SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

INFORMATION, SYSTEMS, COMMUNICATION

Dottorato di Ricerca in / PhD program COMPUTER SCIENCE Ciclo / Cycle XXXIII

Curriculum in (se presente / if it is)

MULTIDIMENSIONAL RELEVANCE IN TASK-SPECIFIC RETRIEVAL

Cognome / Surname PUTRI Nome / Name DIVI GALIH PRASETYO

Matricola / Registration number 827248

Tutore / Tutor: Prof. Stefania Bandini

Cotutore / Co-tutor:
(se presente / if there is one)

Supervisor: Prof. Gabriella Pasi
(se presente / if there is one)

Coordinatore / Coordinator: Prof. Leonardo Mariani

ANNO ACCADEMICO / ACADEMIC YEAR 2020/2021

Universita degli Studi di Milano-Bicocca,
Dipartimento di Informatica Sistemistica e Comunicazione

MULTIDIMENSIONAL RELEVANCE IN TASK-SPECIFIC RETRIEVAL

Ph.D dissertation of

Divi Galih Prasetyo Putri

Supervisor: Prof. Gabriella Pasi

Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Computer Science,

2021

Abstract Relevance is the core notion in Information Retrieval. Several criteria of relevance have been proposed in the literature. Relevance criteria are strongly related to the search task. Thus, it is important to employ the criteria that are useful for the considered search task. This research explores the concept of multidimensional relevance in a specific search-task. In the first phase of this PhD thesis, we aim to investigate the relationship between the search tasks and the considered relevance dimensions. We performed an exploratory study on different search tasks in the Microblog search context, and we identify some related relevance dimensions. Our findings show that there is a relation between a task and specific relevance dimensions. This suggests that in different search-tasks, some relevance dimensions should be prioritized while others should not be considered. In the second part, we propose an approach that can be used to combine more than one relevance dimension. In particular, given that recent advancements in deep neural networks enable several learning tasks to be solved simultaneously, we examine the possibility of modeling multidimensional relevance by jointly solving a retrieval task, to learn topical relevance, and a classification task, to learn additional relevance dimensions. To instantiate and evaluate the proposed model, we consider three query-independent relevance dimensions beyond topicality, i.e., readability, trustworthiness, and credibility. The findings show that the proposed joint modeling can improve the performance of the retrieval task.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Overview

We live in an era where the information available on the web is overwhelming. Web articles and social media contents, including texts, photos, and videos contribute to the vast amount of data created in the digital sphere. The information available on the web is beyond the user's ability to process and manage. However, the abundance and diversified information allow users to fulfill any information needs. Various type of information is available on the web from recent news, people's reviews of a movie or restaurant, even information related to treatment for a health problem. Access to information has a significant role for internet users in their decision-making processes and outcomes. Using a search engine, users can find desired information or usually called relevant documents[1] with respect to the information need without going through non-relevant web pages. The focus of an information retrieval system is the so-called *relevance*. Relevance, as the core notion in Information retrieval, is a complex subject. The multidimensionality of relevance has been discussed for a long time by the research community. Topicality is the basic relevance criterion, but it constitutes just one facet of relevance. In some cases, a topically relevant document is not useful for the user because it is hard to read or

[1] In Information Retrieval, documents can indicate various things, including web pages, videos, pictures, or music

not credible. The overall relevance score is based on the computation of the considered criteria or the relevance dimensions. Many aspects can impact the considered criteria, such as context and situation. Thus, it is essential to take into account the task that a user needs to accomplish in the information searching process.

The development of a task-specific retrieval system requires a more detailed analysis of the search objectives compared to a general-purpose retrieval system to accommodate the task at hand. The characteristics of the content and the behaviour of the user should be taken into consideration when designing a task-specific retrieval system. Thus, the main objective of the first phase of the current PhD is to study the relationship between search tasks and the considered criteria of relevance. We investigate the impact of the identified relevance dimensions on the system effectiveness, with respect to the considered search tasks. The second part of the research work is related to the aggregation problem of the criteria. Specifically, we address the approach that can be used to combine the considered relevance dimensions for task-based retrieval. We investigate the possibility of incorporating several relevance dimensions beyond topicality in the neural ranking approach. All proposed approaches is evaluated on a case study associated with a specific search task.

1.2 Context and Motivation

The objective of an information retrieval system is to provide the user with an easier way to access the information related to their interest [Baeza-Yates et al., 1999]. The system evaluates the expected utility of a document, usually called relevance, with respect to the user's query representing the information needs. Then, the system will return the user a ranked list of documents in the order of their relevance. The information should have a utility which means it can solve the given problems and fit into the goal of the information searching process. Hence, relevance is considered as the core of a retrieval system, yet it is a complex notion.

A lot of research have described the definition of relevance (See Section 2.1.2 for

details) [Bates, 1979, Mizzaro, 1997a, Saracevic, 1975, 1996]. It has been addressed in numerous studies concluding that relevance is a (i) multidimensional concept that has to be considered from the user's perspectives and search situation (ii) dynamic concept which may change over time and involves some selection (iii) complex concept but measurable. The relevance judgement of a document made by the user can be based on several criteria. Thus, the importance of the dynamic and multidimensional nature of relevance becomes the motivation of this PhD Thesis. Commonly, the topical matching signals of document concerning the user's information need are used as criteria to assess relevance, while other criteria are being excluded from the information seeking task. However, the other criteria beside topical relevance can be crucial in some cases. These relevance criteria are also called relevance dimensions. A wide variety of relevance dimensions has been introduced in the literature (See Section 2.1.2 for details) such as understandability, novelty and reliability. In different search contexts or domains, researchers used different sets of criteria based on what criteria they consider to be effective in solving their problem. The relevance criteria may also be different for each user. The information searching process is related to some aspects, including the user's situations, goals, and problems [Belkin and Kwaśnik, 1986], that are diverse and require different types of information. It is necessary to focus on the task behind an information searching process to facilitate user's situations, help achieve their goals, and solve their problems. Thus, by considering the search task, we can build a more effective and suitable retrieval system. Specific search tasks with specific goals require a deeper analysis of the problems than the general retrieval system. Thus, the search task also serves as the motivation of the present PhD theses. In this thesis, we focus on the concept of multidimensional relevance in specific search tasks. We study the impact of distinct relevance dimensions on different search tasks and propose an approach to combine several relevance criteria beyond relevance.

1.3 Research Objectives and Challenges

The objectives reported in this section represent the main challenges of this thesis. The objectives are as follows, (i) to study the relevance criteria and their importance in different search tasks and (ii) to explore the method to combine several relevance dimensions.

The first phase of the PhD is concerned with the exploratory analysis of relevance criteria in a task-based retrieval. The objective of this work is to study and investigate the correlation between relevance dimension and the effectiveness of the retrieval system in relation to the search tasks. The challenge in this work is related to identifying the impact of a distinct relevance dimension on different search tasks. The system has to provide the user with relevant documents related to the goal of the specific task. Another challenge is to handle the peculiar characteristic of the document. As a case study, we have considered the task of Microblog Search where the User-Generated-Content (UGC) has unique characteristics. The characteristics of the content and the behaviour of the user also affect the choice of relevance dimensions. In fact, microblog posts are short texts and often contain non-homogeneous linguistic quality as users tend to employ more abbreviations, acronyms, remove vowels and articles. Thus, preprocessing the data while maintaining useful information becomes a challenge. Besides, rumours and misinformation that are intentionally created to mislead others can be spread rapidly on social media. The retrieval system should also handle this issue to provide the user with credible and relevant documents.

The second phase of the thesis is focused on the issue of aggregation in multidimensional relevance, specifically on producing a single final relevance score based on the considered relevance criteria. The objective of the work of this phase was to model multidimensional relevance in the neural ranking approach. In the research area of neural ranking, existing works mainly focused on the matching or the interaction between query and document, which related to topical relevance. Thus, the challenge is to incorporate the relevance dimensions beyond topicality in the neural network. However, neural models are data-hungry and their performance depends on the size of data training. For this

challenge, we propose to adopt multi-task approach that jointly learns between retrieval task and classification task of the considered relevance dimensions. For this work, we consider the task of *Consumer Health Search* as a case study. Another challenge is related to the assessment of the relevance criteria. One of the processes in some works in this thesis is to compute the score of the considered relevance dimension such as credibility score and readability score. Since the assessment issue is not the objective of this study, we have adopted existing works that specifically address the task of evaluating a particular relevance dimension.

1.4 Research Contributions

The contributions in this PhD thesis are related to the challenges and objectives explained in Section 1.3. We explain the contributions of this thesis which we investigate, respectively in Section 3 and 4. The primary contribution of this thesis can be concluded as follows:

- The first contribution of this thesis is an exploratory study of the relationship between relevance criteria and search tasks. We investigate the impact of distinct criteria on different tasks. We consider two Microblog search tasks as the case study: Disaster-related search task and Opinion search task. For each of the above tasks, four relevance dimensions have been considered beyond topicality, including informativeness, interestingness, opinionatedness, and credibility. We demonstrate that specific search tasks in a microblogging context are impacted differently by distinct relevance dimensions.

- We proposed a neural model that jointly learns between retrieval task and classification task of the considered relevance dimension. In, the multi-task approach, we adopt the *hard-parameter sharing* where several hidden layers are shared between tasks, while the output layers are task-specific. We perform a series of experiments with respect to distinct relevance dimension including readability, trustworthiness, and credibility. The evaluation results show performance improvement over single

task that considers only the retrieval task (and, hence, only topicality).

1.5 List of Publications

1. Divi Galih Prasetyo Putri, Marco Viviani, Gabriella Pasi. A Multi-Task Learning Model for Multidimensional Relevance Assessment. In *The CLEF Association-12th International Conference, CLEF 2021*

2. Divi Galih Prasetyo Putri. Multidimensional Relevance in Task-Specific Retrieval. In *Advances in Information Retrieval - Doctoral Consortium at 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1.*, pages 677-681, 2021.

3. Divi Galih Prasetyo Putri, Marco Viviani, Gabriella Pasi. Social Search and Task-Related Relevance Dimensions in Microblogging Sites. In *Social Informatics - 12th International Conference, SocInfo 2020, Pisa, Italy, October 6-9, 2020, Proceedings.*, pages 297-311, 2020.

4. Divi Galih Prasetyo Putri. Multidimensional Relevance in Microblog Search. In *Proceedings of the 9th PhD Symposium on Future Directions in Information Access co-located with 12th European Summer School in Information Retrieval (ESSIR 2019), Milan, Italy, July 17-18, 2019.*, pages 36-41, 2019.

5. Divi Galih Prasetyo Putri, Gabriella Pasi. Exploring Relevance in Microblog Search. In *Proceedings of the 10th Italian Information Retrieval Workshop, Padova, Italy, September 16-18, 2019.*, pages 8-9, 2019.

1.6 Thesis Outline

Besides the introduction presented in this Section, the rest of the thesis is organized into 5 parts. The rest of the thesis is organized as follows:

- Section 2 contains the related works of this thesis. We briefly explain the field of information retrieval, the concept of relevance, and search task in information retrieval. Then, we provide an overview of the state-of-the-art research related to multidimensional relevance, task-based information retrieval, and several concepts related to the case studies exploited in this thesis.

- In Section 3, we present the study of the relationship between search task and relevance dimension. We explain in detail the definition and method to assess each of the considered relevance dimension. We focus on exploring the impact of distinct relevance dimension in different tasks. We report the proposed approach and the findings.

- In Section 4, we present the approach that we propose to model multidimensional relevance. We attempt to incorporate multidimensional relevance on the neural ranking model. We introduce the background information on the task and present the proposed model. Lastly, we show the results and findings.

- Finally, in Section 5 we conclude the works presented in this thesis. We also draw some future works direction and opportunities.

Chapter 2

Literature Review and Background

In this section, we discuss an overview of the research area related to this thesis. Firstly in Section 2.1, we explain information retrieval and the aspects by which information can be managed and processed to provide the user with relevant information related to their information needs. In Section 2.1.2, we present the concept of relevance in information retrieval and summarize the existing works especially regarding the multidimensionality of relevance, which is the core of the works in this thesis. Next, we describe the different retrieval models used in this thesis. As explained in Section 1, the focus of this PhD thesis is to explore the relationship between relevance criteria and search task. Thus, in Section 2.1.4 we explain the background information of task and task-based information retrieval. Finally, in Section 2.3 we present the evaluation measures used in this thesis to evaluate the effectiveness of the retrieval systems.

2.1 Information Retrieval

Schütze et al. [2008] defined Information Retrieval (IR) as follows:

"IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)"

The field of IR is related to the representation, storage, organization, access, analysis of information [Salton, 1968]. Information Retrieval system often handles textual documents, but multimedia documents can be managed as well. In this PhD work, we focus on text retrieval in IR. The difference between an IR system and a Database Management System is that the textual documents are commonly unstructured. Some examples of textual documents are articles, web pages, textual user-generated content, emails, etc. In a classical IR system (also named *search engine*), a user motivated by an information need uses the search engine to access relevant documents from the stored collection of documents. An information need is defined by Croft et al. [2010] as *" the underlying cause of the query that a person submits to a search engine".* It is expressed by means of a formal query, usually denoted by a few keywords. The objective of an IR system is to assess the relevance of documents with respect to a user query and to provide the user with relevant documents from the repository. The retrieval system's effectiveness is evaluated based on its capability to satisfy the searcher and retrieve relevant documents. Hence, the main research issues in this field are related to the *document and query representation*, *relevance assessment*, and *evaluation*. Figure 2.1 displays the basic processes of an information retrieval system including: (i) representing the document, (ii) representing the user's information need, and (iii) comparing both representation. It is necessary to define a formal surrogate of a text in order to efficiently access and process the documents that are stored in the data collection. An IR system commonly indexes the documents in advance. The index creation process is an offline process that does not directly involve the end-user. The user's information need will be transformed into a structured query representation and processed by the search engine. This process is called the *query formulation process*. In the matching process, the formal representation of the document and the query are compared to assessed the document with respect to the query and to determine the documents that will be returned to the user. The determination of a document's relevance is the ultimate goal of any search system, mainly relaying on the satisfaction of the topical criterion. However, relevance is a multidimensional concept [Schamber and Eisenberg, 1988]. It is based on the factors that affect the user's judge-

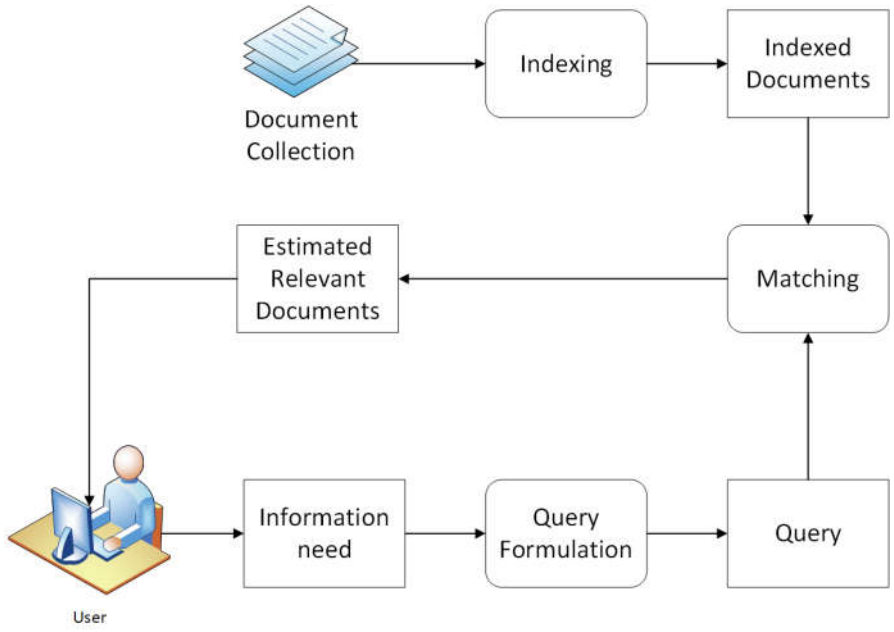

Figure 2.1: Information Retrieval System.

ment when determining whether a document is relevant or not. These factors have to be considered when designing a model to assess the overall relevance of a document. As a result, the matching process produces the final relevance score of each document and a ranked list of the document will be presented to the user. The user will try to infer how well the retrieved documents can support the search task completion. Hence, it is necessary to consider the search task in order to understand the user's information need and design an effective retrieval system. The role of multiple relevance criteria in relation to a specific search task being done is the main topic of interests of this thesis: an IR system should provide documents that can help users to solve their problem related to the task. The development of a task-specific retrieval system is challenging as it requires a deeper understanding and analysis of the factors that affect the relevance judgment. In Section 2.1.2, we will introduce the definitions of multiple relevance criteria. The background information related to the search task is presented in Section 2.1.4.

2.1.1 Indexing

"The usual meaning of indexing is to build a data structure that will allow quick searching of the text" [Baeza-Yates et al., 1999].

It is the process of generating a formal representation of the document. This task involves several processes including data acquisition, data transformation, and index creation [Croft et al., 2010]. First, the documents that will be indexed should be acquired. The documents can be obtained from an existing collections or by crawling the web. It is necessary in order to obtain the features that are useful to describe the document contents, which are called the *index terms*. *Index term* is the part of a document stored in the index and used in the search process, which commonly is the word. However, not all of the words or phrases are good index terms and preprocessing steps have to be applied to the document. The preprocessing steps usually include tokenization, stopwords removal and stemming. Later in the query processing step, the same transformation will be applied to the query. Tokenization is the process of transforming the texts into tokens, which in many cases are constituted by words. The challenges are how to deal with the special characters and the different syntax of the document. The stopword removal is the process of eliminating common words from the stream of tokens. By removing these words, we can reduce the size of the index and also help to improve the performance of the retrieval system. Lastly, stemming is the process of reducing the words to their word stem. The most popular data structure in IR is the inverted index. The input of the indexing creation process is a list of pairs of document identifiers and tokens for each document. The following procedure involves merging terms that appear more than one in a document and grouping terms that appear in several documents. The data structure in an inverted index contains a *dictionary* of terms and *posting lists*. For each term in the *dictionary*, we have a pointer to a *posting list* that contains the document identifier in which the term occurs.

2.1.2 Relevance in Information Retrieval

In information retrieval, the concept of relevance has been investigated extensively since the early years. There are two definitions of relevance based on previous studies [Saracevic, 1975, Swanson, 1986, Harter, 1992]: system-based relevance (objective) and user-based relevance (subjective). System-based relevance is an objective concept based on the objective assessment of topical relevance. In contrast, user-based relevance is a subjective notion based on the user's perspective of the information and situation. The entities of relevance can be broken down into several components, including topic, task, and context [Mizzaro, 1997b]. Topic component indicates the topic that user interested in, a task is the intention of the user while doing the search. Other factors that affecting the search result is represented in the context component. For example, a user wants to retrieve documents that he/she never read before or only the documents that are useful for some specific task. The core relevance dimensions is *topicality*, but it is insufficient to define relevance. [Schamber and Eisenberg, 1988] defined *topical relevance* as the topical matching between information retrieved by the system and the request. The term *aboutness* also has been used instead of *topicality* or *topical relevance* [Bruza et al., 2000]. Cooper [1973] suggested that topical relevance is encompassed in the concept of *utility*. Utility is a *"catch-all-concept"* and a *"cover term for whatever the user find to be the value of the system output, whether its usefulness, its entertaintment, or esthetic value, or anything else"*. The study pointed out several criteria based on document properties including novelty, informativeness, and credibility. Some works addressed that the retrieved document should be 'on-topic' before considering other criteria [Wang and Soergel, 1998, Crystal and Greenberg, 2006]. Schamber and Eisenberg [1988] defined relevance as: *"a multidimensional concept based on the human judgment process; it is dependent on both internal (cognitive) and external (situational) factors; and it is intersubjective but nevertheless systematic and measurable"*. Early studies on relevance dimensions mostly worked on identifying, analyzing, and clustering the criteria into categories. Several works performed a user study and showed that user's relevance judgement is related to much more diverse criteria than just topical relevance [Barry, 1994, Park, 1993, Schamber, 1991].

Some of the criteria identified in the literature have shown consistency across IR tasks including: *accuracy*, *effectiveness*, *amount of information*. *authority*, *novelty*, *recency*, *source quality*, *time constraint*, and *understandability*.

Some works tried to explore the factors that are affecting the choice of relevance criteria [Tombros et al., 2005, Taylor et al., 2007, Taylor, 2012, 2013]. All of them conducted a user study to address the factors related the information retrieval process itself such as search task, stage, or session. Tombros et al. [2005] investigated the criteria and aspects affecting user's relevance judgement specifically related to two situational variables: task type and task stage. The study showed the variation of features importance and criteria across tasks. In [Taylor et al., 2007, Taylor, 2012, 2013], the authors specifically studied user relevance judgement and choice of relevance criteria concerning some aspects of search. These studies found that user choice of relevance criteria is related to the task, stage, or progress of the information search process. The criteria of relevance can differ in distinct contexts, such as e-commerce [Alonso and Mizzaro, 2009], legal search [Van Opijnen and Santos, 2017] and health informtion search [Palotti et al., 2016]. Each of them tends to favor different criteria of relevance. In [Alonso and Mizzaro, 2009], the authors found that most e-commerce users consider accuracy and availability when assessing relevance. Besides topical relevance, six additional criteria are addressed in legal search, including algorithmic relevance, bibliographic relevance, cognitive relevance, situational relevance, and domain relevance [Van Opijnen and Santos, 2017]. In Microblog retrieval, several criteria of relevance have been introduced; they include credibility, informativeness, and interestingness. While in health information search, credibility and understandability have been considered in the literature as important criteria of relevance. This PhD thesis focuses on Microblog and health information on the web (Consumer Health Search) as the study cases. We present a detailed review of the works that address the task of Microblog search and Consumer Health search, respectively, in Section 2.2 and 4.2. We describe the relevance dimensions that have been addressed in both tasks.

2.1.2.1 Multidimensional Relevance

When we consider several criteria of relevance, we need to assess each relevance dimension and then combine the scores to produce a final ranking score or *Retrieval Status Value (RSV)* for each documents. Several works perform a linear combination of different criteria scores to generate the final RSV [Kang and Kim, 2003, Craswell et al., 2005, Bendersky et al., 2011]. Linear combination is a simple yet effective approach. Kang and Kim [2003] linearly combined several features including content information, link information, and URL information for web document retrieval. Craswell et al. [2005] focuses on exploring several weighting functions to transform query independent features into relevance weights. These weights will be linearly combined with BM25. They consider several features including PageRank, URL length, and click Distance. Instead of *link-analysis*, Bendersky et al. [2011] focuses on incorporating the document's content quality, which is calculated based on the content-related features, readability, and trustworthiness of the content. The authors include the document quality features into the query-document features via linear combination. In Gerani et al. [2012], the authors addresses the issue of the incompatibility between relevance scores associated with topicality and *opinion* scores in a linear combination function and compared several score transformation approaches.

In such a multidimensional context, the global relevance score can be obtained based on the aggregation of the distinct numerical assessments associated with the considered criteria. Thus, it can be seen as an instance of a Multi-Criteria Decision Making (MCDM) problem. Some research has proposed to adopt MCDM approaches in the IR context [da Costa Pereira et al., 2009b,a, 2012, Moulahi et al., 2014a,b]. An overview of the works using *the aggregation approach* in the context of IR can be found in [Marrara et al., 2017]. da Costa Pereira et al. [2009b,a, 2012] propose a prioritized aggregation scheme to combine several dimensions in the personalized IR task. The authors consider four relevance dimensions; such as *aboutness*, *coverage*, *appropriateness*, and *reliability*. In this case, the weight of each criteria is computed based on the order of user's preference over the relevance criteria. The authors proposed two aggregation operators called the

"scoring" and *"and"* operator. Several works propose *Copula-based* [Eickhoff et al., 2013, Eickhoff and de Vries, 2014] and *Choquet-based* [Moulahi et al., 2014a,b], respectively to accentuate, the non-linear and the inter dependencies among the considered relevance criteria.

Learning-to-rank (L2R) is a method based on machine learning widely used in IR with the objective of optimizing documents ranking. It combines multiple features including query-dependent features (for example BM25 score) and query-independent score (for example PageRank score). L2R approach also has been used to handle multidimensional relevance aggregation problems by considering different relevance dimensions as features. In addition to the features derived from traditional retrieval model, [Palotti et al., 2016] and [Palotti et al., 2019a] propose to include some features related to the understandability of the document into feature matrix representation. L2R exhibits the best performance compared to other methods used to integrate understandability into retrieval methods including re-ranking and ranking fusion. Another study tried to optimize the importance of the considered relevance dimension by applying a multi-objective technique in the learning-to-rank setting [van Doorn et al., 2016]. The results show that learning a set of rankers that offer the best available trade-offs between considered relevance criteria can improve the performance of the retrieval system. Their approach has been evaluated within two IR tasks, namely the Consumer Health Search task and Web Track diversity task, by balancing topicality with readability and diversity. Recent works used L2R and exploited some features related to the considered relevance dimensions, including *interest*, *habit*, *reliability*, *novelty*, *topicality*, *scope*, *understandability* [Li et al., 2017, Uprety et al., 2018, Su et al., 2018]. In [Uprety et al., 2018], the authors proposed a model based on quantum theory to capture the dynamic aspect of relevance dimensions as the search session progresses. They used LambdaMART to produce seven relevance scores, one for each relevance criteria for each query-document pairs. A set of features related to a specific relevance dimension are extracted and integrated into the model. The features adopted from the previous study [Li et al., 2017].

2.1.3 Retrieval Models

One of the primary goals of research on IR field is to learn how human decides that a piece of text is relevant to her information need. Researchers have proposed mathematical retrieval models to reason and understand the behaviour of information searchers. The model should provide results that correspond to the human decision on relevance. In another way, a retrieval model control how the relevance of a document to a user query is defined. The core of IR model is the assessment of topical relevance between query and document.

2.1.3.1 Traditional Retrieval Models

In this section, we introduce the basic IR models.

- Boolean Model

  The *boolean retrieval model* is the very first retrieval model; it is based on set theory and boolean algebra [Jones and Willett, 1997], also known as *exact-match retrieval*. The query is in the form of boolean expressions of terms that used operators from classical logic to combine the terms. The retrieval model assesses relevance as a binary property of documents with respect to a query, and as such it does not produce a ranking. For example, the query **sport** will return all documents that are indexed with the index term **sport**. All of the retrieved documents are considered equally without considering the term frequency. Using the logical product operator (**AND**), the user might want to narrow the search scope. The query **sport AND movie** will retrieve a set of documents indexed with both the terms **sport** and **movie**. If we use the operator **OR**, we will obtain a set of documents indexed with either the term **sport** or **movie**. This model is very straightforward and can be easily interpreted. It gives the user control over the system. It is evident how such results are produced based on the query. If the results are not as expected, the user can easily extend the query with proximity operators and wildcard. However, this model comes with several disadvantages. It does not provide the results in ranked

order. Thus, it does not fit the need for a retrieval system. Besides, users have to use faceted queries to retrieve documents containing similar terms or synonyms.

- Vector Space Model

  The *Vector Space Model* is based on linear algebra and it was introduced by Salton et al. [1975]. It enables term weighting and ranking. In the model, each document d and query q are represented in an n-dimensional vector space:

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{in}), Q = (q_1, q_2, \ldots, q_n) \tag{2.1}$$

  where n is the total number of index terms and d_{ij} is the weight of the ith term. One of the weighting functions is the *term-frequency (tf)* and *inverse document frequency (idf)*. Relevance is modelled on the proximity between both vector representations. Proximity can be measured in terms of similarity. To quantify the similarity, we can use the magnitude of the vector difference between the two vectors. However, the query and document length might affect the result. For example, when a relevant document is much longer than the query the difference of the two vectors can be large. The most commonly used similarity measure is cosine similarity which is based on the cosine of the angle between the query and the document vectors:

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{n} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{n} d_{ij}^2} \cdot \sqrt{\sum_{j=1}^{n} q_j^2}} \tag{2.2}$$

  where the numerator is dot product of the two vectors. The denominator normalizes the score by dividing by the product of their euclidean lengths.

- Probabilistic Models

  Probabilistic models in IR have been proposed to handle the uncertainty of relevance between query and document. The model is originated by the Probability Ranking Principle [Robertson, 1977]:

  "If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately

as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

Thus, the ideal ranking of document should be in the order of the probability of relevance with respect to the information need. The basic probabilistic model is called Binary Independence Model (BIM). It is the first ranking model based on the probability theory that considers the terms in a document as independent with each other.

– BM25

BM25 is one of the most popular retrieval models which extend the BIM with additional features including *tf.idf* weighting and document length normalization [Robertson and Zaragoza, 2009]. The query-document scoring function of BM25 can be formally defined as follows [Croft et al., 2010]:

$$BM25(q,d) = \sum_{i \in Q} log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i)/(N - n_i - R + r_i + 0.5)} \cdot \frac{(k_1 - 1)f_i}{k + f_i} \cdot \frac{(k_2 + 1)qf_i}{k_2 + qf_i} \tag{2.3}$$

where N is the total number of documents in the collection, R is the total number of relevant documents for the query, r_i is the number of relevant documents containing term i, and n_i is the number of documents containing term i. If there is no relevant information, R and r_i are set to zero. f_i and qf_i are the frequency of term i, respectively in the document and query. k_1 is a constant value that controls the term weight. The term frequency will be ignored if k_1 is zero. Otherwise, if the k_1 is large, the term weight will be increased. K is a parameter that normalized the document length. It is formally defined as:

$$K = k_1((1 - b) + b \cdot \frac{dl}{avdl}) \tag{2.4}$$

where b controls the impact of length normalization, No normalization will be done if the parameter b is set to zero.

– Language Models

Statistical language model is another IR model that is widely used in the literature. In the late 1990's, the language model started to be applied in the field of IR [Ponte and Croft, 1998]. A language model is built for each document. Query likelihood model is an approach based on the language model. It ranks the documents based on the posterior probability of the document $P(d|q)$, which can be interpreted as the likelihood of a document relevant to a query [Schütze et al., 2008]. Based on the Bayes rule, $P(d|q)$ is formally defined as:

$$P(d|q) = P(q|d)P(d)/P(q) \tag{2.5}$$

where $P(d)$ is the prior probability of a document that is usually considered the same across all documents. $P(q)$ is also uniform in all of the documents. So, both $P(d)$ and $P(q)$ could be ignored. However, some documents might not contain the query terms. A smoothing technique is essential to avoid zero probability as it assigns a non-zero probability for unseen terms. Some smoothing techniques used in the literature are Jelinek-Mercer and Dirichlet smoothing [Zhai and Lafferty, 2004].

2.1.3.2 Neural Ranking Models

Deep Neural Networks (DNN) have been used in many tasks, including adhoc-retrieval, question answering, community question answering, and automatic conversation. Generally, the primary process of document ranking includes generating the query and document representations and the matching process between two representations to estimate relevance. The existing neural ranking architecture can be divided into two categories: *representation-focused* and *interaction-focused architecture* [Mitra and Craswell, 2017, Guo et al., 2020]. In representation-focused architecture, the main objective is to design a function to obtain the best representations of the input sequences. As illustrated in Figure 2.2, the models in this architecture build a representation function to produce a high-level representations of both inputs. Commonly, the model uses a deep

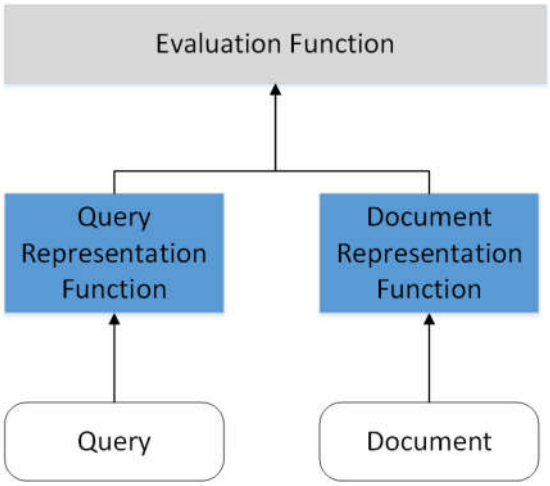

Figure 2.2: Representation-focused architecture

neural network as the representation function and a simple matching function (such as cosine similarity) to compute the final relevance score. Some existing models are designed based on the representation-focused architecture, such as DSSM [Huang et al., 2013], Arc-I [Hu et al., 2014], CNTN [Qiu and Huang, 2015], CLSM Shen et al. [2014], LSTM-RNN [Palangi et al., 2016], and MV-LSTM [Wan et al., 2016]. These models implemented different deep network structures. A fully-connected network is used in DSSM as the representation function. Some works proposed to implement convolutional network, including Arc-I, CNTN, CLSM. Besides, recurrent networks also have been used to generate the input representations. Compared to the representation-focused architecture, interaction-focused architecture aims to obtain the matching signals at a lower level. In this type of architecture, the first step is to generate interaction between the query and document as the network's input via an interaction function (as seen in Figure 2.3). Then, a deep neural network is used as the matching function to learn the interaction and produce the final relevance score. There are two categories of interaction functions that have been introduced in the literature, called *non-parametric* and *parametric* interaction functions. In the *non-parametric interaction functions*, the function does not involve learnable parameters. Some functions evaluate the interaction between the input word vectors, such as cosine similarity function [Pang et al., 2017], dot-product [Fan et al.,

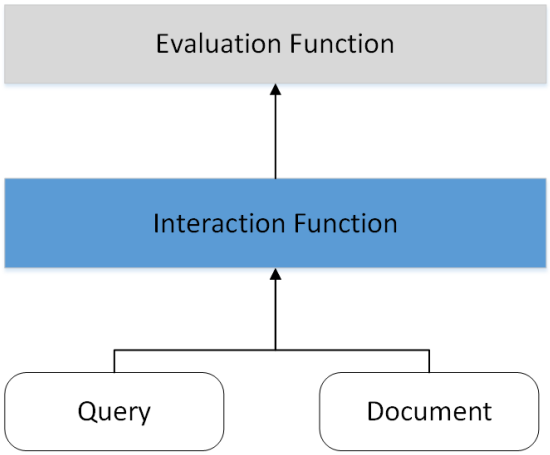

Figure 2.3: Interaction-focused architecture

2018], and radial-basis function [Pang et al., 2016]. Other functions are determined between a word vector and a set of word vectors, e.g. the matching histogram mapping [Guo et al., 2016] and the kernel pooling layer [Xiong et al., 2017]. In the *parametric interaction function* category, the interaction is learned from the data. For example, some works employed 1D Convolutional layer [Hu et al., 2014] and BERT-based model [Yang et al., 2019].

In [Guo et al., 2016, Rao et al., 2019], in particular, the authors introduced the concepts of *semantic matching* and *relevance matching*. The deep learning model based on *semantic matching*, which assess the semantic meaning and matching degree between two texts, has proven to be ineffective for the IR task. The authors suggested that the neural ranking model should incorporate *relevance matching*. The focus is to produce a ranking of retrieved documents that are relevant to a user's information need. In relevance matching, several factors must be taken into account by the neural model, including the exact matching between query and document terms, and the query terms importance. Moreover, in Information Retrieval, the query is relatively short and the document might only partially match the query.

2.1.4 Task-Based Information Retrieval

A search engine is expected to help the user find and identify useful information to help its user to complete their tasks. A task, as addressed in [Hackos and Redish, 1998, Hansen, 1999, Vakkari, 2003], is *"an activity to be performed to accomplish a goal."* Task can be considered as an abstract concept and may include more specific and smaller sub-tasks. From a functional point of view, a task consists of several actions taken to achieve a goal. Researchers have introduced several classification schemes to classify tasks [Marchionini, 1989, Qiu, 1993, Ingwersen and Järvelin, 2006]. For example, Qiu [1993] and Marchionini [1989] categorize task based on the product or the output of the information seeking process. Qiu [1993] focuses on general task and specific task. While Ingwersen and Järvelin [2006] takes a different perspective of the task by taking into account the subjective task complexity and classify tasks as routine, normal, and genuine search task. However, each of these studies focuses on specific aspects of task. Li and Belkin [2008] reviews and analyse the classification schemes introduced in the previous studies. The authors propose to develop a generally applicable classification of task based on the task's facets. Several facets are defined by considering the task's external characteristics including *source of the task*, *task doer*, *time*, *process*, *product*, and *goal*. One can consider the internal characteristics of the task depending on the complexity, difficulty, interdependence, degree of structure, salience, degree of urgency, and knowledge. In this PhD thesis, we pay particular attention to the goal of the IR task. Based on the quantity, a task can have one goal (single-goal) or more than one goal (multi-goal). A task can also be classified based on the goal's quality. It can be task with explicit and concrete goals (specific), abstract goals (amorphous), and a combination of both (mixed). We consider several case studies which is related to a specific search task with a concrete goals. First, we consider the search tasks in Microblog Search. As the topic and type of contents in Microblog are extremely varied, research of IR in Microblog is not only limited to general tweet or content retrieval. Several works focused on specific types of content such as user opinion [Luo et al., 2015] [Giachanou et al., 2016], question-answering [Herrera et al., 2018], disaster-related retrieval [Basu et al., 2020], and cultural

event retrieval [Goeuriot et al., 2018].

2.2 Microblog Search and Related Relevance Dimensions

Microblog is a service that is widely used by both personal and organization. It allows the user of the service to write any content as a post including news, personal opinion, or event updates. Since 2008, the Twitter team has been focused on developing an efficient search engine for Twitter data and providing users with the ability to search for the most relevant Tweets.[1] In Microblog, user search motivation is different compared to a web search [Teevan et al., 2011]. Users tend to use Microblog to find time-related information (such as news or event) and social-related information (such as user's information or trend). The concept of multidimensional relevance also applies to Microblog search. The UGC peculiar characteristics and the search motivation influence the choice of relevance criteria. In Table 2.1, we summarize relevance criteria studied related to Microblog search in addition to *topicality*. Beside *topicality*, there are seven dimensions has been studied in the previous studies namely *informativeness*, *interestingness*, *temporal aspect*, *spatial aspect*, *credibility/authority*, *opinionatedness*, and *popularity*.

1. Temporal Aspect

   We can see from the table that a lot of works try to incorporate temporal aspects in their system. Time is one of the information carried by each Microblog post. The temporal aspect is important for searching real-time information in Microblog streams. The *timestamp* of the content itself was used as a feature in the classification task [Vosecky et al., 2012]. In some tasks or even queries, *recency* becomes essential because user more likely favors recent content. Previous studies tried to incorporate *recency* in the query-likelihood language modeling approach as the prior probability of a document [Massoudi et al., 2011, Efron and Golovchinsky, 2011]. In another work, *recency* is exploited as one of the features in the machine

[1] https://blog.twitter.com/engineering/en_us/a/2011/the-engineering-behind-twitter-s-new-search-exper
html

Table 2.1: Relevance Dimensions in Microblog Retrieval

Dimensions	**References**
Informativeness	Choi et al. [2012], Huang et al. [2012]
Interestingness	Alhadi et al. [2011], Naveed et al. [2011b],Vosecky et al. [2012]
Temporal Aspect	Derczynski et al. [2013], Lin and Efron [2013], Massoudi et al. [2011], Efron et al. [2014], Damak et al. [2013], Magnani et al. [2012], Vosecky et al. [2012], Herrera et al. [2018], Efron and Golovchinsky [2011], Choi and Croft [2012], Chen et al. [2018], Rao et al. [2017]
Spatial Aspect	Derczynski et al. [2013], Kotov et al. [2013] Kotov et al. [2015]
Credibility	Massoudi et al. [2011], Ravikumar et al. [2012]
Opinionatedness	Luo et al. [2015], Giachanou et al. [2016]
Popularity	Damak et al. [2013], Duan et al. [2010], Magnani et al. [2012]

learning approach [Damak et al., 2013]. The results showed that recency is one of the best features to define relevance. However, not all of the queries consider a recent document to be more relevant than other documents. Several studies focused on the time distribution of relevant tweets [Choi and Croft, 2012, Lin and Efron, 2013, Efron et al., 2014] because the *temporal distribution* of relevant tweets tend to group in the same time range. Choi and Croft [2012] hypothesize that user behaviour such as retweets can be used to define the relevant time of a query. The retrieval system is enhanced using a query expansion method in a pseudo relevance feedback setting by exploiting the documents retrieved in the relevant period. Chen et al. [2018] observed the query trend of Microblog TREC 2011 data and defined three temporal classes of relevant documents, time uniform, time recency, and time variant. For some queries, relevant documents are found uniformly over time (time uniform). While some only exist near query time (time recency) or clustered in a specific period (time variant). Furthermore, retrieval system that focuses on user conversation or thread considers the *time interval* and/or *average time* from the

first until the last tweet as a feature in the learning approach [Magnani et al., 2012, Herrera et al., 2018].

2. Interestingness

   Microblog contains many contents that are only interesting for the author or the authors' friends [Java et al., 2007]. A tweet is considered to match with user needs if the tweet is interesting to the user. Therefore, some studies try to exploit interestingness in the relevance assessment. Retweet has been considered an indication of interestingness [Naveed et al., 2011b, Alhadi et al., 2011]. The authors used a machine learning approach to estimate the probability of a tweet being retweeted. Naveed et al. [2011b] performed re-ranking on the top-100 retrieved relevant tweets based on the interestingness score. Besides re-ranking, Alhadi et al. [2011] also tried to incorporate interestingness in the retrieval system by filtering out tweets with scores less than the threshold. From both works, it is proven that incorporating interestingness as the quality indicator of a document can benefit the retrieval system, especially on short queries.

3. Informativeness

   Informativeness can be defined as *"the extent to which a tweet meets the general interest of people involved with or tracking the event"* [Huang et al., 2012]. Choi et al. [2012] adopted similar definition of interestingness as in [Alonso et al., 2010] as *"specific information that people might care about"*. The relation of relevance and informativeness has been studied by Choi et al. [2012]. They performed human annotation to assess whether a document is relevant or not and whether it is informative or not. The results from the analysis showed the relationship between informativeness and relevance. Moreover, retweeted tweets are tend to be more informative. Based on that, they used the probability of a document being retweeted as the indicator of informativeness. The authors argue that informativeness score represent the quality of a document. They employed a learning approach using logistic regression and exploited more than 20 features including *web-specific*

content-based feature, *microblog-specific content-based feature*, and non-RT score. Then, the score is assigned as the prior probability in a language model to combine informativeness and topical relevance in the retrieval system. Other studies define informativeness using three hypotheses: (i) 'An informative tweets are more likely posted by credible user', (ii)'tweets involving many users are more likely to be informative', and (iii) 'tweets aligned with contents of web document are more likely to be informative' [Huang et al., 2012]. First, they eliminated non-informative tweets. Then, they constructed a heterogeneous network consisting of user-tweet and web-tweet networks based on HITS. The result showed that credible users tend to post informative tweets. This information can help improve the ranking quality.

4. Credibility

   Credibility has been used in blog post retrieval [Weerkamp and De Rijke, 2008, Weerkamp and de Rijke, 2012] and web retrieval [Lewandowski, 2013]. In Microblog search, the credibility of content is considered as the indicator of document quality. Massoudi et al. [2011] adopted the quality indicators proposed by Weerkamp and De Rijke [2008] and added several Microblog features such as repost, follower, and recency. Then, the credibility score is used as the prior probability of a document in a language model approach. Ravikumar et al. [2012] propose to consider trustworthiness and popularity in the tweet ranking task. The authors build a three-layer graph by exploiting user network, the semantic similarity between tweets (content agreement), hyperlinks to and between web pages. A tweet is said to be trustworthy if more than one tweet has an agreement on the content.

5. Spatial Aspect

   Social media content mostly contains location information, whether it is from the GPS coordinates or written in the text. Derczynski et al. [2013] discussed the challenges related to spatial and temporal context of search in Twitter. The spatial aspect is essential to handle spatial keyword and local-intent searches. Spatial data in Microblog is hard to process because most of them are implicit in the

textual content. However, spatial content in Microblog can help to improve the effectiveness of the retrieval system. To incorporate spatial aspect in ranking, one can combine the textual similarity and the location proximity using a weighted sum method. Other research tried to incorporate spatial aspect in the retrieval task [Kotov et al., 2013, 2015]. The authors use the Latent Dirichlet Allocation (LDA) method to group the document into geo-specific topics. Then, the topic model is incorporated into the Language Model for retrieval.

6. Opinionatedness

   Microblog is a valuable source of public opinion on various topics. Opinionatedness as a relevance dimension is related to the opinion retrieval task. It can be defined as the likelihood of a document to express an opinion about a topic. Previous research implemented the lexicon-based approach to estimate the opinionatedness of tweets [Luo et al., 2015]. They calculated the average opinion score of certain terms. Later, the score will be combined with social features using L2R to generate document ranking. Another study estimates the opinionatedness of a document using the average opinion score and added a stylistic-based opinion score [Giachanou et al., 2016]. The final ranking of tweets is produced by combining the score from any existing IR models and the opinionatedness score of the document.

7. Popularity

   In the Microblog domain, the indicator of popularity is the retweet. The score is calculated using retweet relation based on the PageRank Algorithm. Then, they combine the popularity score with other scores to retrieve relevant documents using the L2R approach [Damak et al., 2013, Duan et al., 2010]. Other studies estimated not only content popularity but also user popularity [Magnani et al., 2012]. The content popularity score is calculated based on retweets, and the user popularity is related to the number of followers. They used a simple aggregation function to combine all the scores, including relevance, user popularity, content popularity, and time-related measures.

2.3 Evaluation in IR

In Information Retrieval, in order to evaluate the effectiveness of search engines, large test collections are generated and made publicly available. Such collections are usually constructed so that, besides the document collection, also a set of queries is provided (generally topics); for each query, the documents deemed relevant by a pool of assessors are also provided. Several evaluation measures that have been used in the literature can be categorized as *set-based* and *rank-based* evaluation measures. In this section, we present the commonly used measures as follows:

Set-based

The *set-based* measures evaluates the IR system based on an unordered set of documents for a query. The aim of the measures in this category is to show how well the system retrieved the relevant documents and filter the non-relevant documents. A general scheme for the measures can be seen in Table 2.2.

Table 2.2: Confusion matrix

	Relevant	**Non Relevant**
Retrieved	true positive (tp)	false positive (fp)
Not Retrieved	false negative (fn)	true negative (tn)

- Recall and Precision

  Precision (P) is the number of retrieved relevant documents divided by by the number of all retrieved documents.

$$P = \frac{\#\text{relevant items retrieved}}{\#\text{retrieved items}} = \frac{tp}{(tp/fp)} \tag{2.6}$$

  Recall (R) is the fraction of relevant documents that are retrieved.

$$R = \frac{\#\text{relevant items retrieved}}{\#\text{relevant items}} = \frac{tp}{(tp/fn)} \tag{2.7}$$

- Accuracy

  Accuracy (ACC) is the number of relevant document retrieved and non relevant

document that are not retrieved divided by total number of documents in the collection.

$$ACC = \frac{(tp + tn)}{(tp + fp + tn + fn)} \tag{2.8}$$

Rank-based

- Precision@k

  Precision is a way of measuring the accuracy of a system. Specifically, it is the fraction of the documents retrieved that are relevant to the user's information needs. In particular, we consider *Precision at k* (P@k), which is defined as the total number R of relevant documents from the top-k retrieved documents. It is formally defined as:

  $$\text{P@}k = \frac{R}{k} \tag{2.9}$$

- Mean Average Precision

  To compute *Mean Average Precision* (MAP), it is first necessary to compute *Average Precision* (AP), which is the mean of the precision scores after each relevant document is retrieved. AP is formally defined as:

  $$\text{AP} = \frac{1}{R} \sum_{k \in \mathcal{R}} \text{P@}k \tag{2.10}$$

  where R is the total number of relevant documents, and $\mathcal{R}$ is the set of the ranks of the relevant documents. MAP is computed as the mean of AP on the entire query set. Formally:

  $$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}_q \tag{2.11}$$

  where Q is the total number of queries in the query set, and AP_q is the average precision for the query q.

- Bpref

  The *bpref* measures considers whether relevant documents are ranked above irrelevant ones [Buckley and Voorhees, 2004, Craswell, 2009]. It is designed to be robust

to missing relevance judgments, since in large collections not all the documents can be judged due to the onerousness of this operation. It is formally defined as:

$$bpref = \frac{1}{R} \sum_{r} \left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)} \right) \tag{2.12}$$

where R is the total number of truly relevant documents, N is the number of known irrelevant documents, r is a relevant document in the ranked result list, and n is a non-relevant document retrieved by the system.

- Discounted Cumulative Gain

  The Discounted Cumulative Gain (DCG) exploit *graded relevance* value [Järvelin and Kekäläinen, 2002]. This measure is based on the assumptions that the document with higher relevance should appear in the higher ranking. The gain is accumulated from the top of ranking and penalized for the relevant document appear in the lower rank. It is formally defined as:

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{2.13}$$

  where rel_i is the relevance value of the document at position i. Since different queries might have different number of relevant document, the DCG value should be normalized ($nDCG$).

$$nDCG = \frac{DCG_k}{IDCG_k} \tag{2.14}$$

  where $IDCG@k = \sum_{i=1}^{|REL_k|} \frac{2^{rel_i}-1}{\log_2(i+1)}$ is the ideal discounted cumulative gain. REL_k denotes the list of relevant document up to rank k.

2.3.1 Multidimensional Evaluation

The available relevance judgements of the data collections, employed as ground truth, are mostly judged by external annotators without explicitly considering the search context or multiple relevance dimensions, but mostly based on topicality [Jiang et al., 2017, Verma et al., 2016]. To consider multidimensional relevance, some test collections provide additional judgements related to other relevance criteria beyond topicality. To better

assess the contribution of the additional relevance criteria to the final rankings, other measures has been introduced in the literature. In this section, we summarize some measures that are used in this thesis. The measures are as follows:

- *Understandability-biased* RBP (*uRBP*)

  uRBP is a measure based on *Rank-Biased Precision* (RBP) [Moffat and Zobel, 2008], and considers both topicality and readability in the evaluation process. The measure is formally defined as [Zuccon, 2016]:

  $$\text{uRBP}(\rho) = (1-\rho) * \sum^{k} \rho^{k-1} * r(d@k) * u(d@k) \tag{2.15}$$

  where $r(d@k)$ and $u(d@k)$ are the gains for retrieving a document at rank k, respectively considering topicality and readability. ρ is the so-called *persistence parameter* of RBP, and indicates the user's persistence in search, or the probability of moving from a document in the rank k to the next document at rank $k+1$.

- MM Framework ($\mathcal{MM}$)

  The MM Framework $\mathcal{MM}$ [Palotti et al., 2018] is computed as the *weighted harmonic mean* of the values produced by two metrics $\mathcal{M}_{rel}$ and $\mathcal{M}_{\kappa}$. Formally:

  $$\mathcal{MM} = 2 * \frac{\mathcal{M}_{rel} * \mathcal{M}_{\kappa}}{\mathcal{M}_{rel} + \mathcal{M}_{\kappa}} \tag{2.16}$$

  where $\mathcal{M}_{rel}$ and $\mathcal{M}_{\kappa}$ are any valid evaluation measures respectively for assessing (topical) relevance, and another additional criteria.

- *Convex Aggregating Measure* (CAM)

  In [Lioma et al., 2017], the authors proposed some measures that consider both the topical relevance and the credibility of the retrieved results. One of them is called *Convex Aggregating Measure* (CAM). CAM is expressed as the *convex sum* of the scores calculated for each relevance dimension, and is formally defined as:

  $$\text{CAM} = \lambda\mathcal{M}_r + (1-\lambda)\mathcal{M}_c \tag{2.17}$$

  where $\mathcal{M}_r$ and $\mathcal{M}_c$ are any valid evaluation measures respectively for assessing (topical) relevance and credibility. λ is a value in the range of [0,1].

Chapter 3

Task-related Relevance Dimensions in Microblogging Search

In general, when developing a *search engine*, the key issue is how to assess the *relevance* of a document (e.g., a textual document, a Web page, a social media post, etc.) to a user *query* that expresses the user's information needs. As explained in Chapter 2.1.2, relevance is a complex notion, which relies on several criteria or dimensions. The choice of relevance dimension depends on the *search task* and the *context* in which search is performed, which concur to define the so-called *situational relevance* [Borlund, 2003]. In this scenario, the relevance dimensions that are effective to the retrieval process are those able to capture the usefulness of the documents in helping the user to understand, make decisions, and solve the problems related to the search task or context underlying the user's information needs. Thus, our hypothesis is that specific search tasks in a microblogging context are impacted differently by distinct relevance dimensions. The main objective on the first phase of this PhD thesis is to explore the relationship between search task and the importance of distinct relevance dimension. To the aim of conducting an exploratory study, we consider two Microblog search tasks: (*i*) the *disaster-related retrieval* task, and (*ii*) the *opinion retrieval* task. For each of the above-mentioned tasks, four relevance dimensions are considered beyond topicality. For example, we can hypotesize that in the

disaster-related retrieval task both informativeness and credibility should play a major role, and that in opinion retrieval, both opinionatedness and credibility should strongly impact the process of tweet retrieval. In the following Section (Section 3.1) we present a synthetic review of the research contributions addressing the task of microblog search. In Section 3.2 we present the proposed methodology. Lastly, in Section 3.4, we present and discuss the results of the study.

3.1 Microblog Search

The spread of social media has made it possible for anyone to generate *User-Generated Content* (UGC), and this has led to the availability of a huge amount of online information (which is, in many cases, misinformation [Livraga and Viviani, 2019, Viviani and Pasi, 2017]). Among the several social platforms available, microblogging services play a fundamental role in content generation. Twitter, in particular, is incredibly popular, with more than 300 million monthly active users in the first quarter of 2019.[1] Microblogs allow users to freely share content in the form of short texts. Microblog posts can be constituted of personal updates or discussions, the so-called *conversational* content, but also by *newsworthy* content, such as news and public events [De Grandis et al., 2019]. This second category of content is of fundamental importance to users if we consider that many people turn to microblogging platforms as their main source of news. For example, from 2013 to 2017, the percentage of Twitter users employing the platform to fulfill their information needs has grown by more than twenty percent.[2] In this scenario, characterized by huge amounts of heterogeneous information, implementing *social search* services to effectively provide users with social media contents satisfying their information needs becomes crucial. Retrieving useful information in microblogs is not an easy task, and some peculiar characteristics of the UGC diffused in these platforms make information processing and analysis a challenging problem. In fact, as previously pointed out, microblog posts are short texts that do not have a homogeneous linguistic quality, as users

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter- users/

[2] https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms- 2017/

tend to employ abbreviations and acronyms, they remove vowels and articles [Gouws et al., 2011], and even shorten URLs [Tao et al., 2013]. Moreover, a lot of users spread rumour, misinformation, hoaxes, and spam. Finally, given that these platforms are used to post content that covers various information needs in different contexts, the purpose for which the search is performed is another fundamental aspect to be taken into consideration [Efron, 2011, Teevan et al., 2011]. Efron [2011] summarized some main problems and potential research opportunities in Microblog search. The main issues are related to the sentiment analysis, entity search, and modelling abstraction. Many peoples use Microblog services to express their opinions or emotions towards a certain topics. Thus, users can use the Microblog search engine to find information related to other people's opinions. For example, they can ask about "What is the current public opinion about Covid19?" In this case, the search engine should incorporate sentiment analysis and emotion detection that will affect the performance of the retrieval system. The second problem is related to the useful units of retrieval. A suitable retrieval unit with respect to user information need has to be defined. A user might be interested in a single tweet, a thread, or even a person. The user-generated metadata, such as social links and hashtags, is valuable information and could improve the retrieval system. Teevan et al. [2011] tried to study the user behaviour during the search process and how it differs from web search. The authors conducted a user study to observe the underlying goals and motivations of the user when performing an information-seeking process on Microblog. They tried to study the user behaviour during the search process and how it differs from the web search. The study showed that even though the primary motivation of microblog search is the same as web search (topically related information), users tend to find the time-related and social information very interesting.

Early work on microblog search simply tried to exploit some *features* specific to microblogging platforms in the ranking process. In [Duan et al., 2010] and [Nagmoti et al., 2010], the authors employ *learning-to-rank* [Liu, 2009] to inject in the search process some features related to the 'authoritativeness' of an account and to the network structure, such as the number of times a user is mentioned in tweets, the number of

her/his followers, the number of lists or groups a user belongs to, a popularity score estimated by the PageRank algorithm based on the retweet network, the presence of URLs in the tweets, etc. However, specific relevance dimensions and the role played by multidimensional relevance in microblog search are not analysed in the above-mentioned works. Some other works have investigated different properties that can be useful in ranking microblog posts (see Section 2.2 for details), beyond topicality; they have been presented as quality indicators, and include *interestingness* [Alhadi et al., 2011, Naveed et al., 2011b, Tao et al., 2012], *informativeness* [Choi et al., 2012, Huang et al., 2012], and *credibility* [Huang et al., 2012, Massoudi et al., 2011, Vosecky et al., 2012]. *Interestingness* has been discussed as a *static* quality measure of a microblog post. In addition to the above-mentioned properties, the so-called *opinionatedness* has been proposed in the *opinion retrieval* context [Giachanou et al., 2016, Luo et al., 2015].

Several studies in the literature focus on a specific task, such as disaster-related retrieval [Basu et al., 2020], opinion retrieval [Giachanou et al., 2016], news retrieval [Suarez et al., 2018], cultural event retrieval [Goeuriot et al., 2017]. Microblog is a valuable source of information during critical events such as disaster. The real-time UGC is very useful to assist the disaster relief and recovery operation. Several shared challenges have been organized in the context of disaster-related retrieval, including FIRE IRMiDiS 2016-2018 [Basu et al., 2020] and ECIR SMERP 2017 [Ghosh et al., 2017]. In both challenges, the focus was information extraction related to critical information during the disaster event. The goal of this retrieval task is to provide a relevant and actionable information the emergency relief operations and preparedness initiatives. The information is very crucical to help effectively conduct post-disaster relief operations. Specifically, the information needs are related to the availability of the resources (foods, shelters, medical resources, and so on), infrastructure damages, and activities of NGOs/ government. The objective of the opinion retrieval task is to understand the opinions of other users with respect to the query topic. Thus, not only topicaly relevant, the retrieved information should also contain opinion with respect to the topic [Giachanou et al., 2016]. Whereas in [Suarez et al., 2018], the authors created a TREC-like data collection that

focus on news retrieval. The task aims to rank tweets related to a published news article. The relevant tweets should contains the topic related to the news or any topics that are related to the main topic. The cultural event retrieval task has been used as the focus of the CLEF MC2 lab since 2016 until 2018[Ferro and Peters, 2019]. The organizers provided a large multilingual microblog stream of The Festival Galleries project.

Although several properties of microblogs and their impact on retrieval have been investigated in the literature, the concept of task-related multidimensional relevance has not been investigated yet. For this reason, in the first phase of this PhD work, we propose an approach to microblog search that relies on the identification of distinct relevance dimensions for different *search tasks*, and we evaluate their impact on the search effectiveness.

3.2 Exploratory Study

We consider four *relevance dimensions* that can have a possible impact on the effectiveness of the search task. In the following sections we provide a detail description on the way in which the four dimensions are modeled (Section 3.2.1) and the method to combine the relevance dimensions (Section 3.2.2) .

3.2.1 Relevance Dimensions

3.2.1.1 Interestingness.

In the literature, *interestingness* has been defined as the extent to which a content significantly arouses reactions [Webberley et al., 2016]. Microblogging services enable users to republish content that could be of interest to other users. Thus, a tweet that received a high number of retweets can be considered as of general interest [Alhadi et al., 2011, Naveed et al., 2011a,b]. We follow the same approach adopted in [Naveed et al., 2011a] to compute the *interestingness score* of a single tweet. This score is computed as the probability of a tweet being retweeted.

To do so, *logistic regression* [Hosmer Jr et al., 2013], as implemented in the `Python`

library `scikit-learn`,[3] is used to train a model on several features related to tweets, such as the presence of URLs, usernames, hashtags, exclamation marks, question marks, emoticons, sentiment. The logistic regression model learns the influence that features have on the probability of a message being retweeted; this way, after this training phase, we are able to compute the probability of a new tweet being retweeted. To train the model, a corpus of tweets is required. Because the datasets used in [Naveed et al., 2011a,b] are not publicly available, we have used the data collection introduced in [Zubiaga, 2018], in which tweets have been collected by using a set of relevant keywords and hashtags associated with 30 real-world events. From this collection, we have extracted around 20 thousand tweets. Specifically, we have selected ten thousand tweets with zero retweets and ten thousand tweets with more than 10 retweets. The tweets with zero retweets are considered as *not interesting*, while the rest as *interesting*.

3.2.1.2 Informativeness.

In [Choi et al., 2012], *informativeness* has been defined as "specific information that people might care about". To assess the informativeness of a tweet, we follow the approach proposed in [Mahata et al., 2015] in the context of crisis events. This approach has proven to be effective not only to assess the informativeness of crisis-related content, but also to other kinds of content [Imran et al., 2013]. To calculate an *informativeness score* for each tweet, we have employed the data collection introduced in [Olteanu et al., 2015]. The dataset contains tweets gathered during 26 crisis events in the period 2012-2013, which are labeled based on their level of informativeness. Labels assume binary values; the 1 value is employed if a tweets is *related and informative*,[4] and the 0 value is associated with tweets that are *related but not informative*, or *not related*. As for the case of interestingness, based on the above-mentioned training data, we have used *logistic regression* to train a model to compute the probability of informativeness of a

[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] In this case, being a *related* tweet means that the tweet discusses the crisis event.

new tweet.Also for informativeness, logistic regression has been performed by employing the model implemented by the `scikit-learn` library [Pedregosa et al., 2011], using the default parameters. We consider the same features proposed in [Mahata et al., 2015] to train the model, such as the presence of URLs, the number of words, stopwords, sentiment-related words, hashtags, user's mentions, the tweet length, and some other features identifying the content style.

3.2.1.3 Opinionatedness.

The concept of *opinionatedness* has been discussed in the context of opinion retrieval. A way to assess opinionatedness for a document is to compute both a *term-based opinion score*, based on the presence of opinionated terms in a document, and a *stylistic-based opinion score*, which focuses on some stylistic aspects of the document under consideration. We have used the definition and the approach proposed in [Giachanou et al., 2016] to compute an *opinionatedness score* associated with a tweet. The authors propose to compute this score as a linear combination of a term-based opinion score and a stylistic-based opinion score. The former is computed as the average opinion score over all terms in the document, as illustrated in [Giachanou et al., 2016], and by employing the AFINN Lexicon to identify opinionated terms [Nielsen, 2011]. The latter is computed by considering the stylistic variations that a tweet contains, as illustrated in [Giachanou et al., 2016]. The following stylistic features have been considered to capture stylistic variations: the presence of emoticons, exclamation marks, emphatic lengthening, and opinionated hashtags.

Since stylistic variations are topic-dependent in capturing opinionatedness, we have performed *topic modelling* by using *Latent Dirichlet Allocation* (LDA), as implemented in `gensim`,[5] to identify topics and related tweets. The importance of each stylistic variation is calculated for each topic based on the tweets assigned to the topic.

[5]https://radimrehurek.com/gensim/models/ldamodel.html

3.2.1.4 Credibility.

In the literature, *credibility* has been referred to as a perceived quality of the information receiver; it involves several factors related to the source of information, the information itself, and the media employed to diffuse information [Fogg and Tseng, 1999]. To assess credibility of microblog contents, we adopt the approach described in [Pasi and Viviani, 2018], which is a model-driven approach based *Multi-Criteria Decision Making* (MCDM). A microblog post (i.e., a tweet) is considered credible if it 'satisfies' several criteria, i.e., some features in the microblogging context, which can be interpreted from a credibility point of view (e.g., a tweet written by a user who has many friends can be considered as more credible w.r.t. a tweet written by a user with no friends). To compute the *credibility score* of a tweet, we exploited some features taken from the literature [Gupta et al., 2014, Mitra and Gilbert, 2015], employed in association with the model presented in [Viviani and Pasi, 2016]. The considered features include, for each tweet: the number of retweets, the number of followers, friends, tweets of the author, the presence of URLs, and the author's account age w.r.t. the date in which the tweet has been posted. Formally, with each tweet, distinct credibility scores $x_1, \ldots, x_n$ are associated, one for each feature, as illustrated in [Viviani and Pasi, 2016]; each score expresses how much the tweet is credible based on the related feature. In order to obtain an overall credibility score $\mathcal{A}(x_1, \ldots, x_n)$ for the tweet, the feature-related credibility scores must be *aggregated* through a suitable function $\mathcal{A}$. As suggested in [Pasi et al., 2020, Viviani and Pasi, 2016], *Ordered Weighted Averaging* (OWA) [Yager, 1988] aggregation operators are employed.

3.2.2 Combining Relevance Dimensions

To evaluate the impact of each relevance dimension on each of the two considered search tasks, we have implemented a simple *retrieval model*, which implements a *linear combination* of the two relevance scores related to *topicality* and an *additional relevance dimension* respectively. For each tweet, the combined relevance score (named *Retrieval*

Status Value - RSV) is then computed as follows:

$$\text{RSV} = \alpha \text{RSV}_t + (1 - \alpha)\text{RSV}_i \tag{3.1}$$

where RSV_t is the topicality score, and RSV_i is the score associated with the additional relevance dimension.

To compute the RSV_t score, we have employed the *query likelihood model* [Manning et al., 2008], which builds a probabilistic language model from each document d, and ranks documents based on the probability of the model to generate the query q. This is interpreted as the likelihood that the document is relevant to the query. Formally:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \tag{3.2}$$

where $P(q|d)$ is the probability of the query q under the language model derived from d, $P(q)$ is the same for all documents, and so can be ignored, and $P(d)$ is the prior probability of a document often treated as uniform across all d and so it can also be ignored. We have used the *Dirichlet prior smoothing* in association with the query likelihood model to compute $P(d|q)$, as illustrated in detail in [Zhai and Lafferty, 2004]. The RSV_i score is obtained by applying the approaches described in Section 3.2.1 for each relevance dimension i. The weight α takes values in the $[0, 1]$ interval, and, in the evaluation process, we tune its value by considering steps of 0.1.

The obtained overall RSV score represents the *estimated relevance* of the tweet to a query (based on multiple relevance dimensions), based on which the tweets are ranked.

3.3 Evaluating the Impact of Relevance Dimensions

We made a comparative evaluation of a baseline microblog search engine computing relevance as topicality only, according to Equation (3.2), and microblog search engines assessing relevance by combining topicality with just an additional dimension at a time, according to Equation (3.1). This section presents the experiments that have been carried out to evaluate the impact of the relevance dimensions presented in Section 3.2.1 on the disaster-related retrieval and the opinion retrieval search tasks. To perform the

evaluations, two publicly available datasets have been employed, which are described in the following section; each dataset is related to one of the two considered search tasks.

3.3.1 Datasets

To evaluate the impact of the considered relevance dimensions on the disaster-related task, a dataset has been used that was introduced at the ECIR 2017 Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017) [Ghosh et al., 2017]. This dataset is composed of around 42,000 tweets and is split into two parts. The SMERP level-1 dataset contains tweets collected on the first day of the disaster (around 28,000 tweets), while the SMERP level-2 dataset contains the tweets that are collected during the second day of the disaster (around 14,000 tweets). In both datasets, four topics are discussed, i.e., available resources, required resources, infrastructure damages, and rescue activities.[6] For the opinion retrieval task, we used the dataset introduced in [Luo et al., 2015]. The original dataset refers to 5,000 tweets and 50 topics collected in 2011. Some of the original tweets could not be crawled anymore because, in the meantime, they have been set as private or permanently deleted. So, the considered dataset contains around 3,100 tweets. In both the considered data collections, tweets are simply labelled as *relevant* or *not relevant*.

All the considered data have been preprocessed by using the Anserini IR open-source toolkit [Yang et al., 2017], and indexed using Apache Lucene.[7] In particular, during the preprocessing phase, the tweets have been tokenized and normalized. Then, stemming has been applied to all normalized tokens by using the Porter Stemmer [Porter, 2008] that is implemented in Lucene.

3.3.2 Experimental Setting

For each search task, we have conducted several experiments by implementing multiple microblog search engines; each search engine implements a model that combines top-

[6] https://www.computing.dcu.ie/~dganguly/smerp2017/

[7] https://lucene.apache.org/

icality with just an additional relevance dimension, based on Equation (3.1). Hence, we have the following 4 search engine configurations, which we denote as: *Topicality + Informativeness*, *Topicality + Interestingness*, *Topicality + Opinionatedness*, and *Topicality + Credibility*. We compare the results produced by each search engine with a *Topicality-based Baseline*, i.e., a search engine that estimates relevance based on topicality only. In order to compute the topicality score, as discussed in Section 3.2.2, we have employed the *query likelihood model* together with the *Dirichlet prior smoothing*, as implemented in Lucene. Since for the disaster-related retrieval task we have considered both datasets illustrated in Section 3.3.1, we have performed 15 evaluations, 5 for the disaster-related retrieval task w.r.t. the SMERP level-1 dataset (i.e., the results produced by the Topicality-based Baseline and by the 4 search engines combining topicality with an additional relevance dimension), 5 for the disaster-related retrieval task w.r.t. the SMERP level-2 dataset, and the remaining 5 for the opinionated tweets dataset, related to the opinion retrieval task.

To comparatively evaluate the effectiveness of the results produced by each search engine for each considered search task, we have employed the *bpref*, *Precision at k* (P@k), and *Mean Average Precision* (MAP) measures, which are the same suggested and employed in [Ghosh et al., 2017] as evaluation metrics.

3.4 Discussion

In this section, we present the results obtained by performing the evaluation strategy illustrated in Section 3.3.2. In Table 3.1, the results obtained in terms of *bpref*, P@20, and MAP are illustrated for each implemented search engine for each task. This allows to evaluate the different impact that each relevance dimension has on the different search tasks.

As explained in Section 3.2.2, we have used a linear combination of the considered relevance dimensions to get the overall relevance score of a tweet. In the performed experiments, we have set different values of α in the linear combination (i.e., Equation

Table 3.1: Results obtained for each microblog search engine w.r.t. different search tasks.

Microblog search engine	**α value**	**Evaluation measure**		
		bpref	**P@20**	**MAP**
Disaster-related Retrieval Task				
Dataset: SMERP level-1				
Topicality-based Baseline	-	0.0300	0.0875	0.0213
Topicality + Interestingness	0.8	**0.0409**	**0.1250**	**0.0251**
Topicality + Informativeness	0.6	**0.0564**	**0.1250**	**0.0247**
Topicality + Opinionatedness	0.9	0.0290	0.0875	0.0200
Topicality + Credibility	0.8	**0.0390**	**0.1000**	0.0138
Dataset: SMERP level-2				
Topicality-based Baseline	-	0.0459	0.1875	0.0215
Topicality + Interestingness	0.8	**0.0562**	0.1250	**0.0275**
Topicality + Informativeness	0.7	**0.0589**	**0.2000**	**0.0290**
Topicality + Opinionatedness	0.9	**0.0479**	0.1375	0.0213
Topicality + Credibility	0.9	**0.0488**	0.1500	**0.0222**
Opinion Retrieval Task				
Topicality-based Baseline	-	0.1398	0.1470	0.2126
Topicality + Interestingness	0.9	0.1250	0.1460	0.2041
Topicality + Informativeness	0.9	0.1227	0.1310	0.1962
Topicality + Opinionatedness	0.4	**0.2049**	**0.1910**	**0.2517**
Topicality + Credibility	0.7	**0.1492**	**0.1670**	**0.2186**

(3.1)), and we have selected, as α value, the one providing the best search engine performance with respect to each specific data collection (i.e., with respect to each search task). For each data collection, and for each search engine, the scores that are indicated in bold in Table 3.1 refer to the results that outperform the *Topicality-based Baseline* for

the considered search task.

3.4.1 Impact of distinct Relevance Dimensions

The obtained results show that despite the method we employ is quite simple, we obtain an indication of the possible impact that each of the considered relevance dimensions may have on the different search tasks. In fact, we observe that the search engines whose retrieval model is based on the combination of *Topicality* + *Interestingness*, *Topicality* + *Informativeness*, and *Topicality* + *Credibility*, perform better than the *Topicality-based Baseline* in the disaster-related retrieval task, for both SMERP level-1 and level-2 datasets. When considering the opinion retrieval task, it can be noticed that neither informativeness nor interestingness as relevance dimensions have a significant impact on the results. On the contrary, we can observe that using credibility and opinionatedness as relevance dimensions is useful for the opinion retrieval task.

These preliminary results confirm our hypothesis that for a given search task some specific dimensions may have an impact while others could be not considered. Thus, the impactful dimensions should be emphasised by the retrieval process to improve the effectiveness of search for a given task.

Chapter 4

A Multi-Task Learning Model for Multidimensional Relevance Assessment

The results of the exploratory study implemented in Section 3 show the distinct importance of a relevance dimension in different search task. The findings prove the relationship between the search task and relevance dimensions. Thus, it is essential to pay attention to the process of defining the considered relevance dimensions in order to design an information retrieval system. The next essential process is to model and aggregate the relevance dimensions. The model has to consider the preference of a specific search task over the relevance dimension.

In Section 3, we employ a simple linear combination strategy. We propose to study other approaches to combine the relevance dimensions and model the importance. Following the success of the neural ranking approaches in information retrieval, we wish to model the concept of multidimensional relevance via a multi-task neural network. Section 4.1 presents relevant background concept related to Multi-task learning and a brief review of prior works on the adoption of the multi-task learning in the Information Retrieval field. In Section 4.2, we summarize recent works related to Consumer Health Search,

which is the case study in this work. Then, we introduce the proposed methodology in Section 4.3. In Section 4.4 and Section 4.5, we describe the experimental evaluation setting and the results of the experiment.

4.1 Background

4.1.1 Multi-task Learning

MTL has proven effective in many fields from computer vision [Vandenhende et al., 2021], speech recognition [Pironkov et al., 2016], to Natural Languange Processing [Worsham and Kalita, 2020]. Multi-task learning (MTL) is a machine learning paradigm introduced by Caruana [1997]. The authors stated that *"MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks"*. The human learning paradigm is the inspiration of Multi-task learning. Human often applies their knowledge to learn similar tasks. For example, we can use our knowledge and skill of riding a bicycle to learn to ride a motorcycle. The main objective of MTL is to jointly solve several tasks by taking advantage of useful information from other related learning tasks. Initially, the motivation of MTL is to handle data sparsity problem where the number of labeled data is insufficient to train an effective classifier. MTL combines the data from all of the tasks to obtain better performance for each task. Thus, it can reduce the cost of data labeling. When a large amount of data is available, more robust and universal representations for multiple tasks can be learned. Thus, MTL enable knowledge sharing among tasks, reducing overfitting to a specific task and improving performance of each task.

In the next section (Section 4.1.1.1), we introduce the MTL methods in Deep Neural Network (DNN). We also summarize recent works on the IR field that applied MTL approaches (Section 4.1.1.2).

4.1.1.1 Parameter sharing

In the DNN context, there are two commonly used approach of multi-task learning called *soft-parameter sharing* and *hard-parameter sharing*.

In **hard-parameter sharing**, several hidden layers are shared between tasks and the

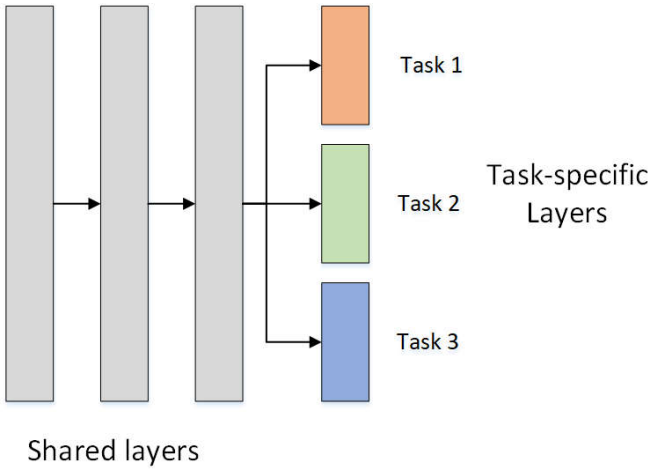

Figure 4.1: Hard parameter sharing.

output layers are set to be task-specific, as seen in Figure 4.1. This approach is the most used approach of MTL in DNN [Kokkinos, 2017, Chennupati et al., 2019]. Most works in IR are also based on this approach (see Section 4.1.1.2 for details). The hard parameter sharing approaches lower the risk of overfitting because the model has to find a representation that covering all of the tasks. Some studies stated that the performance of such an approach depends on the relative weighting between each task's loss [Kendall et al., 2018, Leang et al., 2020]. Thus, the authors aim to dynamically weight the loss of each task during training.

Whereas in **soft-parameter sharing**, the model of each tasks is implemented separately and hold its own parameter. Figure 4.2 shows the architecture of this approach. The models are attached either by information sharing or by requiring parameters to be similar. Commonly, the distance between the parameters is regularized to make the parameters similar. For example, Duong et al. [2015] employed ℓ_2 distance for regular-

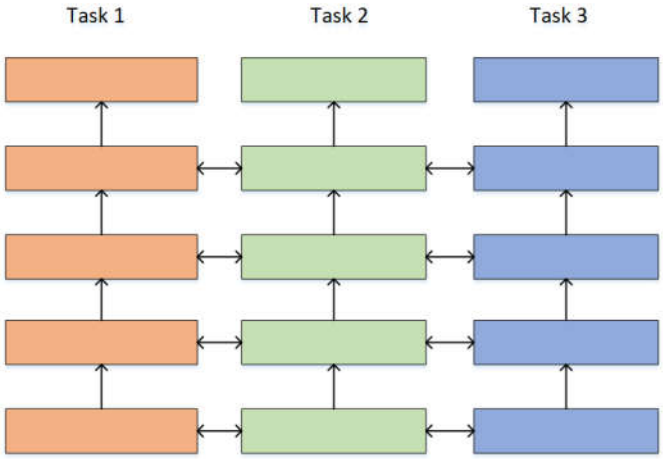

Figure 4.2: Soft parameter sharing.

ization.

4.1.1.2 Multi-task learning in IR

In Information Retrieval, several works have tried to improve the effectiveness of the retrieval task using the MTL approach [Ahmad et al., 2018, 2019, Liu et al., 2015, Zamani and Croft, 2018, 2020]. Some works proposed to exploit the information related to the query to increase retrieval effectiveness [Liu et al., 2015, Ahmad et al., 2018, 2019]. Liu et al. [2015] proposed to unify the representation learning process for query classification task and Web search task. The total loss function is used to optimize both classification loss and ranking loss. The authors build two models based on hard-sharing parameter architecture: (i) the representation layer of both document and query inputs are shared among all tasks (ii) only the query part is shared. The first proposed model is able to achieve better performance for the query classification task but not for the web search task. The second model demonstrated strong results on the web search task. Other works, proposed a MTL model for the *query suggestion* in search sessions and document ranking task [Ahmad et al., 2019]. This model also shared the query representation layers between task. The model encodes the current query, updates corresponding session-level recurrent state, and and uses a greedy decoding algorithm to suggest the next

query. For the document ranker, the representation of the current query, session, and document is concatenated. Then, a sigmoid function is used to produce the final ranking scores. The results of the experiments prove that the proposed method can improve the performance of both tasks. In [Zamani and Croft, 2018] and [Zamani and Croft, 2020], the authors showed the benefit of jointly learning relevance by performing together the *item retrieval task* and the *item recommendation task*. In both works, the models learn the item representation simultaneously. Zamani and Croft [2018] performed a preliminary experiment by using a simple fully connected feed-forward networks. The results proved that there is a potential benefit of jointly learn both tasks. Instead of using query-document relevance information to train the MTL model, Zamani and Croft [2020] tried to use user-item information to train their proposed model. The experiment on four datasets showed that the model outperform competitive baselines in the retrieval task and state-of-the-art collaborative filtering models in the recommendation task. However, no MTL-based solution has focused on modeling multidimensional relevance to date.

4.2 Consumer Health Search and Related Relevance Dimensions

People often seek health-related information on the web. Stankova et al. [2020] performed a user study to understand the use of internet as an information source for health related information. The study showed that 71% of the participants use internet to search for health information. The information may have a significant impact on user's healthcare decisions and outcomes. Low-quality health information can harm the information seeker if the information is used to make health-related decision. Thus, it is crucial to develop a search engine that not only provide relevant information with respect to the query. The system should also promotes information with high quality. This problem has been the background and the motivation of several shared challenges related to patient-centered

retrieval, including CLEF eHealth[1] and TREC Health Misinformation[2]. CLEF eHealth has been organizing evaluation labs related to layperson and professional information extraction, management, and retrieval since 2013. One of the subtask is adhoc search that aim to retrieve information relevant to people's health information need. The organizers provides an additional relevance assessment on *readability*, *trustworthiness* (CLEF 2018), and *credibility* (CLEF 2020). Since 2019, TREC Health Misinformation also focuses on the domain of consumer health search. The main aim of the task is to build an IR system that can help the users make correct decisions by retrieving *relevant*, *credible*, and *correct* information.

Relevance Dimensions

In consumer health search domain, several *readability* and *credibility* have been considered as the criteria of relevance [Hersh, 2008].

1. Understandability

   Understandability, as defined by Xu and Chen [2006], is *"the extent to which the content of a retrieved document is perceived by the user as easy to read and understand"*. In health information, Shoemaker et al. [2014] defined a piece of information as understandable *"when consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages"*. Readability is commonly used in the literature to substitute understandability. Although readability is not completely the same as understandability, readability is one of the factors that affect the understanding of a text [Xu and Chen, 2006]. Thus, some studies employed features related to readability to estimate understandability. The most basic feature are the *traditional readability formula* [DuBay, 2004], including Flesch-Kincaid Index, Automated Readibility Index, Coleman-Liau Index (CLI), Dale-Chall Index (DCI), Flesch Reading Ease (FRE), Gunning Fog Index (GFI), Lasbarhetsindex (LIX), and Simple Measure of Gobbledygook (SMOG).

   Incorporating readability or understandability in the relevance assessment process

[1] https://clefehealth.imag.fr/

[2] https://trec-health-misinfo.github.io/

has been proven to be beneficial for the CHS task [Palotti et al., 2016, van Doorn et al., 2016, Palotti et al., 2019a]. Palotti et al. [2016] applied learning-to-rank approach and exploited features related to document's readability such as the *traditional readability formulas*, *lexical features* and features based on medical dictionaries. van Doorn et al. [2016] addressed the optimization problem over the relevance criteria to find an optimal solution for a different trade-off between relevance criteria. Some experiments is performed on the CHS task considering topicality and understandability. To predict the understandbility score of a document, SVM is used to train a model on several features related to traditional readability formulas and document length. In addition, they consider the frequency of medical terms in the document as the feature. Palotti et al. [2019a] explored three different approaches to combine understandability and topicality including re-ranking, ranking fusion, and learning-to-rank. They used XGBRegressor [Chen and Guestrin, 2016] as the understandability estimator. The considered features include, the *traditional readability formulas*; lexial and syntactic features; features derived from general medical vocabulary, consumer medial vocabulary, expert medical vocabulary; Natural language features; HTML features; and word frequency features. The results showed that learning-to-rank has the best performance compared to the other proposed approaches.

2. Credibility

   Another essential aspect in the consumer health search is the *credibility* of the information [Sbaffi and Rowley, 2017]. As explained in Section 3.2.1.4,credibility is related to the preceived quality of the information receiver. In the literature, some works considered *trustworthiness*[Jimmy et al., 2018]. Although the concepts of trustworthiness and credibility are only partially overlapping, they are closely interdependent [Self, 2008]. Trustworthiness is one of the key aspect related to *credibility*[Viviani and Pasi, 2017]. Several works tried to incorporate *credibility/trustworthiness* in the health-related information retrieval system [Bondarenko et al., Abualsaud et al., 2019, Lima et al., Fernández-Pichel et al., TAO and SAKAI].

All of them re-ranking approach upon the top-k documents produced in the initial search result obtained using BM25. To obtain the credibility score, each study adopted different approaches. Bondarenko et al. manually labeled each document based on the domain category defined in the TREC guidelines using the information from OpenDNS[3]. Some works used a supervised classification approaches to predict the credibility scores. The features are varied, including n-gram [Abualsaud et al., 2019]; html and PageRank features [Lima et al.], link-based features, commercial features, and word features [Fernández-Pichel et al.]. TAO and SAKAI proposed to assess credibility by calculating the *majority score*. The authors hypothesised that *"the more similar a document is to others, the more likely the document is credible"*.

4.3 Methodology

In this section we illustrate the proposed Multi-Task Learning model for learning multi-dimensional relevance in the presence of both relevance judgements (which refer to the topical relevance of documents w.r.t. queries) and other labelled data with respect to additional relevance criteria such as readability, trustworthiness, and credibility. Each of these additional relevance criteria is considered separately along with topicality in the proposed model.

Relevance judgements are employed as training data of the *retrieval task* that is performed by means of an existing neural model based on a *representation-focused architecture* [Guo et al., 2020], as detailed in Section 4.3.1.The use of more advanced neural ranking and classification models will be investigated in the future. The other labelled data, referred to either readability, trustworthiness, or credibility of the considered documents, are employed as training data of another simple neural model that performs the *classification task*, as detailed in Section 4.3.2. The two models are employed together in the proposed Multi-Task Learning model, whose high-level architecture is illustrated

[3] https://community.opendns.com/domaintagging/

in Figure 4.3. As it is shown in the figure, the model consists of a *ranking model* and a

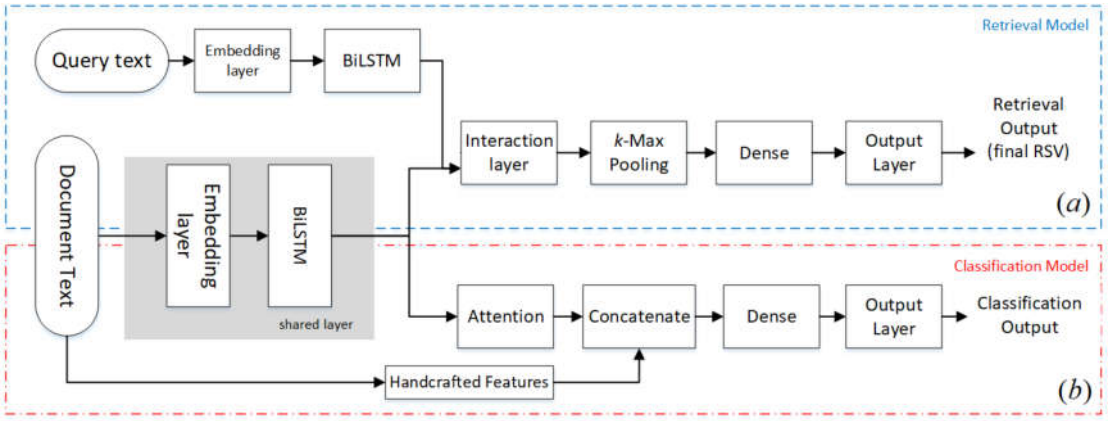

Figure 4.3: The proposed Multi-Task Learning model for ranking, jointly performing the *retrieval task* and the *classification task*.

classification model, which operate simultaneously and share parameters in order to generate an assessment of relevance that is multidimensional, by taking into account both topical relevance and another relevance criterion at a time.

To make this possible, we adopted one of the most common approaches in Multi-Task Learning, called *hard-parameter sharing* [Caruana, 1997]. Generally, in this approach, several hidden layers are shared between tasks, while the output layers are task-specific. The tasks are usually related and have common aspects. Such related tasks might also share a common underlying representation, as in the case of the proposed model.

As illustrated in Figure 4.3, there are two inputs in the model: the *query*, for the retrieval task, and the *document*, for both the retrieval and the classification tasks. In the model, the shared layers consist of the embedding and the Bi-LSTM layers, of which the document's semantic representation is learned across tasks. The technical choices behind embedding representations and, in general, the ranking and classification models will be detailed in Sections 4.3.1 and 4.3.2. The weights associated with the embedding representations in such layers are initialized using pre-trained word vectors and the weights are updated during the training process. the pre-trained word vectors are pre-trained on large text corpora, as it will be detailed in Section 4.4. In this way, learning

parameters related to the retrieval and the classification task together allows to consider additional relevance dimension simultaneously, because it enables the obtained semantic representations capturing information from both tasks.

The non-shared layers are task-specific, and perform the tasks separately without sharing any parameters. Each of the task-specific layers returns separate outputs. However, the output of the retrieval task produces the final *Retrieval Status Value* (RSV) of the document, which takes into account also the additional relevance dimension beyond topicality due to the embedding and Bi-LSTM layers shared with the classification task. In fact, the Multi-Task Learning model is trained by minimizing a *global loss function* L, which is expressed as the weighted sum of the losses of both tasks. Formally:

$$L = \lambda_r L_r + \lambda_c L_c$$

where λ_r and λ_c are the weight parameters respectively for the loss L_r of the retrieval task and the loss L_c of the classification task. For both tasks, the *cross-entropy loss function* is applied.

4.3.1 The Retrieval Model

We followed the approach proposed in [Wan et al., 2016] to perform the retrieval task. As previosly introduced, it is performed by employing a neural ranking model based on a so-called *representation-focused architecture*, which, given a *query* and a *document* as input, builds their word embedding representations and generate relevance assessments via a simple matching function such as the *cosine similarity*. The architecture of the retrieval model is illustrated in Figure 4.3 (a). To transform the word sequences into fixed-length vector representations, a *Recurrent Neural Network* (RNN) with *Bidirectional Long-Short Term Memory* (Bi-LSTM) to encode the input texts are adopted. The Bi-LSTM layer in particular, provides as output a positional text representation of both the query and document. Then, the interactions between the two positional text representations are assessed by means of the cosine similarity function.We mean, by positional representation, a representation in which, for a textual document, the position of each word in the

document is kept; we mean, by interaction between two positional representations the matching between words at the same position in the documents. The top-k strong matching interactions are extracted by using k-Maxpooling, and aggregated by a multi-layer perception to output the final Retrieval Status Value. The sigmoid function is employed as activation function in the output layer.

4.3.2 The Classification Model

The classification model is used to learn the labels that can be associated with documents w.r.t. the additional relevance criteria considered in this article. As explained earlier, these relevance criteria are query-independent, and relate to a document's readability, trustworthiness, and credibility. By performing classification based on such criteria, it is possible, in fact, assign a document to suitable readability, trustworthiness and credibility classes (in a single-label multi-class way). A single classification task is performed w.r.t. each relevance criterion.

As illustrated in Figure 4.3 (b), the classification model is constituted by a simple neural model widely used in Natural Language Processing tasks such as credibility assessment [Giachanou et al., 2019], but it can be generalized to other tasks. In the classification task, the only input is constituted by the *document*. The input pass through an embedding and a Bidirectional LSTM layer, and is treated by an attention layer that is introduced to increment the importance (the weight) of words identified as crucial, and decrease the importance (the weight) of less-crucial words. This is followed by a dense layer with a ReLU activation function. We used ReLu as it is considered as the most preferable activation function in the literature [Giachanou et al., 2019, Wadawadagi and Pagi, 2020].The last layer employs a sigmoid function as the output activation function. To improve the effectiveness of the model, is also possible to incorporate additional hand-crafted features after the attention layer. Such solution has been employed both in [Rashkin et al., 2017], where the output of the Bi-LSTM layer is concatenated with linguistic feature vectors, and in [Giachanou et al., 2019], where emotional features are concatenated in the neural model. In our approach, we consider distinct groups of ad-

ditional features that are suitable w.r.t. the relevance criteria taken into account, as it will be detailed in Section 4.4.2.

4.4 Experimental Evaluations

This section is devoted to presenting the experimental evaluation setting to assess the effectiveness of the proposed approach. In particular, we describe the datasets, the technical implementation details, the baselines and evaluation measures taken into consideration. Results will be illustrated and discussed in Section 4.5.

4.4.1 Datasets

At present, there are not many publicly available data collections to tackle the Information Retrieval task that also contain additional assessments on other relevance criteria beyond topicality. To both implement and evaluate the proposed approach, we used the 2018 and 2020 data collections from the CLEF eHealth - Consumer Health Search (CHS) task.[4] In particular, we focused on the *ad-hoc retrieval* subtask. The data consist of Web pages crawled by means of CommonCrawl,[5] related to the health-related domain. The data collections consider 50 topics/queries and associated documents. Besides relevance judgements, based on topical relevance, assessments on other relevance dimensions are also available in the collections. In the CLEF eHealth 2018 collection, the additional assessments are related to the document's readability and trustworthiness. In the CLEF eHealth 2020 collection, the additional relevance assessments consider the readability and credibility of the Web page. We used the readability assessments from both the 2018 and 2020 collections, the trustworthiness assessments from the 2018 collection, and the credibility assessments from the 2020 collection. Each assessment is a score in the $[0, 1]$ interval.

[4] https://clefehealth.imag.fr/

[5] http://commoncrawl.org/

4.4.2 Additional Features in the Classification Model

As explained in Section 4.3.2, the classification task is based only on document's characteristics. Such task is modelled to perform classification w.r.t. a single relevance criterion at a time. In addition to the textual representation features, we propose to consider feature sets related to each additional relevance criterion.

4.4.2.1 Readability.

To assess readability by considering the deep neural approach, we took into account eight readability features implemented in previous works [Deutsch et al., 2020, Palotti et al., 2019b]. Such features are computed by means of well-known readability indexes,[6] in particular, the

1. Flesch-Kincaid Index,
2. Automated Readibility Index,
3. Coleman-Liau Index (CLI),
4. Dale-Chall Index (DCI),
5. Flesch Reading Ease (FRE),
6. Gunning Fog Index (GFI),
7. Lasbarhetsindex (LIX), and
8. Simple Measure of Gobbledygook (SMOG).

References to each of these indexes are provided in [Palotti et al., 2019b].

4.4.2.2 Trustworthiness and Credibility.

To assess both the trustworthiness and credibility of a document by means of the neural model, we incorporated some features proposed in [Sondhi et al., 2012] to assess the

[6] By means of `https://pypi.org/project/ReadabilityCalculator`, the *ReadibilityCalculator* tool.

reliability of medical Web pages. However, it will be necessary in the future to provide more specific features relating to the two different criteria for relevance. The features are as follows:

1. *Link-based features*: the presence of links is considered an indicator of reliability. Reliable Websites are more likely to contain internal links, while less reliable Websites have more external links. Besides, the existence of privacy policy information and contact link are also taken into account.

2. *Commercial-based features*: the presence of commercial interests in a Website is a sign of low reliability. Therefore, the frequency of commercial words and of commercial links are used as features.

4.4.3 Experimental Setup

Experimental evaluations were performed by taking as a baseline the single retrieval task performed by the neural model shown in Figure 4.3 (a), and comparing its results with those obtained by the Multi-Task Learning model when considering the three distinct additional relevance criteria separately in the classification model. Since the relevance judgements in the collections were not available for all the query-document pairs, we only considered the documents with an associated topical relevance assessment in the collections. However, some documents does not have label for trustworthiness and credibility. To handle this issue, we excluded the loss on the document with empty labels from the total loss calculation.

4.4.3.1 Implementation Details.

From a technical point of view, we implemented all the models including the baseline using Keras.[7] For word embedding, we employed GloVe (trained on the Wiki+Gigaword dataset) with a vector size of 200.[8] We performed 5-fold cross-validation to avoid overfitting, because of the collections quite small sizes. We tested the Adam optimizer with

[7] https://keras.io/

[8] https://nlp.stanford.edu/projects/glove/

different learning rates and selected the optimal one.The selected learning rate was from {1E - 5, 1E - 4, 1E - 3}. To assess the effectiveness of the approach, we used *TrecTools*.[9]

4.4.3.2 Evaluation Metrics.

We evaluated the obtained rankings by considering the following evaluation metrics: *Mean Average Precision* (MAP), *Precision@k* (with $k = 10$ and $k = 20$), and nDCG@10. Furthermore, to better assess the contribution of the additional relevance criteria to the final rankings, we considered other metrics, such as *understandability-biased* RBP (uRBP), $\mathcal{MM}$ [Palotti et al., 2018], and the *Convex Aggregating Measure* (CAM). The details of these metrics can be seen on Section 2.3 uRBP is a measure based on *Rank-Biased Precision* (RBP) [Moffat and Zobel, 2008], and considers both topicality and readability in the evaluation process. The use of this metrics is also suggested in the CLEF eHealth Evaluation Lab, with a ρ value equal to 0.8, the same employed in this work. To assess the influence of trustworthiness and credibility, we considered two further metrics. The first metric is denoted as $\mathcal{MM}$. As illustrated in [Roberts et al., 2019, Zuccon, 2016], in this work we consider nDCG. The second considered metric is the *Convex Aggregating Measure* (CAM). As proposed in [Roberts et al., 2019], also in this case we used nDCG for both $\mathcal{M}_{rel}$ and $\mathcal{M}_{\kappa}$, and set the λ value to 0.5. In both $\mathcal{MM}$ and CAM cases, is possible to use nDCG also for trustworthiness and credibility since we deal with graded scores and we can consider the ranking of trustworthy/credible information based on such scores.

4.5 Results and Discussion

In this section, we report the results of the experimental evaluation against the evaluation measures outlined in the previous section. In Table 4.1 and 4.2, in particular, are reported the results obtained w.r.t. the baseline (**STL**), which only considers the retrieval task (and, hence, only topicality), and to the Multi-Task Learning model con-

[9] https://github.com/joaopalotti/trectools

sidering readability, denoted as **MTL(R)**, trustworthiness, denoted as **MTL(T)**, and credibility, denoted as **MTL(C)**. In the tables, the asterisk symbol (i.e., *) denotes a p-value < 0.05, by using the *two-tailed t-test* for assessing statistical significance. The scores that are indicated in bold refer to the results that outperform the baseline (**STL**).

Table 4.1: Experiment results w.r.t. MAP, Precision, and nDCG.

System	**Experiment Results**			
	MAP	**P@10**	**P@20**	**nDCG@10**
STL	0.484	0.448	0.443	0.325
MTL(R)	**0.498**	**0.526***	**0.531***	**0.405***
MTL(T)	**0.491**	**0.488**	**0.51***	**0.364**
MTL(C)	**0.485**	**0.454**	**0.459**	**0.345**

In particular, in Table 4.1, we report the values of the standard retrieval evaluation measures including MAP, P@10. P@20. and nDCG@10. Overall, the performance of the Multi-Task Learning model improves w.r.t. the single-task model for all of the experiments; hence, joint learning of retrieval and classification tasks proved beneficial.

Table 4.2: Experiment results for the readability-biased, trustworthiness-biased and credibility-biased evaluations.

Model	**Readability-biased evaluation**			**Trust.-/Cred.-biased evaluation**			
	uRBP	uRBP@10	uRBP@20	$\mathcal{MM}_t$	CAM$_t$	$\mathcal{MM}_c$	CAM$_c$
STL	0.221	0.196	0.218	0.824	0.826	0.727	0.733
MTL(R)	**0.271***	**0.241**	**0.268***	**0.825**	**0.827**	**0.732***	**0.737***
MTL(T)	0.212	0.183	0.209	**0.829***	**0.830***	**0.734***	**0.739***
MTL(C)	0.205	0.181	0.202	**0.828***	**0.830***	**0.728**	**0.734**

Table 4.2 reports the results of the so-called *readability-biased*, *trustworthiness-biased*, and *credibility-biased* evaluations. For each considered metric, i.e., uRBP, $\mathcal{MM}$, and CAM, we evaluate the results by considering exactly one additional relevance dimension

beyond topicality. Specifically, $\mathcal{MM}_t$ and $\mathcal{MM}_c$ denote the $\mathcal{MM}$ metric when applied to trustworthiness and credibility. The same holds for CAM_t and CAM_c. By observing the results in Table 4.2, we may conclude that the proposed MTL model also performs better than the single-task model in all experiments, also when we have the possibility to investigate the effectiveness of additional relevance criteria by means of criteria-biased evaluation metrics.

The results of MTL considering the classification task applied to readability (i.e., **MTL(R)**) outperforms the single-task baseline w.r.t. almost all of the metrics. In particular, the uRBP scores denote a statistically significant improvement over the baseline when considering the top-20 ranked documents. The effectiveness of the two latter models can best be appreciated by considering measures that are specifically designed to take trustworthiness and credibility into account. Both the **MTL(T)** and **MTL(C)** models increase the performance w.r.t. the single-task model when considering $\mathcal{MM}$ and CAM scores, in all configurations. With respect to these latter results, however, notwithstanding the positive contribution of such criteria, it is necessary to say that in the datasets used for experimentation the two concepts were rather overlapping, and that the number of credibility labels was much lower than the number of readability and trustworthiness labels, so the results obtained do not allow a clear disambiguation of their individual contribution. What can be affirmed is that, considering such reliability indicators has a positive effect on ranking effectiveness.

Chapter 5

Conclusions and Future Work

This chapter concludes the overall thesis, by providing the final remarks based on the experimental results, which will be presented in Section 5.1, and also providing several possible future works in this research line in Section 5.2.

5.1 Conclusion

In this PhD thesis we have investigated several research issues related to multidimensional relevance assessment in task-based retrieval. In the first phase, we conducted an exploratory analysis of relevance dimension in specific-task retrieval. Specifically, we have tackled the problem of assessing the impact of multidimensional relevance in task-based microblog search. In particular, we have analyzed the impact of different relevance dimensions on different search tasks, by combining a single relevance dimension in addition to topicality in the retrieval model of different search engines. Although in this preliminary work we have only used a simple linear combination of relevance dimensions, the obtained results show that there is indeed a relationship between a search task and specific relevance dimensions. This suggests that, for different search tasks, some relevance dimensions should be prioritized in the computation of an overall Retrieval Status Value based on which documents are ranked.

The second phase of the thesis is focused on the issue of aggregation in multidimen-

sional relevance. We proposed a deep neural model that incorporates more than one relevance criterion besides topicality in the context of Information Retrieval, by using supervised Multi-Task Learning (MTL). In particular, we focused on the possibility to jointly model the retrieval task, to learn topical relevance, and a classification task, to learn the assessments of an additional query-independent relevance criterion at a time. The proposed approach was based on the intuition that the joint optimization of some parameters during retrieval and classification tasks has an impact on the overall relevance value, which, in this way, is affected not only by topical relevance. To verify this intuition, we have performed a set of experiments by considering readability, trustworthiness, and credibility as additional relevance criteria beyond topicality. We observed a substantial improvement compared to the single-task baseline, based on a simple neural model devoted to the retrieval task. We believe that these results may be useful for the subsequent investigation of other aspects related to the multidimensional nature of relevance in retrieval systems that rely on neural approaches.

5.2 Future Works

Based on our experiment in this PhD thesis, we observed several open challenges that we would like to address in the future. First, to deeply understand the relationship between a task behind the search process and the user's choice of relevance dimension, we aim at performing user study. As a future development of the proposed methods, we plan to perform extensive experiments on other search tasks. Moreover, we also plan to explore more advanced neural models and considering the interplay between multiple criteria of relevance together. However, there are not many publicly available data on task-based retrieval. Moreover, the relevance judgement has to consider more than just the topical relevance between query and document. Thus, we intend to construct ad hoc datasets related to task-based retrieval, in order to be release labelled data where relevance is assessed w.r.t. multiple aspects.

Bibliography

M. Abualsaud, F. C. Beyluniogl, M. D. Smucker, and P. R. Duimering. Uwaterloomds at the trec 2019 decision track. In *TREC*, 2019.

W. U. Ahmad, K. Chang, and H. Wang. Multi-task learning for document ranking and query suggestion. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

W. U. Ahmad, K.-W. Chang, and H. Wang. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 385–394, 2019.

A. C. Alhadi, T. Gottron, J. Kunegis, and N. Naveed. Livetweet: Microblog retrieval based on interestingness and an adaptation of the vector space model. In *TREC*, 2011.

O. Alonso and S. Mizzaro. Relevance criteria for e-commerce: a crowdsourcing-based experimental analysis. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 760–761, 2009.

O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

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

C. L. Barry. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159, 1994.

M. Basu, K. Ghosh, and S. Ghosh. Information retrieval from microblogs during disasters: In the light of irmidis task. *SN Computer Science*, 1(1):1–10, 2020.

M. J. Bates. Information search tactics. *J. Am. Soc. Inf. Sci.*, 30(4):205–214, 1979. doi: 10.1002/asi.4630300406. URL https://doi.org/10.1002/asi.4630300406.

N. J. Belkin and B. Kwaśnik. Using structural representation of anomalous states of knowledge for choosing document retrieval strategies. In *Proceedings of the 9th annual international ACM SIGIR conference on research and development in information retrieval*, pages 11–22, 1986.

M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 95–104, 2011.

A. Bondarenko, M. Fröbe, V. Kasturia, M. Völske, B. Stein, and M. hias Hagen. Webis at trec 2019: Decision track.

P. Borlund. The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10):913–925, 2003.

P. D. Bruza, D. W. Song, and K.-F. Wong. Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51(12):1090–1105, 2000.

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Q. Chen, Q. Hu, J. Huang, and L. He. Taker: Fine-grained time-aware microblog search with kernel density estimation. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

S. Chennupati, G. Sistu, S. Yogamani, and S. A Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

J. Choi and W. B. Croft. Temporal models for microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2491–2494. ACM, 2012.

J. Choi, W. B. Croft, and J. Y. Kim. Quality models for microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1834–1838. ACM, 2012.

W. S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973.

N. Craswell. Bpref. In *Encyclopedia of Database Systems*, pages 266–267. Springer US, Boston, MA, 2009.

N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, 2005.

W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.

A. Crystal and J. Greenberg. Relevance criteria identified by health information users during web searches. *Journal of the American Society for Information Science and Technology*, 57(10):1368–1382, 2006.

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: A new aggregation criterion. In *European Conference on Information Retrieval*, pages 264–275. Springer, 2009a.

C. da Costa Pereira, M. Dragoni, and G. Pasi. A prioritized "and" aggregation operator for multidimensional relevance assessment. In *Congress of the Italian Association for Artificial Intelligence*, pages 72–81. Springer, 2009b.

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information processing & management*, 48(2):340–357, 2012.

F. Damak, K. Pinel-Sauvagnat, M. Boughanem, and G. Cabanac. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 914–919. ACM, 2013.

M. De Grandis, G. Pasi, and M. Viviani. Fake news detection in microblogging through quantifier-guided aggregation. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 64–76. Springer, 2019.

L. R. Derczynski, B. Yang, and C. S. Jensen. Towards context-aware search and analysis on social media data. In *Proceedings of the 16th international conference on extending database technology*, pages 137–142. ACM, 2013.

T. Deutsch, M. Jasbi, and S. M. Shieber. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, 2020.

Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010.

W. H. DuBay. The principles of readability. *Online Submission*, 2004.

L. Duong, T. Cohn, S. Bird, and P. Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850, 2015.

M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 495–504. ACM, 2011.

M. Efron, J. Lin, J. He, and A. De Vries. Temporal feedback for tweet search with non-parametric density estimation. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 33–42. ACM, 2014.

C. Eickhoff and A. P. de Vries. Modelling complex relevance spaces with copulas. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management*, pages 1831–1834, 2014.

C. Eickhoff, A. P. de Vries, and K. Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 663–672, 2013.

Y. Fan, J. Guo, Y. Lan, J. Xu, C. Zhai, and X. Cheng. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 375–384, 2018.

M. Fernández-Pichel, D. E. Losada, J. C. Pichel, and D. Elsweiler. Citius at the trec 2020 health misinformation track.

N. Ferro and C. Peters. From multilingual to multimodal: the evolution of clef over two decades. In *Information Retrieval Evaluation in a Changing World*, pages 3–44. Springer, 2019.

B. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87. ACM, 1999.

S. Gerani, C. Zhai, and F. Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *European Conference on Information Retrieval*, pages 256–267. Springer, 2012.

S. Ghosh, K. Ghosh, D. Ganguly, T. Chakraborty, G. J. Jones, and M.-F. Moens. ECIR 2017 workshop on exploitation of social media for emergency relief and preparedness (SMERP 2017). In *ACM SIGIR Forum*, volume 51, pages 36–41. ACM, 2017.

A. Giachanou, M. Harvey, and F. Crestani. Topic-specific stylistic variations for opinion retrieval on twitter. In *European Conference on Information Retrieval*, pages 466–478. Springer, 2016.

A. Giachanou, P. Rosso, and F. Crestani. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 877–880, 2019.

L. Goeuriot, P. Mulhem, and E. SanJuan. Clef 2017 mc2 search and time line tasks overview. In *CLEF (Working Notes)*, 2017.

L. Goeuriot, J. Mothe, P. Mulhem, and E. SanJuan. Building evaluation datasets for cultural microblog retrieval. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, pages 20–29. Association for Computational Linguistics, 2011.

J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 55–64, 2016.

J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.

A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on Twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

J. T. Hackos and J. Redish. *User and task analysis for interface design*, volume 1. Wiley New York, 1998.

P. Hansen. User interface design for ir interaction. a task-oriented approach. In *CoLIS3: Third International Conference on the Conceptions of the Library and Information Science, 23-26 May 1999, Dubrovnik, Croatia*, pages 191–205, 1999.

S. P. Harter. Psychological relevance and information science. *Journal of the American Society for information Science*, 43(9):602–615, 1992.

J. Herrera, B. Poblete, and D. Parra. Learning to leverage microblog information for qa retrieval. In *European Conference on Information Retrieval*, pages 507–520. Springer, 2018.

W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media, 2008.

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. Le, T. Abdelzaher, J. Han, A. Leung, J. Hancock, et al. Tweet ranking based on heterogeneous networks. *Proceedings of COLING 2012*, pages 1239–1256, 2012.

P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM, 2013.

P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer Science & Business Media, 2006.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

J. Jiang, D. He, D. Kelly, and J. Allan. Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 137–146, 2017.

J. Jimmy, G. Zuccon, J. Palotti, L. Goeuriot, and L. Kelly. Overview of the clef 2018 consumer health search task. 2018.

K. S. Jones and P. Willett. *Readings in information retrieval*. Morgan Kaufmann, 1997.

I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71, 2003.

A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017.

A. Kotov, Y. Wang, and E. Agichtein. Leveraging geographical metadata to improve search over social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 151–152, 2013.

A. Kotov, V. Rakesh, E. Agichtein, and C. K. Reddy. Geographical latent variable models for microblog retrieval. In *European Conference on Information Retrieval*, pages 635–647. Springer, 2015.

I. Leang, G. Sistu, F. Bürger, A. Bursuc, and S. Yogamani. Dynamic task weighting methods for multi-task networks in autonomous driving systems. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2020.

D. Lewandowski. Credibility in web search engines. In *Online credibility and digital ethos: Evaluating computer-mediated communication*, pages 131–146. IGI Global, 2013.

J. Li, P. Zhang, D. Song, and Y. Wu. Understanding an enriched multidimensional user relevance model by analyzing query logs. *Journal of the Association for Information Science and Technology*, 68(12):2743–2754, 2017.

Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manag.*, 44(6):1822–1837, 2008. doi: 10.1016/j.ipm.2008.07.005. URL https://doi.org/10.1016/j.ipm.2008.07.005.

L. C. Lima, D. B. Wright, I. Augenstein, and M. Maistro. University of copenhagen participation in trec health misinformation track 2020.

J. Lin and M. Efron. Temporal relevance profiles for tweet search. In *SIGIR Workshop on Time-aware Information Access*, 2013.

C. Lioma, J. G. Simonsen, and B. Larsen. Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 91–98, 2017.

T.-Y. Liu. Learning to rank for information retrieval. *Foundations and trends in information retrieval*, 3(3):225–331, 2009.

X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-y. Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, 2015.

G. Livraga and M. Viviani. Data confidentiality and information credibility in on-line ecosystems. In *Proceedings of the 11th International Conference on Management of Digital EcoSystems*, pages 191–198, 2019.

Z. Luo, M. Osborne, and T. Wang. An effective approach to tweets opinion retrieval. *World Wide Web*, 18(3):545–566, 2015.

M. Magnani, D. Montesi, and L. Rossi. Conversation retrieval for microblogging sites. *Information retrieval*, 15(3-4):354–372, 2012.

D. Mahata, J. R. Talburt, and V. K. Singh. From chirps to whistles: discovering event-specific informative content from Twitter. In *Proceedings of the ACM web science conference*, page 17. ACM, 2015.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

G. Marchionini. Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1):54–66, 1989.

S. Marrara, G. Pasi, and M. Viviani. Aggregation operators in information retrieval. *Fuzzy Sets and Systems*, 324:3–19, 2017.

K. Massoudi, M. Tsagkias, M. De Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *European Conference on Information Retrieval*, pages 362–367. Springer, 2011.

B. Mitra and N. Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.

T. Mitra and E. Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

S. Mizzaro. Relevance: The whole history. *J. Am. Soc. Inf. Sci.*, 48(9):810–832, 1997a. doi: 10.1002/(SICI)1097-4571(199709)48:9\<810::AID-ASI6\>3.0.CO;2-U. URL https://doi.org/10.1002/(SICI)1097-4571(199709)48:9\<810::AID-ASI6\mskip\medmuskip3.0.CO;2-U.

S. Mizzaro. Relevance: The whole history. *Journal of the American society for information science*, 48(9):810–832, 1997b.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.

B. Moulahi, L. Tamine, and S. B. Yahia. i a ggregator: Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, 65(10):2062–2083, 2014a.

B. Moulahi, L. Tamine, and S. B. Yahia. Toward a personalized approach for combining document relevance estimates. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 158–170. Springer, 2014b.

R. Nagmoti, A. Teredesai, and M. De Cock. Ranking approaches for microblog search. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 153–157. IEEE Computer Society, 2010.

N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd international web science conference*, pages 1–7, 2011a.

N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 183–188. ACM, 2011b.

F. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages 718 in CEUR Workshop Proceedings, Heraklion*, 2011.

A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009. ACM, 2015.

H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707, 2016.

J. Palotti, L. Goeuriot, G. Zuccon, and A. Hanbury. Ranking health web pages with relevance and understandability. In *Proceedings of the 39th International ACM SIGIR*

conference on Research and Development in Information Retrieval, pages 965–968, 2016.

J. Palotti, G. Zuccon, and A. Hanbury. MM: a new framework for multidimensional evaluation of search engines. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1699–1702, 2018.

J. Palotti, H. Scells, and G. Zuccon. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM, 2019a.

J. Palotti, G. Zuccon, and A. Hanbury. Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. *Journal of medical Internet research*, 21(1):e10986, 2019b.

L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, and X. Cheng. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 257–266, 2017.

T. K. Park. The nature of relevance in information retrieval: An empirical study. *The library quarterly*, 63(3):318–351, 1993.

G. Pasi and M. Viviani. Application of aggregation operators to assess the credibility of user-generated content in social media. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 342–353. Springer, 2018.

G. Pasi, M. De Grandis, and M. Viviani. Decision making over multiple criteria to assess news credibility in microblogging sites. In *IEEE World Congress on Computational Intelligence (WCCI) 2020, Proceedings*. IEEE, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

G. Pironkov, S. Dupont, and T. Dutoit. Multi-task learning for speech recognition: an overview. In *ESANN*, 2016.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, 1998.

M. Porter. The Porter stemming algorithm, 2005. *URL http://www. tartarus. org/martin/PorterStemmer/index. html*, 2008.

L. Qiu. Analytical searching vs. browsing in hypertext information retrieval systems. *Canadian Journal of Information and Library Science*, 18(4):1–13, 1993.

X. Qiu and X. Huang. Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth international joint conference on artificial intelligence*, 2015.

J. Rao, H. He, H. Zhang, F. Ture, R. Sequiera, S. Mohammed, and J. Lin. Integrating lexical and temporal signals in neural ranking models for searching social media streams. *arXiv preprint arXiv:1707.07792*, 2017.

J. Rao, L. Liu, Y. Tay, W. Yang, P. Shi, and J. Lin. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5373–5384, 2019.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the*

2017 conference on empirical methods in natural language processing, pages 2931–2937, 2017.

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

K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, and F. Meric-Bernstam. Overview of the TREC 2019 precision medicine track. In E. M. Voorhees and A. Ellis, editors, *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2019. URL https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.PM.pdf.

S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

S. E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 1977.

G. Salton. Automatic information organization and retrieval. 1968.

G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

T. Saracevic. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.*, 26(6):321–343, 1975. doi: 10.1002/asi.4630260604. URL https://doi.org/10.1002/asi.4630260604.

T. Saracevic. Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, volume 201, page 18. ACM New York, 1996.

L. Sbaffi and J. Rowley. Trust and credibility in web-based health information: a review and agenda for future research. *Journal of medical Internet research*, 19(6):e218, 2017.

L. Schamber. Users' criteria for evaluation in a multimedia environment. In *Proceedings of the ASIS Annual Meeting*, volume 28, pages 126–33. ERIC, 1991.

L. Schamber and M. Eisenberg. Relevance: The search for a definition. 1988.

H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

C. C. Self. Credibility. In *An Integrated Approach to Communication Theory and Research, 2nd Edition.*, page 22. Taylor & Francis, 2008.

Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 101–110, 2014.

S. J. Shoemaker, M. S. Wolf, and C. Brach. Development of the patient education materials assessment tool (pemat): a new measure of understandability and actionability for print and audiovisual patient information. *Patient education and counseling*, 96(3): 395–403, 2014.

P. Sondhi, V. V. Vydiswaran, and C. Zhai. Reliability prediction of webpages in the medical domain. In *European conference on information retrieval*, pages 219–231. Springer, 2012.

M. Stankova, P. Mihova, F. Andonov, and T. Datchev. Health information and cam online search. *Procedia Computer Science*, 176:2794–2801, 2020.

Y. Su, J. Li, D. Song, P. Zhang, and Y. Zhang. Investigating the dynamic decision mechanisms of users' relevance judgment for information retrieval via log analysis. In *Pacific Rim International Conference on Artificial Intelligence*, pages 968–979. Springer, 2018.

A. Suarez, D. Albakour, D. Corney, M. Martinez, and J. Esquivel. A data collection for evaluating the retrieval of related tweets to news articles. In *European Conference on Information Retrieval*, pages 780–786. Springer, 2018.

D. R. Swanson. Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, 56(4):389–398, 1986.

K. Tao, F. Abel, C. Hauff, and G.-J. Houben. Twinder: a search engine for twitter streams. In *International Conference on Web Engineering*, pages 153–168. Springer, 2012.

K. Tao, C. Hauff, F. Abel, and G.-J. Houben. *Information Retrieval for Twitter Data*, pages 195–206. Digital Formations. Peter Lang, 11 2013. ISBN 978-1-4331-2170-8.

S. TAO and T. SAKAI. Realsakailab at the trec 2020 health misinformation track.

A. Taylor. User relevance criteria choices and the information search process. *Information Processing & Management*, 48(1):136–153, 2012.

A. Taylor. Examination of work task and criteria choices for the relevance judgment process. *Journal of Documentation*, 2013.

A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio. Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management*, 43(4):1071–1084, 2007.

J. Teevan, D. Ramage, and M. R. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *Journal of the American society for Information Science and Technology*, 56 (4):327–344, 2005.

S. Uprety, Y. Su, D. Song, and J. Li. Modeling multidimensional user relevance in ir using vector spaces. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 993–996, 2018.

P. Vakkari. Task-based information searching. *Annual review of information science and technology*, 37(1):413–464, 2003.

J. van Doorn, D. Odijk, D. M. Roijers, and M. de Rijke. Balancing relevance criteria through multi-objective optimization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 769–772, 2016.

M. Van Opijnen and C. Santos. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1):65–87, 2017.

S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

M. Verma, E. Yilmaz, and N. Craswell. On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 277–286. ACM, 2016.

M. Viviani and G. Pasi. A multi-criteria decision making approach for the assessment of information credibility in social media. In *International Workshop on Fuzzy Logic and Applications*, pages 197–207. Springer, 2016.

M. Viviani and G. Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5):e1209, 2017.

J. Vosecky, K. W.-T. Leung, and W. Ng. Searching for quality microblog posts:f filtering and ranking based on content analysis and implicit links. In *International Conference on Database Systems for Advanced Applications*, pages 397–413. Springer, 2012.

R. Wadawadagi and V. Pagi. Sentiment analysis with deep neural networks: comparative study and performance assessment. *Artificial Intelligence Review*, 53:6155–6195, 2020.

S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

P. Wang and D. Soergel. A cognitive model of document use during a research project. study i. document selection. *Journal of the American Society for Information Science*, 49(2):115–133, 1998.

W. M. Webberley, S. M. Allen, and R. M. Whitaker. Retweeting beyond expectation: Inferring interestingness in Twitter. *Computer Communications*, 73:229–235, 2016.

W. Weerkamp and M. De Rijke. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, pages 923–931, 2008.

W. Weerkamp and M. de Rijke. Credibility-inspired ranking for blog post retrieval. *Information retrieval*, 15(3-4):243–277, 2012.

J. Worsham and J. Kalita. Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters*, 136:120–126, 2020.

C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.

Y. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190, 1988.

P. Yang, H. Fang, and J. Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256. ACM, 2017.

W. Yang, H. Zhang, and J. Lin. Simple applications of BERT for ad hoc document retrieval. *CoRR*, abs/1903.10972, 2019. URL http://arxiv.org/abs/1903.10972.

H. Zamani and W. B. Croft. Joint modeling and optimization of search and recommendation. In O. Alonso and G. Silvello, editors, *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018*, volume 2167 of *CEUR Workshop Proceedings*, pages 36–41. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2167/paper2.pdf.

H. Zamani and W. B. Croft. Learning a joint search and recommendation model from user-item interactions. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 717–725, 2020.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2): 179–214, 2004.

A. Zubiaga. A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984, 2018.

G. Zuccon. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval*, pages 280–292. Springer, 2016.