason is that an increasing number of search platforms are introducing new features with a focus on

integrating and assessing the credibility criteria while ranking the results.

As an example of these works, Weerkamp and de Rijke [2012] investigated a ranking method of blog posts based on credibility indicators. These indicators are derived from the blog's post such as spelling, timeliness and document length, and some others are derived from the blog itself like regularity, expertise and comments. The hypothesis that was confirmed in this work is that such indicators will improve the precision of the search results.

Zhang et al. [2012] is another work that implemented a model integrating the credibility of reviewers into the task of ranking products. The authors of this research followed a strategy to increase or decrease the credibility of a reviewer based on the ratings and reviews given by them to other products.

Ravikumar et al. [2012] proposed a ranking method for Twitter messages that considers both trust and relevance. Authors have used a three-layer graph model for deriving trust of system components consisting of users, tweets and web pages. Using the implicit linkages between tweets, their proposed model achieved strong computational performance and great accuracy in judging the credibility of tweets.

As we can see, there are various ways and factors that can define and influence the credibility of social content. Network-based, user-based, content-based, or a combination of multiple factors are just some examples of these definitions. Regardless of the type of credibility that is considered for the IR task, our main focus in this area is to integrate this information into the rankings to enhance the quality of the search results. In contrast to the previous works that emphasised the credibility dimension as a separate parameter, our research will address this issue in conjunction with other factors of social content.

2.2.2 Contradictions

Contradictions in social content is also one issue that needs to be addressed when using such data in IR tasks. When the contributors to a particular article have different opinions regarding the issue discussed, it means that their opinions are not consistent and thus it can create a conflict regarding their content.

Such a contradiction in social content creates a challenge for automatic IR systems as it prevents these systems from extracting useful and informative

content out of them or at least generates confusing and incorrect results. Contradiction is an essential concept in IR tasks and can serve various purposes, such as improving knowledge discovery, increasing the utility of information, etc. However, the use of such information in IR tasks comes with its own challenges, which can be addressed using different techniques depending on the nature of the task.

This theme was discussed a lot in the literature from different points of view. Many works such as [Dori-Hacohen and Allan 2015; Badache et al. 2018; Garimella et al. 2018] focused on the techniques for detecting contradictions or finding contradictory topics in social content. On the other hand, other works have utilised the contradictions in various ways. For example, Ali et al. [2021] made use of contradictions to enhance the prediction of product quality. Also, Könsgen et al. [2018] analysed the effects of companies' reviews' contradictions on the intention of job seekers, and performed different experiments to understand the nature of such contradictions.

Although classical IR has extensively studied the issue of contradictions, the majority of works on this topic rely on term-based approaches for detecting contradictions. However, these approaches have limitations in identifying contradictions, as they only focus on the terms used in the text and fail to take into account the meaning and sentiment expressed in contradictory statements. For instance, consider the two statements "The reception staff was very friendly" and "The man at the reception could have been more welcoming". Although these statements do not contain any contradictory terms, the underlying logic in these statements is contradictory. To overcome this limitation, logical models offer a broader framework that surpasses the mere reliance on terms. In these models, text can be categorised into logical statements that can be evaluated based on their logical value in relation to a specific topic. Furthermore, logical statements take into account the logical relationships between explicit terms and implicit indications within the context in which they are used. Consequently, the use of logical statements in IR systems provides a more comprehensive approach to analysing text and identifying contradictions, particularly in social contexts.

In our work, we have utilised two logic-based approaches to model the IR task. Contradictions in these models originate due to logical statements conflicting with each other. As there is no way to identify arbitrary statements in general, we limit our focus to one specific case which is product reviews. So, we extract and identify logical statements and their orientation toward predefined topics or aspects of the domain. In both approaches we used, contradictions between contributions are an inevitable outcome of the fusion of

logical statements representing the differing opinions in the reviews. During the results ranking process, contradictions play either a direct or indirect role along with other logical values, depending on the ranking scenario being followed.

2.2.3 Missing Information

Missing information is also one of the challenges in IR systems. The concept is more frequent to appear in review-based systems. The review of a product is often incomplete, which can lead to incorrect conclusions about the product. That happens usually as most reviewers are reviewing a limited number of the product's aspects or features (e.g. price, performance), without taking into account the other aspects and features of the product that may influence the decision-making of the consumers.

Such an issue was widely discussed in the literature of information systems research, and several solution approaches have been proposed to tackle the issue. The common solution to this issue is to fill in the missing data with some valid data such as the most frequent feature values, uniformly randomised values, or values generated randomly according to the observed distribution. Furthermore, there exist some works using auto-encoders [Asadi and Regan 2019] or KNN-based [Pujianto et al. 2019] methods to conduct missing data estimation. However, most of these methods require a large amount of training data in order to generate a reasonable approximation of the missing data and hence may not be feasible for real-world applications. Moreover, filling in missing data may be inappropriate for some fields such as social content since such data may contain opinions and subjective assessments rather than objective facts which can be automatically inferred.

On the other hand, the problem of missing information can be addressed through the concept of the Open World Assumption (OWA). OWA is a philosophical and computational principle that assumes that the information held by a system is incomplete and that there may be additional, unknown information. In contrast, the Closed World Assumption (CWA) assumes that the information held by a system is complete and that there is no additional, unknown information. An example of OWA is a search engine, where the search results represent a limited sample of the available information, and not a comprehensive or exhaustive list. For instance, a Google search doesn't guarantee the inclusion of every relevant website, but instead presents a sample of the most relevant websites based on the search query. The search en-

gine operates under OWA because it assumes that there is more information available beyond what is being shown in the search results. An example of CWA is a database management system, where all data is stored in the database and any information not stored in the database is considered false. For instance, in a database of university students, if a student is not recorded, the database management system assumes that the student is not enrolled in the university. The database operates under CWA because it assumes that all relevant information is stored in the database, and any information not stored in the database can be considered false.

However, the OWA is not widely considered in IR, as most approaches do not deal with negative information, rendering the distinction between false and unknown irrelevant. In our research, we treat the texts of the search documents as logical statements that can take not only a true or false stance towards the search topic, but also an uncertain one. This uncertainty is represented by the "unknown" value in the multi-valued logical models we use. This way, our modelling process takes into account the uncertainty associated with missing values and attempts to interpret them in various ways. A more in-depth discussion on how we treat missing values will be presented in Chapter 3.

2.2.4 Search as An Aspect-Oriented Task

While traditional IR approaches are effective for dealing with text within traditional web documents, they may be inadequate when it comes to handling social content, especially under the consideration of the mentioned issues. One of the main problems is that the relationship between terms and entities in social content is complex and dynamic, with terms evolving over time and users expressing complex relationships. As traditional IR approaches are mostly keyword based, they may not provide a good representation of social content.

Based on the nature of social platforms, users often mention multiple aspects or features of the entities they are commenting on. In order to make use of these mentions in IR research, it is necessary to carefully extract the relevant aspects of these comments and transform them into semantic entities that can be analysed further. Consequently, an efficient IR task in this space should be treated as an aspect-oriented instead of a keyword-based search.

Calculating quantitative ranking scores is a common approach for utilising

data in social contributions. This includes overall average ratings and black-box-based recommendations, which are frequently found in review-based platforms. Nevertheless, there is a wide range of works in the literature that focus on the orientation of the IR task toward particular aspects of the contributed objects. Works in this field are mostly recognised in the summarization, faceted search, and opinion retrieval categories. In this section, we will briefly go over related works that present various aspects of aspect-oriented IR tasks in each of these categories.

A number of research studies have explored extracting summaries and using them as parts of user-oriented descriptions. Comments summarising research in this field has focused on topic-specific summaries. Major opinion summarization methods work on the classification of the overall sentiments of contributions into positive or negative orientations [Bahrainian and Dengel 2013; Raut and Londhe 2014; Abdi et al. 2018]. In contrast to sentiment classification methods, textual or quantitative-based evaluations have integrated summarisation techniques to help users in obtaining topic-related descriptions of social content [Blair-Goldensohn et al. 2008; Jmal and Faiz 2013].

Faceted search [Tunkelang 2009] allows users to explore and find information that they need by filtering or navigating with the help of some predetermined facets. These facets might be related to the target object, to the description of it or to other criteria like time of creation or date of the last update. In the area of social content, it was noticed that such resources are important for supporting the aspect-oriented search tasks [Adriaans et al. 2011].

Another area of IR research that makes advantage of social contributions is opinion retrieval. Applications for retrieving opinions are designed to analyse and obtain relevant documents based on their significance in relation to a certain opinion. The core of opinion retrieval systems is usually based on traditional retrieval techniques. However, these approaches can be enhanced by analysing the social resources [Eirinaki et al. 2012; Gozuacik et al. 2021].

While our work is similar in that it focuses on aspect-oriented IR tasks, there are some key differences that make our task unique. Specifically, we focus on the process of aggregating user opinions toward certain aspects of objects of interest and representing this information in a form of multiple values. Unlike most of the previous works, we are providing not only the positive and the negative representation of the opinions but also include information about missing data and contradicting opinions.

2.3 Reviews as A Case Study of Social Content

As we have seen, the area of social content is wide. It can be divided into different types, such as social networking sites, blogs, microblogs, product or service reviews and many others. This variety has driven the IR research community to develop different models and approaches for the analysis of social content, each according to its characteristics, type of users and type of content. In this thesis, we are considering user reviews as the main source of social data for our studies.

2.3.1 User Reviews: A Source for Product Ranking

Reviews are one of the most popular ways for consumers to exchange their experiences on products [Curien et al. 2006; Dellarocas 2006]. A typical product review consists of ratings and comments by users of the product being reviewed. These comments are a rich source of information that can be used to generate insights that can help both companies to improve products or services, or use them for marketing purposes, and consumers to make better purchase decisions.

Expert reviews tend to offer highly structured and detailed opinions about the product, including pros and cons, as well as recommendations. While this type of analysis can be very useful for decision-making purposes, it is not always possible to obtain this kind of information from experts. Moreover, an expert's opinion may be biased and may not reflect the opinion of the user population as a whole [Vollaard and van Ours 2022]. Therefore, while relying primarily on expert reviews for product analysis may not be a bad idea in some cases, it is generally not the best approach and has its limitations.

On the other hand, user-generated reviews tend to be less structured and more informal. However, they also tend to be more honest because users feel more personally involved in their evaluation and are more likely to point out potential problems or drawbacks that may not be mentioned in expert reviews. Moreover, they can be collected from a larger number of users over a longer period of time, leading to more overall data points that can be used to draw more accurate conclusions about the products. At the same time, user reviews represent a sub-sample of social content and suffer from the same problems that other types of user-generated content face. They represent mostly subjective opinions of strangers [Burgess et al. 2011]. Moreover, some reviews might be erroneous, or intentionally misleading, e.g. for commercial

reasons. They are also biased in many cases to the negative experience of the users [Aithal and Tan 2021]. They can be inconsistent and contradict each other. Also, user reviews have mostly incomplete evaluations of product aspects. All these limitations raise serious concerns about the credibility, contradictions and missing information issues in user data.

Analysing user reviews is an important process for both users and companies that are interested in improving their products or services. Positive and well-reasoned reviews have a positive influence on the likelihood of customer purchase decision [Park et al. 2007]. However, negative reviews are not always of negative influence. For instance, Sorensen and Rasmussen [2004] found that not only positive reviews have a positive impact on sales, but also negative ones. Sun [2012] showed that negative reviews are more powerful in reducing sales than positive ones in increasing them.

Building frameworks for IR tasks based on user reviews can offer users an efficient way to evaluate products. The difficulty lies in the development of effective frameworks that can handle the issues that were discussed above. In the next session, we explore some of the techniques and problems that arise when building such systems.

2.4 Review-Based IR

Because online reviews are textual data, methods for evaluating products based on online reviews often require three primary processes: the first is product feature extraction, followed by sentiment analysis, and finally product ranking. However, for the purposes of this thesis, we will be concentrating mainly on ranking strategies or other relevant methods that also make use of reviews in order to propose products or assist consumers in making selections. Methods based on fuzzy set theory extensions, directed weighted graphs, product recommendation, customer satisfaction, and other unclassified studies are included in this section. Research in these fields is discussed in the subsections that follow.

2.4.1 Fuzzy Set Theory Extensions

Intuitionistic fuzzy set theory [Atanassov 1986] is an extension of fuzzy set theory. It is one of the often utilised approaches in the literature for aggre-

gating information from reviews. Unlike classical set theory, which specifies element membership on a binary level, the intuitionistic fuzzy set theory provides degrees of membership for the elements. In this field of study, degrees of membership have been used generally in terms of sentiment levels of product attributes.

For instance, Liu et al. [2017*b*] suggested an approach that utilised both sentiment analysis and intuitionistic fuzzy set theory. The sentiment analysis of reviews is used to calculate an intuitionistic fuzzy number for each product feature. They have used a special kind of weighted averaging operators [Xu 2007] to combine the intuitionistic fuzzy number and the sentiment weight of each product with respect to different features. In another work, Bi, Liu and Fan [2019] used interval type-2 fuzzy numbers [Mendel et al. 2014] to express review sentiment analysis results. Fuzzy numbers are then aggregated to determine the final product rankings.

Liu et al. [2017*a*] proposed a method for ranking products based on the similarity to the ideal solution. A method based on intuitionistic fuzzy theory and TOPSIS (technique for order preference by similarity to an ideal solution) [Yoon 1980] is employed to determine a ranking of the products. Finally, an expanded TODIM approach (interactive and multi-criteria decision-making) [Gomes and Lima 1991] based on online reviews was suggested by Zhang, Li and Wu [2020]. The approach makes use of an intuitionistic fuzzy set to figure out how customers feel about a product's sentiment and how strong that feeling is. Additionally, a case study was conducted to confirm the efficacy of the suggested approach. In summary, methods based on intuitionistic fuzzy theory for aggregating information offer an advantage in handling uncertain information towards the sentiment of what reviews mentioned. Aside from that, similar techniques were used to weight product features, therefore enhancing the quality of the rankings.

Data aggregation strategies based on hesitant fuzzy theory, a kind of fuzzy set extensions, have been studied. The use of hesitant fuzzy theory in aggregate calculation has various benefits. The membership degree of several items is permitted in this kind of fuzzy set [Wen et al. 2019]. However, the research on the topic of product rankings is uncommon. One of them is Zhang, Wu and Liu [2020], who presented a method to rank products using reviews based on the hesitant fuzzy set and sentiment analysis. Product feature sentiments were utilised to determine the overall performance of each feature using a hesitant fuzzy set. Based on how much attention each feature received, rankings were created. A comparison study was conducted to verify the suggested method's performance.

2.4.2 Weighted Directed Graphs

On the basis of weighted directed graphs, certain research on information aggregation is also noticeable in the literature. Kong et al. [2011] classified the features of products in reviews in categories. Features according to [Kong et al. 2011] are mentioned in either positive or negative sentiments in a direct or indirect (comparative) statement. Based on this, they established a weighted graph to aggregate reviews information and to finally obtain the result of ranking products. Zhang et al. [2010] and Guo et al. [2018] used similar techniques for product ranking. The method is based on constructing a weighted directed graph to aggregate the information of reviews. The graph is built on top of product features that are classified into subjective (or personalised) and comparative categories. A "page-rank" like algorithm based on the created graphs is used to obtain the final rankings. Li et al. [2011] constructed a unified graph model to integrate the comparison among products. The comparison is built on top of relations mined from user reviews and community-based question-answer pairs containing product information. Finally, Yang et al. [2016] used a graph structure to aggregate numeric ratings and text sentiments of reviews with other comparative content like statements and votes of reviews, to obtain the overall ranking score of the products.

Using weighted directed graphs to aggregate data has several benefits [Yang et al. 2016]. One of these benefits is the capacity to integrate heterogeneous information. In actuality, online evaluations include not only descriptive information such as written descriptions and digital ratings, but also comparison information. As a result, certain data aggregation algorithms based on the weighted directed graphs may clearly represent the comparison network and comparative advantages across alternative products, potentially improving product ranking quality [Kong et al. 2011].

2.4.3 Product Recommendation

Since the first attempts of creating recommendations based on user reviews [Aciar et al. 2006*a*;*b*], this topic is now widely studied in the literature. As an example, Siering et al. [2018] investigated in a study the relations between the consumer recommendations and the content of the reviews. Based on the key factors that determine this relation, they proposed a method for automatically predicting recommendations in the domain of airlines. Serrano-Guerrero et al. [2020] created personalised recommendations based

on weighted aspects and a T1OWA-based mechanism [Zhou et al. 2008] for characterising the user of being more influenced by negative or positive opinions. Alexandridis et al. [2019] proposed a recommendation method by incorporating reviews in combination with the rating scores into a collaborative filtering matrix factorisation algorithm. Guerreiro and Rita [2020] investigated the determinants of customers' explicit recommendations. The findings indicated that customers are more likely to give explicit recommendations when the product or service represents positive sentiments of customers. Stavrianou and Brun [2015] proposed a recommendation method based on comparative relations between products. Such relations are extracted with the help of user reviews and expert recommendations extracted from them.

2.4.4 Customer Satisfaction

Analysing customer satisfaction regarding a product or service is one of the fields that makes use of user reviews for determining the quality of the products and services. This topic was investigated in the literature as an alternative field of product recommendation and ranking.

Zhao et al. [2019] evaluated overall customer satisfaction by analysing the technical attributes of reviews and consumer participation in the review community. They discovered that a lengthier review wording will decrease customer satisfaction. Additionally, they discovered that the diversity and sentiment orientations of reviews would support enhancing customer satisfaction. Bi, Liu, Fan and Cambria [2019] created a model of customer satisfaction by analysing reviews. The suggested model is built on a neural network that takes into account customer satisfaction dimensions that were obtained by LDA. The findings demonstrated that the model was capable of accurately predicting customer satisfaction.

Kang and Park [2014] proposed a method for assessing customer satisfaction in the domain of mobile services based on reviews. Attributes of the services are weighted based on the sentiment score and the frequency of attributes. They use the VIKOR method [Opricovic 1998] for measuring the level of customer satisfaction regarding the service attributes. Also Dina et al. [2021] used VIKOR for studying the user satisfaction of educational services based on reviews. Wang et al. [2018] developed a model to analyse the influence of product attributes on customer satisfaction in the washing machines domain. Similarly, Xu [2020] studied whether the determinants of customer satisfac-

tion are reflected truly in the reviews. The results of both studies showed that the relevance of product attributes on customer satisfaction is different [Wang et al. 2018] and not all reviews can significantly affect customer satisfaction [Xu 2020].

2.4.5 Other Works

Najmi et al. [2015] provided a method for product ranking based on reviews. Review scores and brand scores are combined together to get the ranking scores. An updated page-rank algorithm was used to get the brand score, and the sentiment and usefulness of each online review were taken into account to determine the review score. Chen et al. [2015] developed a market structure visualisation method based on reviews. The method for ranking products is based on the TOPSIS (technique for order preference by similarity to an ideal solution) [Yoon 1980]. Fan et al. [2017] created a decision-based methodology to support customers in making purchases based on product attributes and reviews. The approach measures consumer preferences based on either product attributes or customer reviews. They also used the TOPSIS approach to generate the ranking scores. Finally, Li et al. [2020] proposed a method for product selection considering consumers' expectations and reviews. Their approach distinguishes between two types of product features based on the availability of quantified expectations which can be provided by customers. The proposed method used different techniques to deal with each type of features in order to select products and support user decisions.

We have discussed in this section some of the major works that have addressed the problem of ranking products based on user reviews. A salient feature of these works is that they rely on sentiment analysis to determine the polarity and level of sentiments related to product features. They used various techniques to synthesise the information from the reviews and calculate a final product score. The works provided personalised or average rankings. In our method, we also provide a sentiment-based product ranking. However, we consider not only the positive, negative, and neutral cases of mentioning product features with different weights, but also the case of not mentioning features. This provides an advantage for the ranking, since this type of information is usually positively weighted, as we have found in our studies. We also use simple IR models to weigh the sentiment scores of product features. However, the system we present is flexible and can be integrated with any method for measuring sentiment levels. The model can be personalised in terms of feature weights and sentiment weights. However, it

is also possible to generalise the model to provide average rankings.

2.5 Logic Based IR

Logic has been employed as a formal language to define models for IR. The idea of employing logic was early discussed in the work of Van Rijsbergen [1986], who modelled IR based on the assumption that both documents and queries are representable as logical sentences. Since then, researchers have explored various logic extensions that enable more accurate and efficient retrieval of information.

Propositional logic is one of the most basic logic extensions that has found extensive use in IR applications. Its ability to represent knowledge about a domain in the form of propositions, which are statements that can be true or false, makes it a valuable tool for structured reasoning. It is particularly useful in ontology-based retrieval systems that require reasoning with structured knowledge. In its basic form, propositional logic uses propositional variables and logical connectives, such as "and", "or", and "not", to construct logical expressions. These expressions can be then used to represent the content of documents and queries, and to match them for retrieval purposes.

However, representing information in IR using binary propositions alone is not enough due to the incompleteness, ambiguity, and contradiction of real-world information. As a result, one of the major challenges of IR was to develop techniques for representing and reasoning with uncertain information. Probabilistic models such as the binary independence retrieval (BIR) and Markov models are one approach to dealing with the issue [Fuhr 1992]. These models allow for the representation of uncertainty as probabilities, which can be updated as new information becomes available. Probabilistic models have been used in various IR tasks, including document retrieval, query expansion, and relevance feedback.

The use of fuzzy logic presents an alternative approach for handling uncertainty in information. It allows for the representation of degrees of truth rather than solely binary propositions. This enables the capture of imprecision inherent in natural language and provides a more accurate reflection of the similarity between a query and a document. Some of the previously described works, such as Liu et al. [2017*a*;*b*]; Atanassov [1986], have employed fuzzy logic or its extensions in IR tasks. However, such an approach faces several challenges, including the lack of transparency as it lacks a direct

connection to empirical data used for parameter estimation. Moreover, the use of fuzzy logic in IR can increase computational complexity and processing time, reducing its practical applicability in real-world settings. Additionally, selecting appropriate linguistic variables and membership functions can be complex and lead to inconsistent or unreliable results, especially when based on incomplete or inaccurate data.

Multi-valued logic (MVL) is a type of logic that has gained considerable attention for its ability to deal with uncertainty. Unlike traditional binary logic, MVL allows for multiple truth values for every predicate, including values such as "unknown" to represent uncertainty or the absence of information, and "inconsistent" to express cases where information was introduced as both true and false simultaneously. This makes it particularly useful for modelling complex information sets with uncertain and contradictory conditions. MVL encompasses different types such as Kleene's Three-Valued logic [Fitting 1994], which includes "true", "false", and "unknown" values; Belnap's Four-Valued Logic [Belnap 1977], which includes similar values in addition to an "inconsistent" value; and Subjective Logic [Jøsang 2002; Jøsang and Hankin 2012; Jøsang 2016], which has similar core values that are named "belief", "disbelief", and "uncertainty".

In our work, we are considering two types of MVL, namely four-valued logic and subjective logic. Four-valued logic was proposed for various retrieval tasks, including its usage for performing retrieval on complex objects and handling contradictions when aggregating information from different sources, as described in Rölleke and Fuhr [1996]; Fuhr and Rölleke [1998]. In addition, Frommholz and Fuhr [2006] utilised it within the framework POLAR (Probabilistic, Object-oriented Logics for Annotation-based Retrieval) for performing retrieval on annotated documents in digital libraries. The second type, subjective logic, is widely discussed and used to fuse information from various sources, particularly in domains such as trust assessment and network security [Koster et al. 2017]. However, it is rarely discussed in the domains of information retrieval or recommender systems [Haydar and Boyer 2017].

2.6 Summary

In this chapter, we conducted a literature review on the use of social data for information retrieval tasks. We described how the structure of web documents has evolved with the existence of social platforms and the emergence

of the field of social information retrieval (SIR). The review discussed the main challenges in SIR and how previous literature tackled them, as well as how we are addressing these challenges. Our solution focuses on recognising text as a set of logical statements describing a set of aspects or topics defined based on the domain of interest, such as hotels or laptops. Therefore, we included a review of IR tasks that also considered a similar task named as aspect-oriented search. We focused our research on the domain of reviews and how to utilise them for product ranking. Hence, we also included a literature review focused on a case study that used user reviews to rank or recommend products. Finally, as our work relies on logical models as the basis of the proposed approach, we included a review of the usage of such models in similar retrieval tasks.

Three common challenges were identified when using social content for IR tasks, including information credibility, contradictions, and missing information. We analysed previous works that addressed these challenges and highlighted the limitations of these works. We found that the previous works often focused on individual challenges, such as credibility or contradictions, and failed to provide an integrated solution for IR modelling in this context. Additionally, most of the previous works focused on keyword-based approaches to handle credibility and contradiction issues, whereas aspect- or topic-based logical statements were rarely utilised. Unlike traditional IR matching, which is typically based on two-valued logic, our work supports matching under multi-valued logic conditions. These are positive or negative sentiments of the logical statement, unknown if the query has no clear answer, and inconsistent if there are contradictions between the logical statements. We restricted our study to user reviews as a specific type of social data with a predefined set of aspects (or queries) because arbitrary statements cannot be identified in general.

The second part of the literature review focused on the use of user reviews for product ranking and recommendation modelling. We presented an expanded description of the topic, including expert and user-generated reviews. We also discussed some common approaches that were used to achieve the task, including works based on fuzzy set extensions, weighted directed graphs, product recommendation, and customer satisfaction. Some of these works offer solutions to deal with the missing information issue, but contradictions and credibility issues were not addressed.

Finally, we included a review of the usage of logical models in similar retrieval tasks, as our work utilises logical models as the basis of the proposed approach. We presented some examples of multi-valued logic models to

demonstrate their potential usefulness in this area, such as the four-valued logic and the subjective logic. This review emphasises the importance of using logical models to handle challenges in social data retrieval, specifically handling contradictions and missing information.

Overall, this chapter provided a comprehensive review of the relevant literature on using social data for IR tasks and highlighted the challenges faced in this field, along with potential solutions. The current study aims to provide an integrated solution for IR modelling in the context of SIR, with a focus on the use of logical models to address the challenges of credibility, contradictions, and missing information in user reviews for product ranking and recommendation modelling.

Chapter 3

Logical Framework for Modeling Review Based IR

3.1 Introduction

In the previous chapter, a comprehensive overview of various approaches for modelling the information retrieval task based on social content was presented. However, the literature review revealed a missing proper handling of the major issues of social content, including information credibility, contradictions, and missing information. To address this gap, our aim is to design a contextual framework for IR that takes into account these issues. The proposed framework will adopt a broader methodology for treating text in social contexts by representing it as a set of logical statements based on aspect-based queries. This will allow us to properly represent user opinions with positive and negative sentiments, instead of treating the text merely as a bag of keywords. Moreover, it is also necessary to handle missing information in a proper manner, following the principles of the Open World Assumption. All these features and conditions lead us to the use of multi-valued logic based approaches in building such a framework.

In this chapter, we present the theoretical basis of building frameworks for a review-based IR task using two multi-valued logical models. In the next section, we model the task using the four-valued logic (4vL). In section 3.3, we present the modelling of the framework using the subjective logic (SL). After that, we show the possible indexing and weighting techniques that can be used with both models.

3.2 Four-valued Logic(4vL)

Belnap's relevance logic [Belnap 1977] is a 4vL designed to aggregate information from multiple information sources, like the different reviews for a product in our case. Belnap complemented the two standard truth values *true* and *false* by *inconsistent* and *unknown*. *Inconsistent* means that we have both *true* and *false* values from different sources (e.g. reviews on one aspect), while *unknown* refers to the fact that we are missing information.

For applying 4vL, we assign truth values to each pair of an aspect a and a review r: *true* if the review talks positively about the aspect, and *false* in the contrary case; *unknown* is assigned if the aspect is not mentioned in the review. Below, we denote these three truth values by t, f, and u, respectively.

In addition, we compute probabilities for the truth values assigned, which reflect the strength of the sentiment of the reviewer's comment on the specific aspect. $P(t|a,r)$ reflects the positivity of r with regard to a and $P(f|a,r)$ the negativity, respectively. Normally, only one of these values will be different from zero, expressing a clearly positive or negative opinion. If both of these values are zeros, the aspect is not mentioned at all in the review, and if both are greater than zero, we would have mixed feelings in a single review (e.g. "brilliant display, but low resolution").

Furthermore, we always have the probability of uncertainty regarding review's r opinion with regard to a. The uncertainty which is also denoted as *unknown*, reflects the complement probability of all what is identified neither *true* nor *false*. It is formulated as follows:

$$P(u|r,a) = 1 - P(t|a,r) - P(f|a,r)$$

This way, our method clearly distinguishes between the case when an aspect is not mentioned in a review, and the case when there are negative comments. This is in stark contrast to standard IR methods which ignore negation, and only distinguish between absence and presence of a term. In 4vL, absence of an aspect in a review leads to $P(u|r,a) = 1$ and $P(t|a,r) = P(f|a,r) = 0$.

Given the probabilities for the three truth values (*true*, *false* and *unknown*) for each review, we need a method to aggregate these values for a set of reviews.

Here we also have to consider the credibility $cr(r)$ of a review, the probability that the claims in r are actually true.

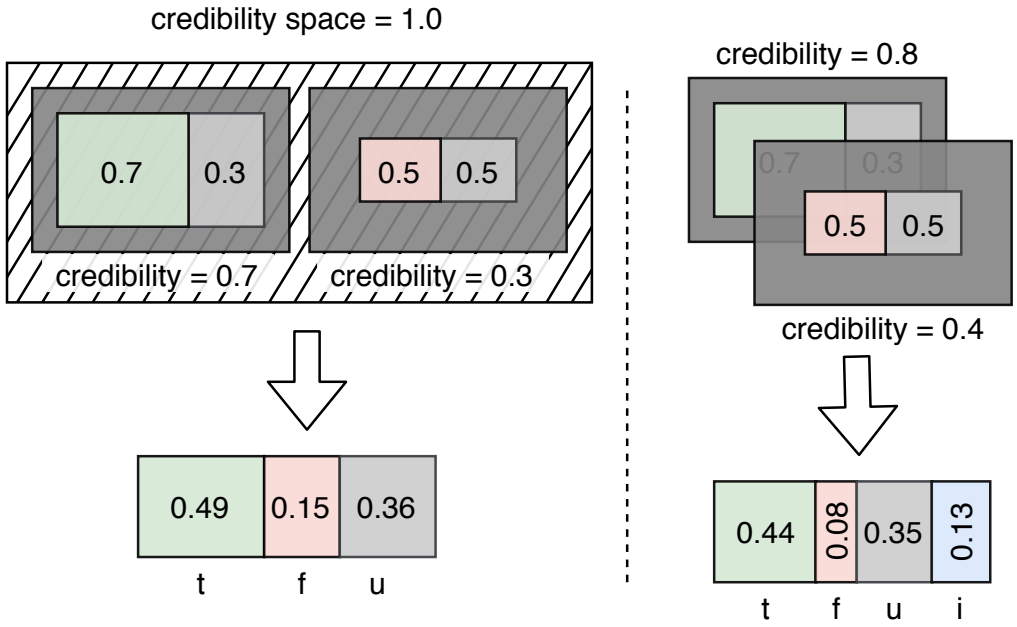


Figure 3.1: Aggregated 4vL truth values (t,f,u,i) for disjoint (left) and independent (right) credibility spaces.

3.2.1 Aggregation Options

For aggregating the reviews for an aspect, we regard two possibilities:

- The reviews actually refer to different instances of an item, e.g. some buyers of a hard disk might complain that they experienced a disk crash after a short time. In case other users have no such problems, there is no contradiction between the reviews – there is just a certain percentage of bad devices. In probabilistic terms, the different reviews can be modelled as *disjoint* events.

- The reviews are regarded as independent comments on the same instance (e.g. the content of a book), and there may be contradictory views. In this case, we regard the reviews as *independent* events which may overlap in event space, and in case they have different truth values, this contradiction leads to the truth value *inconsistent*.

Figure 3.1 illustrates the aggregation of probabilities from two reviews for both cases, for which we now give the precise definitions.

Disjoint case The default credibility space in this case is 1. Therefore we have to transform the original credibility values of individual reviews so that their sum does not exceed 1. The case when credibility space or the sum of credibility is less than 1 reflects an incomplete trust in the reviews, for example, when there are only a few of them. In such a case, the model will treat this kind of information (i.e. incredible) as an *unknown*.

Let R denote the set of reviews for a specific item, and β the overall trust in reviews, then we compute:

$$cr_d(r) = \beta \cdot cr(r) / \sum_{r' \in R} cr(r') \tag{3.1}$$

Assuming disjointness of reviews, we can now compute the truth values for an aspect of an item by aggregating over the set of reviews R in the following way:

$$P(t|a,R) = \sum_{r \in R} P(t|a,r) \cdot cr_d(r) \tag{3.2}$$

$$P(f|a,R) = \sum_{r \in R} P(f|a,r) \cdot cr_d(r) \tag{3.3}$$

$$P(u|a,R) = 1 - P(t|a,R) - P(f|a,R) \tag{3.4}$$

Independent case Here we have to consider all possible combinations of the truth values of the reviews. For two reviews, we get the aggregated truth values as a result of the following combinations:

- overall *true*: for the combinations (t,t), (t,u) or (u,t)
- overall *false*:from (f,f), (f,u) or (u,f)
- *inconsistent*: from (t,f) or (f,t)
- *unknown*: from (u,u)

As a simple example, assume that for some aspect, we have a positive review r_1 with $P(t|a,r_1) = 0.6$ and a negative one r_2 with $P(f|a,r_2) = 0.7$. Assuming that the reviews are independent events, we would get $P(t|a,R) = 0.6 \cdot (1-0.3)$, $P(f|a,R) = (1-0.6) \cdot 0.7$, $P(i|a,R) = 0.6 \cdot 0.7$ and $P(u|a,R) = (1-0.6) \cdot (1-0.7)$ (assuming that both reviews have credibility 1).

For the general case, we need to define paths representing combinations of truth values of all reviews. We have to regard all paths and classify them

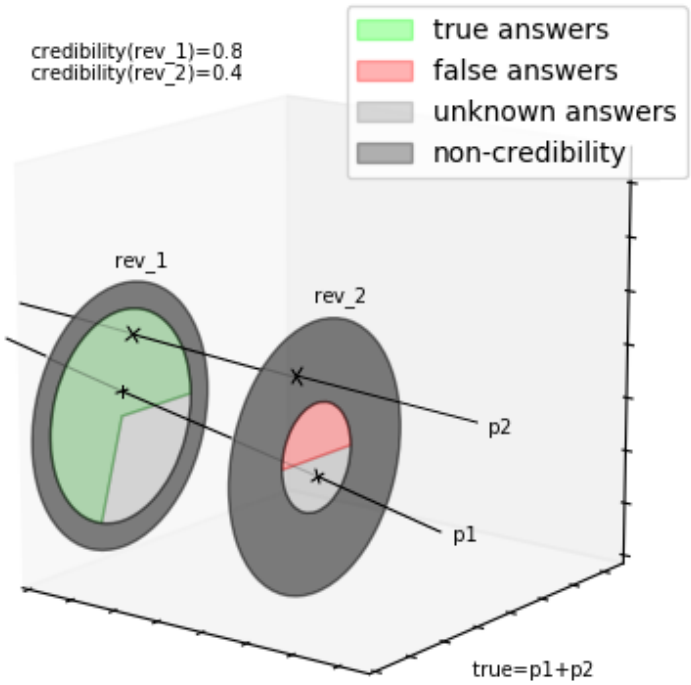


Figure 3.2: Example paths generation that lead to *true* (Accumulated *true* is not refined yet as mentioned in the final step of independent aggregation).

according to their output truth values, and then sum up the probabilities of all paths for a specific truth value.

The paths for the four truth values are defined according to the following rules:

- *True*: At least one *true* value, no *false* values, and zero or more *unknowns*.
- *False*: At least one *false* value, no *true* values, and zero or more *unknowns*.
- *Unknown*: All are *unknown* values.
- *Inconsistent*: At least one *true* and one *false* values, in addition to zero or more *unknowns*.

Figure 3.2 demonstrates the generation of all possible paths leading to true values. The graph shows two reviews (*rev_1* and *rev_2*) with different credibility values. According to the generation of paths for aggregation in the

independent case described previously, nine paths will be generated by aggregating the reviews. This number represents all possible combinations of truth values contained in reviews. These are (t,u,u) from *rev_1* and (f,u,u) from *rev_2* (Note that non-credible information has been treated as *unknown*). $p1$ and $p2$ are only two of the nine paths that would produce $true$ values. In this example, $p1$ and $p2$ form the overall $true$ value when they are summed together (before refining as described next in the last step).

In the final step, the probabilities are modified by removing the accumulated *unknown* information from the accumulated *true* and *false* reviews. The following formulas summarise the calculation probabilities in the independent case:

$$P(u|a,R) = \prod_{r \in R} 1 - (P(t|a,r) \cdot cr(r) + P(f|a,r) \cdot cr(r)) \tag{3.5}$$

$$P(t|a,R) = \left(\prod_{r \in R} 1 - (P(f|a,r) \cdot cr(r)) \right) - P(u|a,R) \tag{3.6}$$

$$P(f|a,R) = \left(\prod_{r \in R} 1 - (P(t|a,r) \cdot cr(r)) \right) - P(u|a,R) \tag{3.7}$$

$$P(i|a,R) = 1 - P(t|a,R) - P(f|a,R) - P(u|a,R) \tag{3.8}$$

3.3 Subjective Logic (SL)

SL [Jøsang 2002; Jøsang and Hankin 2012; Jøsang 2016] is a multi valued probabilistic logic that considers uncertainty and subjective opinions. It provides definitions for binomial and multinomial cases. For information retrieval tasks, query matching is regarded as a binomial case (i.e. relevant and irrelevant).

In SL, the truth of a binomial opinion about proposition x is defined as a tuple $\omega_x = (b,d,u,a)$, with $b,d,u,a \in [0,1]$ and $b+d+u = 1$. Values b,d,u,a are identified as:

- b represents the belief mass in support of x being true.
- d is the disbelief mass in support of x being false.

- u is the uncertainty mass about the probability of x.
- a is the prior probability of x.

The probability projection (a.k.a. expected probability) of a binomial proposition x is defined as $P(x) = b + a \cdot u$. This value represents the degree of certainty wrt. the truth of proposition x.

3.3.1 Fusion Operators

To combine opinions from various sources, SL provides many fusion operators for binomial opinions. Here we make use of two common operators: *cumulative* and *averaging fusion* which are comparable to the independent and disjoint cases in 4vL, respectively. If the observations (i.e. opinions) are about the same state of an object, cumulative fusion is used. Averaging fusion should be applied if the observations are about different states of an object (like e.g. reviewing different instances of a device, as mentioned above).

Cumulative fusion Let x be a proposition and $\omega_x^A = (b^A, d^A, u^A, a^A)$ and $\omega_x^B = (b^B, d^B, u^B, a^B)$ be source A and B's respective opinions over the same proposition x. The cumulative opinion $\omega_x^{A \oplus B} = \left(b^{A \oplus B}, d^{A \oplus B}, u^{A \oplus B}, a^{A \oplus B}\right)$ is calculated as:

For $u^A \neq 0$ or $u^B \neq 0$:

$$\omega_x^{(A \oplus B)} = \begin{cases} b^{A \oplus B} = & \frac{b^A u^B + b^B u^A}{u^A + u^B - u^A u^B} \\ d^{A \oplus B} = & \frac{d^A u^B + d^B u^A}{u^A + u^B - u^A u^B} \\ u^{A \oplus B} = & \frac{u^A u^B}{u^A + u^B - u^A u^B} \\ a^{A \oplus B} = & \frac{a^A u^B + a^B u^A - (a^A + a^B) u^A u^B}{u^A + u^B - 2u^A u^B} \end{cases}, \tag{3.9}$$

Averaging fusion Let x be a proposition and $\omega_x^A = (b^A, d^A, u^A, a^A)$ and $\omega_x^B = (b^B, d^B, u^B, a^B)$ be source A and B's respective opinions over the same proposition x. The average opinion $\omega_x^{A \varnothing B} = \left(b^{A \varnothing B}, d^{A \varnothing B}, u^{A \varnothing B}, a^{A \varnothing B}\right)$ is calculated as[1]:

[1] The original literature uses symbol '$\underline{\oplus}$' instead of '$\varnothing$' for denoting averaging fusion operator.

For $u^A \neq 0$ or $u^B \neq 0$:

$$\omega_x^{(A\varnothing B)} = \begin{cases} b^{A\varnothing B} = & \frac{b^A u^B + b^B u^A}{u^A + u^B} \\ d^{A\varnothing B} = & \frac{d^A u^B + d^B u^A}{u^A + u^B} \\ u^{A\varnothing B} = & \frac{2u^A u^B}{u^A + u^B} \\ a^{A\varnothing B} = & \frac{a^A + a^B}{2} \end{cases}, \tag{3.10}$$

In the absence of uncertainty ($u^A = 0$ and $u^B = 0$), different formulas are employed to handle dogmatic opinions. As IR rarely deals with certain information, we do not regard this case here and refer to the original papers mentioned below.

Cumulative fusion is commutative, associative and non-idempotent. On the other hand, Averaging fusion is commutative and idempotent, but not associative. Although fusion operators are always applicable in two-sources cases, the missing of the associative property of the averaging fusion makes this operator not well-defined in the case of multi-source fusion. This challenge and other multi-source fusion related issues have been addressed in recent work Van Der Heijden et al. [2018]. For formulas, justifications and other details about the operators, we refer to the original papers Jøsang [2002]; Jøsang and Hankin [2012].

As a concrete example for these fusion operators, let us assume a user is searching for a high-performance laptop. One of the offered laptops has two reviews. Review X reports the high-performance of the laptop with 0.9 confidence, while review Y reports the low-performance with 0.7 confidence. Assuming the prior knowledge about this aspect to be 0.5, cumulative fusion yields $\omega^{X\oplus Y}_{high-performance} = (0.73, 0.19, 0.08, 0.5)$; on the other hand, averaging fusion would lead to $\omega^{X\varnothing Y}_{high-performance} = (0.67, 0.17, 0.15, 0.5)$. .

3.3.2 Trust networks

For creating a more reliable analysis of information that considers the trustworthiness of information, SL has proposed the trust networks [Josang et al. 2006] which presents a method for deriving information through a complex network of referral trusts and functional trusts. A referral trust is a relation between the information seeker and the information owner. Functional trust

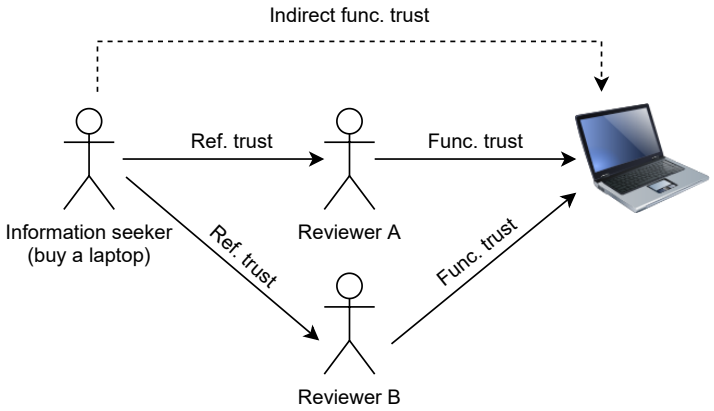


Figure 3.3: SL trust network of an IR task based on reviews.

represents a relation between the information owner and the sought target (see figure 3.3).

In trust networks, both referral and functional trusts are represented by subjective opinion. SL has proposed an operator called *trust discounting* which is used to derive trust through the so-called functional and referral trust. Here we only regard referral trust for modelling a user's trust in a review. In the case of complex networks, different fusion operators are used for aggregating referral and functional trust opinions into one opinion as described in Škorić et al. [2016].

In SL terminology, an agent A's referral trust about agent B (i.e. B's credibility in the eyes of A) is represented as a subjective opinion and denoted by ω_B^A. The projected probability of ω_B^A is defined as $P_B^A = b_B^A + u_B^A \cdot a_B^A$. The opinion ω_x^B is B's opinion on the proposition x (functional trust) as x is recommended by B to A. The function that yields the trust-discounted opinion $\omega_x^{[A;B]} = \omega_B^A \otimes \omega_x^B$ is defined with the following components:

$$\omega_x^{[A;B]} = \begin{cases} b_x^{[A;B]} = & P_B^A b_x^B \\ d_x^{[A;B]} = & P_B^A d_x^B \\ u_x^{[A;B]} = & 1 - b_x^{[A;B]} - d_x^{[A;B]} \\ a_x^{[A;B]} = & a_x^B \end{cases} \tag{3.11}$$

As a follow-up of the *high-performance* laptop example, let us assume that user A searching for a laptop trusts the opinion of reviewer X by

70% as a "belief" and the remaining 30% as "unsure". Here, the user's overall trust of the information of review X is $\omega_X^A \otimes \omega_{high-performance}^X = (0.7, 0, 0.3, 0.5) \otimes (0.9, 0, 0.1, 0.5) = (0.765, 0, 0.235, 0.5)$.

3.3.3 Conflicts handling

SL has also proposed operators for handling the possible conflicts between the opinions under the case of uncertainty.

Let A and B be two agents that have their respective opinions ω_x^A and ω_x^B about proposition x in domain X. Let PD denote the projected distance between opinions of agents A and B is defined as:

$$PD(\omega_X^A, \omega_X^B) = \frac{\sum_{x \in X} |P_X^A(x) - P_X^B(x)|}{2}$$

.

Furthermore, let CC be the conjunctive certainty of opinions of agents A and B defined as:

$$CC(\omega_X^A, \omega_X^B) = (1 - u_X^A)(1 - u_X^B)$$

The degree of conflict DC between opinions of agents A and B is defined as:

$$DC(\omega_X^A, \omega_X^B) = PD(\omega_X^A, \omega_X^B) \cdot CC(\omega_X^A, \omega_X^B) \tag{3.12}$$

On the same *high-performance* example, to discover the degree of conflict between opinions of X and Y, the DC operator has to be applied, $DC(\omega_{high-performance}^X, \omega_{high-performance}^Y) = 0.504$.

3.4 Differences between 4vL and SL

Most of the features of the two types of logics are quite similar, but there are a few key differences that should be noted when discussing them. On the similarities side, both logics deal with propositional logical statements in an uncertainty-supportive environment. Both types also support aggregation operators for different aggregation types.

One of the major differences between the two approaches is the handling of missing information. In four-valued logic (4vL), this is modelled via an explicit truth value for the aggregated reviews, which can be made transparent to the end user. In contrast, subjective logic (SL) handles missing information via the prior beliefs for positive and negative opinions. For items with belief values in the medium range, it is unclear if these scores are supported by actual comments in the reviews or if they are mainly the result of prior beliefs.

The calculation of contradictions/inconsistency is also a difference between the two logics. In 4vL, it is calculated as the complement product of probabilities of true, false, and unknown information, whereas SL measures it by the distance between the projected probabilities of the SL opinions.

To illustrate this difference, let's consider two logical statements representing opinions extracted from different user reviews regarding the cleanliness of a hotel. In opinion A, a user reflects their positive impression about the cleanliness with a degree of 80%. In opinion B, on the other hand, a user sees the hotel as not clean with a degree of 40%. By representing opinions with 4vL, we obtain probabilities of $P(t|cleanliness,A) = 0.8$, $P(u|cleanliness,A) = 0.2$, $P(f|cleanliness,B) = 0.4$, and $P(u|cleanliness,B) = 0.6$ with no predefined prior beliefs. Representation using SL, on the other hand, results in $\omega^{A}_{cleanliness} = (0.8, 0, 0.2, 0.5)$ and $\omega^{B}_{cleanliness} = (0, 0.4, 0.6, 0.5)$ given a predefined prior belief of 0.5. By fusing the opinions under the conditions of the independence (or cumulative) aggregation of the events, we end up with a quadruple of probabilities of $(t, f, u, i) = (0.48, 0.08, 0.12, 0.32)$ in 4vL and $(b, d, u, a) = (0.7, 0.12, 0.18, 0.5)$ in SL. Note that the value a in SL is the prior probability of the resulted aggregated opinion, whereas to find the conflict degree (DC) as an alternative of the value i in 4vL, we use the equation 3.12, yielding $DC = 0.19$.

From the previous example, it is clear that the different calculation methods specifically affected the t/b and i/DC values. This is because missing information in SL is treated as true under the prior probability, resulting in an overestimation of the positive scores. To reduce the effect of prior beliefs in SL, we suggested two separate query propositions to support the positive and negative assumptions of the unknown information. Further information can be found in section 3.5.2.

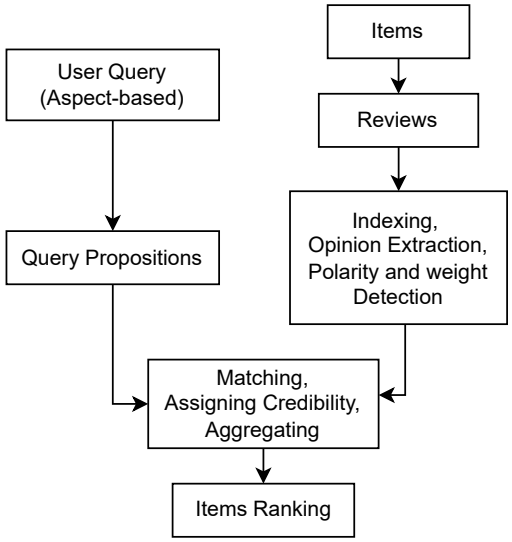


Figure 3.4: An abstract overview of the IR task based on the reviews

3.5 Complete Review-Based Retrieval Task

After having presented the logical foundations of our models, we focus now on the IR task. This task consists in retrieving relevant products to a user query based on the opinions expressed in a set of product reviews. While the task sounds simple, it presents a number of interesting challenges such as how to aggregate the opinions, how to deal with vague or ambiguous information, how to deal with the credibility and contradictions issues appearing in the reviews and how to assign weights to the opinions in order to compute accurate ratings or ranking scores for the products. In our work, we present a method to tackle all these issues and describe the methods employed to solve the task.

Before starting, it is important to indicate the intended meanings of the terms used throughout this and next chapters. Following the common terminology used in IR, we use the term items or documents to refer to a specific type of web documents that consist of reviews on a particular product (e.g. hotels, laptops). A query is referred here to the request made by the user to retrieve these documents. This takes the form of an aspect of the items of the domain of interest (e.g. cleanliness of a hotel or performance of a laptop). Opinions are pieces of information extracted from reviews. They represent the abstract comparable form of queries.

The retrieval task we present is performed through one of the two types of

probabilistic logic i.e 4vL and SL. The ranking of the items is depending on the values we create by aggregating the opinions from its reviews. Figure 3.4 shows an overview of the task. A general pipeline of a retrieval task consists in general of the following phases:

- A user is submitting an aspect-based query to search for items based on their reviews.
- Query propositions are introduced. The proposition of a query represents the positive and negative sentiments of the reviews toward the user query.
- User reviews are processed and indexed. This includes aspects definition, opinions extraction and sentiment polarity and weight detection.
- A credibility value is assigned for each review depending on the aggregation type. By aggregating the indexed opinions from all reviews, yielding a triple or quadruple of probabilities for the truth values that form the matching status between the query and the items.
- Items are ranked based on the generated truth values.

3.5.1 Review Indexing

Indexing is a vital component in any information retrieval system. In social media information retrieval, we are concerned with capturing and indexing logical statements that reflect users' opinions about aspects of the domain of interest. Traditional information retrieval systems typically treat this as a keyword-based task, but this approach falls short when dealing with the unique challenges present in social media contexts. To address these challenges, we require a different indexing method that can capture and represent social data at the level of individual opinions.

The core concept of this type of indexing approach assumes to have a well-defined product ontology, in which we categorise products and define a set of aspects for each category. To extract logical statements from product reviews, three crucial components are necessary. One component must be able to recognise when a particular aspect has been mentioned in a review. The second component must determine the polarity of the sentiment expressed towards that aspect. Finally, as a probabilistic approach, it is also required to determine the weight of the opinions expressed in each review.

The first step in the process involves recognising words or phrases in the reviews that describe a specific aspect of the product. This involves mapping these terms to the user's opinion on that aspect. For example, if a review about a hotel mentions "very clean," this expression would be interpreted as a positive sentiment regarding the aspect of cleanliness. Our approach uses word embeddings to convert the words in the reviews into numerical vectors, which allows us to easily calculate the similarity between words based on their vector representation. We start with a set of seed terms that represent the predefined aspects in the product ontology and then iteratively expand the list of terms to cover the aspects in the relevant domain.

The second component of our task requires determining the sentiment orientation of the mentioned aspects. To do this, we must conduct sentiment analysis on the opinion phrases or sentences that discuss the aspects. Finding the right phrases is crucial because sentiment analysis can make incorrect conclusions if the phrases are misidentified. Our approach involves using an improved sentence level tokenisation technique[2] that involves correcting punctuation and spellings. Although this method is simple, it has demonstrated better results compared to other state-of-the-art tools[3] for this domain. For sentiment analysis, we use a pre-trained tool [Sun et al. 2019] based on the SemEval 2014 dataset [Wagner et al. 2014], that contains data on hotel and laptop domains.

The determination of the weight of the opinions is a crucial matter to address. While it is possible to calculate the overall positivity or negativity of an opinion for each aspect using sentiment analysis, we sought to explore an alternative method in our experiments to test the viability of using multivalued logic for IR in review-based environments. To this end, we employed a method that considers the relevance score of the opinions and, more importantly, enables a fair comparison with conventional IR methods. Specifically, we evaluated the relevance of the opinions based on popular IR models, such as TF, TFIDF, and Okapi BM25.

This is a summary of the methods we utilised for extracting the logical statements present in the reviews. Other techniques discussed in the literature can also be implemented here. For instance, Dalila et al. [2018] demonstrates a survey of relevant approaches. In our work, we have partially[4] or

[2]fastPunct: `https://pypi.org/project/fastpunct/` (last accessed 12.12.2022)

[3]DeepSegment: `https://github.com/notAI-tech/deepsegment` (last accessed 12.12.2022)

[4]In the hotel domain, sentiment analysis was not necessary as the reviewers explicitly expressed their positive or negative sentiments.

fully applied these methods, as we will see in the upcoming chapters.

So far, we have accomplished the following steps: 1) defined the aspects of the targeted domain, 2) paired each aspect with a collection of relevant keywords, 3) gathered opinions from the product reviews, and 4) assigned both a sentiment and a weight to each opinion. By combining all of these elements, we can form logical statements that express the opinions of reviewers, which can then be indexed as follows:

Each aspect a is represented by a set of keywords K_a terms. For a specific review r, let r^+ denote the set of terms occurring with positive sentiment, and r^- those with negative sentiment. Furthermore, $w(k|r)$ is the term weight of keyword k in r. Then we can compute the positive and negative opinion's sentiment of r wrt. aspect a as follows:

$$\begin{aligned} sl^+(a,r) = &\quad \alpha \textstyle\sum_{k \in K_a \cap r^+} w(k|r) \\ sl^-(a,r) = &\quad \alpha \textstyle\sum_{k \in K_a \cap r^-} w(k|r) \end{aligned} \tag{3.13}$$

α is a normalisation constant depending on the actual term weighting method used, which ensures that both sl^+ and sl^- can be interpreted as probabilities such that $0 \leq sl^+(a,r) + sl^-(a,r) \leq 1$.

For $w(k|r)$, the following weighting methods were regarded *term frequencies* (tf), *tf-idf* and *Okapi's BM25*. Weights are defined as follows:

$$\begin{aligned} tf(k,r) = &\quad \tfrac{f(k,r)}{|r|} \\ tf-idf(k,r) = &\quad tf(k,r) \cdot \log(R/(f(k)+1)) \\ bm25(k,r) = &\quad \log(R/(rf+1)) \cdot \tfrac{f(k,r)\cdot(a+1)}{f(t)+a\cdot(1-b+b\cdot|r|/r_{avg}))} \end{aligned} \tag{3.14}$$

where $f(r,t)$ is the number of times term t occurs in review r. R is the number of elements of the reviews set. $|r|$ is the number of words in review r. r_{avg} is the average number of words per review. a and b are hyperparameters for BM25.

We also store a unique ID of the object ($ID_{obj}(r)$) that is the subject of a review r in order to be used later for the aggregation procedure. The index representation of reviews set R with a predefined aspect set A is shown in table 3.1.

Document	Index
r_1	$[\{ID_{obj}(r_1), sl^+(a_1, r_1), sl^-(a_1, r_1)\}, ...\{ID_{obj}(r_1), sl^+(a_k, r_1), sl^-(a_k, r_1)\}]$
	...
r_k	$[\{ID_{obj}(r_k), sl^+(a_1, r_k), sl^-(a_1, r_k)\}, ...\{ID_{obj}(r_k), sl^+(a_k, r_k), sl^-(a_k, r_k)\}]$

Table 3.1: Reviews index

3.5.2 Query propositions

The matching procedure of the user query and the reviews is based on query propositions. In a traditional retrieval task, a user query has only one proposition that reflects positive matching of the query and related to the presence of the query terms. Here with the logical models, a query can be interpreted as two propositions. The first proposition represents the positive side of the user query while the second represents the negative one.

In general, positive and negative supportive propositions are associated with the sl^+ and sl^- values respectively. However, there are differences between the logics on the handling of the missing or unknown information.

In the 4vL case, we consider for each review the probabilities $P(t|a,r) = sl^+(a,r)$ and $P(f|a,r) = sl^-(a,r)$ as supportive to the positive and negative query propositions respectively. For the unknown probability, there will be no need to handle it as an independent query proposition. It will be handled as a complement of the positive and negative probabilities $P(u|r,a) = 1 - P(t|a,r) - P(f|a,r)$. Also inconsistency probability is implicitly considered in the positive and negative propositions, however, the value will be set to $P(i|r,a) = 0$ as a review cannot contradict itself.

For SL, we extract for each review a positive supportive opinion $\omega_P = (sl^+(a,r), 0, 1 - sl^+(a,r), 0.5)$ and a negative supportive opinion $\omega_N = (sl^-(a,r), 0, 1 - sl^-(a,r), 0.5)$ and assign them to the positive and negative query propositions respectively. The uncertainty about the propositions will be automatically included within these two opinions.

3.5.3 Credibility Assignment

The model we propose is agnostic to the details of the credibility model, and we henceforth assume that it has been utilised, and the trustworthiness of each review and each piece of information in the review is known. We embed the credibility information of reviews into the logical models.

Assignment of credibility is varying depending on the aggregation method considered. In disjoint and averaging fusion cases, credibility values have to be normalised so that all credibility values sum up to 1.0. In independent and cumulative fusion cases, credibility values are in the range from 0 to 1.0.

In 4vL, the credibility of a review is simply multiplied by probabilities of the truth values to create credible probabilities $P(t|a,r) \cdot cr(r)$ and $P(f|a,r) \cdot cr(r)$ for the positive and negative query propositions respectively. The incredible information is handled as *unknown* information.

In SL, acquiring credible information based on user reviews can be achieved via a network of different trust types. First, the information seeker who has the user query, trusts the reviewers indirectly. Such trust is the *referral trust*. Second, the reviewers have another kind of trust based on a personal experience, i.e. *functional trust*. For these trust types, we assume that a trust model should be used to estimate the trust value of the information written by the reviewer. For estimating the trust in trust networks, the operator *Trust discounting* is used. For each review, a trust opinion $\omega_{cr} = (cr(r), 0, 1-cr(r), 0.5)$ is discounted from the SL opinions that represent the positive ω_P and negative ω_N query propositions. Trust types and trust operator in SL were discussed in section 3.3.2.

3.5.4 Aggregation

This operation aims to fuse the information mentioned in the reviews into one single piece of information. Information fusion can be done in various ways, each having an impact on how the specific piece of information is interpreted. It is often challenging to determine the correct or the most appropriate fusion method for a specific situation. In general, it is more appropriate to apply the *independent case* in 4vL or *cumulative fusion* in SL when the reviews are representing observations about the same state of the evaluated object. In contrast, the *disjoint case* or *averaging fusion* are more suitable if the reviewers are observing different states or instances of the evaluated object e.g. refurbished or used products.

As a result of fusing information of reviews into one piece of information, the user query is now represented by a concrete answer based on the user reviews and shows the matching level between the searched items and the user query. In 4vL, the answer to the query propositions is formatted in a shape of quadruple of aggregated logical values (t, f, u, i). In SL, two opinions ω_P and

ω_N are resulted of the aggregation representing the positive and the negative query propositions respectively. Each of the opinions is projected later into a single probability value ($P_P = b_P + u_P.a_P$ and $P_N = b_N + u_N.a_N$).

After aggregating reviews, contradictions are already addressed with the 4vL. For explicit contradictory value, *inconsistency* probability is used to indicate it. In SL, the DC (degree of conflict) operator has to be used in order to find the inconsistency between the fused positive and negative supportive opinions ω_P and ω_N.

3.5.5 Ranking

Ranking in IR can generally be divided into one-dimensional and multi-dimensional rankings. Due to its simplicity, computational efficiency, and user-friendly interfaces, most IR systems use a single value to rank the results in a list with the most relevant results appearing at the top. This single value typically represents the relevancy information in different ways, depending on the ranking model used. For instance, it could be a traditional score such as TFIDF, BM25, or PageRank, a combination of multiple relevance criteria, or the outcome of a nontransparent black-box model trained on vast amounts of Clicked-Query-Document pairs.

On the other hand, there are IR methods that present outcomes on multiple dimensions [Federkeil et al. 2012]. These methods can rank search results based on various criteria, such as relevance, popularity, and others. Similarly, some methods employ one-dimensional ranking, but they can also support multi-aspect ranking, such as performance versus price. These approaches typically use a range of diverse visualisation techniques to enable users to quickly comprehend the results and make informed decisions. Examples of visualisation techniques used in this area include 2D and 3D graphs, heatmaps, scatter plots, histograms, and others.

In multi-valued logic models, results' scores are represented either as a quadruple (t, f, u, i) in 4vL or as two values P_P and P_N in SL. Our work has utilised transparent models to convert these values into a single score for each rankable item. The process of converting multi-valued logical values into single scores is discussed in section 4.2. Integrating the multi-valued logical models to visualise a multi-dimensional search task is an idea worth exploring in future work. Examples of this are demonstrated in figure 3.5.

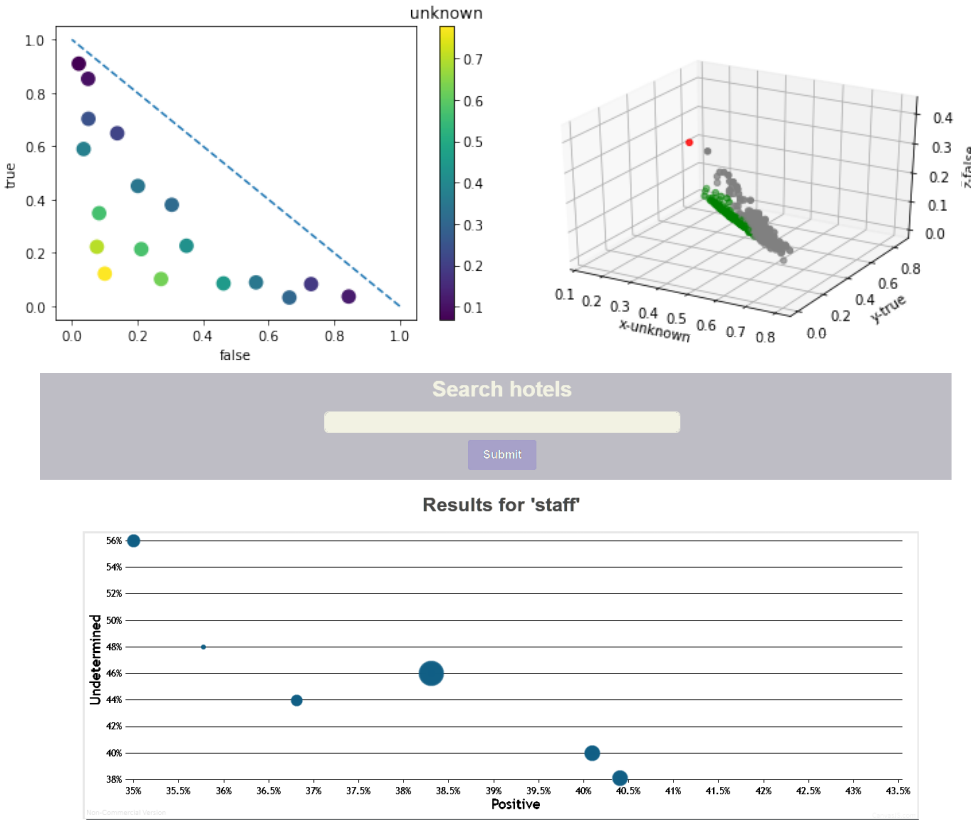


Figure 3.5: Examples of using 4vL values to display search results on 2D, 3D, or interactive interfaces.

3.6 Summary

In this chapter, theoretical foundations of the proposed models based on four valued logic and subjective logic were presented. Our Modelling procedure focused on the issues of the thesis scope: addressing credibility, contradictions and information omissions in a review-based IR task. To the end, an ideal task based on the proposed models is shown in multiple phases. Each of the phases shows an implementation of the models' theories on a set of reviews. Parts of the work presented here have also been previously published in Sabbah [2019]; Sabbah and Fuhr [2021].

Part II

System- & User-Oriented Evaluations

Chapter 4

Example Application I: Aspect-oriented Hotel Evaluation

The last chapter has discussed the foundations of the multi-valued logical models. In this chapter, we will start the process of assessment with a system oriented approach. We apply the proposed models in an actual retrieval task. For that, we have considered the hotel domain. A general pipeline for such a task consists of: 1) aspect extraction and indexing, 2) credibility assignment, 3) opinion aggregation and 4) ranking hotels based on the given user query (i.e. aspects). Figure 4.1 demonstrates an overview of the complete retrieval task. In the next sections, we describe the process step by step and present the evaluation of the performance of the proposed models by comparing them with baseline models.

4.1 Dataset

In order to evaluate the aspect-specific weighting of items based on reviews, we need a dataset that includes aspect scores for products. Fortunately, there is a hotel booking site (Booking.com) containing these values. We used a data collection from Feuerbach et al. [2017] for our evaluation. This dataset contains 839K reviews of 11.5K hotels in Berlin, Brussels, Barcelona, London and Rome. Each review consists of an overall rating score, a title/summary, a section of positive and a section of negative points (see figure 4.2). For our experiments described in the following, we only considered the latter two parts. This allows us to do a proper evaluation of the logic-based weighting formulas, which are the focus of this chapter; otherwise, if we had to apply sentiment analysis, the intrinsic difficulties of this method (e.g., 'low price' vs. 'low comfort') could have had an unknown effect on the experimental

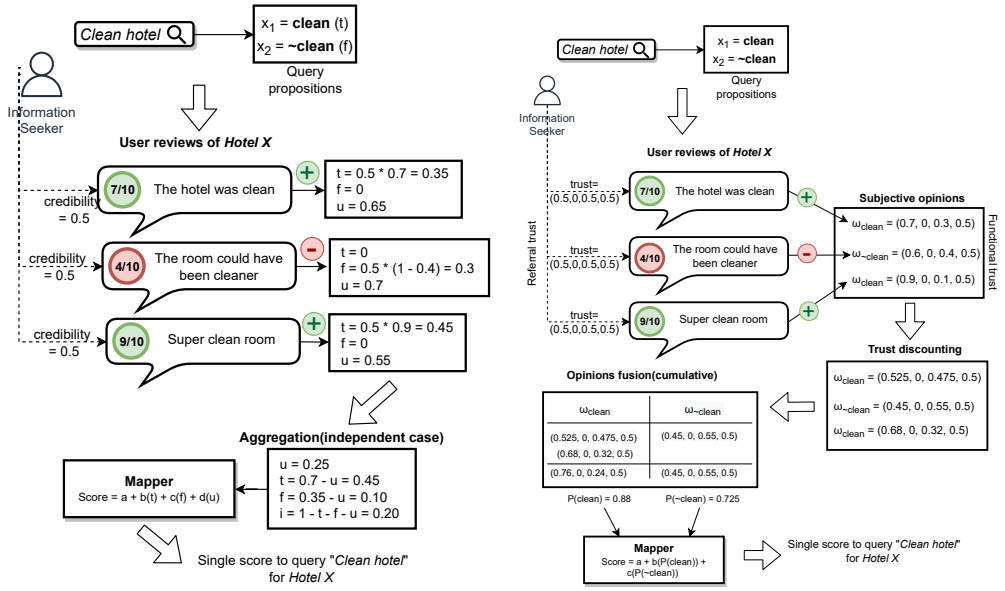


Figure 4.1: An overview of the retrieval task through 4vL and SL approaches in hotel domain.

results.

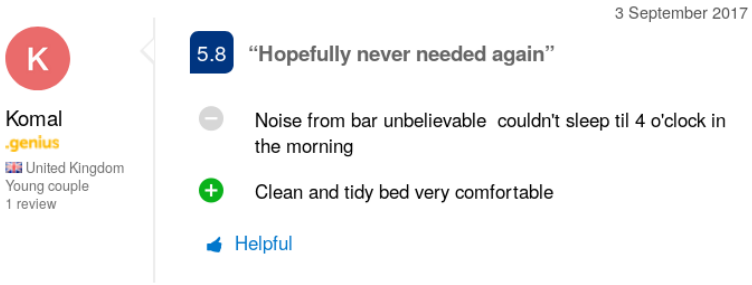


Figure 4.2: Sample hotel review from Booking.com

In addition to the scores and textual contents of the review described above, the review form in Booking.com (figure 4.3) asks the users to rate the following seven aspects of the hotel they stayed in: cleanliness, comfort, staff, value for money, location, wifi, and facilities. Users rate each of these aspects on a four-point likert scale, but these individual judgements are not available in the dataset; for each aspect, we only have the overall score aggregated over all reviews of a hotel (as shown in figure 4.4[1]). In our experiments, we use

[1]This screenshot of the review form does not show the wifi aspect. It is possible that the form was changed when this screenshot was taken.

these aspect ratings as ground truth for the aspect scores to be estimated by the methods proposed in this thesis.

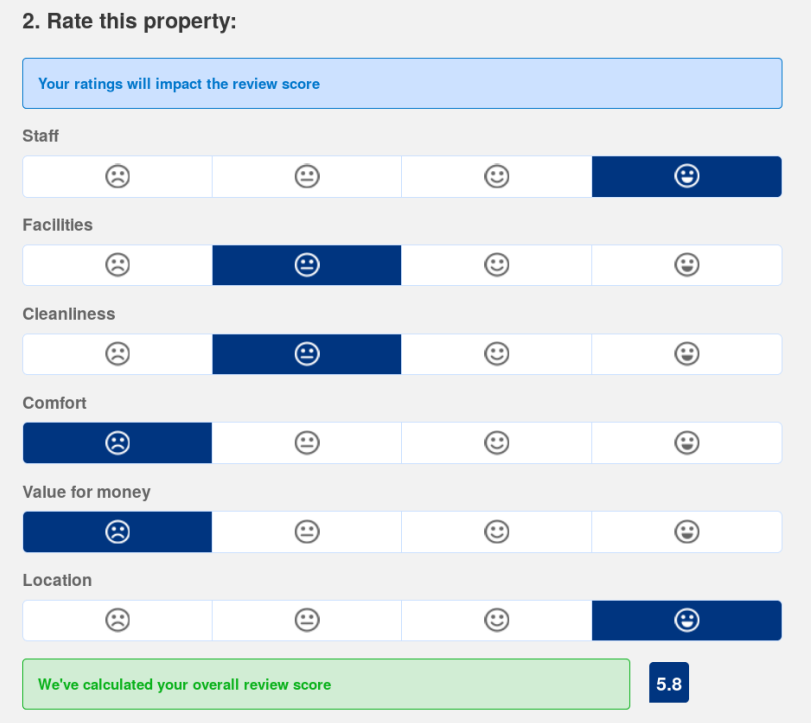


Figure 4.3: An example hotel review-form in Booking.com. User inputs are mapped to numerical values from 2.5 to 10 on a 2.5 step.

Figure 4.4: An overall aspect scores of an example hotel in Booking.com

4.2 Applying the Proposed Approaches

Before applying the logical models, we first need to process the reviews. This requires defining the relevant terms for each of the seven mentioned hotel aspects. Then, assigning positive and negative weights for each review with regard to each aspect. At this point, an index of review data can be established. In the retrieval, we need to simulate seven user queries, each representing one of the hotel aspects. Once the retrieval system gets any of these queries, it is required to determine the aggregation method (i.e. disjoint or independent); according to that, the credibility of the reviews is assigned. The logical models aggregate the review data in the index and create an output for each hotel in the dataset. Details on these steps are provided below.

Mapping Reviews to Aspect-Sentiment Pairs For mapping the texts from the positive and negative sections onto aspects, a word2vec model was trained using the text from the reviews. This method generated a set of related terms for each aspect label (see table 4.1). The full list of the aspect keywords is attached in appendix A.1.

Term Weighting Methods To weigh the terms of a review with regard to the searched aspect, we use traditional term weighting methods of IR. The following weighting methods were regarded in our experiments: tf, tf-idf and Okapi's BM25. There will be two weights; one for the terms of the positive section of the review and one for the negative section. Weights are calculated and normalised as described in equation 3.13.

Creating A Review Index By mapping reviews to aspects and finding their weights, we can now start to build an index. Each entry in the index will be a tuple containing a review ID, hotel ID, and aspect's positive and negative weights. Multiple entries may be linked to the same review if multiple aspects were satisfied simultaneously. An index of sample reviews given above is shown in table 4.2

Credibility Assignment The logic-based models regarded here all consider review-specific credibility values. However, Booking.com allows only confirmed customers to write reviews. It also verifies the authenticity of reviews

Aspect	Keywords (including misspelled terms)
cleanliness	clean, unclean, smelly, cleaning, neat, clen, dirty, ...
location	lokation, locatio, locstion, centrality, situated, ...
staff	reciption, fiendly, staff, owner, managers, crew, ...
	...
facilities	gym, bathroom, cooking, shower, garden, entrance, ...

(a)

Review			Mapped aspects	
ID	Positives	Negatives	Positives	Negatives
R1	Man on **reception** was smiley and very helpful. Breakfast was ok.	**Bathroom smelt** like sewers had to keep the door closed. Ants in the room.	staff	facilities, cleanliness
R2	**Staff** were great.	Could do with a small **gym** on site.	staff	facilities

(b)

Table 4.1: Sample reviews (b) mapped to aspects based on an aspects dictionary (a).

Review ID	Hotel ID	Aspect	Positive Weight	Negative Weight
R1	H1	staff	0.3	0
		facilities	0	0.2
		cleanliness	0	0.2
R2	H1	staff	0.2	0
		facilities	0	0.1

Table 4.2: Example reviews index. Weights are not real values.

before publishing them[2]. Thus, we assume the credibility values are equal for all reviews: For 4vL, we choose a credibility of 1.0 in the independent case, and distribute a value of 1.0 equally over all reviews in the disjoint case. For SL, we use a fixed value of 1.0 for trust of each proposition, which is represented by a subjective logic opinion $trust = (1.0, 0, 0, 0.5)$. In this case, no trust discounting operation is necessary, as the trust is neutralised.

[2] `https://partner.booking.com/en-us/help/guest-reviews/what-are-guest-reviews-and-who-can-write-one` (last accessed on 12.12.2022)

Logical Values to Numeric Single Value Mapping The aggregation of reviews is performed using one of the logical models. Both methods yield a vector of probabilities for the different truth values of a hotel-aspect pair. As we want to relate these estimates to the corresponding aspect scores in the dataset, we have to map the probability vectors onto a single value (i.ewe want to predict the corresponding aspect ratings). For this purpose, we tested linear regression, kNN, SVR and random forest, and found that simple linear regression gave the best results; so we used this method for all experiments described in the following, and applied it in the same 10-fold cross validation setup for all methods tested.

4.3 Baselines and Evaluation Metric

We compare our logic-based methods to different baselines where in each case, we index the positive and the negative sections of each review separately. Then, we make use of the term weights of traditional approaches (tf, tfidf and BM25) to compute positive and negative scores for each hotel. These scores are computed as follows:

$$Sc^{+}(h,a) = \frac{\sum_{r\in R}\sum_{k\in K_a\cap r^{+}} w(k^{+}|r)}{N_a^{+}} \tag{4.1}$$

$$Sc^{-}(h,a) = \frac{\sum_{r\in R}\sum_{k\in K_a\cap r^{-}} w(k^{-}|r)}{N_a^{-}} \tag{4.2}$$

Here $Sc^{+}(h,a)$ and $Sc^{-}(h,a)$ are the positive and negative scores of a hotel h for an aspect a. R is the set of reviews of hotel h, and N_a^{+} and N_a^{-} are the numbers of reviews with positive/negative comments on aspect a, respectively. K_a, r^{+}, r^{-} as well as the term weighting function $w(.)$ are defined as in equation 3.13.

We used these scores and the number of positive and negative reviews in a feature vector and then applied linear regression for predicting the aspect ratings of hotels, in the same setting as with the logic-based methods.

For measuring the quality of the prediction, we adopt the well-known Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3}$$

Here $\hat{y}_i$ is the predicted value of a (hotel,aspect) pair i, and y_i is the corresponding ground truth value. $\bar{y}$ is the mean of all ground truth values.

4.4 Results and Discussion

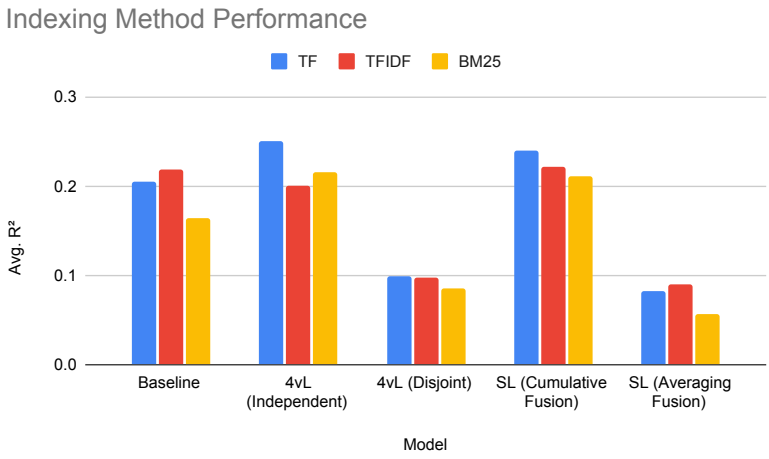


Figure 4.5: Indexing method performance measured by the average R^2 scores of all aspects.

Figure 4.5 shows the average aspect R^2 scores of all tested models for each indexing method. As we can see, the most appropriate indexing method for this task appears to be term frequency alone. This may be because the aspect-oriented word embedding method used to determine meaningful terms cannot be further improved by IDF weighting. It is also notable that in most cases, TFIDF indexing performed better than BM25, which suggests that the impact of review length and term frequency importance are not major factors in this task. For more detailed results, please refer to appendix A.2.

Figures in 4.6 present the R^2 scores of the linear models in each aspect. The results show that 4vL in the independent case is outperforming the baseline approaches in all aspects. Subjective logic in the cumulative fusion case also achieves close results. However, when examining the results of the 4vL (disjoint) and SL (averaging fusion) cases, it becomes apparent that their performance is not as good as that of the baselines. There is a noticeable difference in performance between the independent 4vL and disjoint 4vL, as well as between cumulative and averaging fusion in SL. This suggests that in each scenario, the former method is more suitable. This can be explained by

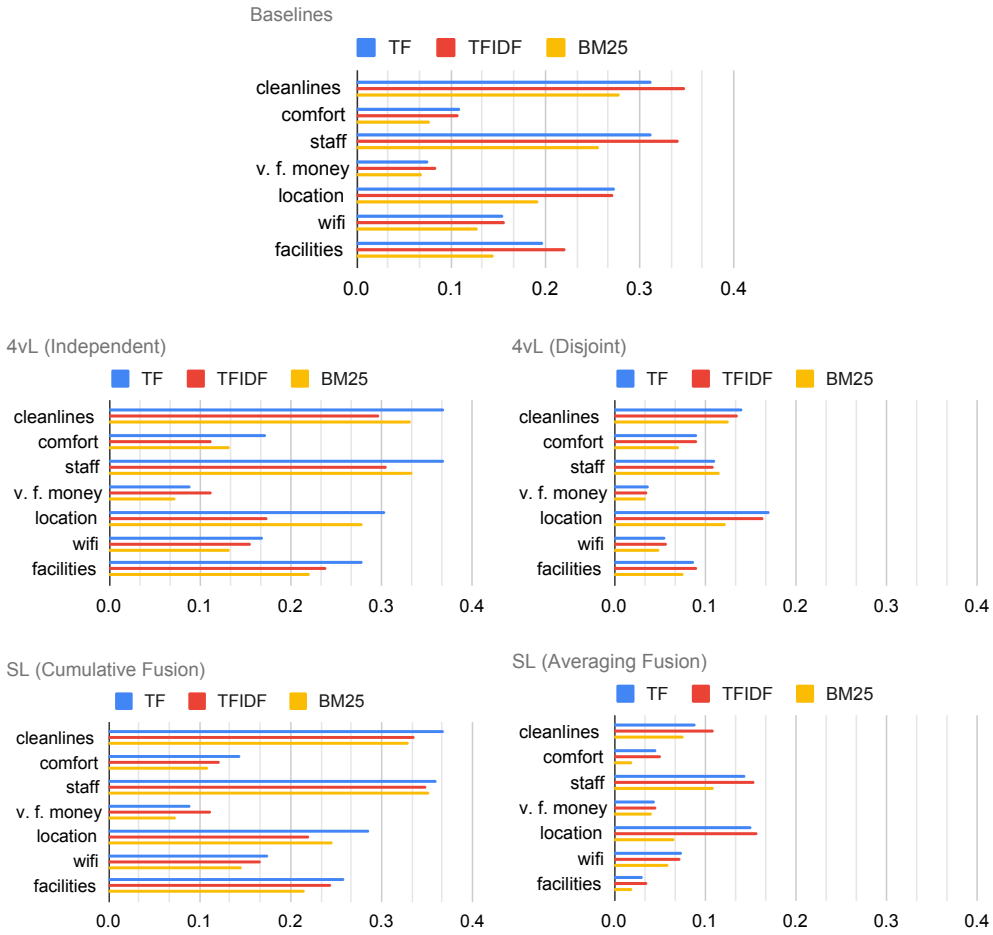


Figure 4.6: R^2 scores of logic-based and baseline approaches for each hotel aspect.

the fact that all users are evaluating essentially the same item - any variations between e.g. rooms or the behaviour of staff on different days appear to be minor and have a negligible impact.

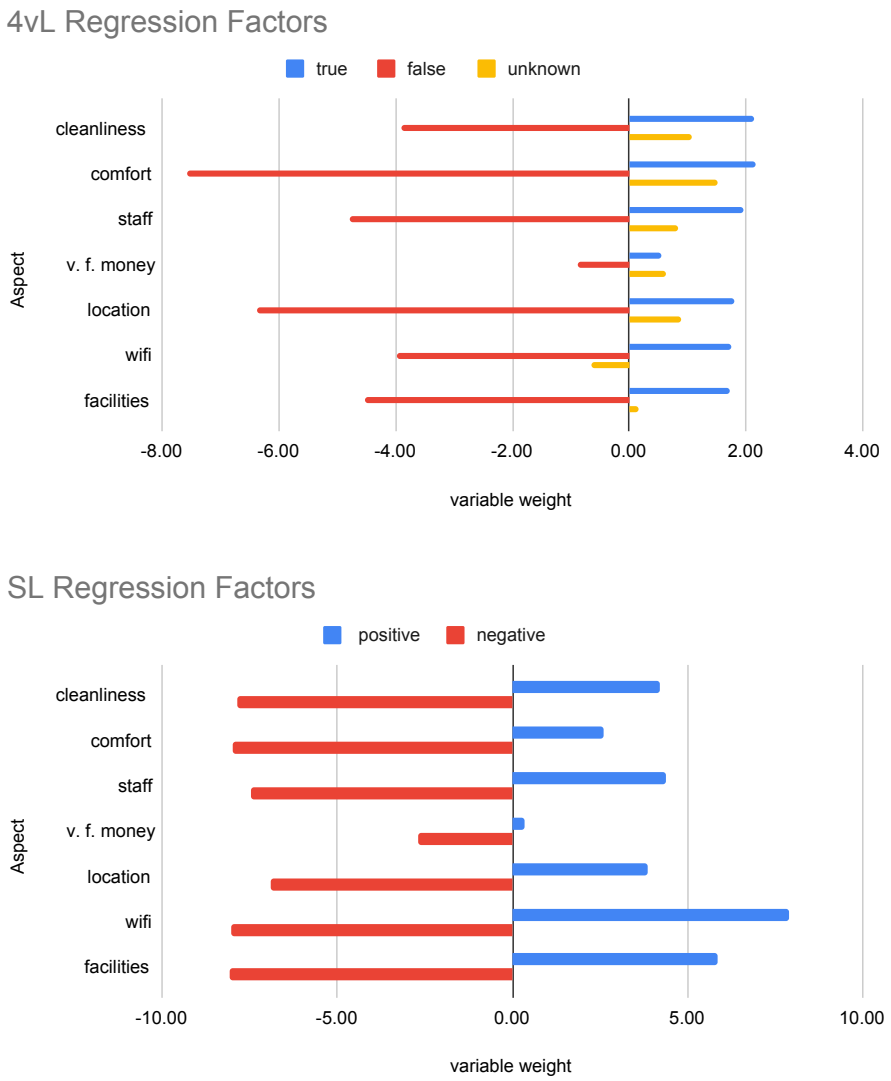


Figure 4.7: Regression factors of the 4vL and SL values for the models based on the TF indexing method.

Experiments here were conducted on two aggregation cases. The reason for this is the difficulty in determining the most appropriate aggregation method beforehand. This refers to the way users evaluate objects and the nature of the evaluated objects. Although in most cases and domains, independent or cumulative cases are considered the most appropriate for aggregation due

	4vL (independent)				SL (cum. fusion)		
Aspect	SL (cum. fusion)	TF (B)	TFIDF (B)	BM25 (B)	TF (B)	TFIDF (B)	BM25 (B)
location	0.671	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
staff	0.749	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
cleanliness	0.494	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
comfort	0.227	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
value for money	0.919	0.451	0.612	0.581	0.513	0.684	0.652
facilities	0.237	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
wifi	0.838	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**

Table 4.3: Post-hoc tests results of the multi-comparison test of the 4vL and SL models compared with baselines (B) TF, TFIDF and BM25.

to the neglected differences in services offered among different customers, disjoint and averaging fusion cases may demonstrate noticeable differences in performance when applied to different domains where changes between objects are noticeable, such as seller/shop reviews, etc. This suggests that the logical models have the advantage of distinguishing between various cases. On the other hand, baseline models only provide an average review for all cases, lacking the ability to differentiate between them.

Looking at the regression factors of the logical models displayed in diagrams 4.7, we can observe that, with the exception of the aspect "value for money" (where overall rating is the best predictor), the weighting factors for all aspects are relatively similar. The false/negative values have a greater weight than the true/positive ones. Additionally, the weight for "unknown" values is mostly positive, which indicates that the default assumption for reviews tends to be positive. It is worth noting that negative reviews have a significant impact on the overall score of an item, aligning with the findings from Sun [2012]. Please note that there are no coefficients for "inconsistent" values, as its probability is directly proportional to the other three probabilities. For more detailed results about the regression factors, please refer to appendix A.3

However, the changes in weights reveal the varying significance among the different aspects. While these numbers indicate the general user opinion about each aspect, it is also important to consider factors that may impact the results. This includes the frequency of mentioning the aspects, which is also dependent on the quality of the topic detection method and the number of keywords defined for each aspect. Additionally, there may be interference between some topics, such as comfort and facilities.

We further conducted a test to determine if the performance of the proposed

models significantly differed from each other. For this purpose, we ran a One-Way ANOVA multiple comparison test on three baseline models (tf, tfidf, and bm25) and the two logical models (4vL and SL). Out of the logical models, we selected the ones built using the tf indexing method, as it demonstrated the best performance according to the R^2 results. The comparison was based on the absolute error, which is calculated as $e = |\hat{y} - y|$, where y is the proposed hotel-aspect score on a 10-point scale and $\hat{y}$ is the golden standard score of the hotel aspect pair as given in the dataset. The significance p-values for all comparisons were adjusted using the Bonferroni method. As can be seen in Table 4.3, both logical models outperformed the baselines on all aspects except for 'value for money'. The results also indicate that there is no significant difference in performance between the 4vL and SL models. Further details about this test can be found in appendix A.4.

A powerful characteristic of our models is that they account for both positive and negative interpretations of the *unknown* information. The 4-valued logic model determines whether the knowledge is *true* or *false* by assuming that at least one of the reviews has been classified as such, while the rest are classified as *unknown*. This *unknown* probability reduces the values of *true* and *false*, offering the model the chance for multiple interpretations of reviews that have been labeled as *unknown*. On the other hand, subjective logic also considers the distinction between positive and negative evidence, and it assumes that the *uncertainty* regarding an event can be viewed as the *belief* probability associated with the aspect being considered.

4.5 Summary

The chapter has focused on a system-oriented experimental study applied to a single specific collection, where we have predefined aspects and aspect ratings for each item. The aim of this study was to examine the general suitability of logic-based approaches for handling credibility, contradictions and omissions in product reviews. For new domains or applications, an adaptation of logic-based models with the required task is straightforward. Aspects can be identified with their identifier keywords. In addition, comments or reviews have to be classified according to the sentiments toward the aspects. The aspect-specific ratings in the Booking.com case are mainly needed for evaluation; here we also used them for tuning the mapping of the probability vectors onto a linear scale. Parts of the work presented here have also been previously published in Sabbah and Fuhr [2021].

Chapter 5

Example Application II: User Support on Products Ranking Based on Reviews

In the previous chapter, we evaluated the proposed logical models using a system-oriented approach with a labelled dataset from the hotel domain. In this chapter, we will take a user-oriented approach and evaluate the models' applicability in situations where a ground truth is not available. We will focus on the task of assisting users in making purchasing decisions, specifically in an online laptop store application. The evaluation will be based on a comparison between instances of the application that utilise one of the suggested logical models and other instances that use a baseline method for product ranking.

5.1 Introduction

When users make online purchases, they need to be confident that the products they are buying meet their needs and expectations. There are different existing methods to helping users make decisions about their purchase choices; one of the most known and effective approaches is star ratings. However, such ratings usually provide a generic overview of products, and they do not take into account the needs and preferences of users in a specific aspect of the product. The star rating may also be inaccurate or misguiding since the reason for a low or high rating is not always known.

In this work, we offer a method that employs the textual content of reviews to create a ranking that matches the user-requested aspects. The approach

relies on a prior work's findings about the average importance of positive and negative information included in reviews as discussed in chapter 4. It also considers the average opinion of required information that was not previously included in user reviews. The approach suggested here is structured in three main phases: 1) finding related products based on specific filter conditions, 2) collecting and processing online reviews, and 3) assessing and comparing the performance of the new ranking method to the performance of alternative baseline methods.

We performed two user studies to evaluate the performance of the suggested approach on a set of predefined user requirements on laptop aspects (Battery, Keyboard/Mouse, Price, Screen/Display, Sound/Audio, and Storage). In the first user study, we ask participants to perform a laptop search operation. We provide them with product ranked using our method as well as two alternative baselines. The second study examines whether the highlighting of relevant passages in reviews influences users' search and buying decisions across different ranking systems.

The remainder of this section is structured as follows. Sections 5.2 and 5.3 provide details of baselines and our ranking method, respectively. In Section 5.4, we illustrate the feasibility of the proposed method by providing a case study in which three laptops are ranked using three ranking methods. In section 5.5, we present the second user study with four different setting combinations of 1) ranking methods and 2) highlighting reviews functionality. Section 5.6 provides an analysis and discussion of the results of both studies. Finally, Section 5.7 discusses the conclusions and limitations of this study, and suggests directions for future research.

5.2 Baseline Models for Ranking Products

In this section, we provide two baselines for ranking products depending on the reviews. One uses a numerical scale, in the form of stars (traditional star ratings). Second, we discuss another method for evaluating products based on the review's textual content. The evaluation of the latter method includes six predefined aspects of laptops mentioned above.

5.2.1 Star Rating Based Ranking Method

In the first baseline, we determine the strength of a product's recommendation based only on the star ratings of the reviews. The method is well-known and very common on many e-commerce platforms. Furthermore, consumers have typically utilised it to get a better sense of their buying decisions [Zhou and Duan 2012; Chen 2017]. For this baseline, we simply take the average star rating of each product's existing reviews and then calculate the ranking based on this number.

5.2.2 Textual Ranking Method Based on the Two Valued Logic (Textual 2vL)

In star rating based ranking, the actual ranking is determined only by the number of stars. Therefore, the text of the reviews has no influence on the ranking. As a result, the ranking cannot be customised to a particular search interest of the user. The method provides only a summary or an average assessment of the products. Consequently, as an alternative baseline for ranking products, we suggest a text-based ranking approach.

The basic idea behind this method is to identify a match between the user's needs, which are represented by a set of desired product features or aspects, and the product reviews. This procedure is similar to how search engines find matches between user queries and document contents. As we will discuss a method based on four-valued logic later, we note here that this method is based on two-valued logic. This means that a match is considered *true*, whereas a non-match is considered *false*. This method will later be referred to as "2vL" for short.

We analyse the set of reviews for each product to see if any of the predefined aspects are mentioned in the reviews. We apply a *word-to-vec* based method to generate a set of terms relevant to the searched aspects. Then, using the *tfidf* weights, we weigh the mention of each searched aspect. As a result, the final weight of an aspect a composed by a set of terms T for a review r is:

$$weight_r(a) = \sum_{t}^{T} tfidf(t)$$

The weights are always normalised in accordance with the length of the re-

view. For a set of reviews R of a product P, aspect a is scored as the following:

$$score_P(a) = \sum_{r}^{R} weight_r(a)$$

Finally, we calculate the normalised score of a set of aspects A of a product P by:

$$score_P(A) = \frac{\sum_{a}^{A} score_P(a)}{number\ of\ aspects\ in\ A}$$

5.3 Four Valued Logic Based Ranking (Textual 4vL)

Using the two-valued logic based method explained in the section above, matching scores are computed for aspects based on their appearance in reviews. However, when applied in the field of reviews, the method encounters two fundamental issues. First, the method creates ratings based only on the explicit mentioning of the aspects and disregards the possible meaning of the aspects' absence in the reviews. Second, although the method considers varying weights for aspect terms according to their frequency, sentiments that accompany these terms are not taken into account. For the first issue, the disappearance of an aspect in the reviews may represent meaningful feelings about that aspect. Regarding the second issue, for a more precise ranking, the ranking method should consider the aspect terms in the manner in which buyers perceive the reviews.

As an alternative method, we will use our four-valued logic based approach (4vL) which addresses both aforementioned issues. The complete details about 4vL are described in chapter 3. Most of the details about the approach usage in this domain are identical to its usage in the first example application as in chapter 4. However, a few things have to be introduced as we are dealing with a different orientation of the IR task (i.e. user oriented task).

First, the mapping between the truth values resulting from an aggregation operation of the reviews is different from what was performed in chapter 4. Previously, we had gold standard values for each domain aspect that formed the basis of the mapping; in this experiment, such values are not available. As a result, we will calculate the average need of users for *posi-*

tive, *negative*, and *unknown* knowledge cases using the weights of the system-oriented experiment. More details about this are given in section 5.3.1.

Second, in chapter 4, each of the domain aspects was handled and evaluated individually; however, in this experiment users can be interested in more than one aspect. Therefore, for visualisation purposes, we introduce a normalised ranking score of a set of selected aspects as described in section 5.3.2.

Finally, in this experiment, we have to handle the problem that users may be affected (or biased) because of the existing star-rating evaluations. To do that, we have performed an additional normalisation of the 4vL and 2vL scores in order to make them as similar as possible to the star-rating scores. Further details are shown in section 5.3.3.

5.3.1 4vL Values Mapping

In this step, we are mapping the scores of 4vL (true, false, unknown) into one value. We are relying on the results of our experiments in the hotel domain as described in chapter 4 to fit the weights of the three scores on a linear model of a 10-points scale.

The results of the fitting have shown that the weights are changing depending on the given aspect. However, most aspects have shown closer weights, therefore, we are taking the mean weights as heuristic values to build up a standard final linear model of all aspects. We ended up with fitted 4vL values weights that also describe the average information need of users toward an aspect a:

$$\begin{aligned} score(a) = \ & 1.73 * true \\ & - 4.58 * false \\ & + 0.64 * unknown \end{aligned} \tag{5.1}$$

5.3.2 Creating a Normalised Score for Multiple Aspects

Although the non-normalised scores will not affect the actual ranking of the products, normalisation is a required step in order to create alternative visualised representations of the star ratings.

In order to create a normalised ranking score regardless of the number of the searched aspects, we calculate the mean value of all scores of aspects set *A*.

$$Normalised\ ranking\ score\ =\ \frac{\sum_{a}^{A} score_a}{number\ of\ aspects\ in\ A}$$

5.3.3 Avoiding the Potential Impact of Star Rating on User Decisions

Star rating is one of the most common methods for evaluating products and services. This rating can significantly influence a consumer's decision-making. Particularly, it plays a role in increasing the perceived information quality of the users and decreasing the cognitive decision efforts for them [Chen 2017].

In our work, we are focusing on textual based methods for ranking. Measuring the quality of the methods involves keeping the cognitive decision of users independent of any factor except reviews' textual content. However, we see that removing the star rating completely from the tools of textual based ranking methods would probably make an unfair comparison with the star rating method since users are used to that method for evaluation. So, instead, we propose to convert the ranking scores of the textual-based methods (i.e. 2vL, 4vL) into five-star scores and replace the traditional star rating with them.

However, this normalisation is not enough to resolve the issue. In order to make a fair comparison between all methods, the distribution of the star rating evaluations has to be as similar as possible for all methods (i.e. 2vL, 4vL and star rating). For that, we imply an additional normalisation function according to the frequency distribution of the star ratings of the entire dataset. Later on, in the search process, when the user finishes the filtration process, the scores of the textual-based methods are adjusted in order to make the frequency distribution of the scores of the search results equivalent to the distribution of the original star ratings. Nevertheless, that process is performed always on a subset of the dataset based on the selected filters. That means the distribution of star ratings among the three methods might be in total not identical but should be similar.

5.4 USER STUDY I: Textual Based vs. Star Rating Based Rankings

The textual 4vL method attempts to be an alternate ranking technique that relies only on the textual content of reviews. To determine if our method can compete with numeric-based star rating, and textual-based 2vL baselines, we conducted a user study to answer the following research question:

RQ1: How well does the 4vL method perform in comparison to the baseline methods from the perspective of the users?

In this preliminary study, we used a within-subject design to compare the three rankings. Here, we investigated whether users are better served by 4vL, 2vL, or star rating based rankings by focusing only on the highest-ranked product.

As we are focusing on an aspect-oriented product search, we assume that the textual-based ranking methods in general perform better than the star rating method, which merely offers a broad average review of the products. Also, we assume that 4vL will perform better than the traditional 2vL, as in 4vL, the information is handled not only based on the existence of aspect-specific terms, but also according to the sentiment of the existence (positive and negative cases) and the absence of terms (unknown case).

We are evaluating the performance of the ranking methods based on multiple variables: buy likelihood and product inspection time. Additionally, we are measuring customer satisfaction with the ranking results based on the likelihood that they will purchase the products.

To formulate our assumptions into hypotheses, we have for each of the measuring variables two hypotheses. The main hypothesis is comparing the textual methods in general with the star rating based method, and the sub-hypothesis which compares the 4vL method with the 2vL method. For the variable 'buy likelihood', we hypothesise the following:

- **H1**: In an aspect-oriented search, the textual based approaches (4vL and 2vL) would provide users with more relevant products based on the information included in the reviews, hence increasing the buy likelihood in comparison to the star rating method.
- **H1a**: In an aspect-oriented search, the 4vL textual-based approach would provide users with more relevant products based on the infor-

mation included in the reviews, hence increasing the buy likelihood in comparison to the textual 2vL method.

Buying decision may be made regardless of the buy likelihood, for example when users are unsure about their decisions or they are unsatisfied with all products. Consequently, we hypothesise:

- **H2**: In an aspect-oriented search, users are more likely to decide to purchase a product that was ranked highest by using textual based approaches as compared to the star rating method.
- **H2a**: In an aspect-oriented search, users are more likely to decide to purchase a product that was ranked highest by using 4vL textual based approach as compared to the 2vL method.

In the star rating method, review content is not considered in the ranking. Therefore, we expect that users would find it difficult to locate the required information about aspects in the reviews. It is assumed, however, that this will not be an issue with the textual based ranking methods, as product reviews are more likely to include such information. Thus, we hypothesise:

- **H3**: In an aspect-oriented search, the textual based approaches will bring the products that are most relevant to the searched aspects to the top, hence reducing the time required to get sufficient product information compared to the star rating based method.
- **H3a**: In an aspect-oriented search, the 4vL textual based approach will bring the products that are most relevant to the searched aspects to the top, with consideration of the sentiment of the searched aspects, hence reducing the time required to get sufficient product information compared to the traditional 2vL method which does not distinguish between the sentiments of mentioning.

5.4.1 Scenario and Task

This user study comprises three phases: 1) Participants were required to provide informed consent and answer basic demographic questions (age, gender, domain knowledge about laptops). 2) They are then redirected to a tool for an online laptop store to do the actual search and buy. Participants use the tool in the following workflow:

- The participants read the instructions and the scenario description, which led them to assume that they need a new laptop to replace their old damaged laptop. Furthermore, the scenario allowed participants to choose the aspects of laptops that they are interested in. The following aspects are available: (Battery, Keyboard/Mouse, Price, Screen/Display, Sound/Audio, and Storage). See image 5.1.

- Participants are asked to filter the laptops according to their needs. The filters are designed on top of vendor-provided laptop attributes. These filters are (Price in Pound sterling, RAM, operating system, Brand, Hard drive type, screen size, processor cores, processor speed, processor manufacturer and average battery life). Filters are always set to default option *Any*, which indicates that any value for that attribute will be included in search results. See image 5.2.

- The participants are then given a list of three laptops, one for each of the three ranking methods (4vL, 2vL, and star rating). See image 5.3. The laptops are placed on the results panel in a random order to reduce the impact of the item's order in the results list.

- Each of the listed laptops is clickable. Participants will see the laptop details and reviews when they click on it (see image 5.4). Participants are asked to read the details and reviews until they feel well-informed about the aspects they selected at the beginning. We've added two buttons to go to the next phase. The first button is located at the top of the page and is accessible only once the participant has spent 15 seconds on it. Alternatively, if the participant scrolls down to the bottom of the details-reviews page, the second button appears. The laptops in the results list all use the same GUI to display product details and reviews.

- After viewing the details and reviews of the laptops, participants are asked to evaluate the likelihood of buying each laptop. Furthermore, we ask participants to explain their decision in an open text box (as shown in image 5.3).

- Following that, participants choose which of the three laptops to buy.

In phase 3), we provide the participants with a post-questionnaire in which we ask them to give further information about the significance of the selected aspects as well as the source of the information they obtained about them, whether it was from reviews, product details, title, or image. Complete study details and surveys can be found in the appendix A.5.

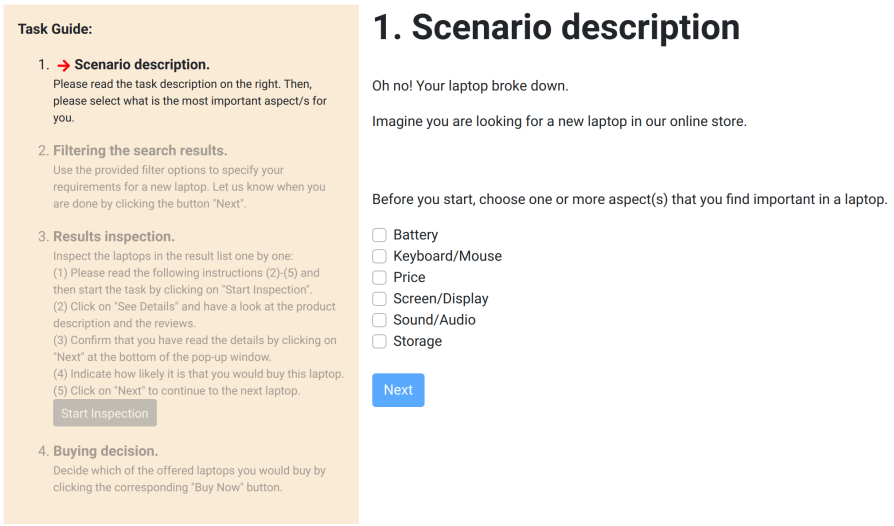


Figure 5.1: A screenshot of the tool shows the task guide and the scenario description.

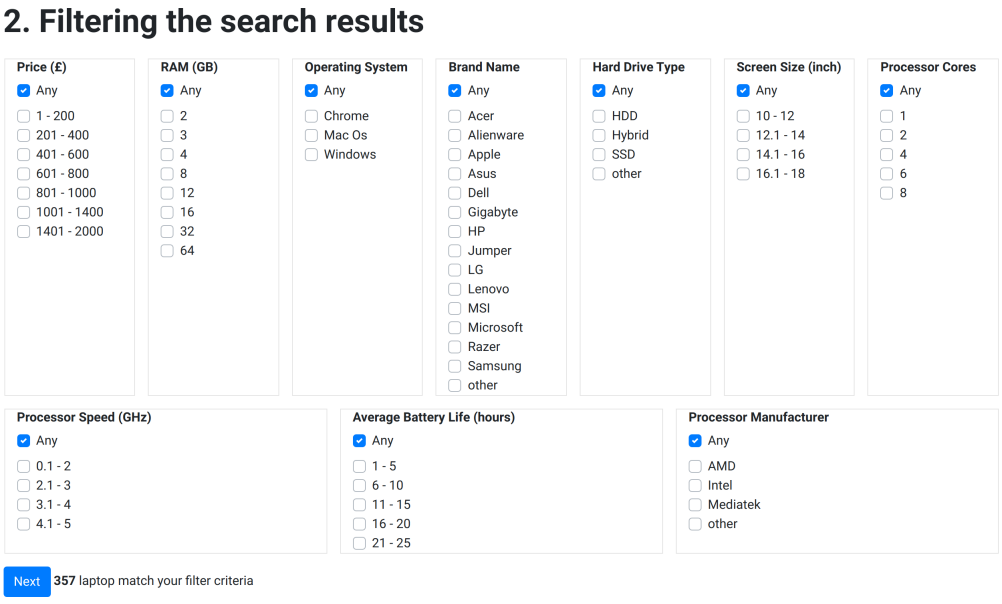


Figure 5.2: A screenshot of the tool shows the vendor-provided attribute based filters.

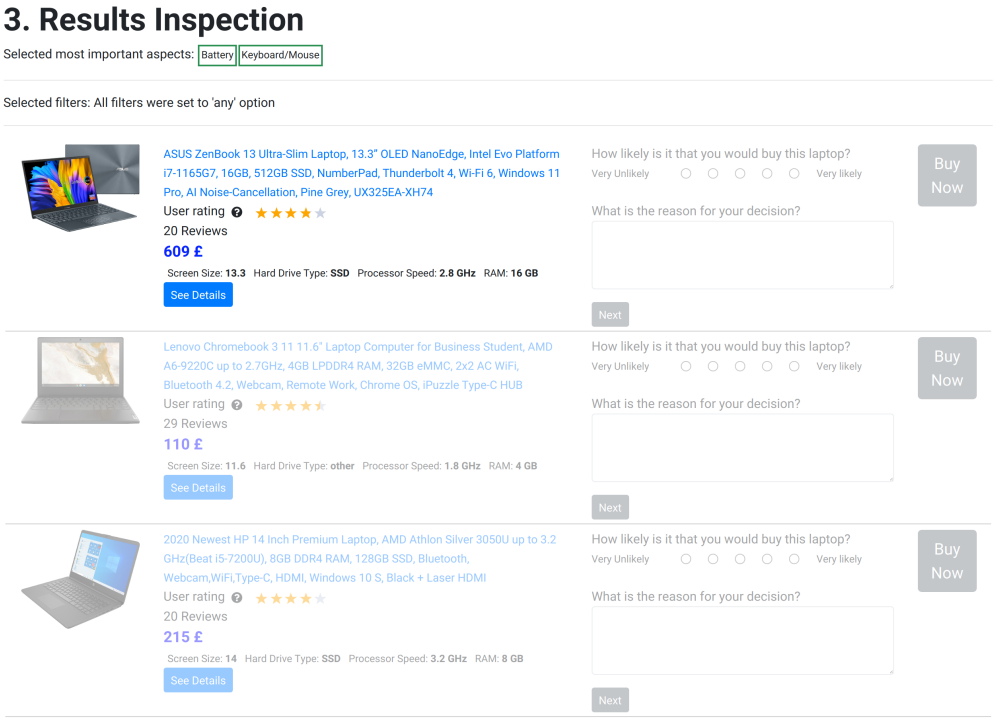


Figure 5.3: A screenshot of the tool shows the results provided by different ranking methods. For each result, an evaluation form is attached.

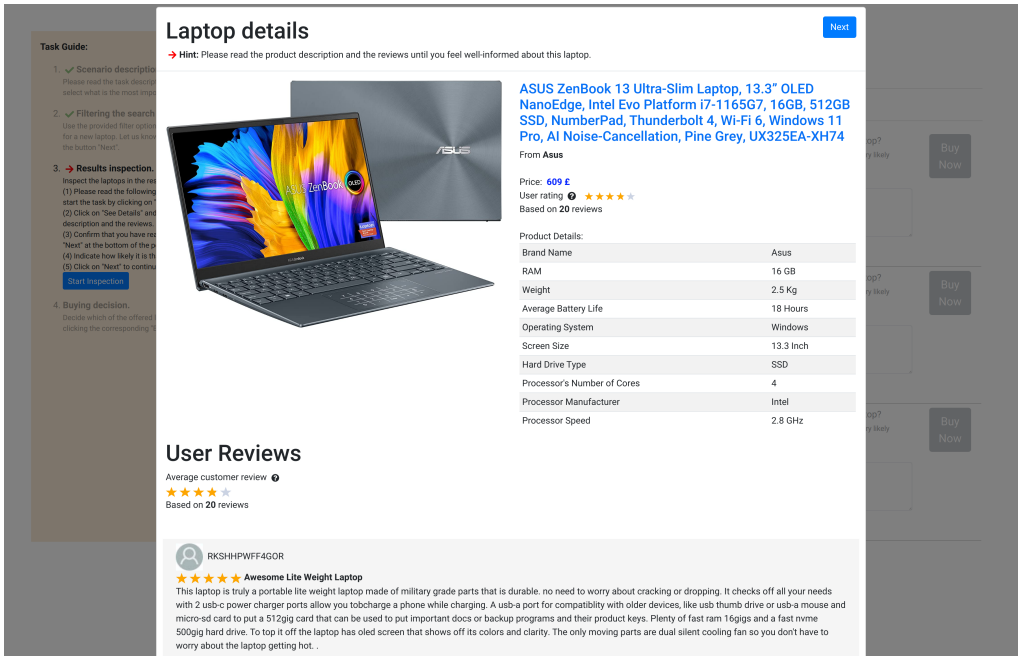


Figure 5.4: A screenshot of the tool shows the details and reviews of the clicked laptop.

5.4.2 Measures

We examined three performance indicators to assess how well the ranking approaches performed for users. (1) Buying Likelihood: Rate each laptop by answering the question "How likely are you to buy this laptop?" On a 5-point likert scale, 1 indicates "very unlikely" and 5 indicates "very likely". (2) Inspection time: A rating of how long it takes to finish the inspection operation of each laptop. This is calculated as the time difference between the moment the user opened and closed the details-reviews page. (3) Laptop Buy decision: the user's final decision in answer to the question, "Which of the given laptops would you buy?". For these three measures, we utilised ANOVA with $alpha = 0.05$ to test whether there were any significant differences between the methods; post-hoc tests with Bonferroni correction were used for between-groups multiple comparisons. We also incorporated a qualitative measure to have a better understanding of how participants make decisions; this is (4) Reason of decision: An open text answer to the question "What is the reason for your decision?" To analyse the data, we employed qualitative coding and grouping approaches for detecting topics and sentiments mentioned in the answers [Wagner et al. 2014; Sun et al. 2019].

5.4.3 Participants

We recruited 200 Prolific users to participate in our user study. We employed pre-screening filters to eliminate the influence of cultural or economic considerations (English native speakers, UK residents, no literacy difficulties). The age distribution (M = 32.1 years, SD = 11.7 years) and gender distribution (100 female, 99 male) were both balanced among the 199 valid submissions (one was excluded because of an unfinished task). Participants indicated medium average domain knowledge of laptops (M = 3.6, SD = 1.1) on a scale of one to five.

5.4.4 Data Collection

We collected a dataset of 2990 laptops from Amazon in April 2020. This includes product information such as technical details, pricing, and user-generated data such as star ratings and reviews. We obtained 30041 reviews for these laptops.

The sentiment and aspects of laptops mentioned in the reviews were extracted. We followed a similar approach as used in the hotel domain for aspect extraction. This approach involved initialising a set of required aspects using seed keywords and using a trained word embedding model to generate a comprehensive list of relevant aspect terms. Some aspects (i.e. Screen/Display, Keyboard/Mouse, and Sound/Audio) had overlapping terms, thus they were treated as single categories. A complete aspect terms list is presented in appendix A.6. To classify the sentiment expressed towards these aspects in the reviews, we utilised a tool [Sun et al. 2019] that was trained on a human annotated dataset in the laptop domain [Wagner et al. 2014].

In the prepossessing, we removed reviews that were not written in English or did not contain enough content (10-250 characters). For laptops, we eliminated items with reviews that did not mention at least one of the studied aspects. Furthermore, to enable a reliable comparison with a star rating based baseline, we only kept products with enough reviews (at least 20) and calculated the average rating score for each laptop. Also, we limited the maximum number of reviews per laptop to 50 in order to make the process of reading or browsing the reviews more reasonable for the participants.

The final dataset contains 319 laptops and 10076 reviews after filtering. Table 5.1 shows statistical information about the dataset.

		April-2020			April-2022		
		N	M	SD	N	M	SD
Laptops		319			357		
Reviews		10076	31.5	8.6	22642	63.4	30.1
Aspects in reviews		15937	1.6	1.2	35687	1.6	1.3
Screen/Display	Positive	1874	0.19	0.39	4298	0.1	0.3
	Negative	1676	0.17	0.37	4237	0.19	0.39
	Neutral	834	0.08	0.28	2204	0.1	0.3
Storage	Positive	793	0.08	0.27	1515	0.07	0.25
	Negative	980	0.1	0.3	1483	0.07	0.25
	Neutral	859	0.09	0.28	1705	0.08	0.26
Keyboard/Mouse	Positive	917	0.09	0.29	2116	0.09	0.29
	Negative	1163	0.12	0.32	2633	0.12	0.32
	Neutral	780	0.08	0.27	1940	0.09	0.28
Price	Positive	2761	0.27	0.45	6807	0.3	0.46
	Negative	2183	0.22	0.41	4932	0.22	0.41
	Neutral	731	0.07	0.26	2024	0.09	0.29
Battery	Positive	895	0.09	0.28	2502	0.11	0.31
	Negative	1401	0.14	0.35	3335	0.15	0.35
	Neutral	431	0.04	0.2	1346	0.06	0.24
Sound/Audio	Positive	343	0.03	0.18	958	0.04	0.2
	Negative	552	0.05	0.23	1357	0.06	0.24
	Neutral	38	0.004	0.06	245	0.01	0.1

Table 5.1: Sample size (N), mean(M) and standard deviation(SD) of the Amazon datasets. April-2020 was used in the first study and April-2022 in the second study.

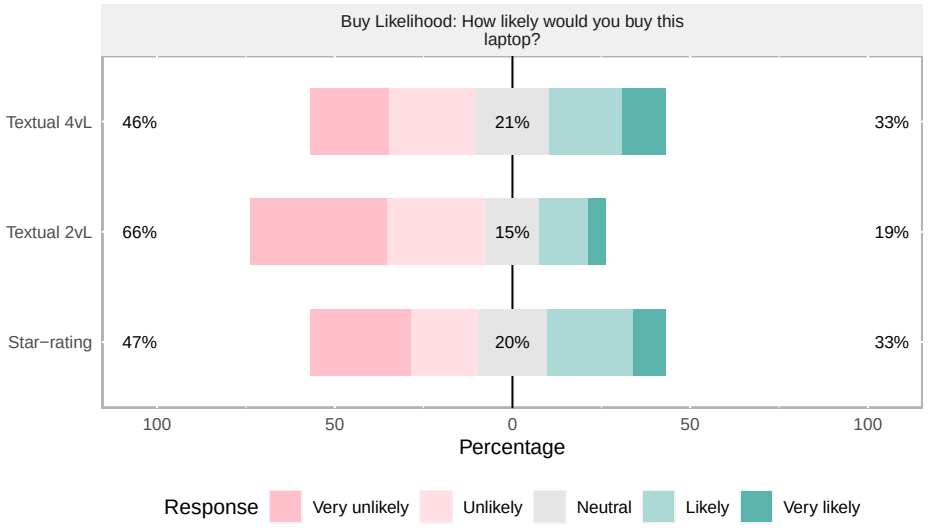
5.4.5 Results

To study how well textual based ranking methods perform for users compared to a star rating based method (RQ1), we first looked at how well ranking methods influence users' buying likelihood. The distribution of the buying likelihood, which was a response to the question following each of the given laptops: "How likely would you buy this laptop?" is shown alongside with other statistical indicators in figure 5.5. As seen in the figure, star rating baseline ($M = 2.67, SD = 1.35$), textual-based 2vL ($M = 2.19, SD = 1.22$) and textual-based 4vL ($M = 2.76, SD = 1.33$) methods have clear differences in their buying likelihood mean values. The difference is significant in favour of two cases, star rating ($p = 0.003$) and 4vL ($p = 0.003$), both when compared to 2vL. The results, however, do not show a significant difference between the star rating and the textual 4vL methods. Regarding hypotheses on the buy likelihood variable, hypothesis H1 is rejected as not both cases (4vL and 2vL) were significantly outperforming the star rating baseline method. However, by looking into the results for 2vL and 4vL methods, we accept hypothesis H1a.

Participants had to make a decision about which of the three given laptops to buy. From all participants, 86 (43%) decided on laptops that were given by the 4vL based method, whereas 72 (36%) selected to buy the items brought by the star rating method, and finally 41 participants (21%) bought the laptops based on the textual 2vL based method. This distribution differs significantly (χ^2 with two degrees of freedom $= 17.594$, $p < 0.001$) from an expected $1/3$ distribution of ranking methods that would draw an equal selection of the methods. As the textual methods both do not outperform the star rating baseline, we reject the hypothesis H2. On the other hand, the difference between 4vL and 2vL is clear and therefore we accept the hypothesis H2a.

We also looked at inspection time as a further measure of how well the ranking methods performed for users. Although participants indicated that the laptop's features and reviews with the textual 4vL ranking require less time to be inspected (as shown in figure 5.6), the significance test results shows that none of the differences between methods is significant. Therefore, the hypotheses H3 and H3a are both rejected here.

In a deeper analysis, we looked at the participants' decisions' clarifications that were given in an open answer. We used a method for terms clustering and sentiment analysis to identify the most frequently discussed topics in the responses. The most often found topics are *price*, *reviews*, *specs*, *screen*,



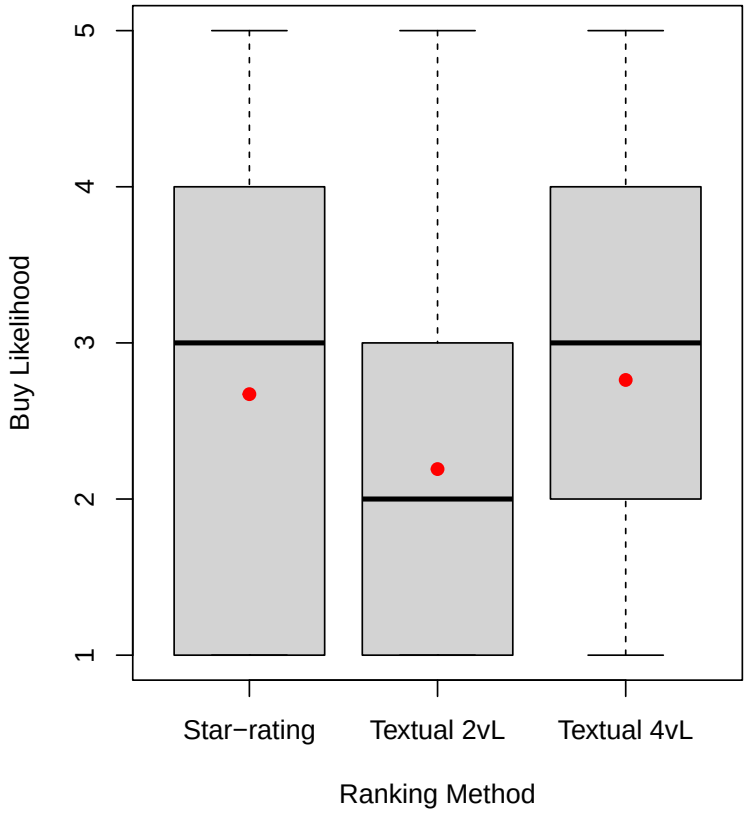


Figure 5.5: Buy likelihood distribution among the three ranking methods.

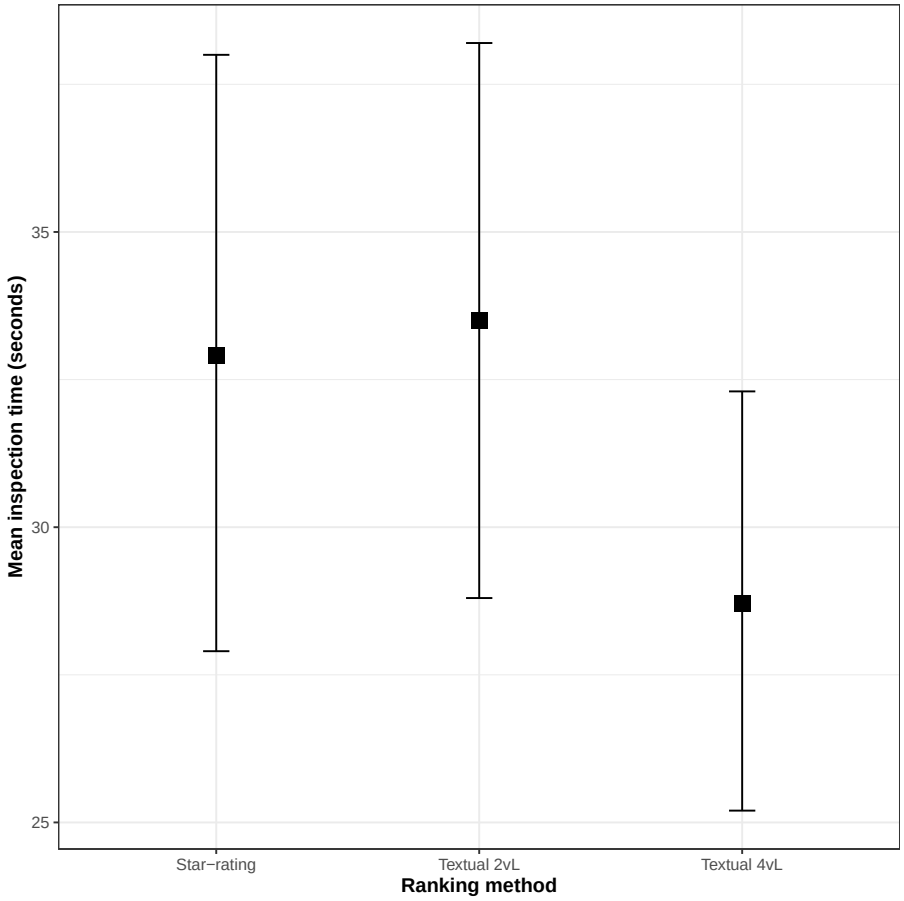


Figure 5.6: Required inspection time of laptops grouped by the ranking method.

	Star rating		Textual 2vL		Textual 4vL		ANOVA		Post Hoc Tests
Topic	M	SD	M	SD	M	SD	F	p value	variable levels (p-value)
price	-0.15	0.779	-0.19	0.656	0.12	0.721	10.703	**<.001**	- 4vL, star rating (<.001) - 4vL, 2vL (<.001)
reviews	0.11	0.541	-0.4	0.595	-0.05	0.659	38.184	**<.001**	- 4vl, star rating (0.023) - 4vl, 2vl (<0.001)
specifications	-0.01	0.349	0.05	0.387	0.08	0.348	2.946	0.053	-
screen	0.02	0.356	-0.06	0.306	-0.01	0.349	2.232	0.108	-
performance	0.02	0.31	-0.04	0.316	-0.05	0.345	2.354	0.096	-
storage	0	0.285	-0.04	0.361	-0.04	0.401	0.868	0.42	-
brand	0.01	0.285	-0.03	0.212	-0.01	0.174	1.194	0.304	-
battery	-0.02	0.2	-0.02	0.213	-0.02	0.276	0.031	0.969	-

Table 5.2: Statistical description and significance testing results of the influence of ranking methods on the most mentioned topics in users' decision reasons. Means state the average sentiment of topic on a scale from -1 (Negative) to +1 (Positive).

performance, *storage*, *brand* and *battery*; figure 5.7 presents a distribution of these mentions. Using the sentiment analyser, we assign one of three ordinal labels to each of these topics: *positive(+1)*, *negative(-1)*, and "*neutral or not discussed*"*(0)*. We investigated if the ranking method had any effect on these topics. Results presented in table 5.2 demonstrate that the topics *price* and *reviews* were shown to be significantly influenced by the ranking method. By looking at the post-hoc testing results, we can observe that the subject *price* was discussed positively ($M = +0.12$) in the open text answers when the item was delivered via the 4vL approach. Whereas subject reviews seem to have a negative influence ($M = -0.15$) on users' decisions. The possible reason for this is that products ranked with textual methods are more often to include negative reviews about the required aspects compared with the star ranking method which does not consider the text at all in the ranking. It is also possible that the star rating (number of stars) is not properly reflected or represented by the textual review content (e.g. a review of 5 stars, but still contains some negative information about the aspects).

We have also investigated whether a specific aspect or a group of aspects in the user queries has an impact on the measures of our user study. Using only a subset of the collected data where an aspect or aspects are always required, the tests are applied to determine if there is a significant difference between the mean values of the measures. Before doing that, we measured what aspects were inquired most in the user queries (As shown in figure 5.8). In order to perform this analysis, we are testing the differences among user groups (i.e. users of each of the 3 ranking methods) in two cases. Firstly, for each of the aspects individually, and secondly, for a group of aspects that were mostly

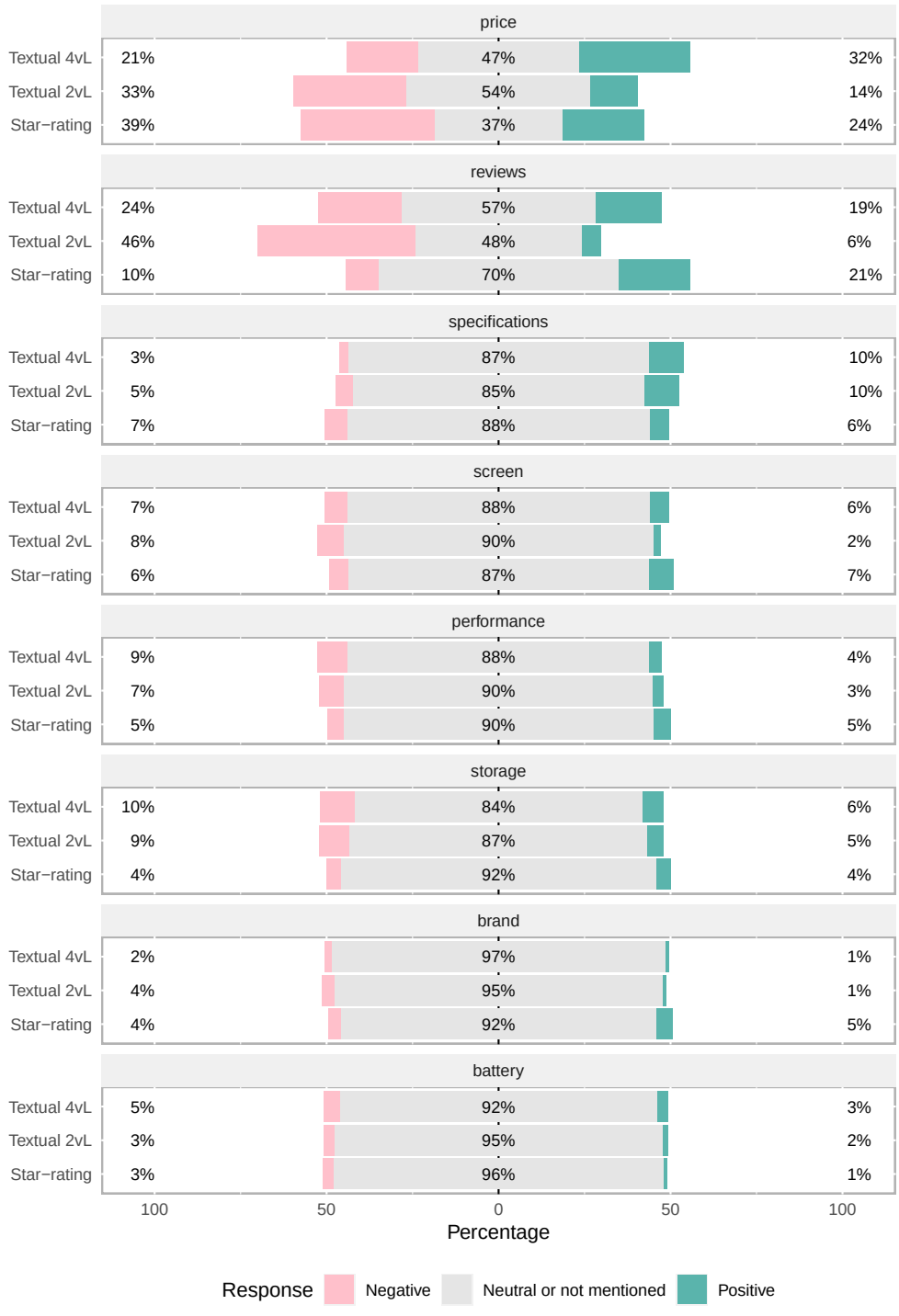


Figure 5.7: Most mentioned topics in the user's decision reasons grouped by ranking methods.

$Query_{ID}$	Occurrence (%)	Query Aspects
22	14.14	Price, Storage, Battery, Screen/Display
17	11.62	Price, Storage, Battery
5	10.10	Price, Storage, Screen/Display
32	8.59	Price, Storage, Battery, Screen/Display, Sound/Audio, Keyboard/Mouse
25	6.57	Price, Storage, Battery, Screen/Display, Sound/Audio
2	5.56	Price, Storage
29	4.55	Price, Storage, Battery, Screen/Display, Keyboard/Mouse
20	4.04	Price, Battery, Screen/Display
0	4.04	Price
19	3.54	Price, Storage, Battery, Sound/Audio

Table 5.3: Top inquired group of aspects in search sessions in user study I.

inquired in the search sessions. We have selected the top 10 most frequent queries (which are shown in table 5.3). Testing results of these two cases are presented in table 5.4. For shorten, results were only included if there are significant differences among the groups. Also only if the subset size of the query/aspect is at least 30 samples. The results shown in this table are along the same lines as the results presented earlier which show in general a slight improvement when 4vL ranking was applied compared with the star rating method. However, we see here some detailed differences depending on particular aspects. For example, it seems that users have benefitted the most from the 4vL method with aspects (Battery, Storage, Price), as shown in the results of these individual aspects and the results of queries ($Query_{ID}$: 22, $Query_{ID}$: 17). On the other hand, the star rating method looks more beneficial with aspects (Keyboard/Mouse, Screen/Display, Sound/Audio). However, 4vL seems to have more influence on the major measures of our study (i.e. buy likelihood and purchase decision) compared with the star rating method which showed less influence on these measures and more on the minor measures (i.e. topics "reviews", "performance" and "screen" in users' decision clarifications).

Additional testing has been conducted to ensure that user decisions and actions were not influenced by undesired factors. This consists of 1) the gender and age of the participants. 2) the location of the laptop on the results page (i.e. first, second or third). The results indicate that these variables do not significantly influence the observed user behaviours.

To conclude, in this experiment we studied the differences between the textual based ranking methods and the star ratings in terms of impact on the user's support on buying decisions. We saw that the textual based 4vL method has value to support users in product ranking tasks, as it allows a better representation of product features, especially for the aspect-oriented search, and

Aspect/Query	Subset size (laptops)	Affected variables	Mean & SD star-rating (S)	2vL	4vL	Significant difference (Post-hoc
Battery*	399	buy likelihood	2.61 (1.33)	2.28 (1.23)	**2.87 (1.36)**	2vL, 4vL (p = 0.001)
		purchased(%)	0.32 (0.46)	0.22 (0.41)	**0.47 (0.5)**	S, 4vL (p = 0.036), 2vL, 4vL (p = 0.0)
		price	-0.17 (0.77)	-0.17 (0.65)	**0.2 (0.72)**	S, 4vL (p = 0.0), 2vL, 4vL (p = 0.0)
		reviews	**0.08 (0.51)**	-0.38 (0.6)	0.0 (0.65)	S, 2vL (p = 0.0), 2vL, 4vL (p = 0.0)
Keyboard/Mouse*	123	purchased(%)	**0.51 (0.5)**	0.2 (0.4)	0.29 (0.45)	S, 2vL (p = 0.007)
		reviews	**0.22 (0.56)**	-0.56 (0.59)	-0.2 (0.71)	S, 2vL (p = 0.0), S, 4vL (p = 0.014), 2vL, 4vL (p = 0.041)
		performance	**0.07 (0.26)**	0.02 (0.27)	-0.07 (0.26)	S, 4vL (p = 0.042)
Screen/Display*	399	buy likelihood	**2.74 (1.35)**	2.19 (1.25)	2.69 (1.32)	S, 2vL (p = 0.002), 2vL, 4vL (p = 0.005)
		purchased(%)	0.38 (0.48)	0.2 (0.4)	**0.43 (0.49)**	S, 2vL (p = 0.003), 2vL, 4vL (p = 0.0)
		reviews	**0.12 (0.52)**	-0.39 (0.59)	-0.05 (0.63)	S, 2vL (p = 0.0), S, 4vL (p = 0.047), 2vL, 4vL (p = 0.0)
Sound/Audio*	168	purchased(%)	0.34 (0.47)	0.18 (0.38)	**0.48 (0.5)**	2vL, 4vL (p = 0.002)
		reviews	**0.07 (0.49)**	-0.34 (0.61)	-0.02 (0.67)	S, 2vL (p = 0.001), 2vL, 4vL (p = 0.028)
		screen	**0.12 (0.33)**	-0.05 (0.23)	-0.02 (0.3)	S, 2vL (p = 0.004)
Storage*	441	buy likelihood	2.63 (1.36)	2.22 (1.19)	**2.82 (1.31)**	S, 2vL (p = 0.018), 2vL, 4vL (p = 0.0)
		purchased(%)	0.33 (0.47)	0.2 (0.4)	**0.46 (0.5)**	S, 2vL (p = 0.037), 2vL, 4vL (p = 0.0)
		price	-0.24 (0.78)	-0.21 (0.66)	**0.16 (0.7)**	S, 4vL (p = 0.0), 2vL, 4vL (p = 0.0)
		reviews	**0.07 (0.53)**	-0.37 (0.59)	-0.02 (0.64)	S, 2vL (p = 0.0), 2vL, 4vL (p = 0.0)
		specifications	-0.03 (0.37)	0.04 (0.4)	**0.1 (0.37)**	S, 4vL (p = 0.016)
Price*	537	buy likelihood	2.63 (1.34)	2.14 (1.18)	**2.83 (1.35)**	S, 2vL (p = 0.001), 2vL, 4vL (p = 0.0)
		purchased(%)	0.35 (0.48)	0.19 (0.39)	**0.46 (0.5)**	S, 2vL (p = 0.002), 2vL, 4vL (p = 0.0)
		price	-0.17 (0.79)	-0.2 (0.65)	**0.14 (0.73)**	S, 4vL (p = 0.001), 2vL, 4vL (p = 0.0)
		reviews	**0.11 (0.53)**	-0.42 (0.58)	-0.02 (0.65)	S, 2vL (p = 0.0), 2vL, 4vL (p = 0.0)
		specifications	-0.02 (0.34)	0.05 (0.37)	**0.09 (0.36)**	S, 4vL (p = 0.013)
		screen	**0.02 (0.35)**	-0.06 (0.3)	-0.01 (0.36)	S, 2vL (p = 0.048)
$Query_{ID}$: 22	84	buy likelihood	2.75 (1.18)	2.11 (0.98)	**3.0 (1.25)**	2vL, 4vL (p = 0.015)
		purchased(%)	0.29 (0.45)	0.11 (0.31)	**0.61 (0.49)**	S, 4vL (p = 0.045), 2vL, 4vL (p = 0.0)
		price	-0.07 (0.75)	-0.21 (0.72)	**0.29 (0.65)**	2vL, 4vL (p = 0.03)
		reviews	**0.11 (0.49)**	-0.39 (0.49)	0.04 (0.68)	S, 2vL (p = 0.001), 2vL, 4vL (p = 0.031)
$Query_{ID}$: 17	69	price	-0.3 (0.86)	-0.17 (0.7)	**0.52 (0.65)**	S, 4vL (p = 0.002), 2vL, 4vL (p = 0.004)
		reviews	0.04 (0.46)	-0.3 (0.62)	**0.22 (0.51)**	2vL, 4vL (p = 0.012)
$Query_{ID}$: 32	51	reviews	**0.12 (0.58)**	-0.47 (0.5)	-0.12 (0.76)	S, 2vL (p = 0.013)
$Query_{ID}$: 25	39	purchased(%)	0.31 (0.46)	0.08 (0.27)	**0.62 (0.49)**	2vL, 4vL (p = 0.008)
$Query_{ID}$: 2	33	reviews	**0.27 (0.62)**	-0.73 (0.45)	-0.18 (0.83)	S, 2vL (p = 0.001)

Table 5.4: Individual aspect and query influence on the measured variables. $*$ indicates that aspect was always selected in the subset, but possibly alongside with one or more aspects.

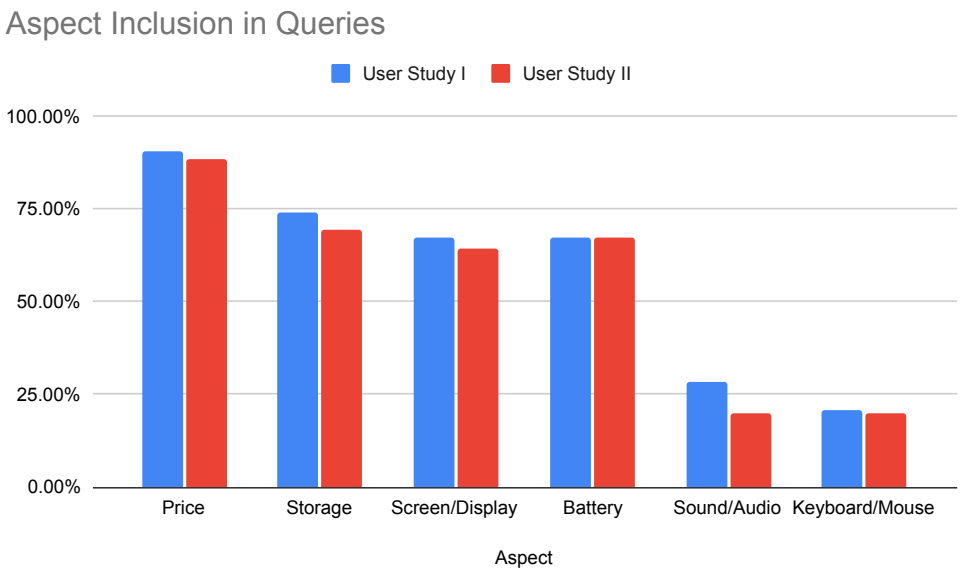


Figure 5.8: Aspect inclusion in user queries.

therefore can encourage users to make better buying decisions. The outperformance of the 4vL over the 2vL method seems to be clear on users' buying likelihood and buying decisions, and ignorable on the required time for inspecting products. However, although 4vL method added clear improvements to the search process, it does not seem to have a generalised significant impact when compared to a star rating method.

Our study is limited by the fact that users were asked to evaluate only the top ranked product for each ranking method. That might raise doubts about the overall relevance and utility of the ranking methods for users who not only need to make use of the top ranked product but also to inspect other ranked products before making a final choice. Another limitation of the study is that we did not measure whether users actually bought the product after reading the product details and reviews before making their choice. For these two problems, we extended the scope of the experiment to the second study.

5.5 USER STUDY II: The Relationship Between Aspects Highlighting in Reviews and the Ranking Method

The textual 4vL method has shown the potential to provide comparable or better rankings compared to the textual 2vL method and the star rating method. In the first user study 5.4, we concentrated on developing rankings based on the review content. However, like with other search tasks, users may be distracted by a variety of factors that affect their evaluation of the ranking's quality. Therefore, we designed a second user study that focused on an action that causes users to concentrate on the review content. In order to do this, we highlight the relevant aspects in reviews when we display the laptop details and reviews. We address the following research question in this user study:

RQ2 What effect does the highlighting function have on a user's buying decisions and search behaviour? To answer this question, we investigate the impact of the highlighting function in two sub-questions:

- **RQ2.1** Does highlighting the reviews support the user generally with buying decisions and search behaviour?
- **RQ2.2** If reviews are highlighted by aspects of user's interest, are users more likely to be attracted to results ranked by the textual 4vL method than to results ranked by the star rating method?

In this study, results attraction is stated particularly by increasing the buy likelihood for supporting the buying decision and decreasing the time needed to investigate product features for supporting the search behaviour.

5.5.1 Study Design

We used a study design and tool similar to the previous user study, but with several differences in the setup and the tool. In a between-subjects experiment, we developed and compared four shops with four different settings. Each setting is a combination of two shop properties: 1) the ranking method and 2) the feature of highlighting the aspect-relevant parts of reviews. The settings are listed in table 5.5.

	Shop properties	
Setting name	Ranking method	Highlighting support
Sn	Star rating	No
SH	Star rating	Yes
4n	4vL	No
4H	4vL	Yes

Table 5.5: Shop settings identification based on two shop properties.

We are concentrating on the goal of determining if the highlighting functionality supports the user in buying decisions and searching behaviour by drawing the user's attention to important portions of the reviews (**RQ2**). To do so, we compare two groups of settings.

In the first group (**Group1**), we address the research question (**RQ2.1**). Does highlighting the reviews support the user in general with buying decisions and search behaviour? To answer this question, we consider the following hypotheses:

- **H1**: Regardless of the ranking method used to rank the products, highlighting functionality gives users more confidence in their buying decisions. It increases the buy-likelihood of products, according to the study's terminology.
- **H2**: Highlighting functionality reduces the time required to investigate products via reviews by drawing users' attention just to the relevant parts of reviews and ignoring the rest.

We are also interested to see whether highlighting affects the position of the product that is chosen to be purchased in the results list. We hypothesise as follows:

- **H3**: The position of the decided-to-buy product on the results list will be closer to the top of the results list because users would be more aware of the product's aspects of interest when they are highlighted in the reviews. We assume that the products at the top of the results list are more relevant than the ones at the bottom.

The three hypotheses H1, H2 and H3 all focus only on the highlighting functionality, regardless of which ranking method is used. We are only interested to see the impact of highlighting.

In the second group of settings (**Group2**), we investigate the research ques-

tion (**RQ2.2**) i.e. when highlighting is enabled, does textual 4vL method offer users more efficient rankings compared with the star rating method? For this question, we hypothesise the following:

- **H4**: Users who use the textual 4vL ranking method have higher confidence in their buying decisions than users who use the star rating ranking method. This is based on the basic highlighting assumption, which states that the highlighted phrases will draw the user's attention.
- **H5**: When using the textual 4vL ranking method, users require less time to investigate the products compared with the case when using the star rating ranking method. We remember from the first user study that, in contrast to the star rating method, which ranks the products based on average user ratings, the textual 4vL ranking method places the most relevant products at the top of the results list based on aspect related passages in the reviews.
- **H6**: When 4vL ranking is used, the position of the decided-to-buy product on the results list will be closer to the top of the results list. This is in comparison to the case when star rating method ranking is used.

For these three hypotheses H4, H5, and H6, highlighting will be always applied, but the ranking method will be changed, as we are only interested to see the differences between 4vL and star rating methods in an enabled-highlighting environment.

5.5.2 Task and Procedure

The task description, procedure, and questions are the same as in the first user study. However, we have included a few additional questions concerning the highlighting functionality. For the tool, in addition to the basic functionality of highlighting the reviews, we have made a change on the number of available items per search session in order to cover the limitations of the first study. In this study, users will get the top five items of each ranking method. Furthermore, we differentiate this time between an item's star rating and the rating of the required aspects. Figure 5.9 shows the details-reviews page after applying the highlighting.

The process of highlighting involves these steps: Reviews are first preprocessed to correct spelling and punctuation errors, but they are later shown to users in their original form. Preprocessing is only required to help the

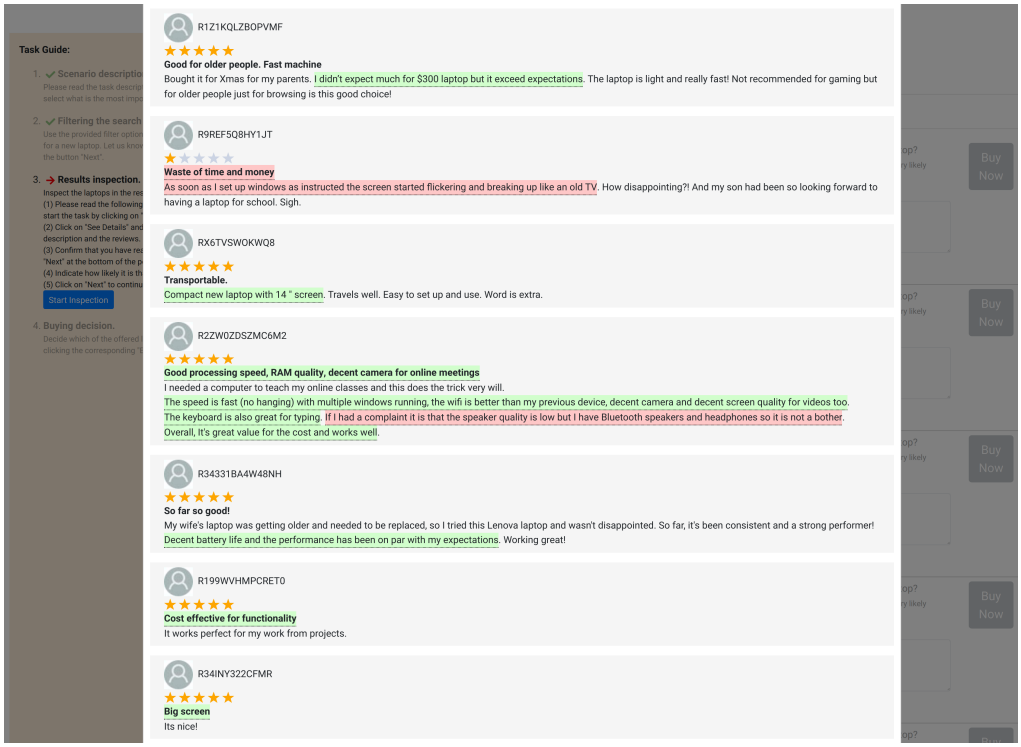


Figure 5.9: A screenshot of the tool shows the reviews highlighted according to aspects Price, Screen/Display, Storage, Battery, Keyboard/Mouse and Sound/Audio and their polarities.

sentence tokenizer perform better. The fastPunct[1] tool is utilised for this pre-processing. Then, the reviews are divided into sentences that each describe one or more aspects. The sentences are then highlighted based on the aspects and the sentiment expressed towards them using the following colour codes: green for positive sentiment, red for negative sentiment, and gray for neutral sentiment. If a sentence mentions one or multiple aspects with mixed sentiments, the colour coding is adjusted as follows: green for positive and neutral sentiments, red for negative and neutral sentiments, and orange for positive and negative sentiments. An example of this can be seen in the sentence "The laptop has a fantastic display but a small screen size" which would be highlighted in orange due to the mixed positive and negative sentiments towards the aspect of Screen/Display.

[1] `https://pypi.org/project/fastpunct/` (last accessed 12.12.2022)

5.5.3 Measures

We concentrated on the same quantitative measurements stated in section 5.4.2 in this study. We performed the one-way ANOVA tests again for the between-subject comparisons. As five items are offered in this study, we choose a measure for the quality of the ranking method based on the position of the purchased product on the results list ($Purchased_{pos}$). All significance tests are assessed at a significance threshold of $\alpha = 0.05$.

5.5.4 Participants

For this study, we repeated the procedures for recruiting and selecting participants as in section 5.4.3. We stopped collecting data after we got around 90 participants in each condition ($N = 87$ in setting *Sn*, $N = 93$ in setting *SH*, $N = 92$ in setting *4n*, and $N = 88$ in setting *4H*), $N = 360$ valid replies in total. The sample was roughly gender balanced (185 females, 175 males). The average age was ($M = 37.5, SD = 13.8$). The average domain knowledge about laptops was ($M = 3.5, SD = 1$) on a 5-point scale.

5.5.5 Dataset

We updated the dataset for this study in order to reduce the possible impact of the product's modernity on user decisions. For comparability to the first dataset, we obtained the new set from Amazon. The same filter rules were applied to products and reviews. After filtering, the final dataset contains 357 laptops and 22642 reviews. Additional information about the dataset is shown in table 5.1.

5.5.6 Results

In order to make it easier to understand and interpret the results, we present the testing results in two sections. Firstly, we show the results of testing whether there are significant differences between setting Sn, SH, 4n and 4H individually (refer to table 5.5 for setting names and details). In the second section, we see the results in the context of the setting groups (Group1 and Group2).

Setting Comparison (Individually)

For the first measure, buy likelihood (figure 5.10), the tests revealed that there was a statistically significant difference in attitudes between at least two groups of attitudes ($p < 0.001$). Post-hoc tests with Bonferroni correction for multiple comparisons revealed that there was a significant difference in the buy likelihood mean value between setting *SH* on one hand and *Sn* ($p = 0.012$), *4n* ($p < 0.001$), and *4H* ($p = 0.016$) on the other hand. However, no significant differences were found between groups *Sn*, *4n*, and *4H*. The mean values of this measure show that users are significantly less likely to buy products when setting SH is applied, i.e. when highlighting is enabled and results are ranked using the star rating method.

For the second measure, inspection time (figure 5.11), tests showed that there was a statistical difference between at least two groups of settings ($p < 0.001$). We found that the differences existed in setting *4n* compared to settings *Sn* ($p = 0.004$), *SH* ($p < 0.001$), and *4H* ($p < 0.001$). There were no significant differences between the remaining settings. In other terms, when users use a system that does not enable review highlighting and ranks the products using the 4vL method (4n), it takes less time to inspect the products compared with any other setting (4H, SH, Sn).

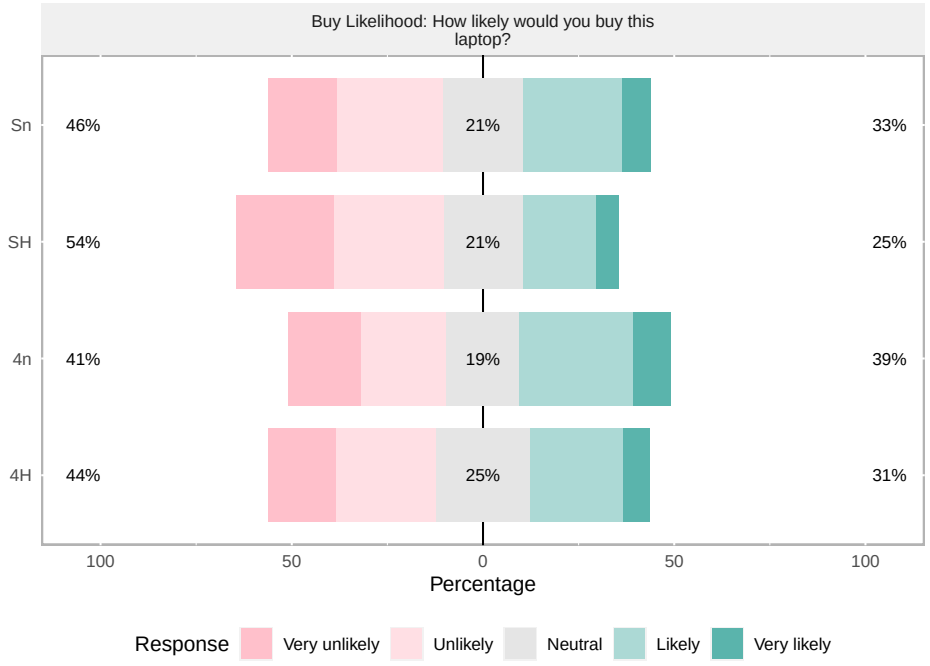
For the third measure, the position of the purchased product on the results list ($Purchased_{pos}$) (figure 5.12), the tests show no significant differences between the groups ($p = 0.082$).

Setting Comparison (Within Groups)

In more detailed results, we conducted the tests for groups (Group1, Group2) mentioned above. For **Group1**, we test the significant differences between two combinations of settings based on the availability of highlighting:

- With and without highlighting when products are ranked using the star rating method (*Sn*, *SH*).
- With and without highlighting when products are ranked using the textual 4vL based method (*4n*, *4H*).

To accept or reject hypotheses H1,H2 and H3, these two combinations of settings should all show significant differences among the settings.



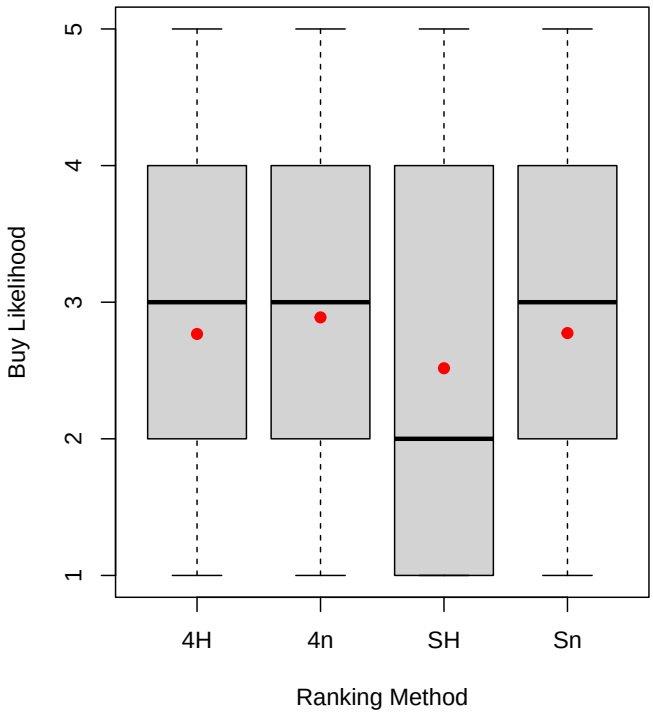


Figure 5.10: Buy likelihood distribution among the four shop settings.

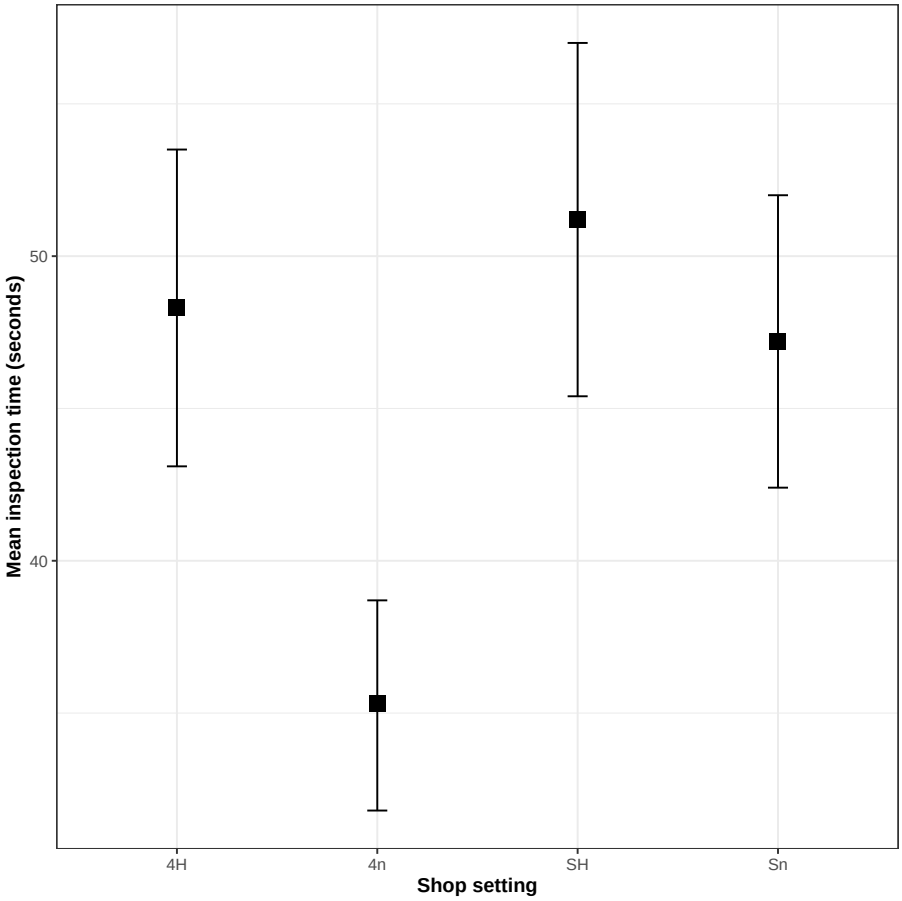


Figure 5.11: Required inspection time of product details and reviews for each of the shop settings.

Figure 5.12: Distribution of decided-to-buy products' positions on results list ($Purchased_{pos}$) among the four shop settings.

For **Group2**, we are interested to see the significant differences between the combinations of settings based on the used ranking method:

- Without highlighting, using two ranking methods i.e. star rating and 4vL (*Sn*, *4n*).
- With highlighting, using two different ranking methods (*SH*, *4H*).

Similarly, to accept or reject H4, H5, and H6, these two combinations of setting should show significant differences when settings are compared with each other.

The statistical description and significance results of the tested settings categorised according to the conditions indicated in groups *Group1* and *Group2* are shown in the tables 5.7. Also figures shown in 5.13 demonstrate the change on each of the measured variables when applying different shop setting.

To answer the research question (**RQ2.1**), we study the influence of highlighting in general using hypotheses (H1, H2 and H3). For (**H1**), when highlighting is applied, the buy likelihood decreases in both ranking systems. However, significance testing reveals that this decrease is significant in only one

Setting	Avg. # of Pos. phrases per review	Avg. # of Neg. phrases per review	Avg. # of Neut. phrases per review
Sn	0.62	0.45	0.27
SH	0.58	0.43	0.26
4n	0.61	0.38	0.22
4H	0.54	0.34	0.2

Table 5.6: Average number of positive, negative and neutral phrases per review within the four tested settings.

of two cases (in star rating ranking case). As a result, we cannot conclude that highlighting reviews will always increase or decrease the likelihood of buying a product. For (**H2**), we discovered that the highlighting feature increases the required average time to investigate the items in both ranking systems; however, significance testing says that the inspection time is significantly reduced only when textual 4vL ranking is applied. As a result, we also cannot conclude that highlighting reviews would always decrease or increase the time required to inspect the products. When highlighting is enabled for (**H3**), the position of the decided-to-buy product moves closer to the top of the results list. However, the difference here is insignificant, therefore the hypothesis is rejected.

Nevertheless, the effect of highlighting on the study measures can also be influenced by the number of highlighted phrases for the selected aspects. Table 5.6 presents the average number of positive, negative, and neutral phrases per review. It can be observed that the changes in the 4vL based settings (4n, 4H) are greater compared to the star rating based settings (Sn, SH). This indicates that the impact of highlighting was greater on the participants in the latter settings as they were exposed to a greater number of highlighted phrases, thereby influencing their behaviour towards the study measures (it should be noted that the phrases were only highlighted in the SH and 4H settings).

For the research question (**RQ2.2**), the investigation focuses on the impact of the highlighting functionality especially when it is combined with different ranking methods. For **H4**, the buy likelihood has increased in general when we use the 4vL ranking. The increase is significant especially when the highlighting functionality is enabled. The increase is not significant when highlighting is not enabled, which is also consistent with the results we got

	no Highlighting (n)						with Highlighting (H)					
	buy likelihood		inspection time		$Purchased_{pos}$		buy likelihood		inspection time		$Purchased_{pos}$	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Star rating (S)	2.77	1.22	47.2	50.21	2.64	1.54	2.52	1.22	51.2	62.57	2.15	1.37
Textual 4vL (4)	2.89	1.28	35.3	36.97	2.63	1.43	2.77	1.2	48.3	54.92	2.59	1.57

(a)

	buy likelihood			inspection time			$Purchased_{pos}$		
	Sn	*SH*	*4H*	*Sn*	*SH*	*4H*	*Sn*	*SH*	*4H*
Sn		**0.012**			1.0			.191	
4n	1.0		.873	**.004**		**.001**	1.0		1.0
4H		**.016**			1.0			.296	

(b)

Table 5.7: Statistical description (a) and post-hoc tests (b) of the tested settings on three measures. Columns in (a) allow for the comparison between settings based on the availability of the review-highlighting feature (Group1). Rows in (a) represent the comparison between the settings on the applied ranking method (Group2). In (b) gray-coloured values refer to (Group1), while purple-coloured refer to (Group2)

in the first user study. We can conclude for **H4**, that 4vL ranking method with an enabled highlighting functionality increases the buy likelihood of products. For (**H5**), the 4vL ranking has reduced the time needed for product inspection, however, the decrease is only significant when the highlighting is disabled. For (**H6**), we found that the position of the decided-to-buy product is moving away from the results list's top in the case of textual 4vL ranking. However, this change is also not significant.

Similar to the first user study, we analysed the answers of the open text questions in order to see whether any of what users commonly mention is being influenced by the setting we use for the ranking and the highlighting (mentioning distribution is presented in figure 5.14). Significance testing results in table 5.8 show that only topic "reviews" was changed between *4n* and *4H* settings. In this scenario, the participants had more negative opinions about *4H*. This is understandable since the negative aspects of the reviews were highlighted in that case, whereas no highlighting was present in *4n*.

For analysing the effect of individual aspects and queries on the study's measures, we found the most required aspects and queries. For individual aspects, participants of this user study included almost similarly the needed aspects as in the previous study (see figure 5.8). There are also many similarities between the top requested queries in this study and the previous one

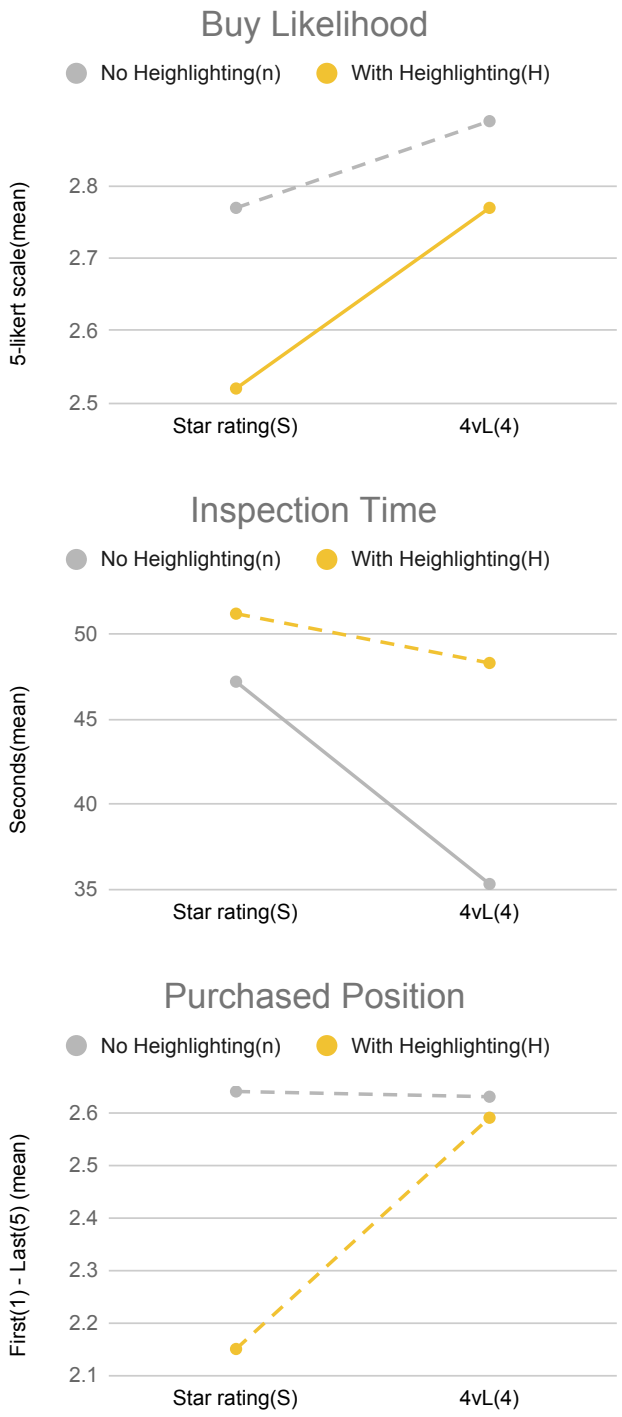


Figure 5.13: Change in means of the three measured variables when different highlighting and rankings settings are applied. Continues lines indicate for significant change, where dashed point to insignificant change.

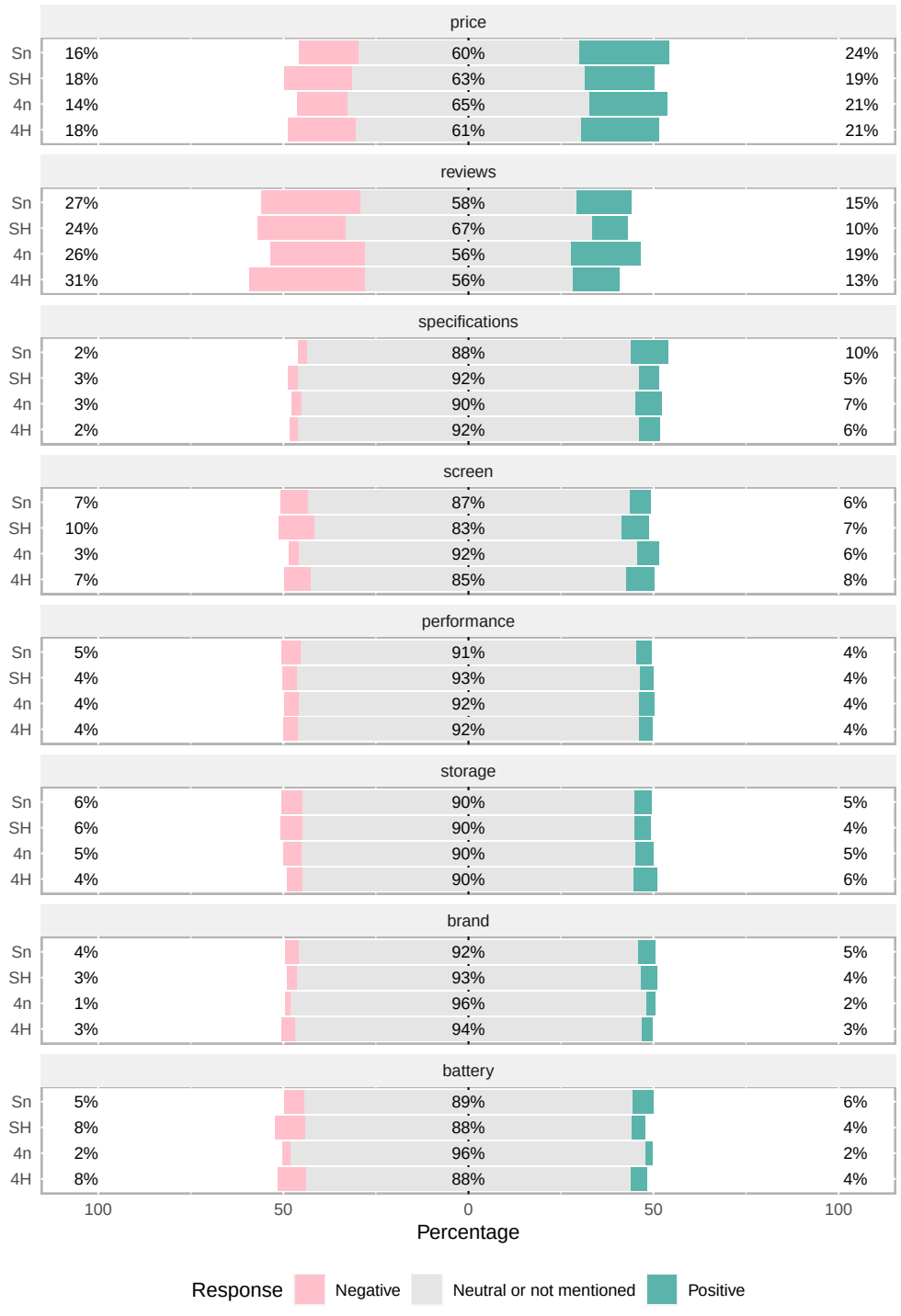


Figure 5.14: Most mentioned topics in the user's decision reasons grouped by ranking methods.

	Sn		*SH*		*4n*		*4H*		ANOVA		Post Hoc Tests
	M	SD	M	SD	M	SD	M	SD	F	p-value	Levels (p-value)
price	0.08	0.63	0	0.608	0.08	0.585	0.03	0.625	1.675	0.17	-
reviews	-0.12	0.635	-0.14	0.561	-0.07	0.662	-0.19	0.635	2.671	**0.046**	*4n*, *4H* (.033)
specifications	0.08	0.345	0.03	0.28	0.04	0.31	0.03	0.275	2.369	0.069	-
screen	-0.02	0.36	-0.02	0.411	0.03	0.289	0.01	0.383	2.111	0.097	-
performance	-0.01	0.303	0	0.27	0	0.283	0	0.277	0.241	0.868	-
storage	-0.01	0.322	-0.01	0.318	0	0.313	0.02	0.323	0.998	0.393	-
brand	0.01	0.287	0.02	0.265	0.01	0.194	-0.01	0.251	0.783	0.503	-
battery	0	0.333	-0.05	0.339	0	0.2	-0.03	0.347	2.43	0.064	-

Table 5.8: Statistical description and significance testing results of the influence of the tested settings on the most mentioned topics in users' decision reasons. Mean values state the average sentiment of a topic on a scale from -1 (Negative) to +1 (Positive).

$Query_{ID}$	Occurrence (%)	Query Aspects
29	15.38	Price, Storage, Battery, Screen/Display
23	13.39	Price, Storage, Battery
10	8.83	Price, Storage, Screen/Display
27	7.98	Price, Battery, Screen/Display
42	7.69	Price, Storage, Battery, Screen/Display, Sound/Audio, Keyboard/Mouse
0	5.98	Price
21	5.41	Price, Battery
28	3.42	Storage, Battery, Screen/Display
33	3.13	Price, Storage, Battery, Screen/Display, Sound/Audio
2	2.85	Price, Storage

Table 5.9: Top inquired groups of aspects in search sessions in user study II.

Aspect/Query	Subset size (laptops)	Affected variables	Mean & SD				Significant difference (Post-hoc)
			Sn	SH	4n	4H	
Battery*	1185	buy likelihood	2.77 (1.22)	2.51 (1.22)	**2.83 (1.29)**	2.66 (1.23)	Sn, SH (p = 0.023)
		inspection time	50.27 (47.54)	53.01 (50.02)	**38.02 (40.88)**	50.86 (54.43)	4n, 4H (p = 0.004), Sn, 4n (p = 0.001)
		reviews	-0.12 (0.65)	-0.12 (0.54)	**-0.07 (0.67)**	-0.22 (0.59)	4n, 4H (p = 0.011)
Keyboard/Mouse*	335	buy likelihood	**3.01 (1.24)**	2.46 (1.29)	2.79 (1.41)	2.56 (1.02)	Sn, SH (p = 0.013)
		inspection time	36.56 (28.19)	71.59 (107.8)	**32.97 (30.64)**	53.12 (35.47)	Sn, SH (p = 0.013), 4n, 4H (p = 0.002)
		screen	0.0 (0.38)	**0.03 (0.4)**	**0.03 (0.28)**	-0.11 (0.38)	4n, 4H (p = 0.044)
Screen/Display*	1120	buy likelihood	2.81 (1.24)	2.48 (1.24)	**2.9 (1.33)**	2.77 (1.16)	Sn, SH (p = 0.006), SH, 4H (p = 0.018)
		inspection time	46.31 (51.61)	49.97 (50.27)	**33.44 (32.23)**	48.18 (55.44)	4n, 4H (p = 0.0), Sn, 4n (p = 0.001)
		reviews	-0.09 (0.59)	-0.16 (0.56)	**-0.04 (0.64)**	-0.2 (0.64)	4n, 4H (p = 0.012)
		specifications	**0.11 (0.38)**	0.03 (0.29)	0.05 (0.33)	0.05 (0.31)	Sn, SH (p = 0.033)
Sound/Audio*	335	specifications	**0.16 (0.46)**	0.08 (0.35)	0.0 (0.34)	0.08 (0.33)	Sn, 4n (p = 0.015)
Storage*	1210	buy likelihood	2.79 (1.23)	2.51 (1.22)	**2.9 (1.26)**	2.78 (1.17)	Sn, SH (p = 0.016), SH, 4H (p = 0.017)
		inspection time	49.32 (53.88)	47.43 (65.36)	**35.82 (36.68)**	53.42 (59.37)	4n, 4H (p = 0.0), Sn, 4n (p = 0.001)
		screen	-0.04 (0.36)	-0.02 (0.44)	**0.04 (0.29)**	0.0 (0.38)	Sn, 4n (p = 0.015)
Price*	1555	buy likelihood	2.79 (1.24)	2.52 (1.23)	**2.91 (1.28)**	2.79 (1.19)	Sn, SH (p = 0.008), SH, 4H (p = 0.006)
		inspection time	48.47 (51.78)	48.48 (61.47)	**35.17 (37.02)**	49.88 (56.41)	4n, 4H (p = 0.0), Sn, 4n (p = 0.0)
		reviews	-0.11 (0.64)	-0.14 (0.56)	**-0.06 (0.66)**	-0.18 (0.64)	4n, 4H (p = 0.037)
$Query_{ID}$: 29	275	inspection time	55.26 (55.74)	39.0 (36.08)	**32.42 (22.94)**	45.0 (44.26)	Sn, 4n (p = 0.004)
$Query_{ID}$: 23	235	inspection time	61.82 (52.61)	46.49 (35.1)	**37.17 (42.53)**	80.39 (83.0)	4n, 4H (p = 0.002), Sn, 4n (p = 0.012), SH, 4H (p = 0.029)
$Query_{ID}$: 10	155	buy likelihood	3.11 (0.98)	2.45 (1.26)	**3.33 (0.98)**	2.98 (1.05)	Sn, SH (p = 0.046)
		inspection time	48.1 (90.67)	**28.78 (21.59)**	30.43 (33.6)	67.7 (84.2)	SH, 4H (p = 0.018)
$Query_{ID}$: 27	140	inspection time	47.33 (34.14)	82.66 (64.95)	**37.89 (25.91)**	38.56 (49.6)	Sn, SH (p = 0.043), SH, 4H (p = 0.014)
		reviews	0.0 (0.68)	**0.08 (0.39)**	0.02 (0.68)	-0.37 (0.54)	4n, 4H (p = 0.019), SH, 4H (p = 0.003)
		performance	-0.1 (0.3)	0.0 (0.28)	**0.04 (0.2)**	0.03 (0.29)	Sn, 4n (p = 0.045)
$Query_{ID}$: 42	135	inspection time	34.26 (33.44)	37.78 (18.57)	**31.41 (30.7)**	55.5 (46.85)	4n, 4H (p = 0.039)
		screen	0.0 (0.48)	**0.08 (0.39)**	0.04 (0.27)	-0.2 (0.51)	4n, 4H (p = 0.038)
$Query_{ID}$: 21	95	buy likelihood	**3.06 (1.19)**	2.34 (1.14)	3.0 (1.26)	2.3 (1.42)	Sn, SH (p = 0.043)
$Query_{ID}$: 33	55	inspection time	**22.95 (13.22)**	43.07 (46.67)	110.75 (77.23)	67.52 (25.39)	Sn, 4n (p = 0.001)

Table 5.10: Individual aspect and query influence on the measured variables. $*$ means that aspect was selected, but not necessarily alone.

(top queries for this study are listed in table 5.9, and table 5.3 for the previous study). The significance testing for these aspects and queries shown in table 5.10 aligns with those obtained from the first study. It seems here as well that the *4n* (i.e. 4vL ranking, no reviews highlighting) setting outperforms the other settings in most of the queries. The noticeable difference with this study is that the inspection time measures started to appear in many queries and aspects with a significant effect, especially in the (4n) setting. Inspection time measure was not reported in the first user study to be significantly affected by the ranking method; it was however on the borderline. This difference may be referred to the fact that users in the second user study had to evaluate five items rather than three as in the first study.

Other variables (age, gender and domain knowledge) have been also taken into consideration whether they play a role on the measures of this study. We could not find any significant changes when these variables change.

5.6 Discussion

In this research, we present the textual 4vL based ranking and analyse its validity as a user-centred ranking. Our objective was to create a ranking that takes into consideration user requirements based on the feedback of prior users. As reading individual reviews is a difficult and time-consuming task, we provide a text-based 4vL based ranking that extracts important information from the reviews and updates the ranking accordingly.

In a first user study in the laptop domain, we compared the 4vL approach to classic 2vL and star rating based methods as baselines, simulating a scenario where a user visits an online store and is offered with an initial ranking of the products. In our experiments, we discovered that participants perceived products ranked using the 4vL method as more likely to be bought and as requiring less inspection time. However, the change of these two measures was significant only when it was compared with the 2vL method. On the buying-decisions measure, we observed that the majority of buying decisions were made when the textual 4vL approach was used. As a consequence of these results, we are able to answer the research question (**RQ1**). The performance of the textual 4vL approach is higher than that of the traditional textual 2vL method. On the other hand, the performance of the textual 4vL method seems to be better or comparable to that of the star rating ranking method.

In the second study, we wanted to ascertain whether users were sufficiently aware of the characteristics of the products mentioned in the reviews and whether this can affect the buying decision when different ranking methods are used. We achieved that by 1) highlighting the products' reviews in order to get users' attention to the positives and the negatives of aspects that they are searching for; and 2) by applying two different methods (star rating and textual 4vL) for product ranking. Through the combinations of these two features, we created four settings i.e. *Sn*, *SH*, *4n* and *4H* (see table 5.5 for abbreviations' meanings). We used the same three measures of the first user study to evaluate the impact of the highlighting in general on the user's buying decisions (**RQ2.1**), and the impact of ranking methods on the user's buying decisions, particularly when highlighting is enabled (**RQ2.2**).

Our hypotheses for (**RQ2.1**) were, that the highlighting will increase the buy likelihood and decrease the inspection time of a product. In the results, we found the opposite effect from our initial hypotheses, i.e. highlighting caused the product to be inspected further before buying it. We also found that

highlighting is only significantly decreasing the buy likelihood when star rating ranking is applied. The differences are not significant in the case when textual 4vL is used. That probably reflects that the quality of the products brought by 4vL is higher when it is compared with products of the star rating method; as the buying likelihood dropped significantly only in the star rating method. The interpretation of this behaviour would be clearer when we look at the results of the (**RQ2.2**) and the general conclusion of both user studies.

For (**RQ2.2**), we supposed that the highlighting will be more effective on the buying likelihood and inspection time of products especially when 4vl ranking method is applied. We see here that the buy likelihood significantly increased only when the 4vL ranking is applied. In the star rating case, differences on the buy likelihood were not significant, but at the same time, they are consistent with the results of buy likelihood of the first user study. However, the effect of a combined setting (i.e. ranking and highlighting features) can only be noticed on the buy likelihood and inspection time measures. For the measure "position of the decided-to-buy product", we could not find significant changes and therefore could not conclude that position will be affected by both setting features.

So, based on the results of questions RQ2.1 and RQ2.2, we see that star rating based ranking works better only when there is no reviews-highlighting. But with the highlighting feature implemented in the system, 4vL based ranking seems to be more effective in terms of increasing the buy likelihood and reducing the time needed for inspecting a product.

A general conclusion can then be stated. By looking at the diagrams of figure 5.13, we notice the general behaviour when applying the 4vL ranking method. It increased the buy likelihood and decreased the inspection time of products. However, this change is affected by the impact of highlighting. This impact can be reasoned in two points: 1) it makes users more aware of product features, and therefore, decreases the buy likelihood (especially when there are negative passages). And 2), it makes users spend more time on seeing reviews (especially because different highlighting colours are applied), as a result, increasing the inspection time.

Some aspects can particularly have an impact on the results, one of them is the star rating. The star rating was always adjusted in the experiments in order to reduce the side effects of it on user's decisions and to keep its distributions similar for all ranking methods (see section 5.3.3). However, applying the normalisation function on a subset of the data (subsets differ based on the applied filters) will not guarantee an equivalent star rating distribution

for all methods among the entire set of the evaluated laptops. To test if the star rating evaluations have an effect on users, we performed a Spearman correlation test to see the impact of star rating on the buy likelihoods. We found that the correlation between the star rating of the laptop and buy likelihood for the entire collected set of evaluated laptops regardless of the used ranking method is significant ($\rho = 0.173$, $p < 0.001$). However, when we examine this correlation for each ranking method, we find a stronger correlation ($\rho = 0.248$, $p < 0.001$) when items are ranked based on star ratings than when they are ranked using 4vL ($\rho = 0.173$, $p = 0.015$). This indicates that the star rating is playing an unequal role on user decisions, and therefore, the results are more probable to be affected by this side factor.

Other aspects that can have an impact on the results are what we could notice in the free-text answers on the buy likelihoods. Participants mentioned the topics "price" and "reviews" most frequently in their answers. The sentiments towards these two topics were identified to be changing when we changed the ranking method in the first user study and the setting in the second user study. Although this can give us a user-perspective look of the quality of the ranking method, it indicates that participants are also affected by the price of the product as an important factor for determining their buying decision.

The side effects of highlighting functionality can also be one of the factors affecting the results. As many participants stated in the post questionnaire of the second user study, highlighting disturbed their browsing routines and caused them to focus less on the task's main goals.

To conclude the results of both studies, textual 4vL helps users to be more confident about their buying decisions. The effect of the textual 4vL ranking is particularly visible when users are able to see the positives and the negatives about the aspects they are interested in (highlighting is enabled). The advantage of textual 4vL ranking over the star rating method comes from the way that the textual 4vL is using for weighting the relevancy of products to user queries. While star rating usually expresses the overall impression of the users about the products, textual 4vL focuses particularly on some parts of the reviews which are identified to be relevant to what users are searching for.

In these two studies, we calculated the scores of 4vL rankings based on the average needs of users toward information that was extracted from the experiment of chapter 4. An improvement of the 4vL method in this domain can be achieved by considering weighted aspects according to the importance of aspects that can be derived from user needs shown in figure 5.8.

5.7 Conclusion

In this chapter, we introduced a 4vL based ranking method as a method for supporting users in the process of ranking products based on textual content of reviews. The method is computing the ranking scores based on the tfidf weights for matching the user needs with the reviews. However, unlike traditional query matching methods which distinguish only between the existence of the match, the 4vL method distinguishes between three types of query matching. 1) a match with a positive sentiment, 2) a match with a negative sentiment, and 3) another type when the query is not matched at all or matched with a neutral sentiment. Using the 4vL, the information from different reviews are fused into one score, which is used as a basis for the ranking.

We aimed in this chapter to review the 4vL method in a user-oriented approach. Therefore, we investigated through two user studies the potential applicability, strengths and weaknesses of the proposed method. Comparing our method with baselines showed that the 4vL method gives the users more confidence about their buying decisions. However, the results show that the method's performance is significantly noticeable when users are provided with a tool to get their attention to what they are looking for in the reviews i.e. highlighting.

The most remarkable feature of this method is its adaptability to the various informational requirements of users. Different weights are assigned to positive, negative, and uncertain matching cases in this study. According to our previous research, the most important factor for users is to avoid purchasing products with negative reviews. After that they are attracted to the positive experiences. Then, they considered the unmentioned information in reviews to be on the positive side, but at the same time with less importance compared to negative and positive experiences. This behaviour has been translated into various weights and incorporated into this 4vL based model. These weights represent the generalised case of users' needs for the information in this domain. Adapting the model to be personalised is achievable by assigning different weights based on the user's needs, which are reflected in the user's perception of some product aspects. For instance, the importance of aspect 'price' can be increased and the impact of negative information about other unimportant aspects can be reduced in order to create a more personalised ranking for that user.

Chapter 6

Conclusion and Outlook

6.1 Conclusion

In this thesis, we have introduced the idea of using multi-valued logic-based models for supporting the users in the task of ranking services or products based on the reviews of previous users.

Reviews are considered one of the most available sources of information that can be used to assess the quality of a service or a product and support the purchasing decision of the consumer. However, as with any other kind of social information, there are challenges involved when using such resources in real-world applications.

The prominent challenges that we addressed in this thesis are 1) The involvement of the credibility of the reviews in the ranking process (i.e. how to deal with the fact that not all reviews are honest and unbiased?); 2) As reviews are personal and often contain subjective elements, they may contain contradictions, what user needs in this case, is that ranking process considers these contradictions. 3) This thesis also discussed the topics of information missing in reviews, and how to handle it effectively in search results.

In order to address these challenges we proposed two multi-valued logic-based models. In the first model, the four-valued logic was used. Within this logic, the information was classified to be true, false, unknown and inconsistent. In the second model, we used subjective logic which introduces the information based on the user's perception of the situation and classifies it to belief, disbelief and uncertainty. The logic provides operators that facilitate the analysis of information within networks of information generators and receivers.

The core idea of the ranking process based on these models includes five main steps. The first step is to identify the required aspects of users' interests and to determine what terms are relevant to them. The second step is to classify the reviews according to the required aspects into positive, negative and unknown categories. Thirdly, propose a measure to weigh the relevance between the reviews and the required aspects e.g. tfidf. Fourthly, aggregate the classified information from different reviews into positive, negative and unknown scores for each aspect. These scores represent the relevance between the required aspects and the reviews. The final ranking score is derived from a relation that merges the positive, negative and unknown scores with each other.

In order to evaluate the proposed models, we applied them within two example applications. In the first application, we followed a system-oriented method for the evaluation. In this application, we used textual reviews for evaluating some of the hotel aspects which were already evaluated separately by the customers of the hotels (i.e. there are gold standard labels). four valued and subjective logic models were employed for the process, and the results were compared to a traditional IR baseline method. The logical models have shown an improved performance compared with the baseline.

In the second example application, we followed a used-oriented method for the evaluation. In this application, we wanted to apply our approach on a real life case, in which users are deciding the quality of the approach (i.e. there is no gold standard labels). We developed therefore a laptop online store that supports users in their purchase decisions by ranking the products depending on the reviews. For this task we used one of the two logical models that performed the best in the first example application (i.e. four-valued logic), we compared it then with another traditional IR model (i.e. two-valued logic) and a baseline that depends on the star rating. In two user studies, we showed that the proposed method is able to support the users in buying decisions by increasing the buying likelihood and decreasing the required time needed to inspect the products.

The prominent feature of the proposed approaches is the ability to process the search results not only by matching specific terms of the search query, but also classify the match into positive and negative cases. Furthermore, it accounts for cases when the searched documents do not have enough specified information for determining the nature of the match.

Within the presented example applications, we inferred and then employed weights for the positive, negative and unknown factors that determine the

relevance between the search query and the products. These obtained weights reflect the average opinion of the users who have expressed their experiences through the reviews. From that average opinion, we can conclude that users' behaviour of receiving information from reviews is that users are negatively affected by the negative reviews more than they are positively affected by the positive ones. In addition, users assume a positive experience if the information they are looking for does not exist in the reviews.

6.2 Outlook

This work has focused on the suitability of the multi-valued logical models in an aspect-oriented search task that bases on the reviews and addresses their issues of credibility contradictions and missing information. Future improvements of our work can be directed in fields like relevancy and sentiment detection improvement, model personalisations, multi-dimensional results presentation and ranking and contradictions-based user support.

The quality of the models is highly relevant to the indexing, weighting and sentiment detection techniques. In order to boost the performance of the current models, either more accurate indexes need to be generated and/or more robust methods for extraction of the opinions and sentiments of reviews should be developed.

Personalisation of the models is a straightforward process, as it requires only computing two types of weights. The first kind is weights that are associated with the user's perception of positive, negative or unknown factors. The second category of weights is relevant to the different importance levels of the searched aspects of the user's interest.

Multi-dimensional results presentation is yet another important development that can be incorporated into future models and ranks the results with respect to the user's perceived relevance. Possible approaches include exploring the use of multi-valued logic values (positive, negative, unknown and inconsistent) as a method to present the search results to users and integrate them into search engines. Additionally, exploring the combination of multiple dimensions into a single one (e.g. positive-negative) or having several dimensions (e.g. positive vs. inconsistent, positive vs. unknown, etc.) will be beneficial since it allows us to explore the data in an alternative manner.

Regarding contradiction-based systems, it relies on the fact that reviews have

contradictory information about the same topic or product. Such inconsistency can be used to better inform users about their searches. Possible application of such an idea is to extract tips for supporting user decisions.

Finally, in this thesis, we analysed user reviews of products as a specific type of social media content. We developed and explored models to extract and classify statements on aspects of the products mentioned in the reviews, as well as to determine the sentiment of the reviewers towards each aspect. Our research focused specifically on this type of social media content, but there are other applications in this field that could benefit from exploring the applicability of our models.

One possible direction for further research is to investigate the models on some basic elements of any social media platform, such as user feeds and search pages. These elements often contain a large amount of content that can be difficult to navigate. By applying our models, we could classify and rank the content, making it easier for users to find the information they need.

Another promising direction for further research is to explore different types of social media content, such as political discussions and trending stories. By analysing the content of social media towards political topics or current events, for instance, we could potentially contribute to improving access to credible information about public issues.

To apply our models to other domains, it will be important to identify appropriate representations for positive and negative statements about the relevant aspects of the domain. This will require a thorough understanding of the domain and its key aspects. By developing suitable representations, we can ensure that our models accurately capture the sentiment of users towards those aspects and provide useful insights for users of the social media platform.

Appendix

A Appendix

A.1 Aspect keyword extraction (hotel domain)

Aspect	Terms
Cleanliness	cleanlines, cleanness, tidy, hygiene, cleaniness, cleaness, brightness, neatness, tidiness, smelly, mouldy, cleanless, dusty, cleaness, cleanness, unclean, hygiene, cleaniness, cleanliness, cleaning, neat, clen, dirty, housekeeping, stained, disgusting, cleaned, clean, damp, filthy, spotless, stinky, immaculate, cleanless
Comfort	convenience, comfortabe, newness, confort, coziness, spacious, comfy, comfotable, quietness, comforts, spaciousness, comfortable, confy, cosiness, comfort, modernness, quiteness, ammenities, tranquility, confortable, warmth, comfey, conveniences, comforable, layout
Staff	reciption, fiendly, helpfullness, kindness, friendlyness, openness, organization, competence, politeness, courtesy, friendlieness, friendless, friendliness, personell, receptionists, stuffs, straff, sevice, employees, staff, stuff, landlady, team, landlord, personnel, personel, stuf, host, staf, owner, managers, crew, members, concierge, staffs, hostess, owners, support, manager, helpfullness, professionalism, helpfulness
Value for money	vfm, affordable, inexpensive, decent, cheap, momey, monies, moeny, cheapish, cheep, monet, money, expensive, mony, miney, maney, monney, price, mone, fair
Location	lokation, locatio, locstion, centrality, situated, museuminsel, locationis, hills, southbank, location, fringes, border, locatin, atmospher, museums, sites, locaion, loction, places, locstion, locations, position, loaction, landmarks, locaton, spot, locatio, foot, lication, lokation, located, markt, ubication, destinations, localisation, neigbourhood, loation, placement, edge, locatiom, attractions, attraction, locality, sights, monuments, place, tramstop, localization, placed
Wifi	wif, computer, wi_fi, wireles, lan, fre, wlan, wifi, internet, wfi, wireless
Facilities	ammenities, necessities, utilities, washer, dishwasher, utensils, services, service, facilities, furnitures, appliances, gadgets, furnishings, commodities, equipments, mashine, furniture, amenities, facility, pans, equiped, cutlery, equipment, forniture, essentials, utilized, cooking, facilites, facilties

Table A.1: Most relevant hotel aspect terms extracted through a word2vec model.

A.2 System oriented experimental results (I)

Table A.2: R^2 scores of logic-based approaches compared with baselines

(a) Baseline models

Aspect	tf	tfidf	bm25
cleanliness	0.312	0.349	0.278
comfort	0.111	0.108	0.078
staff	0.313	0.342	0.257
value for money	0.076	0.085	0.07
location	0.274	0.272	0.193
wifi	0.155	0.157	0.128
facilities	0.197	0.221	0.145
avg. score	0.206	0.219	0.164

(b) 4vL based models

	independent			disjoint		
Aspect	tf	tfidf	bm25	tf	tfidf	bm25
cleanliness	**0.369**	0.298	0.333	0.141	0.136	0.127
comfort	**0.173**	0.113	0.132	0.092	0.091	0.072
staff	**0.368**	0.306	0.335	0.111	0.111	0.116
value for money	0.09	**0.114**	0.073	0.038	0.037	0.034
location	**0.305**	0.174	0.279	0.172	0.164	0.124
wifi	0.169	0.156	0.133	0.057	0.058	0.05
facilities	**0.279**	0.239	0.22	0.088	0.091	0.077
avg. score	**0.25**	0.2	0.215	0.1	0.098	0.086

(c) SL based models

	cumulative fusion			averaging fusion		
Aspect	tf	tfidf	bm25	tf	tfidf	bm25
cleanliness	**0.369**	0.336	0.329	0.09	0.11	0.076
comfort	0.145	0.123	0.109	0.047	0.051	0.021
staff	0.36	0.349	0.352	0.145	0.154	0.109
value for money	0.089	0.113	0.074	0.044	0.046	0.041
location	0.286	0.22	0.247	0.151	0.158	0.067
wifi	**0.175**	0.168	0.147	0.074	0.073	0.06
facilities	0.259	0.244	0.216	0.031	0.036	0.02
avg. score	0.241	0.222	0.211	0.083	0.09	0.056

A.3 System oriented experimental results (II)

Table A.4: Regression factors of the two logics for independent/cumulative fusion with tf weights.

	4vL			SL(b+ua)	
Aspect	true	false	unknown	positive	negative
cleanliness	2.15	-3.91	1.08	4.21	-7.86
comfort	2.17	-7.57	1.51	2.62	-7.97
staff	1.97	-4.78	0.83	4.39	-7.49
value for money	0.54	-0.89	0.63	0.35	-2.68
location	1.81	-6.37	0.88	3.85	-6.89
wifi	1.75	-3.99	-0.64	7.90	-8.02
facilities	1.73	-4.54	0.17	5.87	-8.06

A.4 Statistical Tests of The System Oriented Evaluation Approach

	tf		tfidf		bm25		4vL(independent)		SL(cumulative fusion)	
Aspect	M	SD	M	SD	M	SD	M	SD	M	SD
location	0.634	0.536	0.645	0.541	0.66	0.553	0.538	0.484	0.542	0.49
staff	0.706	0.595	0.701	0.592	0.717	0.604	0.573	0.513	0.576	0.517
cleanliness	0.77	0.644	0.766	0.641	0.786	0.661	0.634	0.564	0.642	0.556
comfort	0.862	0.672	0.863	0.672	0.88	0.679	0.796	0.64	0.812	0.646
value for money	0.643	0.578	0.64	0.574	0.641	0.575	0.634	0.567	0.636	0.567
facilities	0.822	0.675	0.813	0.671	0.841	0.681	0.714	0.603	0.729	0.604
wifi	1.111	1.387	1.118	1.378	1.144	1.395	1.004	1.346	1.01	1.336

Table A.5: Absolute error statistical description of the tested baselines and logical models.

A.5 User Surveys (USI & USII)

Product Search Survey

Dear participant,

We invite you to take part in our study about **a product search system**.
The survey is part of a research project at the University of Duisburg-Essen and the GESIS - Leibniz Institute of the Social Sciences.

Participating in the survey will take about **15 minutes**. You will be asked to search for a laptop on our website and answer some questions about how you perceived the system.

We **do not collect sensitive personal data** such as your name or address, but we will ask for your Prolific-ID to ensure your payment. All data will be **anonymised**, i.e. your Prolific-ID will be deleted from the dataset upon payment. We may make the anonymised dataset publicly available, and may use the anonymised data for publication purposes.

Your participation is **voluntary**. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analysed.

For questions, please do not hesitate to contact:

Firas Sabbah
~ sabbah@is.inf.uni-due.de
Faculty of Engineering Sciences
Universität Duisburg-Essen

For participating in the study, I confirm that:

- ☐ I have read the above-mentioned conditions.
- ☐ I am 18 years or older.
- ☐ My participation is voluntary and I know that I can abandon the survey at any point.

Product Search Survey

Before we start, we have some questions about your background.

1. What is your Prolific ID?

Prolific IDs have 24 alphanumeric characters.

2. Please rate the following statements:

	do not agree at all	do not agree	undecided	rather agree	strongly agree
I consider my knowledge of **laptops** to be very good.	○	○	○	○	○
Friends and family often come to me for advice on **laptops**.	○	○	○	○	○

3. What is your age?

☐ years

Product Search Survey

You will now visit our website.

Please follow the instructions on the website.

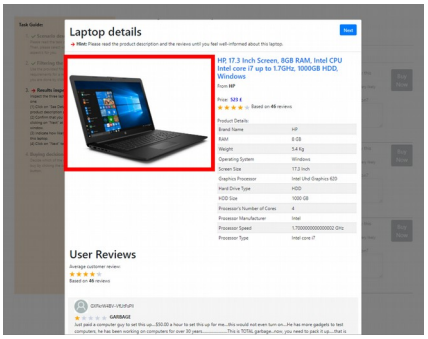
Questionnaire

We now have some questions about your search.

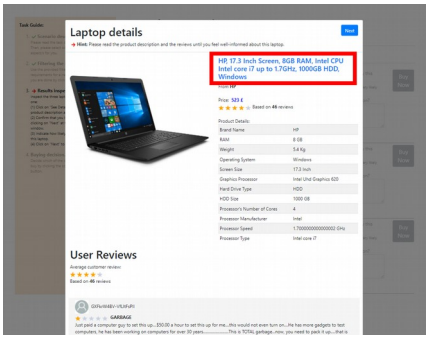
Before your search, you selected some aspects that were important for your search.
You selected the following aspects:

The following question(s) will ask about where you found useful information about the laptops.
We will ask about:

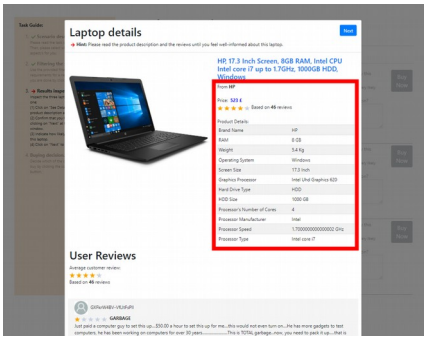
picture



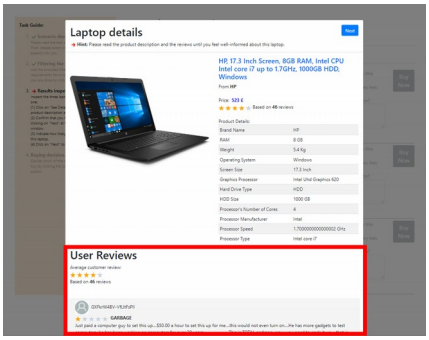
title



details



reviews



5. Where did you find information about the screen / display?

Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

6. Where did you find information about the storage?
Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

7. Where did you find information about the keyboard / mouse?
Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

8. Where did you find information about the price?
Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

9. Where did you find information about the battery?
Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

10. Where did you find information about the sound / audio?
Please select all options that apply.

- [ ] Laptop picture
- [ ] Laptop title
- [ ] Laptop details
- [ ] Laptop reviews

- [ ] None of the above

Questionnaire

We now have some questions about your search.

Before your search, you selected some aspects that were important for your search.
You selected the following aspects:

11. Please tell us how much you agree or disagree with the following statements.

	strongly disagree 1	2	3	4	strongly agree 5
I feel well-informed about the products concerning the aspects I selected.	○	○	○	○	○
I do not have enough information about the products concerning the aspects I selected.	○	○	○	○	○
I am unsure about my buying decision.	○	○	○	○	○
I feel confident about my buying decision.	○	○	○	○	○

12. Please tell us how much you agree or disagree with the following statements.

	strongly disagree 1	2	3	4	strongly agree 5
For me, the reviews were an important source of information about the aspects.	○	○	○	○	○
It was difficult to find relevant information about the aspects in the reviews.	○	○	○	○	○
It was easy to find information about the aspects in the reviews.	○	○	○	○	○
I found the reviews informative concerning the aspects.	○	○	○	○	○
I did not use the reviews to get informed about the aspects.	○	○	○	○	○

Questionnaire (Only applies to USII)

We now have some questions about your search.

Before your search, you selected some aspects that were important for your search.
You selected the following aspects:

13. How well did the highlighting in the reviews support you in finding information on the aspects?

○	○	○	○	○
1 (very poor support)	2	3	4	5 (very good support)

14. What did you like about the highlighting?

Please write 2-3 phrases / sentences.

15. What did you dislike about the highlighting?

Please write 2-3 phrases / sentences.

Page 7

Questionnaire

16. (Optional) Did you experience any technical difficulties during the study?

- ○ no
- ○ yes (please specify):

17. (Optional) Do you have any comments on the survey?

Last Page

Thank you for completing this questionnaire!

In this study, we are testing two different search algorithms. These algorithms select laptops that are a good fit for your wishes.
Other than existing algorithms, our algorithms also take into account which laptop aspects are important to the user.
You were asked to indicate for each laptop in the result list how likely it is that you would buy the laptop.
In this study, we investigate how well our algorithms perform. If they perform well, users like the laptops in the result list.
We are very grateful that you took your time to help us with our research!

If you have questions about this study, please contact:

Firas Sabbah
~ sabbah@is.inf.uni-due.de
Faculty of Engineering Sciences
Universität Duisburg-Essen

You have reached the end of the questionnaire.

Click here to return to Prolific.

Your Prolific completion code is:

A.6 Aspect keyword extraction(laptop domain)

Aspect	Terms
Storage	hhd, harddrives, disk, ram, drive, hdd, dram, mircosd, storage, hardrive, boot, boots, space, memory, mem, seagate, ssd, ddr4, ddr3, gb, 1gb, 2gb, 4gb, 8gb, 16gb, 32gb, 64gb, 256gb, 512gb, 1tb, 2tb, tb, sd, gig, nvme, nvm, boot, bootup
Price	price, cost, priced, competitors, dollar, dollars, pricing, usd, bucks, buck, competition, prices, understated, budget, deal, priced, affordable, penny, inexpensive, expensive, money, cheap, cheaper, sale, sales, cost, worth, pay, paying, paid, vfm, affordable, inexpensive, cheap, momey, monies, moeny, cheapish, cheep, monet, money, expensive, mony, miney, maney, monney, price, mone, fair, pricey, spend, spent, purchase, investment, pay, priced, 2200usd, usd550, sub-1k, buy
Screen/Display	screen, glide, touchscreen, touch, haptic, screen, display, displays, monitor, fhd, colours, colorful, colors, glow, contrast, lid, video, nvidia, nvidea, optimus, gfx, radeon, nvida, graphics, geforce, hd, brightness, resolution, bright, crisp, graphic, responsive, backlight, screens, displays, screen, monitors, display, monitor, pixel
Keyboard/Mouse	layout, kb, tactile, key, keyboard, keypad, trackpad, keys, clicky, numpad, chiclet, chicklet, keyboards, button, numlock, arrow, typing, touchpad, pad, press, spacebar, pen, bar, track-pad, backspace, letters, numberpad, fingerprint, pads, mousepad, trac, mouse, track, touchpad, logitech, trackball, mice, mouses, receiver, curser
Sound/Audio	audio, stereo, sounds, subwoofer, sound, speakers, voices, speaker, bass, tinny, audiophile, sounding, microphone, loud, mic, headphone, audible, volume
Battery	bettery, battery, batterly, prolong, expectancy, batter, shortens, batery, battary, batt, battery, batter, bettery, batt, batteries, batery, unplugged, standby, powersave, power, unplugging, unplugged, energy, charging, charger, charge, duration

Table A.6: Most relevant laptop aspect terms extracted through a word2vec based model.

Indexes

List of Figures

List of Tables

Bibliography

Abdi, A., Shamsuddin, S. M. and Aliguliyev, R. M. [2018], 'Qmos: Query-based multi-documents opinion-oriented summarization', *Information Processing & Management* **54**(2), 318–338.

Aciar, S., Zhang, D., Simoff, S. and Debenham, J. [2006*a*], Informed recommender agent: Utilizing consumer product reviews through text mining, *in* '2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops', IEEE, pp. 37–40.

Aciar, S., Zhang, D., Simoff, S. and Debenham, J. [2006*b*], Recommender system based on consumer product reviews, *in* '2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)', IEEE, pp. 719–723.

Adriaans, F., Kamps, J. and Koolen, M. [2011], The importance of document ranking and user-generated content for faceted search and book suggestions, *in* 'International Workshop of the Initiative for the Evaluation of XML Retrieval', Springer, pp. 30–44.

Aithal, M. and Tan, C. [2021], 'On positivity bias in negative reviews', *arXiv preprint arXiv:2106.12056* .

Aladhadh, S., Zhang, X. and Sanderson, M. [2018], 'Location impact on source and linguistic features for information credibility of social media', *Online Information Review* .

Alexandridis, G., Tagaris, T., Siolas, G. and Stafylopatis, A. [2019], From free-text user reviews to product recommendation using paragraph vectors and matrix factorization, *in* 'Companion Proceedings of the 2019 World Wide Web Conference', pp. 335–343.

Ali, N., Fatima, A., Shahzadi, H., Khan, N. and Polat, K. [2021], 'Online reviews & ratings inter-contradiction based product's quality-prediction through hybrid neural network', *Journal of the Institute of Electronics and Computer* **3**(1), 24–52.

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C. and Vilares, J. [2021], 'Sentiment analysis for fake news detection', *Electronics* **10**(11), 1348.

Asadi, R. and Regan, A. [2019], 'A convolution recurrent autoencoder for spatio-temporal missing data imputation', *arXiv preprint arXiv:1904.12413* .

Atanassov, K. [1986], 'Intuitionistic fuzzy sets', *Fuzzy Sets Syst.* **20**, 87–96.

Badache, I., Fournier, S. and Chifu, A.-G. [2018], Contradiction in reviews: is it strong or low?, *in* '40th European Conference on Information Retrieval, ECIR 2018-BroDyn: Workshop on Analysis of Broad Dynamic Topics over Social Media'.

Baeza-Yates, R., Ribeiro-Neto, B. et al. [1999], *Modern information retrieval*, Vol. 463, ACM press New York.

Bahrainian, S.-A. and Dengel, A. [2013], Sentiment analysis and summarization of twitter data, *in* '2013 IEEE 16th International Conference on Computational Science and Engineering', IEEE, pp. 227–234.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B. and Su, Z. [2007], Optimizing web search using social annotations, *in* 'Proceedings of the 16th international conference on World Wide Web', pp. 501–510.

Belnap, N. D. [1977], A useful four-valued logic, *in* 'Modern uses of multiple-valued logic', Springer, pp. 5–37.

Bi, J.-W., Liu, Y. and Fan, Z.-P. [2019], 'Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking', *Information Sciences* **504**, 293–307.

Bi, J.-W., Liu, Y., Fan, Z.-P. and Cambria, E. [2019], 'Modelling customer satisfaction from online reviews using ensemble neural network and effect-based kano model', *International Journal of Production Research* **57**(22), 7068–7088.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. and Reynar, J. [2008], 'Building a sentiment summarizer for local service reviews'.

Burgess, S., Sellitto, C., Cox, C. and Buultjens, J. [2011], 'Trust perceptions of online travel information by different content creators: Some social and legal implications', *Information systems frontiers* **13**(2), 221–235.

Carmel, D., Roitman, H. and Yom-Tov, E. [2010], 'Social bookmark weighting for search and recommendation', *The VLDB Journal* **19**(6), 761–775.

Chen, C.-W. [2017], 'Five-star or thumbs-up? the influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and continuance intention', *Internet Research* **27**(3), 478–494.

Chen, K., Kou, G., Shang, J. and Chen, Y. [2015], 'Visualizing market structure through online product reviews: Integrate topic modeling, topsis, and multi-dimensional scaling approaches', *Electronic Commerce Research and Applications* **14**(1), 58–74.

Culpepper, J. S., Diaz, F. and Smucker, M. D. [2018], Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018), *in* 'ACM SIGIR Forum', Vol. 52, ACM New York, NY, USA, pp. 34–90.

Curien, N., Fauchart, E., Laffond, G. and Moreau, F. [2006], *Online consumer communities: Escaping the tragedy of the digital commons*, na.

Dalila, B., Mohamed, A. and Bendjanna, H. [2018], A review of recent aspect extraction techniques for opinion mining systems, *in* '2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)', IEEE, pp. 1–6.

Dellarocas, C. [2006], 'Strategic manipulation of internet opinion forums: Implications for consumers and firms', *Management science* **52**(10), 1577–1593.

Dina, N. Z., Yunardi, R. T., Firdaus, A. A. and Juniarta, N. [2021], 'Measuring user satisfaction of educational service applications using text mining and multicriteria decision-making approach.', *International Journal of Emerging Technologies in Learning* **16**(17).

Do, N., Rahayu, W. and Torabi, T. [2016], 'A query expansion approach for social media data extraction', *International Journal of Web and Grid Services* **12**(4), 418–441.

Dori-Hacohen, S. and Allan, J. [2015], Automated controversy detection on the web, *in* 'European Conference on Information Retrieval', Springer, pp. 423–434.

Eirinaki, M., Pisal, S. and Singh, J. [2012], 'Feature-based opinion mining and ranking', *Journal of Computer and System Sciences* **78**(4), 1175–1184.

Fan, Z.-P., Xi, Y. and Li, Y. [2017], 'Supporting the purchase decisions of consumers: a comprehensive method for selecting desirable online products', *Kybernetes* .

Federkeil, G., File, J., Kaiser, F., van Vught, F. A. and Ziegele, F. [2012], 'An interactive multidimensional ranking web tool', *Multidimensional ranking: The design and development of U-Multirank* pp. 167–177.

Feuerbach, J., Loepp, B., Barbu, C.-M. and Ziegler, J. [2017], 'Enhancing an interactive recommendation system with review-based information filtering', *IntRS* **17**, 2–9.

Fitting, M. [1994], 'Kleene's three valued logics and their children', *Fundamenta informaticae* **20**(1, 2, 3), 113–131.

Frommholz, I. and Fuhr, N. [2006], Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries, *in* 'Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries', pp. 55–64.

Fuhr, N. [1992], 'Probabilistic models in information retrieval', *The computer journal* **35**(3), 243–255.

Fuhr, N. and Rölleke, T. [1998], Hyspirit-a probabilistic inference engine for hypermedia retrieval in large databases, *in* 'International Conference on Extending Database Technology', Springer, pp. 24–38.

Garimella, K., Morales, G. D. F., Gionis, A. and Mathioudakis, M. [2018], 'Quantifying controversy on social media', *ACM Transactions on Social Computing* **1**(1), 1–27.

Goker, A. and Davies, J. [2009], *Information retrieval: Searching in the 21st century*, John Wiley & Sons.

Gomes, L. and Lima, M. [1991], 'Todim: Basics and apllication to multicriteria ranking of projects with environmental impacts', *Foundations of Control Engineering* **Vol. 16**, 113–127.

Gozuacik, N., Sakar, C. O. and Ozcan, S. [2021], 'Social media-based opinion retrieval for product analysis using multi-task deep neural networks', *Expert Systems with Applications* **183**, 115388.

Guerreiro, J. and Rita, P. [2020], 'How to predict explicit recommendations in online reviews using text mining and sentiment analysis', *Journal of Hospitality and Tourism Management* **43**, 269–272.

Guo, C., Du, Z. and Kou, X. [2018], 'Products ranking through aspect-based sentiment analysis of online heterogeneous reviews', *Journal of Systems Science and Systems Engineering* **27**(5), 542–558.

Haydar, C. and Boyer, A. [2017], A new statistical density clustering algorithm based on mutual vote and subjective logic applied to recommender systems, *in* 'Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization', pp. 59–66.

Jmal, J. and Faiz, R. [2013], Customer review summarization approach using twitter and sentiwordnet, *in* 'Proceedings of the 3rd international conference on web intelligence, mining and semantics', pp. 1–8.

Jøsang, A. [2002], 'The consensus operator for combining beliefs', *Artificial Intelligence* **141**(1-2), 157–170.

Jøsang, A. [2016], *Subjective logic*, Springer.

Jøsang, A. and Hankin, R. [2012], Interpretation and fusion of hyper opinions in subjective logic, *in* '2012 15th International Conference on Information Fusion', IEEE, pp. 1225–1232.

Josang, A., Hayward, R. F. and Pope, S. [2006], 'Trust network analysis with subjective logic'.

Kang, D. and Park, Y. [2014], 'based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach', *Expert Systems with Applications* **41**(4), 1041–1050.

Kleinberg, J. M. [1999], 'Authoritative sources in a hyperlinked environment', *Journal of the ACM (JACM)* **46**(5), 604–632.

Kong, R., Wang, Y., Xin, W., Yang, T., Hu, J. and Chen, Z. [2011], Customer reviews for individual product feature-based ranking, *in* '2011 first international conference on instrumentation, measurement, computer, communication and control', IEEE, pp. 449–453.

Könsgen, R., Schaarschmidt, M., Ivens, S. and Munzel, A. [2018], 'Finding meaning in contradiction on employee review sites—effects of discrepant online reviews on job application intentions', *Journal of Interactive Marketing* **43**, 165–177.

Koster, A., Bazzan, A. L. and De Souza, M. [2017], 'Liar liar, pants on fire; or how to use subjective logic and argumentation to evaluate information from untrustworthy sources', *Artificial Intelligence Review* **48**(2), 219–235.

Lee, K., Kim, H. G. and Kim, H.-J. [2012], 'A social inverted index for social-tagging-based information retrieval', *Journal of Information Science* **38**, 313 – 332.

Li, J., Lu, K., Huang, Z. and Shen, H. T. [2019], 'On both cold-start and long-tail recommendation with social data', *IEEE Transactions on Knowledge and Data Engineering* **33**(1), 194–208.

Li, M.-Y., Zhao, X.-J., Zhang, L., Ye, X. and Li, B. [2020], 'Method for product selection considering consumer's expectations and online reviews', *Kybernetes* .

Li, S., Zha, Z.-J., Ming, Z., Wang, M., Chua, T.-S., Guo, J. and Xu, W. [2011], Product comparison using comparative relations, *in* 'Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval', pp. 1151–1152.

Liu, Y., Bi, J.-W. and Fan, Z.-P. [2017*a*], 'A method for ranking products through online reviews based on sentiment classification and interval-valued intuitionistic fuzzy topsis', *International Journal of Information Technology & Decision Making* **16**(06), 1497–1522.

Liu, Y., Bi, J.-W. and Fan, Z.-P. [2017*b*], 'Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory', *Information Fusion* **36**, 149–161.

Massoudi, K., Tsagkias, M., Rijke, M. d. and Weerkamp, W. [2011], Incorporating query expansion and quality indicators in searching microblog posts, *in* 'European Conference on Information Retrieval', Springer, pp. 362–367.

Mendel, J., Hagras, H., Tan, W.-W., Melek, W. W. and Ying, H. [2014], *Introduction to type-2 fuzzy logic control: theory and applications*, John Wiley & Sons.

Najmi, E., Hashmi, K., Malik, Z., Rezgui, A. and Khan, H. U. [2015], 'Capra: a comprehensive approach to product ranking using customer reviews', *Computing* **97**(8), 843–867.

Opricovic, S. [1998], 'Multicriteria optimization of civil engineering systems', *Faculty of civil engineering, Belgrade* **2**(1), 5–21.

Park, D.-H., Lee, J. and Han, I. [2007], 'The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement', *International journal of electronic commerce* **11**(4), 125–148.

Poongodi, M., Vijayakumar, V., Rawal, B., Bhardwaj, V., Agarwal, T., Jain, A., Ramanathan, L. and Sriram, V. [2019], 'Recommendation model based on trust relations & user credibility', *Journal of Intelligent & Fuzzy Systems* **36**(5), 4057–4064.

Pujianto, U., Wibawa, A. P., Akbar, M. I. et al. [2019], K-nearest neighbor (k-nn) based missing data imputation, *in* '2019 5th International Conference on Science in Information Technology (ICSITech)', IEEE, pp. 83–88.

Raut, V. B. and Londhe, D. [2014], Opinion mining and summarization of hotel reviews, *in* '2014 International Conference on Computational Intelligence and Communication Networks', IEEE, pp. 556–559.

Ravikumar, S., Balakrishnan, R. and Kambhampati, S. [2012], Ranking tweets considering trust and relevance, *in* 'Proceedings of the Ninth International Workshop on Information Integration on the Web', pp. 1–4.

Robertson, S. E. and Jones, K. S. [1976], 'Relevance weighting of search terms', *Journal of the American Society for Information science* **27**(3), 129–146.

Rölleke, T. and Fuhr, N. [1996], Retrieval of complex objects using a four-valued logic, *in* 'SIGIR', ACM, pp. 206–214.

Sabbah, F. [2019], Logic-based models for social media retrieval, *in* 'European Conference on Information Retrieval', Springer, pp. 348–352.

Sabbah, F. and Fuhr, N. [2021], A transparent logical framework for aspect-oriented product ranking based on user reviews, *in* 'European Conference on Information Retrieval', Springer, pp. 558–571.

Salton, G. [1971], *The SMART retrieval system—experiments in automatic document processing*, Prentice-Hall, Inc.

Sedhain, S., Menon, A., Sanner, S., Xie, L. and Braziunas, D. [2017], Low-rank linear cold-start recommendation from social data, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 31.

Sedhain, S., Sanner, S., Braziunas, D., Xie, L. and Christensen, J. [2014], Social collaborative filtering for cold-start recommendations, *in* 'Proceedings of the 8th ACM Conference on Recommender systems', pp. 345–348.

Serrano-Guerrero, J., Olivas, J. A. and Romero, F. P. [2020], 'A t1owa and aspect-based model for customizing recommendations on ecommerce', *Applied Soft Computing* **97**, 106768.

Siering, M., Deokar, A. V. and Janze, C. [2018], 'Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews', *Decision Support Systems* **107**, 52–63.

Škorić, B., de Hoogh, S. J. and Zannone, N. [2016], 'Flow-based reputation with uncertainty: evidence-based subjective logic', *International Journal of Information Security* **15**(4), 381–402.

Sorensen, A. T. and Rasmussen, S. J. [2004], 'Is any publicity good publicity? a note on the impact of book reviews', *NBER Working paper, Stanford University* .

Stavrianou, A. and Brun, C. [2015], 'Expert recommendations based on opinion mining of user-generated product reviews', *Computational Intelligence* **31**(1), 165–183.

Sun, C., Huang, L. and Qiu, X. [2019], 'Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence', *arXiv preprint arXiv:1903.09588* .

Sun, M. [2012], 'How does the variance of product ratings matter?', *Management Science* **58**(4), 696–707.

Takahashi, T. and Kitagawa, H. [2009], A ranking method for web search using social bookmarks, *in* 'International Conference on Database Systems for Advanced Applications', Springer, pp. 585–589.

Tunkelang, D. [2009], 'Faceted search', *Synthesis lectures on information concepts, retrieval, and services* **1**(1), 1–80.

Van Der Heijden, R. W., Kopp, H. and Kargl, F. [2018], Multi-source fusion operations in subjective logic, *in* '2018 21st International Conference on Information Fusion (FUSION)', IEEE, pp. 1990–1997.

Van Rijsbergen, C. J. [1986], 'A non-classical logic for information retrieval', *The computer journal* **29**(6), 481–485.

Vollaard, B. and van Ours, J. C. [2022], 'Bias in expert product reviews', *Journal of Economic Behavior & Organization* **202**, 105–118.

Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J. and Tounsi, L. [2014], Dcu: Aspect-based polarity classification for semeval task 4., *in* 'SemEval@ COLING', pp. 223–229.

Wang, Y., Lu, X. and Tan, Y. [2018], 'Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines', *Electronic Commerce Research and Applications* **29**, 1–11.

Wawer, A., Nielek, R. and Wierzbicki, A. [2014], Predicting webpage credibility using linguistic features, *in* 'Proceedings of the 23rd international conference on world wide web', pp. 1135–1140.

Weerkamp, W. and de Rijke, M. [2012], 'Credibility-inspired ranking for blog post retrieval', *Information retrieval* **15**(3), 243–277.

Wen, M., Zhao, H. and Xu, Z. [2019], 'Hesitant fuzzy lukasiewicz implication operation and its application to alternatives' sorting and clustering analysis', *Soft Computing* **23**(2), 393–405.

Xu, X. [2020], 'Examining an asymmetric effect between online customer reviews emphasis and overall satisfaction determinants', *Journal of Business Research* **106**, 196–210.

Xu, Z. [2007], 'Intuitionistic fuzzy aggregation operators', *IEEE Transactions on fuzzy systems* **15**(6), 1179–1187.

Yang, X., Yang, G. and Wu, J. [2016], 'Integrating rich and heterogeneous information to design a ranking system for multiple products', *Decision Support Systems* **84**, 117–133.

Yoon, K. [1980], *Systems selection by multiple attribute decision making*, Kansas State University.

Zhang, D., Li, Y. and Wu, C. [2020], 'An extended todim method to rank products with online reviews under intuitionistic fuzzy environment', *Journal of the Operational Research Society* **71**(2), 322–334.

Zhang, D., Wu, C. and Liu, J. [2020], 'Ranking products with online reviews: A novel method based on hesitant fuzzy set and sentiment word framework', *Journal of the Operational Research Society* **71**(3), 528–542.

Zhang, K., Narayanan, R. and Choudhary, A. [2010], Voice of the customers: mining online customer reviews for product feature-based ranking, *in* '3rd Workshop on Online Social Networks (WOSN 2010)'.

Zhang, R., Sha, C. F., Zhou, M. Q. and Zhou, A. Y. [2012], Credibility-based product ranking for c2c transactions, *in* 'Proceedings of the 21st ACM international conference on Information and knowledge management', pp. 2149–2153.

Zhao, Y., Xu, X. and Wang, M. [2019], 'Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews', *International Journal of Hospitality Management* **76**, 111–121.

Zhou, D., Lawless, S. and Wade, V. [2012], 'Improving search via personalized query expansion using social media', *Information retrieval* **15**(3), 218–242.

Zhou, F., Jin, J., Du, X., Zhang, B. and Yin, X. [2017], A calculation method for social network user credibility, *in* '2017 IEEE International Conference on Communications (ICC)', IEEE, pp. 1–6.

Zhou, S.-M., Chiclana, F., John, R. I. and Garibaldi, J. M. [2008], On properties of type-1 owa operators in aggregating uncertain information for soft decision making, *in* 'Proceedings of IPMU', Vol. 8, p. 1369.

Zhou, W. and Duan, W. [2012], 'Online user reviews, product variety, and the long tail: An empirical investigation on online software downloads', *Electronic Commerce Research and Applications* **11**(3), 275–289.

DOI: 10.17185/duepublico/78830
URN: urn:nbn:de:hbz:465-20230821-152927-1