QATAR UNIVERSITY

COLLEGE OF ENGINEERING

# ENABLING EFFECTIVE ARABIC INFORMATION RETRIEVAL ON THE WEB AND SOCIAL MEDIA

BY

MARAM GHANEM HASANAIN

A Dissertation Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Doctorate of Philosophy in Computer Science

June 2022

© 2022. Maram Ghanem Hasanain. All Rights Reserved.

# COMMITTEE PAGE

The members of the Committee approve the Dissertation of

Maram Ghanem Hasanain defended on 22/05/2022.

Dr. Tamer Elsayed
Dissertation Supervisor

Prof. Iadh Ounis
Committee Member

Prof. Abdelaziz Bouras
Committee Member

Dr. Abdelkarim Erradi
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

# ABSTRACT

Hasanain, Maram, G., Doctorate : June : 2022, Doctorate of Philosophy in Computer Science
Title: Enabling Effective Arabic Information Retrieval on the Web and Social Media
Supervisor of Dissertation: Dr. Tamer Elsayed.
Arabic is one of the most dominant languages on the Web and social media. The huge and ever-growing Arabic user generated content, further motivated by the ongoing political unrest in the region, created an immense need for Information Retrieval (IR) systems to support users in consuming and analyzing Arabic content at such scale. In the past decade, tasks like ad hoc retrieval, event detection, document summarization, and fake news detection became of great importance to Arab users. However, research on developing IR systems for these tasks over Arabic content is severely lacking, as compared to higher-resource languages like English.

This dissertation makes an argument that the main reason behind the slow progress in the development of Arabic IR systems is the lack of language resources. In particular, there is a severe shortage of standardized, large-scale, and representative test collections and annotated datasets, needed for system training and evaluation. The main goal of this dissertation is to motivate research on Arabic IR by providing necessary evaluation resources, baseline systems, and alternative approaches to training and evaluation of IR systems. To that end, two IR tasks were identified as important and underdeveloped for Arabic content, namely, ad hoc retrieval, and misinformation detection. Each task was investigated over two domains: the Web, and social media (Twitter in particular).

For the ad hoc retrieval task, an approach for constructing test collections without the need for a shared-task evaluation campaign is proposed. As a result, two large-scale and manually annotated test collections were constructed starting from recent snapshots of each of the Arabic Web and Arabic Twittersphere. Moreover, state-of-the-art retrieval models that were previously tested over English content, were benchmarked over the new test collections, providing baseline performance for future systems. The constructed test collections were proved to include high quality annotations, motivating creation of similar test collections for other problems and domains, with relatively low cost.

As for the misinformation detection problem, I focus on two components that are usually part of the claim verification pipeline followed to address this problem. In particular, this work tackles two problems: (1) claim check-worthiness identification, and (2) evidence retrieval for verification. Claim check-worthiness detection is the problem of identifying claims that should be prioritized for verification. Once a claim is identified to be verified, evidence retrieval involves searching for documents that contain information supporting or denying the claim. This thesis describes the process of creating the first Arabic annotated datasets for the two tasks. Furthermore, for claim check-worthiness detection, studied within the social media domain, I extensively study whether we can avoid creating a dedicated Arabic training dataset to train an effective system for the task. To achieve that, I consider cross-lingual transfer learning, where a supervised model trained on non-Arabic data is applied to an Arabic test set. The study demonstrated that cross-lingual transfer learning from some languages to Arabic is comparable to monolingual models exclusively trained on Arabic. For evidence retrieval, I study the suitability of relying on topical relevance as the main approach to evaluate the task in the Web domain. Moreover, I run an extended study on the effectiveness of Web search systems in retrieving documents containing evidence

as opposed to topically-relevant documents to a claim. My study shows that pages (retrieved by a commercial search engine) that are topically-relevant to a claim are not always useful for verifying it. Given the aforementioned finding, I investigate and identify characteristics or features specific to evidential pages. Furthermore, preliminary experiments show that effectiveness of a supervised evidential pages retrieval model that employs them has a 5.3% increased recall of evidential pages over the search engine.

# ACKNOWLEDGMENTS

I would like to express my gratitude to many people, who generously and continuously supported me during the course of my PhD journey.

My deepest gratitude goes to my supervisor Dr. Tamer Elsayed, who mentored and guided me through both MSc and PhD degrees. I learnt from him how to conduct high-quality research and to always aim for the best. His selfless contribution of time and guidance helped shape the researcher and person I am today. I appreciate his insightful critiques, and all invaluable comments and advice. I deeply thank him for believing in me and continuously encouraging me to excel, even when my enthusiasm and hope were fading away. Thank you Dr. Tamer for being a mentor and a friend.

A special thank you goes to my colleagues and friends from my research group, bigIR, at Qatar University including: Reem, Mrs. Rana, Fatima, Marwa, and Yassmine for their friendship, emotional support, collaboration and the great discussions we had. They helped me learn to appreciate my life and the work I am doing.

I am greatly thankful to Dr. Walid Magdy at University of Edinburgh, and Dr. Abdulaziz Al-Ali at Qatar University for their valuable feedback and direction during the critical stage of writing this dissertation. I am also very grateful to Dr. Mucahid Kutlu at TOBB University for the great collaboration opportunities we had during his time at Qatar University.

My utmost appreciation goes to my family for always being there for me. I thank my parents Amnah and Ghanim, and siblings Mayess, Bara', Rawan, Anas, and Mo'men for their continuous love, support, and prayers that helped make this dissertation a reality. A special mention goes to my 4-year old niece Alyaa' who gave me unconditional love and positive energy; she even encouraged me and prayed for the acceptance of my papers even though she doesn't understand yet what a research paper is.

I thank Qatar University for the two student grants that facilitated my work during one year of my PhD, including Student Grants # QUST-1-CENG-2019-32 and # QUST-2-CENG-2019-20.

This work was made possible by NPRP grant # NPRP11S-1204-170060 and # NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

Recent estimates have ranked Arabic language as the fourth most used language on the Web.[1] Per the 2017 Arab Social Media Report [1], Facebook is estimated to have more than 156 million Arab users, with Arabic used in 55% of their activities on the platform. Twitter [1] witnessed similar prevalence of Arabic content as the Arab world is estimated to generate an average of 27.4 million tweets daily; 72% of them are Arabic.

Modern Standard Arabic (MSA) is the standard language in communication and generally understood by majority of Arabs. For example, formal content like news articles is usually written in MSA. The rise of social media and user-generated content has popularized online communication using the wide variety of Arabic dialects, that can be viewed as different languages in some cases [2]. The huge growth of Arabic content generated a pressing need for effective Information Retrieval (IR) systems through which the user can provide an information need (represented by a free-text query) and the system is expected to return a set of documents satisfying that need. Automatic processing, analysis, and information extraction from Arabic content is now faced by the following modern challenges.

- Code-switching between Arabic and other languages (e.g., French) [3]–[5].
- Transliteration where Arabic content is written partially or entirely using Latin characters (i.e., "Arabizi") [5].
- Word decorations, elongation, repeated characters, and abbreviations [2], [6].
- Variety of spoken dialects, even within the same geographic region, that leaked into written content, especially in social media platforms. Dialects as written now have their own rules, morphology, and syntax as opposed to MSA [7].

Moreover, Arabic text has other features that can affect retrieval of information in response to a user need. For example [2].

- Some letters can have different forms such as the "hamza" that can appear in three forms (ء, ؤ and ئ). This is problematic as the hamza form used in a word affects its meaning. For example, the word بُرُؤٌ meaning "healed" or "cured", can have the different meaning of "innocent" only by the change of the hamza form to become بَرِيءٌ.
- Diacritics might appear on letters, especially in MSA content. As with different letter forms, change of diacritics can change the meaning of a word. The word الجَدّ meaning "grandfather" can become الجِدّ which means "seriousness" by changing a single diacritic.
- Few Arabic letters are often used interchangeably due to varying orthographic conventions or spelling mistakes. This commonly occurs with letter pairs (ي, ى), and (ه, ة).

A simple solution that can come to mind when handling the previously mentioned issues is to simply normalize different letter forms to one form, and remove diacritics. However, this can lead to further ambiguity during retrieval [2] since, as already shown

[1] https://www.internetworldstats.com/stats7.htm

by examples above, two words with the same letters can have completely different meaning based on the diacritics or letter form used. Furthermore, some of the modern challenges described earlier do not have effective solutions so far. Thus, it is essential to take these features in consideration while developing IR systems for Arabic content.

One might stop here and ask: *why do we actually need specific IR systems for Arabic content?* Firstly, the previous discussion on Arabic linguistic features provided a very strong motivation. Mere re-application of existing IR systems tested on high resource languages like English, or language-independent systems, might lead to sub-optimal performance. Secondly, users are the motivators for designing IR systems; effectiveness of an IR system is measured by its ability to satisfy the user information need [8]. As content publishers, especially through social media, user groups such as Arab versus American users, are expected to be interested in different topics and might express their thoughts and opinions in varying patterns. Moreover, some types of documents might be more widely-spread across one group of users as opposed to another. For example, the relative size of Arabic forums content on the Arabic Web is proportionally large compared to the English one, while Arabic Wikipedia is still relatively very small on the Arabic Web compared to English Wikipedia [2]. Therefore, an IR system should be able to model its users such that it can effectively address their needs.

Fortunately, recent years have witnessed big leaps forward in Arabic natural language processing (NLP) [9], with many effective tools proposed to process, normalize, and tokenize Arabic text. Such tools are essential for IR systems such as ad hoc retrieval[2] and a variety of text classification systems. For example, pre-processing text by stemming and applying stopwords removal has become the standard for many IR applications [10]. However, with such advances in NLP, we still observe slow research and commercial progress in developing IR systems targeting Arabic content [9]. A scan of existing literature on Arabic IR reveals one key obstacle. It is the lack of modern, representative, large-scale and publicly-available document collections providing evaluation and training resources for Arabic IR systems [2], [9], [11]–[13]. Filling this gap is the main motivation of this dissertation.

In my work, I[3] identified two critical IR problems that are severely lacking evaluation test collections and annotated datasets, while being very important and timely problems. The first is the typical problem of ***ad hoc retrieval***, which is the main service provided by search engines over the whole Web, in social media platforms, and even within individual websites such as e-commerce websites. Moreover, ad hoc retrieval is usually a component of larger solutions to other IR problems, such as question answering [14]. Typically, evaluating the task follows the Cranfield paradigm [15]. The paradigm requires a test collection composed of (1) a set of topics representing users' information needs, (2) a large set of documents (e.g., Web pages, passages, or tweets) from which the system will retrieve potentially relevant content, and (3) a subset of the documents annotated by their relevance to the information needs. Development of large scale test collections is very expensive since it requires humans to judge or annotate documents [16]. Thus, majority of such collections were constructed through shared-

[2]Also called ad hoc search, ad-hoc search, and ad-hoc retrieval

[3]In this dissertation, I present my contributions, research and development efforts at two scales: 1) activities I was entirely responsible for, and 2) activities I had a major and dominant role in as part of a collaborative work. To differentiate between the two, "I" and "my" are used when describing the first, while "we" and "our" are used in the latter case.

task evaluation campaigns such as those held by the Text Retrieval Conference (TREC). A shared-evaluation campaign involves many participating retrieval systems addressing the same task and consequently producing sets of documents retrieved from the document collection in response to a common set of information needs. The organizers construct pools of documents to judge from these submitted documents, and expert judges annotate the documents by relevance to the information needs. In fact, TREC is the source of the most commonly used Arabic test collection for the ad hoc search task, which is the two-decades old TREC-2001/2002 Cross-Language Information Retrieval (CLIR) dataset, limited to news articles [17], [18].

The second task targeted by this dissertation is ***misinformation detection***. The buzzword "fake news" has been constantly circulating over the past five years over all types of media across the globe. In the Arab world, the problem of misinformation spread is very sever due to the decade-old ongoing political and military conflicts that were ignited starting from the so called "Arab Spring" in 2011 [19]. Moreover, the leak and wide spread of fake news is also considered the spark for the 2017 Qatar-Gulf crisis that lasted over 3 years,[4] leading to catastrophic consequences on the region. With the start of the COVID-19 pandemic, misinformation circulation escalated to become an infodemic in both mainstream and social media [20], [21]. With the recency of the problem and the fact that several sub-problems fall under it, Arabic datasets for training and evaluating systems for this problem are also scarce. In fact, the first Arabic dataset for claim verification [22] was released as recently as 2018.

The two tasks of interest in this research are not isolated from each other. From the misinformation detection angle, ad hoc retrieval usually serves multiple steps in the process of verifying a claim's accuracy. When processing a stream of claims to detect misinformation, ad hoc retrieval can be used to retrieve similar claims to an input claim as an approach to check whether the claim has already been verified [23]. It can also serve as a tool to retrieve textual sources against which the claim can be verified [24]. From the ad hoc retrieval perspective, many studies on Web search, for example, have showed that users tend to trust the results returned by the search engine [25]. With this trust, a huge responsibility falls on the retrieval system to return good quality results, not only in relevance, but in accuracy too. For example, recent studies (e.g., [26]) on a critical domain like health search have found that misinformation in search results can lead to harmful decisions taken by users. Thus, search systems can be designed to consider the document's accuracy as part of its retrieval function, such that it provides more factually-accurate results to its users.

## 1.1. Problem and Research Questions

This dissertation aims at solving the overarching problem of limited Arabic IR evaluation resources over the Web and social media for two problems: ad hoc retrieval and misinformation detection. Figure 1.1 summarizes the work done, and contributions for each of the two problems, over each of the Web and Social Media (Twitter specifically).

[4] https://www.aljazeera.com/features/2020/6/5/qatar-gulf-crisis-your-questions-answered

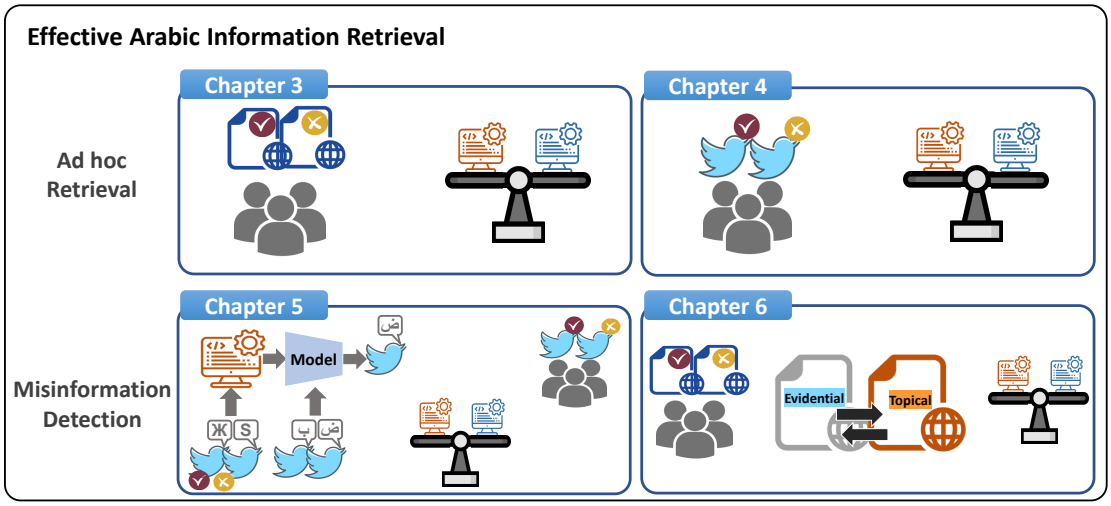



Figure 1.1. The IR problems targeted by this dissertation along with contributions in each.

### *1.1.1. Ad hoc Retrieval*

To facilitate and push system development for this task, it is necessary to provide a large-scale and representative test collection. Constructing such collection through a shared-task campaign is not feasible for small research teams with limited access to human judges; innovative solutions are needed to overcome this obstacle. Moreover, such a test collection is required for every target domain (e.g., The Web or social media). Additionally, before implementing new retrieval systems over Arabic collections, it is necessary to investigate the effectiveness of existing state-of-the-art systems that were not previously tested over Arabic. To achieve these goals for each of the Web and social media, the following research questions will be answered by this dissertation:

- How can we construct a high-quality and large-scale test collection without requiring a shared-task evaluation campaign (e.g., as done through TREC)?
- How effective are state-of-the-art retrieval models over an Arabic test collection?

### *1.1.2. Misinformation Detection*

Automatically detecting misinformation is usually tackled as a pipeline of multiple sub-tasks [21] as shown in Figure 1.2. In my work, I tackle the first and third sub-problems from this pipeline, as explained next.

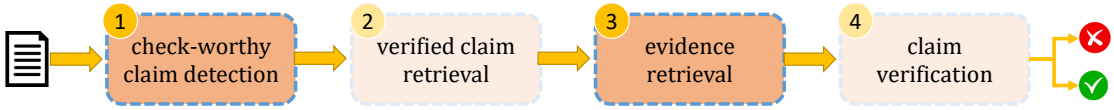



Figure 1.2. Fact checking pipeline.

#### *1.1.2.1. Check-worthy Claim Detection over Social Media*

This stage in the pipeline is very critical due to the enormous number of claims posted daily over social media. Verifying all incoming claims about a topic is not possible through manual efforts or even automatic systems. Thus, we need an automatic system to identify claims that are most critical in terms of future consequences to a person or organization (i.e., check-worthy claims) [21]. Current solutions to the problem employ supervised machine learning techniques [27], inducing a need for annotated datasets to train these systems and evaluate their effectiveness. This task has only been part of the goals of automatic fact checking for the past few years; expectedly, a limited number of annotated datasets exist for any language. To overcome this limitation for Arabic, this work hypothesizes that constructing a language-specific (specifically, Arabic-specific) dataset can be avoided all together. To verify that hypothesis, I aim to study the possibility and effectiveness of cross-lingual transfer learning [28] from a *source* language to *Arabic* for the task of check-worthy claim identification over tweets. In this dissertation, I answer the following research questions.

- Given labeled data in a source language, how effective is zero-shot cross-lingual check-worthiness prediction on Arabic?
- Does translation between source languages and Arabic improve the performance?
- How much improvement can be achieved by adding few labeled Arabic examples to the examples in the source language (i.e., few-shot transfer learning)?
- Can the performance be improved if transfer is done from multiple source languages to Arabic?
- How effective is cross-lingual transfer compared to the state-of-the-art models?

#### *1.1.2.2. Evidence Retrieval over the Web*

Once the check-worthy claims are identified, the actual verification process can start, usually with evidence retrieval. This task can be defined as the process of retrieving information sources (e.g., Web pages) against which a claim can be verified [24]. In long documents like Web pages, only a portion of the page (e.g., one ore more paragraphs) actually includes the information that can help verify a claim. Identifying this useful portion of a Web page (i.e., the *evidence*) is not a straightforward task since the evidence should not only be relevant to the topic of the claim, but should also hold information supporting or denying it. Existing evidence retrieval systems usually start by retrieving long documents like Web or Wikipedia pages for a given claim using an ad hoc retrieval model, and then identifying evidence within them (such as the systems participating in the FEVER challenge [29]). Quantifying the effectiveness of such systems in retrieving documents containing evidence (i.e., evidential documents) for claim verification has been overlooked. Moreover, identifying the textual features distinguishing evidential documents from those that are merely on-topic of the claim (i.e., topical documents) was not previously-studied. My work aims at addressing both concerns, while also providing a first-of-its-kind Arabic evaluation dataset for the task of evidence retrieval over the Arabic Web. Specifically, my work answers the following research questions.

- To what extent topical pages are evidential, and how correlated is the effectiveness of retrieving these two types of pages?

- What types of evidence can be found in evidential pages?
- What textual features distinguish evidential and non-evidential pages?
- How effective are existing systems in retrieving evidential pages?

## 1.2. Contributions

In addressing the above research questions, this dissertation makes the following contributions.

### *1.2.1. Ad hoc Retrieval over the Web*

In Chapter 3, I present an approach to construct a public large-scale test collection (ArTest) over the largest available Arabic Web collection called ArabicWeb16 [30]. To construct the test collection, we hire and train in-house annotators to construct topics reflecting information needs of real users over ArabicWeb16. Annotators were also asked to create multiple queries representing each topic. We then used these queries to retrieve the pools of documents to judge per topic, eliminating the need for a shared-task campaign. The contributions of this work are as follows.

- We develop and share ArTest, the *first* test collection for the evaluation of Web search over the *Arabic* Web.[5] The collection includes 50 topics (and the queries used to develop them), and an associated set of 10,529 judged document-topic pairs.
- I demonstrate the usability of ArTest by evaluating existing state-of-the-art neural retrieval models using the collection. The resulting performance scores constitute reference baselines for future studies.

### *1.2.2. Ad hoc Retrieval over Social Media*

Chapter 4 describes the process of constructing EveTAR, a public large-scale Arabic tweets test collection for the evaluation of ad hoc search systems. The collection includes events as the topics and similar to ArTest, we use query variations to construct the pools of tweets to judge. We handle the scale of annotations to be done by hiring crowdworkers to annotate tweets. Relevance annotations for events made it possible to evaluate both event detection and ad hoc retrieval systems over EveTAR. We further extend the annotations by annotating the relevant tweets by novelty, allowing for the evaluation of two more IR tasks. The contributions on this problem are 3-fold as follows.

- We introduce a novel language-neutral approach for multi-task test collection construction, without requiring a shared-task evaluation campaign. My main contribution was in the formalization and design of the approach. Moreover, I had a key role in annotation tasks design and implementation.
- We introduce and release[6] EveTAR, the *first* large-scale test collection over *Arabic* tweets that supports event detection, ad hoc search, timeline generation,

[5] http://qufaculty.qu.edu.qa/telsayed/datasets/
[6] http://qufaculty.qu.edu.qa/telsayed/evetar

and real-time summarization IR tasks. The collection contains the ids of 355M Arabic tweets, 50 events and 62K relevance judgments, novelty annotations, inter-annotator agreements, queries used to identify potentially-relevant tweets for the events, and documented design of the crowdsourcing tasks. We also release the annotations per tweet to support studies on crowdsourcing in IR.

- I demonstrate the usability of EveTAR by evaluating existing techniques for three of the supported tasks. The resulting performance scores constitute reference baselines for future studies.

### *1.2.3. Misinformation Detection: Check-worthy Claim Detection over Social Media*

In Chapter 5, I describe a proposed approach for constructing an annotated Arabic tweet dataset for the claim check-worthiness detection task. The approach resulted in the CT20–CWT–AR collection of tweets used to evaluate participating systems at the CheckThat! shared-task lab at CLEF2020 [31], [32] which I later extend for CLEF2021 [33], [34]. The latter lab edition also resulted in similar datasets in four languages enabling my extensive study of the effectiveness of cross-lingual check-worthiness prediction for Arabic. My contributions for this problem are detailed next.

- I extensively explore and benchmark diverse methods to train cross-lingual check-worthiness prediction models including zero-shot, few-shot, and translation-based approaches. Existing studies for the task have not provided such a large-scale comparative study with different variants.
- The work demonstrated that cross-lingual transfer learning from some languages to Arabic is comparable to monolingual models exclusively trained on the target language (i.e., Arabic).
- The study offers benchmarking experiments comparing cross-lingual models, state-of-the-art models and strong baselines tested over CT21–CWT–AR. This provides future research on the same dataset with necessary baseline results.
- Create and release two versions of the first of its kind dataset for evaluation of claim check-worthiness estimation and ranking over Arabic tweets. The latest version of the dataset includes 14 topics, and an associated set of 4,705 annotated Arabic tweets among which 1,270 are labelled as check-worthy.

### *1.2.4. Misinformation Detection: Evidence Retrieval over the Web*

Chapter 6 is concerned with understating how current retrieval models perform in retrieving documents that are topically relevant to a claim as opposed to those that contain evidence to verify it. The study is carried over the Arabic Web using a dataset we constructed for this purpose. I also investigate and identify characteristics or features specific to evidential Web pages. Finally, the question on whether these features are actually informative in evidential pages retrieval is examined through the development and evaluation of a supervised model. The contributions for this retrieval problem are listed next.

- I conducted the first in-depth comparative study of the performance of Web search for the tasks of retrieving topically-relevant vs. evidential pages for verifying a given claim, showing that the two tasks are inherently different.

- The study provides a thorough analysis of distinguishing characteristics of *evidence* appearing in *evidential* Web pages, which is rarely studied in existing literature. Furthermore, it shows that the identified characteristics, when leveraged in a supervised evidential pages retrieval model, lead to promising results.
- The study establishes benchmarking results over the given dataset and quantifies the potential performance gain Web search systems can attain to better support the task of retrieving evidential pages for fact-checking.
- We release an annotated dataset for the task of re-ranking of Web pages by usefulness for claim verification.[7] The dataset includes 2,641 Web pages that are potentially-relevant to 59 claims and annotated by both dimensions of relevance (i.e., topical and evidential) compared in this study.

## 1.3. Major Publications

This dissertation is based on the following research publications:

- Journal Articles
  - **M. Hasanain**, R. Suwaileh, T. Elsayed, M. Kutlu, and H. Almerekhi, "EveTAR: building a large-scale multi-task test collection over Arabic tweets," *Information Retrieval Journal*, vol. 21, no. 4, pp. 307–336, 2018
  - **M. Hasanain** and T. Elsayed, "Studying effectiveness of web search for fact checking," *Journal of the Association for Information Science and Technology*, Oct. 2021. eprint: `https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24577`
- Conference Papers
  - **M. Hasanain**, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, and P. Nakov, "Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality," L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds., ser. CEUR Workshop Proceedings, 2019
  - **M. Hasanain**, Y. Barkallah, R. Suwaileh, M. Kutlu, and T. Elsayed, "ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2017–2020
  - **M. Hasanain**, F. Haouari, R. Suwaileh, Z. Ali, B. Hamdan, T. Elsayed, A. Barrón-Cedeño, G. Da San Martino, and P. Nakov, "Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020
  - **M. Hasanain** and T. Elsayed, "bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020

[7] `http://qufaculty.qu.edu.qa/telsayed/datasets`

– S. Shaar, **M. Hasanain**, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, J. Beltrán, T. Elsayed, and P. Nakov, "Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021

## 1.4. Thesis Organization

The remainder of this dissertation is organized as follows. Chapter 2 presents an essential background on evaluation of IR systems (Section 2.1), followed by a survey of the literature on ad hoc retrieval and misinformation detection. Chapter 3 presents the work on ad hoc retrieval over the Web, including dataset construction, and experiments with neural retrieval models. Chapter 4 details the dataset construction approach, and benchmarking experiments over social media. In Chapter 5, I describe the process to construct a claim check-worthiness dataset from Arabic tweets, and an extensive study of cross-lingual transfer learning for the problem. Chapter 6 details the approach to the dataset constructed for evidence retrieval evaluation, followed by the analytical study of evidential retrieval versus topical, and a comparison of systems for the task. Finally, Chapter 7 offers some concluding thoughts, provides directions for future work on Arabic IR, and includes a detailed list of my publications relevant to this dissertation.

# CHAPTER 2: RELATED WORK

This chapter summarizes necessary concepts and discusses relevant studies to my work, focusing on evaluation approaches and resources following two main tasks: a) typical ad hoc retrieval over both Web and Social Media and b) Misinformation detection.

## 2.1. Background: Evaluation of IR Systems

We first start by setting the stage on IR systems evaluation by discussing evaluation using test collections and annotated datasets in general. Moreover, the section discusses evaluation approaches.

### *2.1.1. Test Collections*

Evaluating information retrieval (IR) systems is usually conducted following the Cranfield paradigm [15] using a test collection composed of (1) a collection of documents, (2) a set of topics representing information needs, and (3) relevance judgments indicating which of the collection documents are relevant to which topic [16]. Many test collections were created over time to evaluate a variety of retrieval tasks, especially through the Text REtreival Conference (TREC) evaluation campaigns. However, with the growing size of the data to be searched in practice (e.g., the Web), reliably evaluating systems that can be used at such scale requires a sufficiently-large set of manual relevance judgments [40], [41]. Under the test collection-based evaluation paradigm, collecting relevance judgments is considered the most time- and resource-consuming component of the process [41] as it requires hiring dedicated humans to judge thousands of documents. Thus, deciding on which document-topic pairs to be judged is not a decision to be taken lightly, since it is not feasible to judge all pairs. Consequently, there is a long track of research on how to build reliable test collections focusing on various issues such as the ideal topic set size [42], selecting search topics [43], [44], selecting document-topics pairs to be judged [45], and collecting relevance judgments from crowdworkers [46].

Although evaluation using test collections is the standard in the IR field, available Arabic collections are scarce, slowing down research on IR systems tackling Arabic content [2], [9], [11]–[13]. This hinders evaluation of systems for basic IR problems (e.g., ad hoc retrieval) and extending to more complex Arabic-specific problems such as cross-dialect retrieval. Majority of the available collections are very old. For example, the TREC-2001/2002 Cross-Language Information Retrieval (CLIR) dataset is two decades old [17], [18]. The more recent collections focus on short documents like social media posts as did our "EveTAR" Twitter collection [35], [47], or on religious text such as "AyaTEC" built to support question answering over the holy Qur'an [48] and "Kunuz" providing evaluation resources for IR and CLIR over Hadith[1] [49].

So far, we have discussed evaluation of systems for typical IR problems like ad hoc retrieval and question answering. Within the IR domain, we observe an increased interest in a related space of problems that fall into the broad problem of text classification [8], which can be defined as follows. Given a set of classes, the aim is to decide the class of an input text piece. Example text classification problems include spam e-mail or Web pages detection, sentiment prediction, and fake news detection. Going back to the ad hoc retrieval problem, it can also be solved using text classification

[1]The traditions or sayings of the Prophet Muhammad (saws)

techniques where for a set of potentially-relevant documents to a topic, a classifier can be used to categorize these documents as either relevant or non-relevant. Automatic text classification is performed using machine learning techniques that attempt to learn the classification criteria from input examples, in hope this criteria can generalize to unseen documents. In this dissertation, I focus on supervised classification where a classifier is trained using a *training* dataset of text pieces (e.g., full documents or short text snippets) annotated by the given classes. Similarly, to evaluate the trained classifier, we need an annotated *test* dataset [8]. As with typical test collections, such datasets are generally constructed with the help of human judges, making them an expensive resource.

Many Arabic annotated datatses exist for a variety of text classification problems. Among the problems that gained most popularity, in terms of dataset construction and system development, is the problem of sentiment analysis and classification [50]–[53]. Generally, as with the majority of Arabic text classification problems, efforts on solving this problem do not agree on a common dataset, and some recent benchmarking studies attempted to alleviate this issue (e.g., the work of Farha and Magdy [54]). Fact-checking is another umbrella under which many text classification problems are emerging. A variety of annotated datasets related to fact-checking were constructed in the past few years. Majority of existing Arabic collections target claim verification, with the annotated set mainly composed of claims annotated by veracity as in "ArCOV19-rumors" [55] and "AraStance" [56]. Other datasets provide multi-labeled documents for several sub-problems in the domain. For example, the "AraCOVID19-MFH" dataset [57] includes tweets labelled by 10 labels, like hate speech and claim veracity. Among the pioneering Arabic collections in this domain are those I lead their development as part of the CheckThat! lab at the Conference and Labs of the Evaluation Forum (CLEF). These Arabic datasets served three problems, namely, claim check-worthiness estimation [31], [34], evidence retrieval for verification [31], [37], and claim verification [31], [37]. These datasets are discussed in details in Chapter 5 and Chapter 6.

### *2.1.2. Evaluation Approaches*

Evaluation of IR systems is "the process of assessing how well a system meets the information needs of its users" [58]. Typically, *relevance* of the retrieved documents to the user's information need has been the main measure of user satisfaction regarding these documents and consequently, the system effectiveness [59]–[63]. However, recent years have witnessed more complex IR tasks and modular systems with multiple components, posing new challenges to the traditional approach to evaluation on basis of relevance [63]. One rising trend in IR evaluation is the consideration of other dimensions of relevance [25], [62], [64]–[67] such as document understandability [66] or credibility [68]. Next, I present studies on one of these dimensions, namely, usefulness since it is core to the resources and evaluation study in Chapter 6.

Vakkari [69] presented an elaborate survey on usefulness evaluation in the IR field. This survey found that research usually agrees on usefulness definition, where usefulness of retrieved results is defined as the extent to which information in retrieved documents contribute to performing a larger task. In my work, the larger task is claim verification, and useful documents are those that give evidence needed to fulfill this task. In a related work, Mao, Liu, Zhou, *et al.* [70] compare relevance to usefulness evaluation and investigate, through a user study, how they correlate to user satisfaction. The study was carried over search tasks (topics) that are generally informational in nature (i.e.,

users trying to find information about a certain topic or for a larger task). Majority of existing work studied perceived usefulness as judged by real users or external annotators. Vakkari, Völske, Potthast, *et al.* [71] took it a step further and studied actual usefulness by asking users to benefit from retrieved documents in doing a writing task. Their focus was on information gathering under some search topics. Differently from existing work, my work focuses on measuring usefulness and contrasting it with topical relevance for a specific task, which is claim verification. Furthermore, I identify some of the features distinguishing content of useful/evidential Web pages.

## 2.2. Ad hoc Retrieval

Ad hoc search is one of main retrieval tasks tackled in this dissertation, thus, it is essential to have a general overview of existing systems designed for the task. Relative to high-resource languages like English, fewer search systems were tested on Arabic content; even fewer considered the special characteristics of Arabic text as part of the system design. Darwish, Magdy, and Mourad [6] observed that several spelling variations are frequent in Arabic tweets, such as word shortening, word elongation using repeated letters, and borrowing of similar-looking letters from other languages (e.g., Farsi). Authors proposed Arabic text normalization approaches to handle these issues for the ultimate goal of improving microblog (tweets) search. The experiments were conducted over a *private* tweets test collection with 112M Arabic tweets, 35 topics and their relevance judgments. Experiments showed that this enhanced normalization, followed by ranking using the Okapi BM25 model [72], [73] resulted in significant improvements over the case when no normalization is done. Almazrua, Almazrua, and Alkhalifa [74] started with a similar goal of improving Arabic tweets retrieval, but investigated the optimal stemmer to use. Experiments compared nine existing stemmers with BM25 as the ranking model. Evaluation was done over our EveTAR test collection; it showed that root-based stemmers resulted in best retrieval performance.

With the rise of neural architectures in the IR field, many neural retrieval models were recently proposed, mainly tested on English datasets. A recent and detailed survey [75] provided in-depth examination and comparisons of majority of existing neural retrieval models. The study demonstrated the effectiveness and robustness of models built on top of BERT (Bidirectional Encoder Representations from Transformers) contextual embedding model [76]. Utilization of neural retrieval models is limited over Arabic content, and existing studies usually utilize them in a cross-language retrieval setup as in [77], [13] and [78]. Therefore, my work demonstrates the development of a large-scale public Arabic Web test collection for ad hoc retrieval, and benchmark some of the most effective neural retrieval models on this collection. Discussing the details of these models is left for Chapter 3

## 2.3. Misinformation Detection

Misinformation on the Web and social media encouraged research on approaches to battle this flood of false information. Due to the volume of proposed systems to solve problems in this domain, several literature surveys already exist in literature targeting this area (e.g., [21], [79]–[83]). Two problems in this area were targeted by majority of existing studies, namely claim detection, and verification. My work also targeted evaluation issues and systems for both problems. Thus, this section presents the most

relevant studies to two problems: claim check-worthiness identification over tweets, and claim verification.

### *2.3.1. Check-worthy Claim Identification*

ClaimBuster is one of the pioneering approaches to the problem [84]. The system computes features for each input sentence such as its sentiment score, length, and part-of-speech tags and trains a supervised model with typical classifiers (e.g., SVM). The model was tested on political debates and tweets but limited to English. More recent systems usually use neural models and specifically, classification architectures based on transformer models (e.g., [85]–[90]). A more recent version of ClaimBuster combines BERT [76] and gradient-based adversarial training to build a more effective model. In this system, perturbations are added to the embeddings generated by BERT for an input sentence, and the final model is fine-tuned minimizing both classification and adversarial losses. The approach was tested over English sentences only. Differently and more comprehensively, my work examines multiple alternatives for cross-lingual transfer learning to Arabic where minimal or no training data in Arabic is required. Moreover, the work is not limited to tweets on one topic (i.e., COVID-19) potentially affecting generalizability of the proposed approaches to new topics.

Among the most prominent efforts to approach the problem of check-worthiness detection are those part of the CLEF CheckThat! lab for the past four years [33]. In the initial two editions of the lab, the problem targeted claims within political debates [91], [92]. In the next editions, the lab focused on the social media domain and specifically, check-worthiness estimation for tweets [31], [34], [93]. The problem was defined as follows: given a stream of tweets on a topic, the participating systems were asked to rank the tweets by check-worthiness for the topic. My work adopts a more general definition of the problem, modeling it as a classification task without a limitation to any topic. That is to say, the aim is to develop a system to detect check-worthy claims in a general stream of tweets. This definition is inline with some of the existing studies [84], [94], [95].

The last lab edition (CheckThat! 2021) offered a first-of-its-kind multilingual dataset (CT21–CWT) for the problem. The dataset contained labelled tweets in five languages: Arabic, Bulgarian, English, Spanish and Turkish. This is the evaluation dataset used in Chapter 5 (further details in Section 5.3.1). Only few systems participating in the lab attempted to benefit from the unique nature of this dataset. Schlicht et al. (team UPV) [96] proposed a transformer-based model jointly trained for two tasks: check-worthiness detection and language identification. The team fine-tunes a multilingual transformer model called sentence-BERT [97] optimizing for both classification tasks. The aim of the language identification task is mitigating bias to any of the training languages. Again, in their study, authors train the model over all five languages in CT21–CWT, however, I focus on cross-lingual transfer. In a very recent study [98], a dataset of English and Arabic tweets about COVID-19 was annotated on several aspects including check-worthiness. The authors fine-tune several transformer models for the task, but train a model for each of Arabic and English independently. In a further study [99], the dataset was augmented with Bulgarian and Dutch tweets, and initial experiments on multilingual classification were conducted. In the proposed system, a multilingual BERT (mBERT) transformer model [76] is fine-tuned using all of the four languages and then tested on each. Uyangodage, Ranasinghe, and Hettiarachchi

[100] follow the exact same approach but considering two datasets: NLP4IF [101] and the CheckThat! 2021 Task 1 dataset [34].

The work of Zengin et al. [102] is the closest to mine. In their study, authors attempted a cross-lingual approach where mBERT is fine-tuned on each pair of the five languages in CT21–CWT, then tested per language. Differently, I examine a wider set of variants for cross-lingual check-worthiness estimation and show how they compare to several existing baselines. In a more recent work by the same authors [103], they test mBERT performance in cross-lingual transfer for three languages (Arabic, English and Turkish). However, I observe a potential source of issues in their evaluation setup since the datasets came from different domains (political debates and tweets) and follow different annotation strategies across the languages, while in my setup, I maintain consistency as much as possible using tweets only.

### *2.3.2. Claim Verification*

Claim verification is a problem that is usually at the core of many applications, such as rumor verification and fake news detection. Due to its importance, many studies exist in literature considering different dimensions to the problem. Some systems are proposed to tackle the dimension of verification efficiency (e.g., [104]), while other studies focus on the dimension of modelling the verification resources for optimal system effectiveness (e.g., [105]). A strong line of research is evolving around a third dimension, where the aim is to integrate interpretability in the system design (e.g., [106]). In this dissertation, the focus is on that dimension, which is also gaining importance in other problems solved through deep or traditional machine learning approaches.

System decision explainability is essential in the fact-checking domain. To trust the verification system's decision or even verify it further, the user is expecting interpretable decisions usually explained in terms of evidence used to make these decisions [107]. Recently, studies demonstrated the value of and need for extracting *evidence* snippets from identified information sources. *Evidence* is essential to justify or explain the system's veracity prediction and provide user with information to make further assessment and decision regarding the claim's veracity [108]–[110]. The following sections summarize relevant studies on this type of systems.

#### *2.3.2.1. Evidence-based Verification Systems*

Ma, Gao, Joty, *et al.* [109] proposed a hierarchical neural network using attention to capture sentence topical coherence and semantic entailment with respect to the claim. The DeClarE system is also based on a neural network that predicts claim veracity given related articles. Attention is used to capture article salient words with respect to the claim and present them as evidence [111].

As part of the popular FEVER challenge on evidence-based fact-checking, several systems have been proposed (e.g., [112], [113]). The task focused on Wikipedia articles only from which systems are required to extract stance-based evidence. Another recent challenge is Task 2.A of the CheckThat! lab at CLEF2019 [37]. Similar to the focus of my work, the task targeted source-based evidence and systems were required to rank pages potentially related to a claim by usefulness. Proposed approaches included a BERT model to rank pages [114], and a learning-to-rank model using page credibility and similarity to the claim as features [115]. All aforementioned works, evaluated systems by effectiveness of performing the required task and did not clearly identify

features most helpful in characterizing evidence. Moreover, most of evidence-based fact checking systems aimed to identify/use stance-based evidence while we are interested in source-based evidence (Further details in Chapter 6).

#### *2.3.2.2. Analysis of Web Pages for Verification*

Several studies in the fact-checking domain analyzed language in documents. Work by Rinott, Dankin, Perez, *et al.* [116] is the closest to mine. In their work, authors designed features that capture types of supporting evidence extracted from articles related to a given topic. Features depended on a manually-crafted lexicon for each evidence type, patterns (e.g., presence of quotes), named entities and subjectivity words. Contrary to my work, the focus was on developing a system for evidence ranking. Moreover, effectiveness of the proposed features in characterizing evidence was not studied. Finally, experiments were limited to Wikipedia articles while I consider the general Web. Wang, Yu, Baumgartner, *et al.* [117] designed features to characterize and classify documents as supporting or refuting a claim. Features were mainly textual similarity and entity-based features in addition to a manually-crafted lexicon to detect contradicting discourse. Experiments were conducted on documents acquired by searching the Web using a commercial search engine. Evaluation of retrieval performance of the engine was done using recall of supporting documents which is a measure also considered in this dissertation. Moreover, the evaluation was focused on system classification accuracy and not on the linguistic characteristics of the retrieved pages. Jiang, Baumgartner, Ittycheriah, *et al.* [118] identified linguistic patterns that can aid in extraction of the claim, claimant and claim veracity from fact-checking articles. Although the work pointed out the importance of extracting evidence from fact-checking articles, identifying this component was left for future work. Additionally, the study was limited to articles from fact-checking websites only, while I consider Web pages in general regardless of their domain.

In a different line of work, several studies focused on identifying linguistic characteristics differentiating trusted and false news (e.g., Jiang and Wilson [119] and Trielli and Diakopoulos [120]). Some studies investigated news articles using word lexicons of subjectivity, sentiment, and informal words [121], [122]. Other studies focused on social media. Volkova, Shaffer, Jang, *et al.* [123] analyzed language of tweets with suspicious and verified news using subjectivity, semantic, bias and psychological lexicons. Jiang and Wilson [119] analyzed language of user comments on social media posts using a manually-crafted lexicon to investigate the relation between post veracity and comments language. Trielli and Diakopoulos [120] also investigated features of sentiment and subjectivity in tweets and news article of different truthfulness levels. A clear difference between my work and these studies is that I aim to identify linguistic cues that can be used to extract evidence from Web pages as opposed to differentiating true and false documents.

## CHAPTER 3: AD HOC RETRIEVAL OVER THE WEB

Test collections are the cornerstone of the evaluation of information retrieval (IR) systems in the Cranfield paradigm [15], enabling experimental comparison of different approaches, and therefore pushing research on building effective IR systems. However, as has been thoroughly discussed in Section 2.1.1, building a reliable test collection usually requires a large budget due to the need for humans to annotate or judge documents. Moreover, critical design decisions need to be made to ensure the collection is reliable, and representative of the task and domain (e.g, the Web). Examples of such decisions include selecting the topic set to include in the collection [42], [44], and the topic-document pairs to judge [45]. Unfortunately, very limited Arabic test collections exist [2], [9]. In particular, even though the Arabic Web is a rich and constantly-growing source of information, publicly-available test collections serving Arabic *Web* search do not exist.

This chapter presents the construction process of ArTest, which is the *first* test collection for Arabic Web search. ArTest is built on top of the largest available Arabic web collection, ArabicWeb16 [30], that includes around 150 million Web pages. With the help of in-house annotators, 50 topics were developed and an average of 211 relevance judgments were made per topic. During the topic development phase, annotators were asked to create multiple queries representing each topic, such that these variations can be used to collect documents to judge; eliminating the need for system variations as in shared-task evaluation campaigns. We also make these queries available for future research on relevant IR problems, such as query generation, or analysis of the impact of user variance on IR systems (e.g., [124]).

To encourage development and evaluation of IR systems over ArTest, I take the first step and implement and evaluate standard and state-of-the-art neural retrieval models over this collection. By following a recent and exhaustive survey of neural retrieval models [75], I observe that the most robust and effective systems benefited from the widely-used and very effective BERT (Bidirectional Encoder Representations from Transformers) [76] contextual embedding model. Therefore, I compare the standard retrieval model BM25 [73], [125] to two neural models built over BERT. I experiment with one of the first BERT-based models called "monoBERT" [126], [127]. Additionally, I test the model "Birch-Passage" [128], which is an adapted version of "Birch" [129], as it showed consistently effective performance over two standard English test collections (Gov2 [130] and Robuts04 [131]). Preliminary experiments showed that the neural retrieval models are underperforming as compared to the standard BM25 model, and further investigation is needed to justify this unexpected behaviour.
The contributions of this work are two-fold:

1. We develop and share ArTest, the *first* test collection for the evaluation of Web search over the *Arabic* Web.[1] The collection includes 50 topics (and the queries used to develop them), and an associated set of 10,529 judged document-topic pairs. As part of this group effort, my main contribution was in designing the annotation tasks needed to implement the test collection construction approach.
2. I demonstrate the usability of ArTest by evaluating existing state-of-the-art neural retrieval models over the collection. The resulting performance scores constitute reference baselines for future studies.

In the following sections, the approach devised to construct the test collection is presented (Section 3.1). Next, I present benchmarking results for multiple IR baselines

[1] `http://qufaculty.qu.edu.qa/telsayed/datasets/`

over the collection (Section 3.2). Finally, Section 3.3 concludes the chapter and gives some guidelines for future work.

## 3.1. Dataset Construction

In this section, we describe the process followed to construct ArTest [38]. The process includes the following three steps:

1. Identifying the Web collection used in building ArTest (Section 3.1.1).
2. Collecting a set of topics belonging to the Web collection (Section 3.1.2).
3. Constructing the qrels set with the help of relevance assessors (Section 3.1.3).

### *3.1.1. Document Collection*

Before constructing ArTest, it was essential to identify a suitable Arabic Web collection to be its backbone, and the source of its documents and topics. After exploring existing Arabic Web collections, we opt to use ArabicWeb16 [30] due to it's very large size with 150M Arabic Web pages. Furthermore, the collection is recent constructed on January 2016 and represents a good snapshot of the Arabic Web. It covers diverse types of Web pages such as forums, news articles, Wikipedia, etc. and includes both modern standard Arabic (MSA) and dialectal pages.

### *3.1.2. Topics*

The second component in a test collection is the set of topics representing a sample of users potential information needs. There are several design decisions that need to be made for a topic set (e.g., size, types of topics, etc.). This section describes these design decisions for ArTest, including setting the number of topics, followed by explaining the topic development and topic selection phases.

#### *3.1.2.1. Topics Set Size*

Deciding the size of a topic set is among the most influential decisions in test collection creation, since it has a direct effect of the collection cost and quality. Extensive research efforts in the IR evaluation community have experimented with the required number of topics that lead to reliable evaluation (e.g., [132]). Among the initial efforts, Jones and Van Rijsbergen [133] suggested a minimum of 75 topics. More recently, Buckley and Voorhees [132] showed that 50 topic result a stable evaluation. This golden number of 50 topics has been the standard for many TREC test collections (e.g., [134]). Thus, ArTest was decided to have a set of 50 search topics.

#### *3.1.2.2. Topics Development*

Before constructing the topic set, we identify the main criteria that the topics should fit. A topic should capture a real user's information need which can have different goals (e.g., informational or navigational) and about diverse subjects (e.g religion, art, politics, etc.). Additionally, the topic set should be constructed considering availability of relevant documents in the document collection since this affects evaluation of IR

systems. Some topics might have very few relevant documents, while others might have excessive number of relevant documents. If the topic set includes many topics belonging to both extremes, reliable evaluation of IR systems can be difficult.

Starting with the above criteria, we first start by recruiting real users to provide us with their information needs over ArabicWeb16. The recruitment process started by advertising the task to the community through social media and contacting interested potential *topic developers*. Topic developers then went through a training session in which we explained the required task (e.g., what's a topic and how to define it) and demonstrated an online user interface we developed for this task (further details in [38]) to allow for interactive search over ArabicWeb16. Participants were also presented with example topics to clarify the concept. After this session, we retained 16 topic developers including university students or graduates from various educational backgrounds such as arts, economics, and engineering.

Then, with the help of a colleague from my research team, I lead and monitored the actual topic development sessions carried in house, which enabled us to provide direct guidance during topic development. In our work, we instructed developers to represent a topic by the commonly-adopted representation in TREC test collections, as follows:

- **Title**: A summary of the topic using few words.
- **Description**: A description of the information need using a couple of sentences.
- **Narrative**: An extended description of the topic specifically clarifying the difference between a relevant document and a non-relevant one, which is useful for the judges in the relevance judgment phase.

During topic development, the online interface allowed judges to re-shape and refine their topics using interactive search by creating and searching using multiple queries representing each topic. For each topic, we instructed developers to judge the relevance of the top 50 Web pages retrieved. This preliminary relevance annotation step is necessary to guide topic selection later and to help the developers better understand and shape their information need, resulting in clearer topic definition.

During topic creation and given the 50 relevance judgments, any topic with very few relevant pages ($<5$) was dropped. Eventually, developers created 62 topics, with an average of $\approx 4$ topics per developer. The topics covered variety of subjects and goals. Some example topics include: "Safety of GMO crops consumption", "Effect of early marriage on young adults", "Drones", and "Devaluation of Egyptian pound". A full example topic is presented in figure 3.1.

#### *3.1.2.3. Topics Selection*

Starting from the set of 62 topics, we applied two filtration steps to reach our goal topic set size of 50. Initially, we observed a notable number of topics about diseases, which can bias the test collection. Therefore , we manually filtered six of them out. Finally, after the full relevance judgments phase for the remaining 56 topics (see Section 3.1.3), I randomly sampled 10% of the judged documents of each topic, and manually labelled them to compare with the topic developers'. I assume that clearly described topics result in higher annotation agreement between the topic developer and

**رقم الموضوع:** 23

**العنوان:** نتائج الحرب العالمية الأولى

**الوصف:** البحث عن النتائج التي انتهت عليها الحرب العالميه الأولى من عدة جهات

**الشرح المفصل:** البحث عن مقالات تتحدث عن نتائج الحرب العالميه الأولى من عده جهات و كون المقال ذا صلة اذا احتوى على النتائج الاقتصادية او النتائج السياسية او الاجتماعية للحرب او تضمن بنود معاهده فرساي.

**Topic ID:** 23

**Title:** Outcomes of World War I

**Description:** Find information on outcomes of the First World War from different aspects.

**Narrative:** Search for Web pages that discuss the results of World War I on different fronts. A page is considered relevant if it discusses the economical, political, or social consequences of the war. Pages discussing the terms of the Treaty of Versailles are also relevant.

Figure 3.1. An example topic in ArTest

a secondary assessor. Thus, our final topic set included the 50 topics with the highest agreement level ($\geq$85%) between the topic developers' judgments and mine.

For each topic, annotators developed $\approx 4$ queries on average during topic development. These queries were also released as part of ArTest to support future research on related problems such as query generation. However, we dropped the initial set of 50 relevance judgments done during topic development, since this phase included topic calibration [135]. Eventually, ArTest includes a set of 50 topics covering various domains such as religion, health, politics & economy, and science & technology, reflecting the different types of user interests.

### *3.1.3. Relevance Judgments*

To complete the construction of ArTest, the final step is to collect relevance judgments for our 50 topics. This phase includes two steps: document selection and the actual relevance judging process, as described in this section.

#### *3.1.3.1. Document Selection*

The first and key challenge in constructing the qrels set for a test collection, is deciding which documents to judge for each topic. The optimal or extreme scenario is to try to find every relevant document from the entire document collection, which can be impossible with big collections like ArabicWeb16. Thus, several approximations have become the state of the art. The most common approach is through shared-task evaluation campaigns. In a shared task, several participating IR systems produce ranked results list for a search topic. Given these lists, the documents to be judged are selected using various methods such as pooling [136]. However, organizing such evaluation campaigns is complex, time-consuming, and its effectiveness in identifying documents to judge is dependent on the number of participants. Therefore, we had to identify a more efficient but effective approach. We adopt an existing one which is based on interactive search [137]. In a recent work, Moffat, Scholer, Thomas, *et al.* [138] demonstrated

that for a single topic, it is possible to acquire document pools through query variations that are as diverse as those resulting from system variations. We resort to this approach instead of using multiple IR systems as in shared-tasks. For each topic, we allow annotators to search for documents using various queries. Retrieval is powered using the BM25 retrieval model (over page title and content).

#### *3.1.3.2. Judging Process*

The creator of a topic is the most knowledgeable of the real information need of that topic, making him/her the best to judge the relevance of documents for it. Therefore, for each topic, we recruited its developer to perform relevance annotation. With the help with an online annotation interface ([38]), judges were free to perform the task outside the lab since the task is very time consuming. During relevance judging, each developer was presented with the full topic description created during topic development. As explained earlier, judges first created several queries for a topic, performed interactive search and judged at least 200 pages per topic.

#### *3.1.3.3. Quality*

To verify annotation quality, we manually annotated a 10% sample of the relevance judgments for each topic (See Section3.1.2.3) for quality control. Then, we compute percentage of agreement between the original labels and our annotations. Eventually, high agreement ($\geq$85%) was observed with all 50 topics in ArTest.

## 3.2. Benchmarking

This section starts by a brief overview of BERT Transformer model, followed by describing the BERT-based retrieval models tested in this work. Then, the experimental setup and results are described.

### *3.2.1. BERT for Ad hoc Retrieval*

BERT is a neural network model that is pre-trained over very large *unlabelled* document collections [76]. Following this pre-training, BERT can provide contextual representations for input text sequences. Starting from a publicly available BERT model, we can train (i.e., fine-tune) the model for downstream tasks such as sentiment classification, by fine-tuning the model parameters using *labelled* data for the task [76]. BERT was pre-trained using input sequences with up to 512 token, making it difficult for the model to handle longer documents. In section 3.2.2, I explain how this limitation was handled.

Due to the proven effectiveness of BERT for many text classification problems, it found its way to the ad hoc retrieval problem. To solve the ranking problem using a classification technique, we can model it as a text classification problem where we aim to learn a model that estimates the probability that a document is relevant to the input query [75]. At inference time, the documents can be ranked using this probability. Starting from this simple re-formulation of the problem, many BERT-based retrieval models were developed following a two-stage ranking architecture as illustrated in Figure 3.2. In this architecture, an unsupervised retrieval model that is based on keyword matching (e.g., BM25) is used to retrieve an initial ranked list of documents

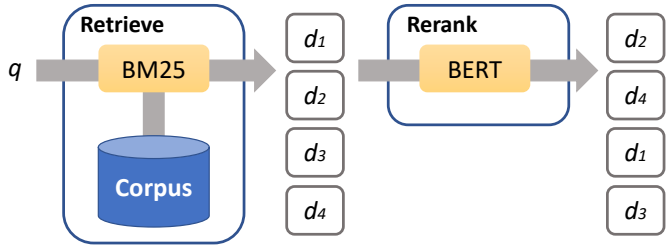



Figure 3.2. Illustration of a two-stage ranking architecture. In the first stage, a keyword search model (BM25 in the figure) retrieves a ranked list of documents from the collection. The second stage uses a neural network based on BERT to re-score and rerank the documents.

for a given query. In the second stage, the fine-tuned BERT model generates a score per document used to rerank this list.

To use BERT for text ranking, we pass the query tokens ($q$), and tokens from the document to be scored ($d_i$) to the model, formatted as follows: [[CLS], $q$, [SEP], $d_i$, [SEP]]. [CLS] and [SEP] are special tokens from BERT. Typically, after training the full architecture (including fine-tuning of the pre-trained model), the hidden state **h** produced by the transformer model for the [CLS] token is used as representation of the input to the remainder of the classification architecture. I continue to use this representation for all neural retrieval models.

Initially, BERT was pre-trained over English data, but recent years witnessed the release of many Arabic BERT variants that were pre-trained over Arabic collections. In this work, I adopt the Arabic model AraBERT [139] as the main model since it showed to be effective for the ad hoc retrieval task in my preliminary experiments. I also show experiments using two other Arabic BERT models: ARBERT and MARBERT [140].

### *3.2.2. Retrieval Models*

Three retrieval models were tested over ArTest as summarized below.

- **BM25** [72] is an unsupervised retrieval model that is based on keyword matching between the query and document. This model is usually used in the first retrieval stage for neural models following two-stage ranking architecture, in addition to serving as a standard baseline [75].

- **monoBERT** [126], [127] follows the exact same multi-stage ranking architecture explained in Section 3.2.1. Starting from a list of candidate documents retrieved by a first retrieval stage, the query $q$ and each candidate document $d_i$ is passed to the BERT model. The resulting contextual representation of the [CLS] token is then passed to a single-layer, fully connected classification neural network (#hidden nodes=768) to acquire the probability that $d_i$ is relevant to $q$. During training, the full architecture, including BERT layers, and the classification network, are trained to minimize cross-entropy loss for the relevance classification task. During training, I train the model on relevance judgments from ArTest. The input to the model is limited to 512 tokens. For each query-page pair, I pass the full query tokens in addition to the Web page truncated such that the total sequence length is 512 tokens (including the special tokens [CLS] and [SEP]).

I hypothesize that this long of a sequence will at least cover the first paragraph of the Web page that usually contains a summary of the page.

- **Birch-Passage** [128] again uses the two-stage ranking architecture. In the reranking stage, the document is split into overlapping passages using a sliding window. Each query-passage pair is passed to BERT, and then inference is applied following the same approach in monoBERT. Finally, the scores of top scoring $n$ passages are aggregated, reflecting the full document's score. In my implementation of this model, I set the passage length to 450 tokens with a stride of 425 tokens. As with the original model, I limit the number of passages extracted from a Web page to 16. I always maintain the first and last passages from a page and randomly sample from the rest till a total of 16 passages is reached. At training time, I assume each passage coming from a relevant Web page is relevant and those from non-relevant pages as non-relevant. This assumption is inline with some of the existing neural retrieval models [75]. The model is trained over these per-passage labels. For inference, for each Web page, the average of the scores of the top $m$ passages is used to represent the page score.

### *3.2.3. Experimental Setup*

For training and evaluation, I follow a 5-fold cross validation approach where folds were created by splitting the relevance judgments set by topics. For each training step, training is done over 40 topics and their associated judged Web pages from ArTest. As for inference, it is applied to 10 topics and their corresponding Web pages coming from the first retrieval stage. Initial retrieval is done through the BM25 model over a Lucene index of the ArabicWeb16; Lucene's default parameters were used to configure the model. In this retrieval stage, I preprocess documents and queries by applying stemming and stopwords removal. The number of retrieved Web pages per query was set to 100.

For the reranking stage, the number of training epochs was set to 2 and the batch size to 8. Systems are evaluated using MAP@100, Precision at Rank 10 (P@10), and Recall at rank 100 (R@100).[2] For the Birch-Passage model, the predicted pages scores were the result of aggregating the score of the top three passages as done in the original work [128].[3] For reranking, I follow the preprocessing used by the underlying BERT model, to ensure the queries and Web pages terms match those in the language model. This preprocessing pipeline applies minimal Arabic text normalisation (e.g., removing elongations and English characters).

### *3.2.4. Results and Discussion*

Table 3.1 compares the performance of the three retrieval models. As the table shows, BM25 is outperforming both neural retrieval models by a big margin. I investigate this unexpected performance by manually inspecting the pages used for training and those reranked. My initial investigation showed that a large percentage of the Web

[2]Not that R@100 for the neural retrieval models is constrained by recall achieved in the first retrieval stage.

[3]During experiments, I found that changing the number of sentences between 1 and 5 had a negligible effect on the system performance.

Table 3.1. Retrieval models performance. Results for best model by $P@10$ are boldfaced.

| Model | *MAP@100* | *P@10* | *R@100* |
|---|---|---|---|
| BM25 | **0.114** | **0.372** | **0.252** |
| monoBERT | 0.103 | 0.310 | 0.252 |
| Birch-Passage | 0.100 | 0.314 | 0.252 |

pages are actually threads from forum pages. This is inline with a similar observation made by the creators of ArabicWeb16 on which we built our test collection. Suwaileh, Kutlu, Fathima, *et al.* [30] reported that more than 50% of the most-common domains in ArabicWeb16 are actually forums. From my observation, parsing the html content of forum pages, in order to process the pages in my experiments, was not effective due to the varied and nonstandard html formatting used with these forums. Moreover, forum pages can contain very long threads of discussions, diluting the portions of the page that actually contain the relevant information. During manual annotation of a forum page, isolating the relevant information is possible by a human since he/she can probably easily understand the structure of forum discussions. However, such task will be very challenging for a machine tackling messy page formatting. This challenge is especially catastrophic for transformer models like BERT, since it can not handle long input sequences, i.e., we can not pass the whole page at once. Training using automatically extracted passages from a page, or truncating it can lead to loss of relevance signal during training. These concerns need further investigation to decide on the optimal approach to train the models under such setup and even to apply the model at inference time.

Next, I investigate whether the performance of the neural retrieval models is affected by the Arabic BERT variant used (Table 3.2). The aim here is to understand whether the performance of these models will improve with other Arabic BERT models. For that purpose, I re-train the monoBERT architecture with two other BERT variants, namely, ARBERT and MARBERT, which are state-of-the-art Arabic BERT models [140]. I focus on monoBERT since both monoBERT and Birch-Passage performed similarly with AraBERT. The table shows that both AraBERT and ARBERT result in comparable performance, while MARBERT is clearly lagging behind. This is an interesting outcome but can be justified by the fact that MARBERT was actually pre-trained using billions of tweets rather than news article or Web pages as in AraBERT and MARBERT. This observation indicates that at least on our test collection and retrieval models, the pre-training domain and test collection domain should be consistent to allow for knowledge transfer. This is not fully consistent with literature over English data, that found that transfer is possible between a BERT-based retrieval model fine-tuned over tweets and the domain of news articles. Such observation conflicting with existing studies motivates future work on analyzing and understanding the behaviour of the different existing neural retrieval models over our Arabic collection.

## 3.3. Conclusions and Future Work

This chapter presented ArTest, which is the first large-scale Arabic Web test collection. ArTest was constructed with the help of in-house annotators who were

Table 3.2. Performance of the monoBERT retrieval model over ArTest while varying the underlying Arabic BERT model used.

| **Model** | ***MAP@100*** | ***P@10*** | ***R@100*** |
|---|---|---|---|
| AraBERT | 0.103 | 0.370 | 0.252 |
| ARBERT | 0.102 | 0.318 | 0.252 |
| MARBERT | 0.076 | 0.226 | 0.252 |

thoroughly trained for topic development and relevance annotation. Annotators were responsible for creating topics reflecting real information needs over a large Arabic collection. Relevance annotations were done without the need for a shared-task evaluation campaign, by engaging annotators in an interactive search approach overArabicWeb16. The resulting judgments had an agreement level that is greater than 85% for all topics, when compared to a random sample of Web pages annotated by the team managing the construction of the test collection. ArTest eventually contains 50 topics (and the queries used to develop them), and the associated set of 10,529 judged document-topic pairs; all are made publicly available.

In this work, I also applied existing effective neural retrieval models, that use BERT transformer model, over ArTest. These experiments have raised interesting questions. Results showed that these models did not beat the typical BM25 model, conflicting with the relative results over English datasets. Indeed, these experiments are still preliminary. With the scale of progress in neural retrieval models, a tremendous experimental and analysis effort is needed to establish the stat-of-the-art over our Arabic test collection.

Experimenting with further neural retrieval models is a clear future direction. Moreover, running in-depth failure analysis on the performance of such models and how affected they are by features specific to Arabic is another necessary exercise.

# CHAPTER 4: AD HOC RETRIEVAL OVER SOCIAL MEDIA

Hundreds of millions of tweets are posted on Twitter daily. Among those, tens of millions are Arabic [1]. Not only the platform is a key source for news sharing and reading, it is heavily used to share updates on real-world events as they progress; sometimes even beating typical mainstream media in publication frequency and breaking news [141], [142]. Furthermore, Twitter discussions about an event, such as the US 2016 presidential elections, can lead to significant impact on the event's consequences [143]. This initiated an intense need for automatic tools that can perform multiple tasks to process event-related Arabic tweets. Examples include automatic detection of events as they emerge [144] (*Event Detection*), or once an event happens, users might be interested in searching for tweets relevant to the event [6] (*Ad hoc Retrieval*), or receiving real-time summaries about an event while it is developing over time [145] (*Real-time Summarization*), or even requesting a summarized timeline of the event once it concludes (*Timeline Generation*).

To evaluate information retrieval (IR) systems serving the aforementioned tasks, test collections are evidently needed. Several Twitter test collections that support a broad range of retrieval tasks are already available [146]–[151]; however none of them is in Arabic. Moreover, shared-task evaluation campaigns were used to create most of these collections. The campaigns included the participation of several research teams resulting in a pool of documents to be judged and later constituting the judgments in the test collection. Additionally, the annotation of a judgment pool usually depends on experienced annotators, making the construction of such collections even more costly and time consuming. Acquiring such resources is not possible for many languages of relatively-low research resources such as Arabic.

This chapter addresses the problem of constructing a large-scale tweets test collection supporting multiple tasks, without conducting a shared-task campaign. We adopted a language-neutral approach with *significant* events at the core of the collection. We define a significant event as a happening that occurs at a particular time in a specific location and is covered by the media (e.g., discussed in an online news article). We elect to focus on significant events for multiple reasons. First, it allows us to develop topics that reflect information needs of Twitter users, since they usually turn to Twitter for timely updates about events [152]–[156]. Second, popular events will probably have rich content. Finally, it inherently allows supporting multiple IR tasks that generally revolve around events.

Applying our approach over *Arabic* tweets resulted in **EveTAR**, our **Eve**nt-centric **T**est Collection of **Ar**abic Tweets. The collection is constructed over a month crawl of 355M Arabic tweets, with topics covering 50 events selected from Wikipedia's Current Events Portal.[1] EveTAR supports four IR tasks: event detection, ad hoc search, timeline generation, and real-time summarization. We construct the judgment pool by running interactive search using multiple manually-crafted queries per topic. To ensure large-scale but high quality judgments, we first recruited crowd-workers to judge tweet *relevance*, then we filtered out the topics with the lowest inter-annotator agreement and dropped inaccessible tweets,[2] reaching a *substantial* agreement level reflected by a Kappa value of 0.71 over 62K judgments. For two of the supported tasks, surfacing novel tweets per topic from those relevant is needed, which was done by in-house annotators to ensure the annotations quality.

I demonstrate the usability of EveTAR by evaluating a number of strong existing

[1] `https://en.wikipedia.org/wiki/Portal:Current_events`

[2] tweets that are no longer accessible due to deletion of tweets or deactivated user accounts

techniques in three of the supported tasks. These benchmarking results provide reference baselines for future studies in the respective research problems.
My contribution in this chapter is 3-fold:

1. We introduce a novel language-neutral approach for multi-task test collection construction, without requiring a shared-task evaluation campaign. My main contribution was in the formalization and design of the approach. Moreover, I had a key role in annotation tasks design and implementation.
2. We introduce and release[3] EveTAR, the *first* large-scale test collection over *Arabic* tweets that supports event detection, ad hoc search, timeline generation, and real-time summarization IR tasks. The collection contains the ids of 355M Arabic tweets, 50 events and 62K relevance judgments, novelty annotations, inter-annotator agreements, queries used to identify potentially-relevant tweets for the events, and documented design of the crowdsourcing tasks. We also release the annotations per tweet to support studies on crowdsourcing in IR.
3. I demonstrate the usability of EveTAR by evaluating existing techniques over three of the supported tasks. The resulting performance scores constitute reference baselines for future studies.

In the following sections, the approach devised to construct a test collection without running a shared-task evaluation campaign is presented (Section 4.1). Next, the process of implementing this approach is detailed (Section 4.2). Then, I present benchmarking results for multiple IR baselines and tasks over the collection (Section 4.3). Finally, Section 4.4 concludes the chapter and gives some guidelines for future work.

## 4.1. Dataset Construction Approach

Constructing a test collection is a challenging task. The document collection should accurately reflect the domain in which IR systems will be used. Furthermore, the search topics should capture real-world information needs. Moreover, selecting documents to judge should be carefully done to achieve reliable evaluation. Another challenge we consider when constructing EveTAR is serving multiple tasks while selecting topics that represent real information needs in all tasks and minimizing manual annotations. Finally, our approach aims at constructing and releasing a large-scale collection.

To address these challenges, we made multiple design decisions and propose an approach to construct EveTAR over Arabic tweets, which supports the following tasks: event detection, ad hoc search, timeline generation, and real-time summarization (defined in Section 4.1.2).

### *4.1.1. Topics*

News and recent events draw a lot of discussion and attention in social media. Thus, we start from popular events as the topics of EveTAR. We refer to each of these events as a *significant* event which is an occurrence at a particular time in a specific location and is covered by the media (e.g., discussed by an online news article). The aforementioned definition of an *event* is similar to existing definitions in [157]; however,

[3] `http://qufaculty.qu.edu.qa/telsayed/evetar`

we emphasize event significance, that is typically overlooked in other definitions [158]–[160]. Significant events reflect common information needs, as users usually search Twitter for information about current events [152], [155]. Moreover, higher volume of tweets is expected around popular events. Finally, it is possible to directly support evaluation of the multiple tasks we consider since they are all associated and needed when interacting with tweet streams about events.

### *4.1.2. Multi-Task Collection*

Starting from events as the topics in EveTAR, our collection naturally supports **Event Detection (ED)** which is the task of detecting events as they evolve over a tweet stream without prior knowledge on what events to expect. We represent each event in our collection by a set of relevant tweets within a time period surrounding the event time. An ED system is required to return a subset of the tweets that are relevant to the events in the collection. To support this task, the collection must include relevance judgments for each event.

Collecting ED relevance judgments inherently enables support for the **Ad hoc Search (AS)** task too. In the IR field, Ad hoc search is a typical search task in which a query (representing a topic of interest to the user) is issued to a search system which then returns a ranked list of documents (i.e., tweets) relevant to the topic over a collection of documents. The query is usually associated with a stamp of its issue time to the system. In EveTAR, we represent a topic by multiple queries that can be used in ad hoc search. An example query is the title of the event.

In recent years, Twitter became an exceptional platform to instantly follow topics (or events) of interest as they evolve. However, this comes with the curse of huge volume of tweets that are not all relevant to the topic, making manual tracking of events almost impossible. This created a pressing need for systems to filter redundant and noisy tweets and recommend relevant updates in real-time about a topic of interest. The **Real-time Summarization (RTS)** TREC 2016 track tackles this exact problem [151]. An RTS system is required to monitor a Twitter stream in real-time to detect relevant but non-redundant tweets for a given topic. The selected tweets represent a developing summary of the topic over time. For each time frame (a day per the TREC track) after the issuance of the query, a new summary is generated and should include relevant tweets that are also not redundant compared to previously created summaries for the topic. To achieve that goal in EveTAR, in addition to relevance judgments, we also need *novelty* judgments. In EveTAR, an RTS system should track an event during its active period only (5 days as explained in Section 4.2).

Having both relevance and novelty judgments for an event, we can also support a retrospective version of the RTS task, called **Tweet Timeline Generation (TTG)** [149]. Given a topic at a specific time, the TTG system is required to return a *set* of relevant but non-redundant (i.e., novel) tweets ordered chronologically (referred to as a *timeline*). The timeline is considered a retrospective summary of the topic at the query issuing time.

Figure 4.1 plots the relationship between the four tasks and time, in addition to the type of targeted tweets for an example event. To evaluate both ED and AS tasks, only the relevant tweets (black- and gray-colored) are needed, while both TTG and RTS require relevant and novel tweets (black-colored). The event detection task starts from the beginning of the collection since the time of events occurrences is not known in

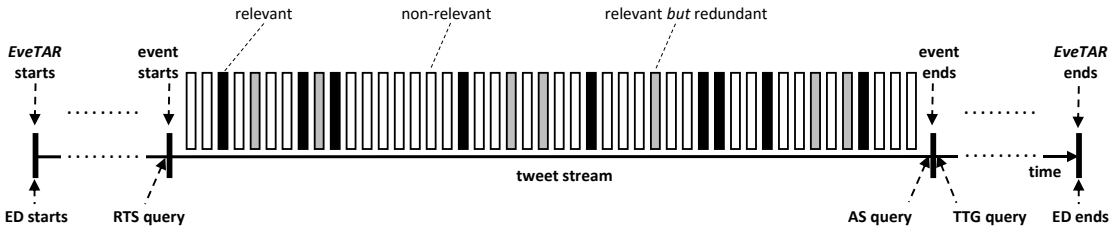



Figure 4.1. The four supported IR tasks over EveTAR. ED and AS tasks target relevant tweets (black and gray) while TTG and RTS tasks target relevant but not redundant tweets (black only)

advance. As for the RTS task, it starts for a topic when the topic becomes of interest, i.e., once the event starts. Differently, queries for the AS and TTG tasks are issued at the end of the event because they require retrospective retrieval of tweets. Table 4.1 compares the four tasks in terms of the nature of the input and output, and what types of judgments are required.

Table 4.1. Supported IR Tasks

| Task | Input Data | Novelty | Expected Output |
|---|---|---|---|
| Event Detection | Tweet stream | ✗ | Events (represented by sets of tweets) |
| Ad hoc Search | Tweet collection | ✗ | Ranked list of tweets |
| Real-time Summarization | Tweet stream | ✓ | 10 tweets per day |
| Timeline Generation | Tweet collection | ✓ | A set of tweets |

### *4.1.3. Large-Scale and Dense Dataset*

With the aim of supporting multiple tasks and the choice of popular events as the topics in EveTAR, the tweet collection should be *large-scale* and *dense*. Enforcing this criteria is necessary to achieve breadth covering a good number of significant events, and depth, to acquire rich relevant content per event.

In order to reach the needed breadth, the dataset has to span a long enough time period to ensure many events are covered. Following preliminary experiments and our own observation of discussions in Twitter (e.g., Arabic Twitter), a month is expected to cover the target number of topics we include in the collection.

To collect a tweet set, Twitter provides a streaming API[4] that typically returns around 1% random sample of tweets posted in Twitter at the time of accessing the stream. These tweets cover all languages used by users in Twitter; however, tweets of non-dominant languages (e.g., Arabic) will be sparsely represented in that sample. Alternatively, we make use of another service of the API where it is possible to track

[4]`https://dev.twitter.com/streaming/reference/get/statuses/sample`

tweets that match a set of keywords. This enables acquiring a focused and language-specific set of tweets. Therefore, constructing this keyword set should be carefully done to maximize the possibility of collecting a non-biased and representative set of tweets of the target language on Twitter.

### *4.1.4. High-Coverage and Diversified Judgment Pool*

As explained in Section 4.1.2, EveTAR should include both relevance and novelty judgments. Judging relevance of the tweets should be done first before deciding which of the relevant ones are novel. Thus, the first step to collect judgments is to construct a judgment pool of potentially-relevant tweets per topic.

Traditionally, several methods have been proposed to select the set of documents to judge, such as pooling [136], statAP [161], and interactive judging [137]. To construct a representative and rich judgment pool, most of these methods need a diverse set of IR systems. Running a shared-task evaluation campaign such as those under TREC is generally among the most common methods to achieve this goal, since many teams participate with their systems. However, organizing a shared-task is very challenging. Recently, an alternative approach has been proposed that depends on query variations rather than system variations. Moffat, Scholer, Thomas, *et al.* [138] demonstrated that query variations for a topic are as effective as system variations in constructing a diverse document pool. Therefore, we benefit from this idea to construct tweet pools to judge by manually creating multiple queries for each topic, using *interactive search* on Twitter's website to refine queries. Resorting to searching Twitter directly rather than our collection will ensure a wider coverage of the topic aspects. The queries can then be used to construct the document judgment pool by searching the collection using an off-the-shelf retrieval system.

### *4.1.5. Reliable Judgments*

After constructing the judgment pools, judging relevance and novelty can then start. There are two issues to consider before deciding on the annotation approach.

1. In selecting the annotators, we note that judging relevance of the tweets for a topic can be done by multiple annotators since each tweet can be judged independently from the others. As for novelty judgments, this is not possible since novelty of a tweet depends on that of the other relevant tweets in the same pool.

2. With popular events as the topics, the pool of tweets to judge is expected to be large for relevancy, requiring a large-scale annotation effort. However, judging novelty will be done on the relevant tweets only, indicating that the judgments pool is expected to be much smaller.

Keeping these issues in mind, we found crowdsourcing to be an effective option for relevance judgements. With crowd judgments, ensuring annotation quality and reliability is essential. To that end, we follow these steps: (1) collect multiple judgments per tweet, (2) hire annotators who have a minimum accuracy level considering their work history in the annotation platform, and (3) implement a qualification test for potential crowd-sourcing annotators and also require maintaining a minimum accuracy throughout the annotation. A final measure we take is filtering out topics with judgments

that exhibit low Kappa inter-annotator agreement [162] after collecting the relevance judgments for all topics.

Judging novelty requires that the annotator goes through the whole set of relevant tweets for a given topic to decide which are novel. Therefore, we resort to hiring and training in-house annotators to ensure reliability and consistency of the annotations. This enables us to train annotators and monitor their annotation effort.

## 4.2. Dataset Construction

I now proceed to explain how we constructed EveTAR following the approach presented in Section 4.1. We implemented our approach in two stages. In the first stage, we constructed and released a preliminary collection [47] with a set of Arabic tweets, topics based on events, and relevance judgments collected through crowdsourcing. In the final stage, our goal was to enhance the quality of the judgments and extend the collection to support other tasks by collecting novelty judgments. The following subsections detail the pipeline of the steps depicted in Figure 4.2.

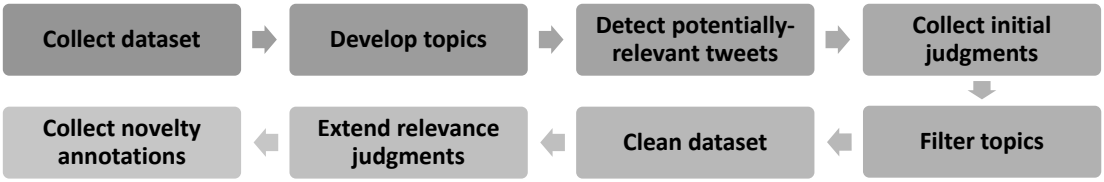



Figure 4.2. Pipeline of steps followed to create EveTAR

### *4.2.1. Collecting the Dataset*

We collected a set of tweets by tracking 400 frequent Arabic words through Twitter's streaming API.[5] The collection spanned the period between December $30^{th}$, 2014 and February $2^{nd}$, 2015. The tracked keywords were the most frequent in a set of 2M Arabic tweets crawled via the 1% random sample of Twitter collected over 10 days starting from April $10^{th}$, 2014. Eventually, the dataset had **590M** Arabic tweets that we indexed using Lucene open-source search library.[6]

### *4.2.2. Developing Topics*

Inspired by the work of McMinn, Moshfeghi, and Jose [157], we identified 357 events that took place in January 2015 according to the English[7] and Arabic[8] Wikipedia's Current Events Portals (WCEP). Each event is represented by a title and a start date. As explained in Section 4.1.1, the topic set only includes significant events. We applied two manual steps to filter out events not meeting our significance criteria. First, we kept only the events that have been discussed by at least one online *Arabic* news article, which resulted in a set of 71 events only. This process was done by manually searching

[5] https://dev.twitter.com/streaming/overview/request-parameters#track
[6] https://lucene.apache.org
[7] https://en.wikipedia.org/wiki/Portal:Current_events
[8] http://bit.ly/2n5TYhY

for articles in popular Arabic news websites such as Aljazeera[9] and CNN Arabic[10] using event titles as queries. In the second step, we only kept events with a minimum of 20 *relevant* Arabic tweets that we found by searching Twitter. For each event, we search through Twitter's advanced search service[11] using several Arabic queries and manually annotate returned tweets. We limited the search for each event to a 5-day period starting 2 days prior to the event date (since some discussions about the event might start prior to it, while most of the relevant tweets can be found on the day of the event or within two days of its occurrence). The final set included 66 significant events. For each event we manually created a rich representation including its title, date of occurrence, location, and category extracted from WCEP. Figure 4.3 shows a translated example of one event in EveTAR. Other examples of events include: "Match of Australia vs Kuwait at the opening of the Asian Cup", "North Korea hacks Sony accounts", and "Suicide bombing in Ibb".

| | |
|---|---|
| **ID** | E12 |
| **Title** | Discovery of tomb of Egyptian queen Khentakawess III |
| **Date** | January 04, 2015 |
| **Location** | Abusir, Egypt |
| **Category** | Arts and Culture |
| **Reference** | `http://cnn.it/106grQK` |
| **Keywords** | Khentakawess, Egyptian queen, archaeologist |
| **Description** | A Czech archaeological team discovered the tomb of an Egyptian queen named Khentakawess III who lived during the $5^{th}$ dynasty. |

Figure 4.3. A translated example of an event as represented in EveTAR

### *4.2.3. Identifying Potentially-Relevant Tweets*

Collecting judgments is the next core step in EveTAR construction process. We create the judgment pool by manually crafting a list of keyword and phrase queries for each topic while performing interactive search using Twitter's search service. For example, some of the (translated) queries for the event in Figure 4.3 are: "Abusir", "Khentakawess Third" and "Egyptian queen".

Starting with an average of 6 queries per topic, we created one long "OR" query comprising all queries per topic, and retrieved 10K tweets per topic using Lucene over the index of our tweet collection. BM25 was the retrieval model employed as it is Lucene's default model for ranking. The search was limited to the event specific time period. Following duplication by tweet-text, the pool had 134K tweets to be judged.

[9] `http://www.aljazeera.net`
[10] `arabic.cnn.com`
[11] `https://twitter.com/search-advanced`

#### 4.2.4. Collecting Initial Relevance Judgments

Relevance judgments were collected by crowdsourcing through Appen (formerly called CrowdFlower) crowdsourcing platform.[12] We created one annotation task per event (i.e., topic) following several pilot studies. For a task, annotators were shown the event title, description, and date, in addition to the content of an Arabic news article discussing the event, providing further context and details.

Annotators were required to be Arabic speakers with an intermediate experience level based on Appen's annotators ranking.[13] Before starting the judging process, judges went through a qualification test in which they were required to correctly label 8 out of 10 gold tweets. Gold tweets are randomly sampled from the set of candidate tweets per event and labeled by either me or a colleague. Annotators were also required to maintain a minimum accuracy of 80% over gold tweets embedded within the actual task.

Each tweet-event pair was judged by 3 annotators to reach a majority label which was used as the final tweet label. Eventually, the relevance judgments set had 134K labeled tweets, out of which 51K tweets were judged as relevant. We measured inter-annotator agreement per event using Fleiss' Kappa [162]. The average value of Kappa over all topics is 0.60. At this point, the test collection including the tweet dataset, 66 topics, and judgments were released as an early version of EveTAR [47].

#### 4.2.5. Filtering Topics

To improve the average judgment quality of EveTAR, we only kept the **50** topics with the highest agreement levels. The distribution of categories of the 50 events is shown in Table 4.2. As the table shows, events in the "Armed conflicts and attacks" category constitutes 64% of the entire set.

Table 4.2. Distribution of events in EveTAR

| **Category (Events)** | **Category (Events)** |
|---|---|
| Armed Conflicts & Attacks (32) | Sports (5) |
| Business & Economy (1) | Arts & Culture (2) |
| International Relations (3) | Law & Crime (1) |
| Disasters & Accidents (2) | Politics & Elections (4) |

#### 4.2.6. Cleaning the Dataset

Before releasing the final version of EveTAR, we had to ensure that all the tweets are accessible for future users of the collection. In December 2016, we verified accessibility of the judged tweets released with the early version of EveTAR [47] and found that around 40% of them have been removed from Twitter. A small number of deletions is due to users deleting their own tweets. However, the major reason is that the whole user account is no longer accessible because either it was suspended,[14], deleted, or made private by the user

[12] `https://appen.com/`

[13] https://success.appen.com/hc/en-us/articles/115000832063-Frequently-Asked-Questions

[14] Shut-down by Twitter; check `https://support.twitter.com/articles/15790`.

Due to the huge size of the dataset (590M tweets), checking availability of every tweet would be very time consuming. Thus, we opted to apply a more efficient solution by removing inaccessible accounts and consequently their tweets. We have automatically checked[15] each of 7.1M unique accounts in our original dataset and only kept the accessible accounts along with their tweets, resulting in around 5.8M accounts and a final dataset of **355M** tweets. This cleaned version of the collection was indexed using Lucene.

### *4.2.7. Extending Relevance Judgments*

Since many of the judged tweets were removed due to their deletion from Twitter, it was decided to extend and further improve the quality of the relevance judgments. In addition to dropping inaccessible tweets, retweets were removed from the judged tweets (since they express no new information). This resulted in a significant drop in the number of relevance judgments. Moreover, we decided to improve linguistic diversity among labeled tweets by collecting and judging dialectal tweets, since Twitter Arab users tend to use dialectal Arabic in their tweets [2].

To meet our goal, 3 to 9 new queries per topic we manually crafted following the same guidelines in Section 4.2.3, while adding dialectal queries whenever possible. In a similar approach to that explained in Section 4.2.3, the developed queries were issued to the index of the cleaned dataset and the top 1K tweets were returned. Before performing relevance annotation, we applied exact-text deduplication and retweet removal, and also excluded the tweets that have been judged before. As explained previously, crowdsourced labels were collected resulting in an additional 180 judgments on average per topic. The relevance judgments were finally propagated to 3,947 exact duplicates from the dataset.

The final relevance judgments set include 61,946 tweets, out of which 24,086 (39%) are relevant. Computing Fleiss' Kappa over the final qrels resulted in an average Kappa of 0.71, which is considered a *substantial* agreement according to a widely-adopted Kappa categorization [163], showing how strong the agreement is.

Figure 4.4 plots the number of relevant tweets per event, color-coded by the event category. The figure shows that the size of the relevant sets spans a large range and that we have events at all difficulty levels, assuming that events with more relevant judgments are easier to handle.

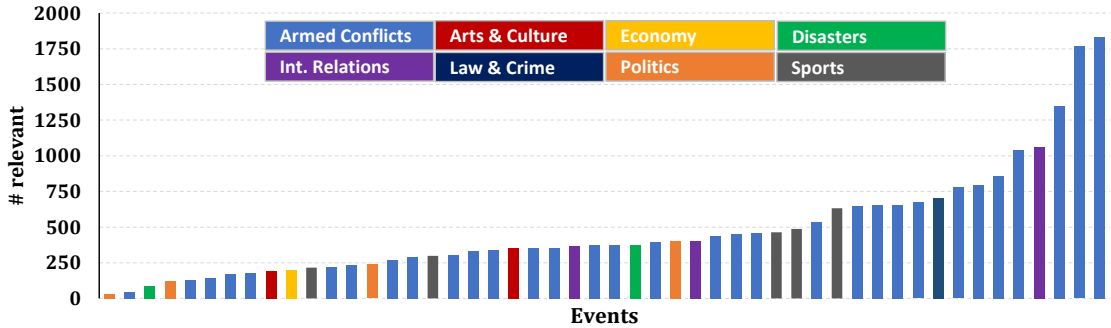



Figure 4.4. Distribution of relevant tweets over EveTAR topics

[15] on December $25^{th}$ 2016.

#### *4.2.8. Collecting Novelty Annotations*

To provide support for the RTS and TTG tasks, we collect novelty annotations for the relevant tweets by hiring 12 in-house annotators. Annotators went through intensive training before preforming the actual task. For each topic, a single annotator read the tweets for that topic in chronological order and manually clustered the tweets by their semantic similarity into multiple clusters such that all tweets in the same cluster convey the same information. This produced 66 clusters per topic on average. The novelty annotation approach is the same as that introduced in TREC 2015 Microblog [150] and TREC 2016 RTS [151] tracks.

### 4.3. Benchmarking

In this section, I demonstrate the usability of EveTAR by evaluating strong existing IR systems for each of three tasks: event detection (ED), Ad hoc Search (AS) and timeline generation (TTG). This produces reference performance results for future systems in these tasks. For evaluation, I mainly depend on evaluation measures used with relevant TREC tracks. Statistical significance of difference between systems is tested using two-tailed paired t-test, with a significance level $\alpha = 0.05$.

#### *4.3.1. Event Detection*

With events as topics in EveTAR, event detection is a primary supported task. To demonstrate the collection usability for this task, I ran two off-the-shelf event detection systems on EveTAR. Namely, I test Peaky Topics [164] and Trending Score [165] algorithms with open-source implementation as part of SONDY platform.[16] Because to the scale of EveTAR, I could not apply the algorithms to the full dataset due to limited computational power. Thus, I run the models over a representative random sample of the collection **EveTAR-S**. The sample includes 15M tweets of EveTAR (i.e., 4% of the full dataset) and includes all judged tweets.

To evaluate the algorithms, I adopted Petrovic's [166] approach that measure effectiveness using recall:

$$recall = \frac{\#\textit{covered events}}{\#\textit{reference events}} \tag{4.1}$$

where $\#$*reference events* is the number of events in EveTAR and $\#$*covered events* is the number of reference events included in the algorithm's detected events (where an event is represented by a list of tweets). A reference event is assumed to be covered by a detected event if at least 50% of the tweets in the detected event belong to that reference event. It should be noted that it is not feasible to automatically compute precision under this setup, since the "complete" list of possible significant events in the collection is not available. As it is difficult to manually compute precision given the huge number of events detected by each algorithm, we only report recall results. Peaky Topics achieve a recall of **0.42**, while Trending Score shows superior recall of **0.52**.

[16] https://github.com/AdrienGuille/SONDY

### 4.3.2. Ad hoc Search

I evaluated four ad hoc search systems: query likelihood (QL), temporal decay (TD), query expansion (QE), and temporal relevance modeling (TRM), which were adopted by one of the top teams [167] in the ad hoc search task in TREC 2014 microblog track [149]. Moreover, these models represent a diverse set of retrieval models usually used in tweet search (i.e., language modelling, temporal and query expansion). To be able to run those systems on Arabic tweets, I applied Arabic-specific preprocessing modules, e.g., stemming and stopwords, emoticons, and diacritics removal, in addition to URL and mentions removal. Systems parameters were set using their reported values in [167]. The ad hoc search per topic is limited to the active period of each event, i.e., the system is applied to tweets posted within the active time period per event. Table 4.3 shows evaluation results, using mean average precision (MAP) and precision at rank 30 (P@30) averaged over all 50 topics as the evaluation measures. This is the typical evaluation setup in TREC for the same task [149].

Table 4.3. Ad hoc search over EveTAR. Best result per evaluation measure is boldfaced. * indicates significant difference over the QL model.

| **Model** | ***MAP*** | ***P@30*** |
|---|---|---|
| QL | 0.3951 | 0.7340 |
| QE | 0.4494* | 0.7180 |
| TD | 0.3926 | 0.7313 |
| TRM | **0.4534*** | **0.7707*** |

Table 4.3 shows that, in both measures, TRM is generally outperforming all other models. As expected, QE is almost as effective as TRM since both systems are query expansion retrieval models that utilize pseudo-relevant tweets to expand the query.

### 4.3.3. Tweet Timeline Generation

Experiments were conducted with two TTG systems, cutoff and incremental clustering, adopted by one of the top teams [167] in the TTG task in TREC 2014 microblog track [149]. Both systems follow a two-step approach to generate a timeline. The system first retrieves a set of candidate tweets in response to a topic using a retrieval model, then applies a summarization technique to the retrieved set to generate a timeline. In the first step, I employ the same four ad hoc retrieval models presented earlier, namely QL, QE, TD, and TRM.

To allow the TTG systems to run on Arabic tweets, I follow a similar approach to the AS systems, and configured both the underlying model and the summarizing system using their reported parameter values in [167]. The TTG system is evaluated using the weighted $F_1$ ($wF_1$) measure, which is the official one used in TREC 2014 [149].

Table 4.4 reports the average $wF_1$ scores over the 50 topics for both TTG systems over EveTAR. Results show that Incremental Clustering is far more effective than Cutoff. Combining Incremental Clustering with the TRM retrieval model achieves the best performance overall. This is consistent with findings of a recent study that showed the

Table 4.4. TTG over EveTAR. The best result per TTG model is boldfaced. * indicates significant difference over the Cutoff model using QL as the underlying retrieval model.

| TTG Model | QL | QE | TD | TRM |
|---|---|---|---|---|
| Cutoff | 0.2164 | 0.2009 | 0.2180 | 0.2043 |
| Incremental Clustering | **0.3649*** | **0.3578*** | **0.3608*** | **0.3809*** |

performance of TTG systems adopting the 2-step approach is highly dependent on the underlying retrieval model [168].

## 4.4. Conclusions and Future Work

This chapter described a language-independent approach to creating a high-quality multi-task tweet test collection without running a shared-task campaign. The approach was applied to Arabic tweets with significant events as the topics and main motivation for the collection design. The proposed Arabic tweet collection, EveTAR, supports four retrieval tasks, namely event detection, ad hoc search, timeline generation, and real-time summarization. The collection includes 355M Arabic tweets, 50 topics, 62K relevance judgments, and novelty annotations. The full collection was released to the research community. I demonstrated one aspect of EveTAR's quality by estimating the inter-rater agreement of relevance annotations which was a substantial agreement (average Kappa score of 0.71), suggesting that we managed to acquire high-quality judgments through crowdsourcing. Finally, I report performance results of multiple effective systems per task which provides reference points for future studies.

Several new ideas can be implemented in the future. To encourage use of EveTAR and research on Arabic IR, a shared-task evaluation campaign around the tasks and data from EveTAR can be carried. A very interesting direction to consider is extending topics and annotations in EveTAR by dialectal content, thus, enabling evaluation of IR systems dedicated to dialect-related tasks such as cross-dialect search.

# CHAPTER 5: MISINFORMATION DETECTION: CHECK-WORTHY CLAIM DETECTION OVER SOCIAL MEDIA

Manual and automated efforts to detect and verify claims are indispensable to protect and inform users, especially in critical times like the current COVID-19 pandemic [20], [21]. While scanning a timeline, a user or a fact-checker is faced by many posts that are potentially false. Verifying all these claims can become cumbersome. Thus, the first step in the process of fact checking is identifying which posts contain claims that are worth verifying [21]. Not all claims are as important; some can have catastrophic impact on a large population, such as the popular claims discouraging COVID-19 vaccination [169]. Other claims might not cause any lasting impact or invoking any action. Figure 5.1 shows examples of tweets containing claims borrowed from Task 1 of the CheckThat! 2021 evaluation lab [34]. The tweet in Figure 5.1a was labelled as containing a check-worthy claim since the news might cause an international political crisis if propagated, making verification of the claim essential before its spread.

A claim check-worthiness is usually defined by its importance to the public and fact-checkers [34], [84], [94]. However, for this work, I adopted a more concrete definition from the check-worthiness estimation task at the CheckThat! lab at CLEF2021 [33], [34]. In the task, a check-worthy sentence is one that: 1) contains a factual claim, 2) is of interest to the public, 3) can potentially cause emotional or physical harm to a person or an organization, 4) a journalist might be interested in covering, and 5) a fact-checker should verify.

Before developing a system tackling the problem over Arabic tweets, a suitable Arabic dataset is necessary for system evaluation. Due to the recency of the problem, the majority of available datasets are English and covering political debates [84], [170], [171]. Therefore, I proposed and led the construction of the first Arabic Twitter dataset for the task. The dataset was constructed manually with the help of locally-hired annotators and was used to evaluate systems in the CheckThat!lab at CLEF2020 [31], [32] and later extended for CLEF2021 [33], [34]. In the 2021 version of the lab, other organizers of the lab created similar tweet datasets but for an additional four languages. The full dataset (CT21–CWT) included tweets in five languages, namely, Arabic, Bulgarian, English, Spanish and Turkish.

We found that creating an evaluation dataset for the problem at hand is not a straightforward task and requires a higher level of expertise. The task requires some knowledge of the fact checking process and general knowledge in terms of understanding the importance and forthcoming consequences behind spread of some claims. This limits the scale of the dataset that can be constructed for any language, and Arabic specifically since such lower-resource languages are not receiving as much resources and annotation effort as other languages (e.g. English). Thus, I propose to benefit from the unique multilingual nature of CT21–CWT and answer the following question: ***can we build an effective supervised model for check-worthiness detection over Arabic tweets without the need for an Arabic training data***? As opposed to focusing on system architecture, my aim is to identify whether we can minimize required annotation efforts to develop an effective system for check-worthiness estimation over Arabic tweets.

In this chapter, I address that question by testing six setups to perform cross-lingual check-worthiness detection over tweets, where a model is trained on data in one ore mode source languages and tested on the target language (Arabic). My work mainly focuses on a well-known setup in related problems, called *zero-shot* transfer learning, where no labeled examples in the target language (e.g., Arabic) are used during model training or fine-tuning. The work starts from the highly effective classification archi-


(a) Check-worthy claim: "(translation) Jordan denies entry of Chinese citizens due to the Corona virus."


(b) Non-check-worthy claim: "(translation) Ghassan Salamé, head of the United Nations support mission in Libya holds a press conference. "

Figure 5.1. An example comparing check-worthy and non-check-worthy Arabic claims (with translation)

tecture based on multilingual BERT (mBERT) [76]. Architectures based on mBERT demonstrated effectiveness in cross-lingual transfer learning in several text classification tasks [172]. Up to my knowledge, this is the first study of this kind and scale for the problem of check-worthiness detection in general and for Arabic specifically. I aim to address the following research questions:

RQ5.1 Given labeled data in a source language, how effective is zero-shot cross-lingual check-worthiness prediction on Arabic?

RQ5.2 Does translation between source languages and Arabic improve the performance?

RQ5.3 How much improvement can be achieved by adding few labeled Arabic examples to the examples in the source language (i.e., few-shot transfer learning)?

RQ5.4 Can the performance be improved if transfer is done from multiple source languages to Arabic?

RQ5.5 How effective is cross-lingual transfer compared to the state of the art models?

My contribution in this work is four-fold:

1. I lead the construction and release of two versions of the first of its kind dataset for evaluation of claim check-worthiness estimation and ranking over Arabic tweets.
2. The study extensively explores and benchmarks diverse methods to train cross-lingual check-worthiness prediction models including zero-shot, few-shot, and translation-based approaches. Existing studies for the task have not provided such a large-scale comparative study with different variants.
3. The work demonstrated that cross-lingual transfer learning from some languages to Arabic is comparable to monolingual models exclusively trained on Arabic.
4. The study offers benchmarking experiments comparing cross-lingual models, state-of-the-art models and strong baselines tested over CT21–CWT–AR. This provides future research on the same dataset with necessary baseline results.

This chapter is organized as follows. First, the full process of dataset construction is presented in Section 5.1. The approach is then motivated and described in Section 5.2. Section 5.3 details the experimental setup and model implementation. In Section 5.4, I present detailed results and discussion, in addition to benchmarking results against state-of-the-art models. The chapter ends with a summary and some concluding remarks in Section 5.5.

## 5.1. Dataset Construction

Since check-worthiness estimation for tweets in general, and for *Arabic* tweets in particular, is a relatively new task, we constructed two versions of a new dataset specifically designed for systems training and evaluation for this task. The dataset versions were constructed over two years with each used for an edition of the CheckThat! evaluation lab at CLEF conference. This section describes the process of constructing both CT20–CWT–AR (Section 5.1.1) and CT21–CWT–AR (Section 5.1.2).

**عنوان الموضوع:** تدخل تركيا في سوريا

**شرح الموضوع:** بعد استمرار الحرب في سوريا ما يُقارِب التسع سنين بعد اشتعال الثورة عام 2011، قررت تركيا الدخول بهدف معلن وهو حماية المدنيين وردع القوات السورية والأجنبية عن توتير الأوضاع وقتل وتشريد المدنيين، وهو ما أثار الرأي العام في العالم. يتحدث هذا الموضوع عن تطورات التدخل التركي العسكري في سوريا على جميع الأصعدة.

**Topic title:** Intervention of Turkey in Syria

**Topic description:** After 9 years of war in Syria since the eruption of the revolution in 2011, Turkey decided to intervene in Syria with the declared aim of protecting civilians and deterring Syrian and foreign forces from aggravating the situation, and killing and displacing civilians, which ignited public opinion in the world. This topic talks about developments related to the Turkish military intervention in Syria on all aspects.

Figure 5.2. Topic CT20-AR-19 from the training subset of CT20–CWT–AR.

### *5.1.1. Constructing Topic-based CT20–CWT–AR Dataset*

We started by creating CT20–CWT–AR at CheckThat!2020. In this version of the dataset, we identified the need for a "context" that affects the check-worthiness of tweets, opting to develop "topics" to represent that context. Topics should reflect real-world issues and occurrences, and should be of interest to a large number of users. To that end, we manually created fifteen topics over the period of several months. These topics were trending at the time among Arab social media users. Each topic was represented using a short title and a much longer text description. Figure 5.2 shows an example topic. Examples of other topic titles include "Coronavirus in the Arab World", "Sudan and normalization", and "Deal of the century".

We augmented each topic with a set of keywords, hashtags, and usernames to track in Twitter. Once a topic is created, we immediately crawled a one-week stream of tweets using the constructed search terms, where we searched Twitter (via the Twitter search API[1]) using each term by the end of each day. We limited the search to original Arabic tweets (i.e., retweets were excluded). We then de-duplicated the tweets and

[1] `https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets`

Table 5.1. Statistics of the CT20–CWT–AR and CT21–CWT–AR datasets.

| Dataset | Task | #Positive | Total |
|---|---|---|---|
| CT20–CWT–AR | Has claim | 2,479 | 6,000 |
| | Check-worthiness | 1,604 | |
| CT21–CWT–AR | Has claim | 1,947 | 4,705 |
| | Check-worthiness | 1,270 | |

dropped those matching a qualification pipeline that excludes tweets containing explicit terms and tweets with more than four hashtags or more than two URLs. Finally, tweets were ranked by popularity (defined by the sum of their retweets and likes), and the top-500 are those that will be annotated. We enforce the popularity condition to increase the chance of identifying check-worthy claims.

The annotation process was conducted in two stages; we first identified the tweets that are relevant to the topic and contain factual claims, then we identified check-worthy tweets among the relevant ones. Table 5.1 summarizes the dataset statistics.

#### *5.1.1.1. Relevance Annotation*

The first stage of annotation was labelling tweets by relevance to the topic. During pilot studies, I found the task to be straight-forward and thus, I elected to have each tweet annotated by a single annotator. In-house annotators were recruited and trained for this task; they were asked to label each tweet by one of three categories:

- Non-relevant to the target topic.
- Relevant but *does not contain any factual claim*, such as a tweet expressing a sentiment or opinion about the topic, reference, or speculation about the future.
- Relevant and contains a factual claim that can be fact-checked by consulting reliable sources. This represents the positive label for this stage.

After finishing this stage, we observed that seven out of the fifteen topics are related to COVID-19 and they cover 60% of all tweets that contain relevant claims. Therefore, we decided to drop three randomly selected topics out of these seven to avoid biasing the dataset to a single topic.

**Annotation Quality.** To verify the quality of relevance annotations and since we have a single annotation per tweet, I randomly sampled 10% of the tweets per topic and re-annotated them. On average, agreement rate is 83.5% with the minimum equal to 80%. We believe this is a reasonable agreement level especially that the tweets go into a second annotation stage where errors (false positives specifically) in this stage can be caught and corrected.

#### *5.1.1.2. Check-worthiness Annotation*

Only relevant tweets with factual claims were labelled for check-worthiness. As with the relevance task, and due to the complexity of annotation of this task, I hired

in-house annotators such that direct training on the task can be provided. Each tweet was first labelled by two annotators and a third *expert* annotator performed disagreement resolution. Due to the subjective nature of check-worthiness, we represent the check-worthiness criteria by several questions, to encourage the annotators to think about different aspects of check-worthiness. The questions were constructed based on our own observation of the problem and studying example check-worthy tweets. The questions were further tuned and re-phrased following pilot studies on example topics. The annotators were asked to answer the following three questions for each tweet, using a scale between 1 and 5:

1. Do you think the claim in the tweet is of interest to the public?

2. To what extent do you think the claim can negatively affect the reputation of an entity, country, etc.?

3. Do you think journalists will be interested in covering the spread of the claim or the information discussed by the claim?

Once the annotator has answered the above questions, s/he is required to answer the following fourth question considering all the ratings given previously:

4. Do you think the claim in the tweet is check-worthy?

This is a yes/no question, and the resulting answer is the label we use to represent check-worthiness in this dataset. For the final set, all tweets (including non-relevant ones) but those labelled as check-worthy were considered not check-worthy. The dataset is made publicly available to the research community.[2]

**Annotation Quality.** We verify the quality of check-worthiness annotations by computing the agreement level between the labels of the two annotators. The average agreement rate is 60%. This is expected due to the difficulty and subjectivity of the task. We should also note here that the disagreements were resolved by an *expert fact-checker* leading to majority labels that are more reflective of check-worthiness among the tweets.

### *5.1.2. Constructing CT21–CWT–AR Dataset*

After the construction of CT20–CWT–AR that was used to evaluate systems in CheckThat!2020, and in order to construct a new test subset to be used in the next edition of the lab (CheckThat!2021), we extended and improved CT20–CWT–AR resulting in CT21–CWT–AR.[3]

The training and the development sets of CT21–CWT–AR include the same 12 topics crawled in January, February, and March 2020 and borrowed from CT20–CWT–AR. As for the annotations for these topics, for each topic, I only kept tweets that were relevant (with and without claims). For tweets with claims, I additionally filter out those that did not have **full** inter-rater agreement on the check-worthiness label.

For the test set, we crawled using two topics in January 2021 and we annotated the resulting tweets as follows. We first recruited one annotator to annotate each tweet for

[2]This section refers to the **test split** as part of the CheckThat!2020 lab and can be found at: `https://gitlab.com/bigirqu/checkthat-ar/-/tree/master/data/2020/task1/testing`

[3]The training and test subsets of this version of the dataset can be found at: `https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1`

its relevance with respect to the target topic. Then, the tweets that were found relevant were labeled for check-worthiness. This second annotation was done by two *expert* annotators. It was followed by a subsequent consolidation step, at which the annotators discussed each disagreement to resolve it. As with CT20–CWT–AR, annotators were asked to answer several questions. For the test set, we added one more annotation question based on further discussions with professional fact-checkers, and changed the answers required to be yes/no for all questions to simplify judgment for annotators. The final set of questions is as follows:

1. **Does the tweet contain a verifiable factual claim?**

I stop here to clarify that, instead of making the above question as part of the relevance annotation task (as done in CT20–CWT–AR), I opted to make it part of the check-worthiness task. This simplifies the first annotation stage, making it easier to be done by non-expert annotators. If the answer to the above question is positive, the annotator is asked to answer the following additional yes/no questions:

2. Does the claim in the tweet appear to be false?
3. Do you think the claim in the tweet is of interest to or would have an impact on the public?
4. To what extent do you think the claim can morally or physically harm an entity, a country, etc.?
5. Do you think that journalists will be interested in covering the spread of the claim or the information discussed by the claim?

Once the annotator has answered the above questions, s/he is further required to answer a final question considering all the answers given previously. The answer to this question is the label that is used to represent the check-worthiness for the target tweet.

6. Do you think that a professional fact-checker or a journalist should verify the claim in the tweet?

In addition to being a valuable resource to train and evaluate check-worthiness estimation systems, the dataset offers valuable evaluation resource to two other types of systems: relevance estimation (e.g., as part of a supervised ad hoc search system over tweets) and claim identification. Furthermore, the approach to check-worthiness annotation where multiple yes/no questions are answered for a tweet provides a great opportunity to train multi-task learning systems to tackle the problem.
**Annotation Quality.** Two subsets of topics and associated annotations constitute CT21–CWT–AR, including: (1) the training subset borrowed from the previous CT20–CWT–AR dataset, and (2) the new test subset created in 2021. Below, the annotation quality of each subset is discussed.

- **Training dataset**: For the relevance labels, it was already established in Section 5.1.1 that the average agreement between the main annotator and a secondary annotator is above 80% on a 10% sample of the annotations per topic. As for the check-worthiness labels, tweets with full agreement on the check-worthiness labels between the two main annotators were used, ensuring a 100% label agreement for all tweets.

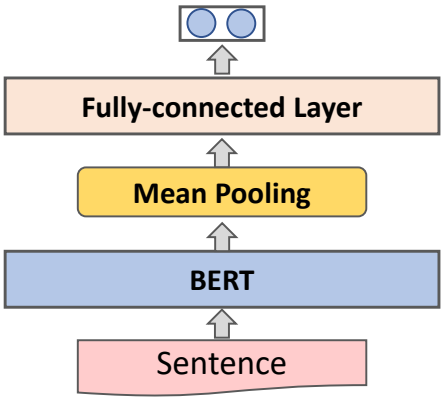



Figure 5.3. Classification architecture. BERT layer represents all BERT-based transformer models used in this work.

- **Test dataset**: A single annotator performed relevance annotation for 1K tweets across the two test topics. I randomly sampled 10% of the tweets per topic, and re-annotated them for relevance. The resulting agreement level was above 82% for both topics. For the check-worthiness annotations, two expert annotators labelled the tweets, and performed a consolidation step, resulting in 100% agreement on the final label.

## 5.2. Approach

This section describes the main architecture used throughout my experiments in this chapter.

### *5.2.1. System Architecture*

The work is motivated by the strong line of research showing the effectiveness of transformer models, such as BERT, for text classification. In the area of fact-checking, architectures based on transformer models are among the best performing for different tasks including check-worthiness prediction [34], [85], [87], [102], claim verification [27] and evidence retrieval [173].

For all experiments, I start from the same BERT-based classification architecture depicted in Figure 5.3. This architecture is constructed based on previous literature using BERT for text classification. Specifically, following BERT layers, I add a feed-forward network with one hidden linear layer (of 256 nodes) with ReLu activation. Softmax activation function is finally applied to the output layer, resulting in two predicted probabilities (one for each of the two classes). The input to the architecture is a single sequence which is the sentence $\boldsymbol{S}$ that I would like to predict its check-worthiness. The input to the model is formatted as follows: [[CLS], $\boldsymbol{S}$, [SEP]]. Typically, after training the full architecture (including fine-tuning of the pre-trained model), the hidden state $\mathbf{h}$ produced by the transformer model for the [CLS] token is used as representation of the input to the remainder of the classification architecture. However, during preliminary experiments on development subsets, I found that using mean pooling over all tokens yields better classification results, thus I adopt this pooled representation. At inference

time, the probability of the positive class determines the predicted label for the input sentence with a 0.5 threshold. The model is trained minimizing cross entropy loss.

### *5.2.2. Cross-lingual Check-worthiness Transfer*

The main aim of the work is to investigate whether check-worthiness learning can be transferred from source languages to Arabic, and then identify potentially effective systems with none or minimal labeled data originally written in Arabic. To that end, I study different strategies for transfer learning from source language *A* to Arabic, starting from the zero-shot transfer learning setup, going through methods that employ minimal labelled Arabic data, and finally, approaches that enrich zero-shot transfer learning with translation. I next describe each of the approaches investigated in this work.

#### *5.2.2.1. Zero-shot Cross-lingual Transfer Learning (**ZS**)*

Given the strong ability of pre-trained models, such as, mBERT in zero-shot cross-lingual transfer over multiple NLP tasks [172], the system described in Section 5.2.1 is fine-tuned on a source language and applied directly at inference time to the Arabic test set. This approach represents the basic ZS model I test in this work; it is a commonly-adopted approach in related cross-lingual transfer learning studies [174].

#### *5.2.2.2. Zero-shot with Translation (**ZS-Tr**)*

In the second research question, I aim to find an improved setup over ZS by translation.[4] Instead of depending on transfer ability of mBERT, I unify the language of both training and test sets using two strategies:

- **ZS-TrSrc**: In this setup, I translate the *training* set of ***source*** language *A* to *Arabic*, then fine-tune the model on the translated data. The model is then directly applied to the Arabic test set.
- **ZS-TrTrg**: This setup shows the second possible translation approach. I first fine-tune the model using the *original* training set of the source language A. I then translate the *Arabic* test set into language *A*, and apply the model on it.

#### *5.2.2.3. Transfer Learning with Few Shots (**FS**)*

In this setup, I experiment with transfer learning *extended* with the addition of few labelled Arabic training examples from the training set. This is different from the translation-based approaches, since I add labeled examples originally written in Arabic, rather than being translated. Few-shot cross-lingual transfer with *two-stage* fine-tuning has gained importance recently, since it generally improves performance with small annotation cost for target language examples (e.g., [175]). In this setup, I fine-tune the model in two stages; first it is fine-tuned over the source language *A*, then further fine-tuned using few Arabic examples. I use random sampling with balanced classes to select few shots for the second stage (details in Section 5.4.3).

[4] I used Google Translation API at `https://cloud.google.com/translate`.

## 5.3. Experimental Setup

This section presents the experimental evaluation setup, designed to answer the research questions, including how CT21–CWT was processed for the experiments, evaluation approach, and implementation details for the classification architecture.

### *5.3.1. Dataset Preparation*

Earlier in this chapter (Section 5.1), I presented the process to construct the first Arabic Twitter dataset for check-worthiness estimation, with two versions and CT21–CWT–AR as the more mature one. During CheckThat!2021 lab, the organizers of the lab have also constructed datasets for the same task but in four other languages: Bulgarian, English, Spanish and Turkish. This resulted in a multilingual dataset (CT21–CWT) that is the basis for this work. Further details on how the other language subsets were created can be found in the official lab overview paper [34].

Before proceeding with CT21–CWT, I first combine the training and development sets per language to acquire a larger training set. I observed a great difference in the training set size across languages, ranging from 962 to 4.1k tweets. More importantly, the class distribution varies significantly with the percentage of positive labels falling between 8% and 38% across languages. Although such class distribution prior might be observed in real-world cases, this can shift the focus of this work from understanding check-worthiness estimation differences across languages to how to best handle this imbalance. Moreover, this imbalance can mask or exaggerate system performance across languages. Such observations were made in previous research concerning systems for cross-lingual transfer [176].

I alleviate the problem of varied dataset sizes across languages by down sampling the training subset per language using a stratified random sampling approach. This ensures that I have the same dataset size and class distribution across languages. I chose the sample size per class based on the minimum number of labels per class across languages. Eventually, for each language except English, I end-up with 300 positive and 1,400 negative examples. In the English dataset, the number of negative examples available was much smaller than 1,400. Thus, the final training set includes 300 positive and only 612 negative *English* examples. As for the Arabic test set, I keep it as released in CheckThat!2021 to enable the benchmarking experiments I perform in Section 5.4.5 and facilitate future comparisons on the same dataset.

### *5.3.2. Implementation and Evaluation Details*

Due to the extent of the experiments and the limited dataset sizes, I unify the model parameters across experiments without hyperparameter tuning (unless otherwise stated). I set the parameters following optimal values identified in the original BERT paper (Appendix A.3 in [76]) and a recent paper that examined mBERT performance for multilingual text classification [177]. I set the training batch size to 32, learning rate to 3e-5, and a maximum sequence length of 128. The model is fine-tuned for three epochs (in line with related work on the same dataset [100]) and model training is repeated five times with different random seeds to account for any randomness in model initialization and training. In this work, I report the average performance over those five re-runs.

The models are evaluated using the $F_1$ score of predicting the positive class. I chose this measure since my aim is to understand the model effectiveness in identifying

Table 5.2. Pre-trained models used in experiments.

| Model | Language | HF Model Name |
|---|---|---|
| mBERT [76] | Multilingual | bert-base-multilingual-cased |
| AraBERTv1 [139] | Arabic | aubmindlab/bert-base-arabert |

check-worthy claims. The statistical significance of difference between systems is tested with two-sided paired t-test over the five re-runs with $\alpha < 0.05$.

For all experiments, I use pre-trained BERT models from the HuggingFace (HF) library (Table 5.2).[5] Base version of the models with 12 layers was used. The monolingual Arabic model used (AraBERTv1) showed superior effectiveness for the task during initial experiments over the development subset, as compared to other recent models like ARBERT and MARBERT described in [140].

## 5.4. Results and Discussion

In this section, I present and discuss the results of the experiments designed to answer each of the research questions.

As a *baseline* for all of my experiments (unless otherwise stated), I report the performance of mBERT when fine-tuned on the Arabic training set.

### *5.4.1. Zero-shot Cross-lingual Transfer Learning*

I start by examining the model effectiveness in zero-shot cross-lingual transfer learning (**RQ5.1**). For this purpose, I train an independent model for each of the five languages, then report the models' performance on the Arabic test set. The model trained on **ar** is the baseline. Figure 5.4 shows the results.

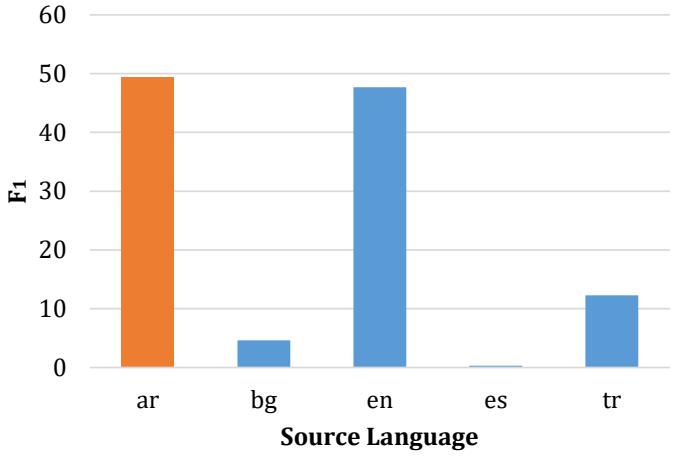



Figure 5.4. ZS results. Arabic (ar) as a source language is the baseline.

The experiment shows promising results that answer **RQ5.1**. The best ZS performance (i.e., **en** as a source language) is comparable to the baseline, with the difference

[5] https://huggingface.co/models

Table 5.3. Effect of translation on transfer learning. %diff is the percentage of difference between the performance in each setup and corresponding ZS result from Figure 5.4. Bold and underlined values represent best and second best per column, respectively. * indicates significant difference from baseline.

| **Source** | **ZS-TrSrc** | | **ZS-TrTrg** | |
|---|---|---|---|---|
| | $F_1$ | %diff | $F_1$ | %diff |
| **ar (baseline)** | **49.5** | | <u>49.5</u> | |
| **bg** | 09.7* | +111% | 05.9* | +28% |
| **en** | <u>45.7</u> | -4% | **52.0** | +9% |
| **es** | 00.0* | -100% | 01.5* | +400% |
| **tr** | 18.9* | +54% | 21.8* | +77% |

not statistically significant. However, the transfer from the other three languages is ineffective. I believe this is due to the following reason, at least for Turkish and Bulgarian: during the pre-training of mBERT, both languages are lower resource languages that were underrepresented compared to English or even Arabic. Such issue has been shown to negatively affect language representation learnt by mBERT and thus affect performance of the model after fine-tuning [178]. *Overall, this experiment showed that for the check-worthiness prediction task, zero-shot transfer is as effective as fine-tuning mBERT over the full Arabic training set at least when English is the source language.*

### *5.4.2. Effect of Translation on ZS*

In **RQ5.2**, I aim to find an improved setup over ZS by translation. In this experiment, Table 5.3 shows check-worthiness estimation performance when translating from each of the two directions: translate the source language (Section 5.4.2.1) or the target language (Section 5.4.2.2).

#### *5.4.2.1. ZS with Source Translation (ZS-TrSrc)*

The results in Table 5.3 show that for **bg** and **tr**, translation of the source language resulted in a notable improvement over original ZS. As for the cases when performance degradation is observed, for **es**, it is negligible since performance of ZS transfer from **es** to Arabic was already close to zero. Moreover, when transferring from **en**, degradation was minimal and performance was still comparable and not significantly different from the baseline. Overall, that indicates slight improvement with translation compared to ZS. A point worth noting here is that the performance changes are also related to the effectiveness of translation system used.

#### *5.4.2.2. ZS with Target Translation (ZS-TrTrg)*

I continue to answer **RQ5.2** by translation of the Arabic test set to match the source language. Differently from ZS-TrSrc, ZS-TrTrg resulted in improved performance over original ZS for all source languages as shown in Table 5.3. Transfer from **en**

specifically showed slight performance increase over the original ZS performance and even the baseline (although not significantly different).

To further understand why translation improved performance over ZS, I dive into the actual predictions made in these two setups, taking transfer from **en** to **ar** as an example. For this experiment, I select the run (out of the five re-runs) that achieved the best performance improvement compared to the corresponding ZS run. I found that 60 tweets were correctly predicted by ZS-TrTrg but miss-classified by ZS. Two thirds of these tweets discuss a single topic "2021 United States Capitol attack". This explains why translation from **ar** test set to **en** helped improve prediction over these tweets that are directly related to U.S. politics. Vocabulary of this topic will more probably appear in English Wikipedia, versus the Arabic version. Wikipedia was used to pre-train the mBERT model with English being the most represented language. Thus, the model is more fitted to the language and topics used in English Wikipedia, making it easier to classify these Arabic tweets once translated to English. *The initial analysis of the tweets in this case raised an important question about the check-worthiness prediction task definition itself. What's the effect of claim topic on model transfer?* Answering this question is left for future work.

### *5.4.3. Transfer Learning with Few Shots (FS)*

I now turn to answer **RQ5.3** concerning the effect of *adding* few shots from the target language. Figure 5.5 shows the results when I *continue* fine-tuning each of the models from Section 5.4.1 using **1%** of the Arabic training set. In more detail, continued training is done using **17** randomly sampled examples, among which **8** are positive.

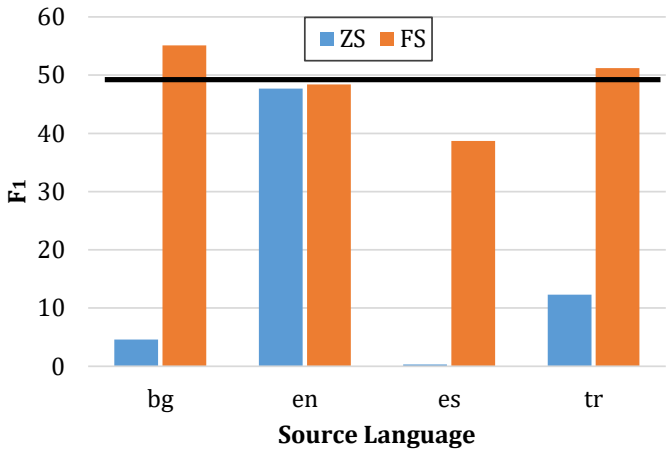



Figure 5.5. Effect of continued fine-tuning using 1% of the target language training set. The black line indicates the baseline when the model is trained on the *full* Arabic training set.

The most prominent observation from Figure 5.5 is that adding as few as 17 Arabic examples to the models, fine-tuned on each of the other languages except Spanish, resulted in comparable performance to the baseline, in which I train on the *whole* Arabic training set ($F_1$ = 49.5). I also observe extreme improvements compared to pure ZS, except when English is the source language; for English, the performance is comparable.

I anticipate this happened because Arabic is written in a very different script (as opposed to remaining languages) making the addition of few target examples useful for mBERT to learn necessary structural and lexical information, which is consistent with observations made in a recent study [174] regarding the effect of writing scripts on FS. These results highlight that with very small annotation effort, I can achieve as good performance as the case where much more annotations in the target language were collected. This is inline with some recent studies on other text classification tasks [175].

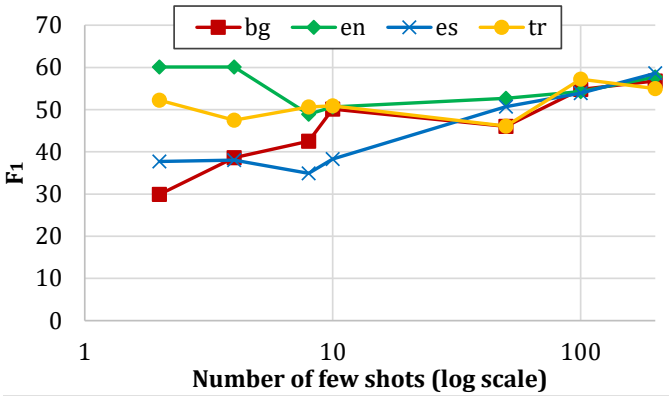



Figure 5.6. Effect of number of few shots on check-worthiness prediction. $x$-axis is in log scale.

What happens when the number of the few shots change? Figure 5.6 shows the effect of continuing the fine-tuning of ZS models with $k$ randomly sampled examples. I experiment with $k \in \{2, 4, 8, 10, 50, 100, 200\}$. The figure indicates several observations. First, an optimal setting of parameter $k$ is needed to get most benefit of few shots learning transfer, but we need to consider the corresponding annotation cost. Second, increasing performance gain is observed with the increase of $k$, at least when **bg** and **es** are the source languages. Third, an interesting pattern emerged, where at the addition of 200 shots, the achieved performance is very similar regardless of the source language. This indicates that the added 200 shots were enough to almost *suppress* the effect of the source language.

*In response to **RQ5.3**, FS with the addition of as little as 1% of the Arabic training set, has resulted in notable improvements over ZS. An important issue to consider is that the effectiveness of FS depends on finding an optimal setting for number of few shots for different source languages.*

### *5.4.4. Multilingual ZS*

I finally address **RQ5.4**: will ZS benefit from multilingual training? Differently from vanilla ZS (Section 5.4.1), I fine-tune the check-worthiness prediction model over multilingual examples excluding the target language. For each language but Arabic, I randomly sample examples with the same fixed class priors, ending up with 1,700 total examples across all four languages, with 300 positive examples and each language is equally represented in the training set. This is equal in total size and distribution to the Arabic training set, to ensure fair comparison.

Figure 5.7 compares results of this model, denoted as $\mathbf{ZS}_{all-\{target\}}$, with two baselines: (1) $\mathbf{mBERT}_{target}$, in which I fine-tune the mBERT model over the training

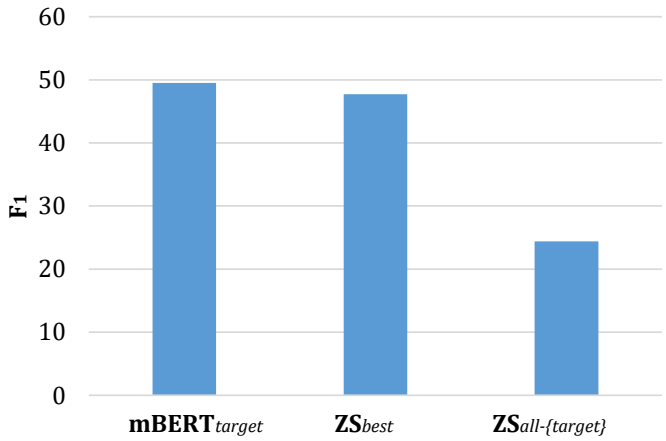



Figure 5.7. Performance of the multilingual ZS model.

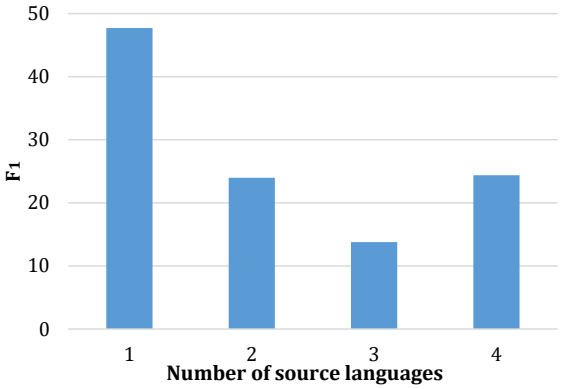
set of the target language, and (2) $\mathbf{ZS}_{best}$, which is the best performing ZS model in Figure 5.4, which is fine-tuned with *one* source language. As can be clearly observed, unified language in training and test sets (i.e., $\text{mBERT}_{target}$) yields the best performance. Surprisingly, this experiment shows that multilinguality in training did not actually help the model achieve better performance compared to the vanilla ZS. I argue that this is due to the *curse of multilinguality* that was observed on multilingual transformer models due to limited model capacity dedicated to each language during pre-training, causing degraded transfer ability [179]. I believe a similar issue is emerging during fine-tuning using multiple languages; with each language being represented by small number of examples, the transfer ability of the model, compared to pure ZS setup, is negatively affected.

I further support my claim by running an experiment in which I incrementally increase the number of source languages (excluding the target language) during fine-tuning, and report the transfer performance over the **ar** test set. As with $\mathbf{ZS}_{all-\{target\}}$, I fix the total training examples to 1.7k and keep each source language equally represented in the training sample. For the case of 1, 2, or 3 source languages, I try all combinations of languages; thus, for those cases, I end up with multiple performance values depending on the combination of source language(s) used during fine-tuning; I report the maximum observed scores in Figure 5.8a. The figure shows the transfer performance generally degrades as I increase the number of languages on which mBERT is fine-tuned. This is consistent with the phenomenon observed in [179] for transformer models during pre-training.

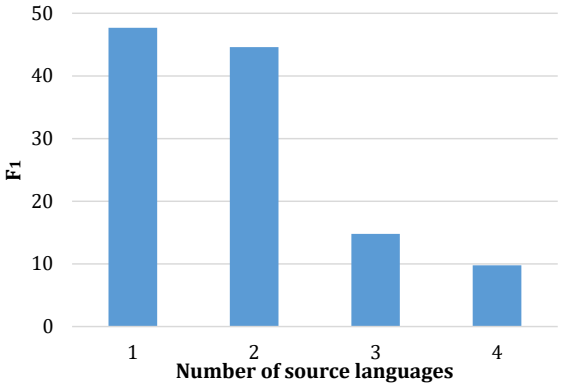
A question arises on whether this effect can be mitigated by increasing the training set size as I increase the number of languages. I repeated the previous experiment, but increased training set size to be 1,700 per source language. Figure 5.8b shows, again, clear degradation in performance. *Overall, experiments indicate that increasing the number of source languages (i.e., introducing multilinguality in training set) is not as effective for our task.*

### *5.4.5. Benchmarking*

To answer **RQ5.5** and to provide benchamrking results on CT21–CWT–AR, I would like now to see how the previous setups compare to baselines that are the state of




(a) Fixed training data size




(b) Increasing training data size

Figure 5.8. Effect of number of source languages used during fine-tuning on transfer performance to Arabic.

the art *on the given test set*. I experiment with the following baselines:

1. **mBERT**$_{target}$, as described earlier.
2. **monoBERT**$_{target}$, in which I fine-tune a monolingual pre-trained BERT model (i.e., AraBERTv1 for Arabic) using the Arabic training set. Such setup is expected to be effective, since the monolingual model is pre-trained on a large corpora in the target language.
3. **CT!2021**$_{best}$, which is the model with best reported performance in CheckThat! 2021 lab [34], [87].[6]
4. **CT!2021**$_{2nd\_best}$, which is the second best reported performance in CheckThat! 2021 lab [34]. I was the developer of this model in which I fine-tuned AraBERTv1 over the Arabic training set and the Turkish one after translating it to Arabic.

[6]I note that the systems were originally evaluated as ranking systems, however, to facilitate the comparison, I re-evaluate the runs as classification systems. To that purpose, I use the prediction of check-worthiness score per tweet as it appeared in the original run files I acquired from the teams. This is possible since the top teams report using a classification model to solve the task and report probability of prediction (between 0 and 1) as the ranking score. I assume any tweet with score above 0.5 to be predicted as check-worthy claim.

Table 5.4. Comparison of performance of best ZS setup and state-of-the-art models. $^*$, † indicate significant difference from mBERT$_{target}$ and monoBERT$_{target}$, respectively.

| Model | ar |
|---|---|
| **CT!2021**$_{best}$ | **60.0** |
| **CT!2021**$_{2nd\_best}$ | <u>57.0</u> |
| **monoBERT**$_{target}$ | 56.3 |
| **mBERT**$_{target}$ | 49.5 |
| **ZS**$_{best}$ | 47.7 |
| **ZS-TrSrc**$_{best}$ | 45.7 |
| **ZS-TrTrg**$_{best}$ | 52.0 |
| **FS**$_{best}$ | 55.1 |

Table 5.4 compares the performance of the above baselines with the best performing models per setup from those presented above, where none or minimal labeled examples in the target language were considered in fine-tuning. The comparison yields a clear observation; all the ZS setups had insignificantly-different performance to both of the BERT-based baselines trained on the full Arabic training set. The experiment also shows that top systems from CheckThat! 2021 achieved the best performance. However, the comparison is somewhat unfair, since both of these systems were trained on a much larger training set (recall that I under-sample the training set from CT!2021 for the experiments). Even with that disadvantage to my models, the models still show performance that is not far from CT!2021$_{best}$ and even comparable to CT!2021$_{2nd_best}$. This demonstrates the strong cross-lingual transfer ability of mBERT for our problem. *Overall, the comparison indicated that it is possible to train an effective cross-lingual check-worthiness model with none or minimal Arabic training examples.*

## 5.5. Conclusions and Future Work

In this chapter, I aimed to investigate and identify optimal setups to facilitate check-worthiness estimation over Arabic tweets without the need for a training set in the same language. My work is motivated by the current dire need to provide an effective model for the problem given the scale of propagating misinformation and scarcity of annotation efforts dedicated to Arabic. The in-depth experiments showed that cross-lingual transfer models result in comparable performance to the monolingual models fully trained on many Arabic examples. Moreover, multilinguality during fine-tuning negatively affected the model transfer performance. I also showed that the proposed models are not far behind the state of the art models on the same Arabic test set. Finally, the addition of few shots showed to be generally helpful compared to zero shot learning transfer setups.

A straightforward extension to this work includes investigating the performance of other multilingual transformer models (e.g., XLM-R). Moreover, experimenting with more training languages and even out-of-domain datasets can result in further improvements.

# CHAPTER 6: MISINFORMATION DETECTION: EVIDENCE RETRIEVAL OVER THE WEB

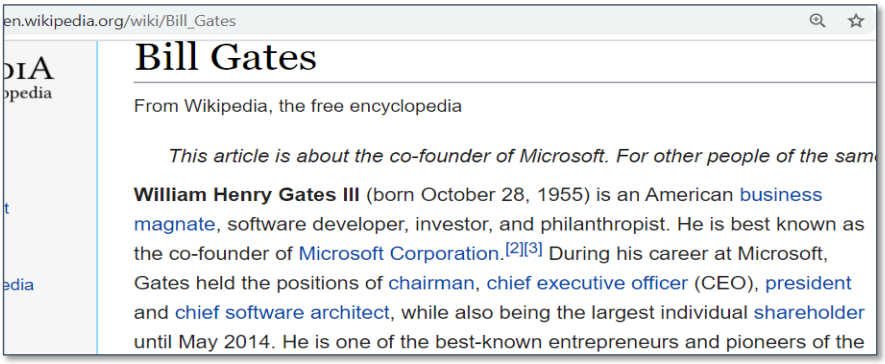
A crucial step in the fact checking pipeline following check-worthy claim identification, is the retrieval of information sources (e.g., Web pages) against which a claim can be verified [24]. Recently, studies demonstrated the value of and need for extracting *evidence* snippets from identified information sources. *Evidence* is essential to justify or explain the system's veracity prediction and provide user with information to make further assessment and decision regarding the claim's veracity [108]–[110]. Despite the recognized importance of evidence-based fact checking and the difficulty of evidence extraction [29], [116], few studies have attempted to characterize such important pieces of information and the documents that contain them [116]. Furthermore, effectiveness of retrieving documents with evidence is rarely characterized or evaluated.

Evaluation of information retrieval (IR) systems usually focused on their ability to retrieve *topically-relevant* documents, i.e., documents that are "about" the "topic" representing the user's information need [59], [60], [62]. However, recent studies argued that other dimensions of relevance (e.g., document understandability) should be considered for better system development and evaluation [25], [62], [64]–[67]. Building on earlier studies that examined the dimension of document utility or usefulness for a user searching the Web (e.g., [70]), this work studies this aspect of relevance within the domain of fact checking.

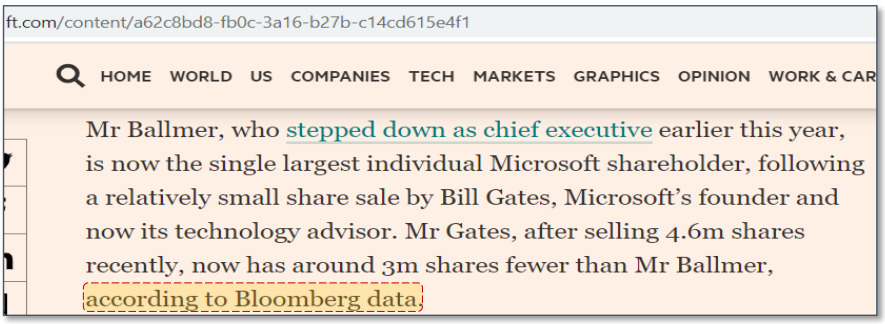
In this chapter, I aim to extensively analyze snippets constituting evidence for fact checking and the documents that contain them, and measure effectiveness of state-of-the-art systems designed to retrieve such documents. The work focuses on Web pages since many fact checking systems rely on searching the Web (e.g., Wikipedia) to extract evidence [180]. Formally, I examine one dimension of page relevance, which is denoted as *usefulness for claim verification*. I refer to a Web page that is useful for claim verification as an "evidential page" defined as *a topically-relevant page that contains at least one objective self-contained source-based evidence*. Examples of evidence can be quotes, statistics, or mentions of sources, to name a few.

In the definition of an evidential page in this work, I focus on one major type of evidence denoted as *source-based evidence* (SRE). "Evidence" in fact-checking research has been typically defined in terms of stance; studies assumed the evidence to be a snippet of text that supports or refutes the claim, e.g., [29], [116], [181]. These *stance-based evidence* (STE) snippets are usually repetitions of the claim itself in addition to making other claims, i.e., STE conveys information not supported by any source or reference but the document itself. Differently, I analyze the SRE type of evidence, which is objective and presented with a clear source of information. Figure 6.1 compares the two evidence types for the claim "Bill Gates was the largest individual shareholder of Microsoft". The STE snippet and claim are borrowed from FEVER dataset [29]. We notice in the SRE snippet that "Bloomberg", a specialized source on the subject, was cited to support the given claim. Such source offers great potential for verification as it can be further consulted, if needed, by the user. For the STE snippet, the page only repeats the claim itself offering no actual evidence other than the statement of the page's author. This situation becomes even trickier when the author is anonymous or unknown to the user, making trusting a mere repetition of the claim unreasonable.

I believe Web search systems tailored for evidence extraction can support two types of users: normal users searching the Web to fact-check a claim, and assistive fact checking systems exploiting the huge amount of information on the Web. Consequently, a major objective of my study is to provide insights on how to improve Web search for




(a) STE




(b) SRE

Figure 6.1. Web pages showing two types of evidence: stance-based (STE) and source-based (SRE) for the claim: "Bill Gates was the largest individual shareholder of Microsoft". Source of evidence in SRE is highlighted.

the task of evidential page retrieval. I address my main objective through answering the following research questions:

RQ6.1 To what extent topical pages are evidential, and how correlated is the effectiveness of retrieving these two types of pages?

RQ6.2 What types of evidence can be found in evidential pages?

RQ6.3 What textual features distinguish evidential and non-evidential pages?

RQ6.4 How effective are existing systems in retrieving evidential pages?

The above questions are answered by analyzing the performance of a commercial search engine in two tasks: topically-relevant pages retrieval and evidential pages retrieval. My study shows that pages (retrieved by a commercial search engine) that are topically-relevant to a claim are not always useful for verifying it, and that the search engine performance is generally weakly correlated across the two tasks using two correlation coefficients (Kendall's $\tau$ and Pearson's $r$). Given the aforementioned finding, I investigate and identify characteristics or features specific to evidential pages. Furthermore, preliminary experiments show that effectiveness of a supervised evidential pages retrieval model that employs them has a 5.3% increased recall of evidential pages over the search engine.
Overall, my contribution in this study is four-fold.

1. I conducted the first in-depth comparative study of the performance of Web search for the tasks of retrieving topically-relevant vs. evidential pages for verifying a given claim, showing that the two tasks are inherently different.
2. The study provides a thorough analysis of distinguishing characteristics of *evidence* appearing in *evidential* Web pages, which is rarely studied in existing literature. Furthermore, it shows that the identified characteristics, when leveraged in a supervised evidential pages retrieval model, lead to promising results.
3. The study establishes benchmarking results over the given dataset and quantifies the potential performance gain Web search systems can attain to better support the task of retrieving evidential pages for fact-checking.
4. We release an annotated dataset for the task of re-ranking of Web pages by usefulness for claim verification.[1] The dataset includes 2,641 Web pages that are potentially-relevant to 59 claims and annotated by both dimensions of relevance (i.e., topical and evidential) compared in this study.

The remainder of this chapter is organized as follows. First, the dataset construction process is described in Section 6.1. Section 6.2 then discusses how the task of evidential pages retrieval is evaluated. I proceed to compare the performance of Web search in the tasks of retrieving topically-relevant vs. evidential pages in Section 6.3. In Section 6.4, I analyze evidential pages and identify distinguishing linguistic characteristics. Effectiveness of those characteristics is then examined in Section 6.5. Before concluding, Section 6.6 offers benchmarking results for state-of-the-art models and demonstrates that a significant improvement in evidential page retrieval is attainable by search engines.

## 6.1. Dataset Construction

Several datasets that tackled evidence-based fact checking can be found in literature; however, they mostly either include artificially-constructed claims such as FEVER [29], provide unlabelled Web pages as evidence such as MultiFC [182] and WIKIFACTCHECK [183], include a small set of claims and evidential pages as in [184], or include pages labelled for stance rather than usefulness such as EMERGENT [185]. Differently, my study aims at understanding how "topical relevance" to a claim is different from "usefulness" for verifying that claim. To achieve this goal, I need a dataset that enables such study. To that end, I conduct my analysis on a dataset we constructed (CT19–T2) designed specifically for the task of predicting page usefulness for claim verification as part of CheckThat! lab at CLEF2019 [37], [186]. CT19–T2 includes general Web pages that are not limited to few domains. Claims were manually curated and coupled with a large set of manual annotations for both topical relevance and usefulness making the effectiveness comparison between the two tasks possible.

In CT19–T2, we define an evidential page (i.e., a page that is *useful* for verification) with respect to a given claim as *a page that is both topically-relevant to the claim and it provides evidence to determine the claim's veracity*. Examples of evidence can be quotes, some statistics, or mentions of sources. The evidence must be supported by a mention to its source in the page. The dataset is comprised of three components: (1) **59** Arabic claims (labelled by veracity); 30 of them are verified as true, and the rest

[1] `http://qufaculty.qu.edu.qa/telsayed/datasets`

are false. (2) **2,641** corresponding Web pages (labelled by usefulness); **661** of them are found evidential, and (3) **1,940** passages resulting from manual splitting of evidential Web pages (labelled also by usefulness); **737** of them are found evidential. This section presents the process of curating each of these components.

### *6.1.1. Claims*

We selected 59 claims from multiple sources including a pre-existing set of Arabic claims [22], a survey in which we asked the public to provide examples of claims they have heard of, and headlines from six Arabic news agencies that we rewrote into claims. The news agencies selected are well-known in the Arab world: AlJazeera, BBC Arabic, CNN Arabic, AlYoum AlSabea, AlArabiya, and RT Arabic. It should be noted that the number of claims used in my study is in range of that reported in many similar datasets (e.g., TREC Web search collections [187]).

We manually categorized the collected claims into topical categories to ensure that the dataset is not skewed towards one type of claims. Table 6.1 demonstrates that the claims cover a variety of topical categories.

Table 6.1. Claims distribution in CT19–T2

| Category | # Claims | Example (translated) claim |
|---|---|---|
| **Politics & Economy** | 21 | CT19–T2-024: Egyptian President El-Sisi proposed expanding the Gaza Strip towards Sinai |
| **Health** | 11 | CT19–T2-055: Excessive consumption of sugar causes the growth and spread of cancer cells in the human body |
| **Science & Technology** | 10 | CT19–T2-029: China announces iPhones sale ban in China starting from iPhone 6 through iPhone X |
| **Arts & Culture** | 6 | CT19–T2-002: *Capernaum* made it to final nominations for the 2019 Golden Globe Awards in the category of Best Foreign Film |
| **Sports** | 4 | CT19–T2-021: Brazilian goalkeeper Alison Baker moved from Italian club Roma to Liverpool |
| **Others** | 4 | CT19–T2-033: Two express trains collided at the Marsandiz Station in Ankara, leaving dozens of deaths and injuries |
| **Social** | 3 | CT19–T2-030: Artist Amr Diab married artist Dina El-Sherbiny |

**Labeling claims.** We acquired the veracity labels for the claims in two steps. First, two graduate students labelled all claims independently. Then, they met to resolve any disagreements, and thus reached consensus on the veracity labels for all claims.

### *6.1.2. Pages and Passages*

We depend on Web search to retrieve potentially-related pages to the claims. We manually formulated a search query representing each claim, and issued it against Google to retrieve the top 50 Web pages for each claim. Retrieved pages that were not Arabic or for which we could not acquire the HTML representation were discarded, leaving us with an average of 45 pages per claim. The pages were labelled following this pipeline:

1. **Topical Relevance**. We first identified topically-relevant pages. In order to speedup this labeling process, I hired two groups of annotators: Amazon Mechanical Turk crowd-workers and in-house annotators. Each page was labeled by *three* annotators, and majority voting determines the final label of the page.
2. **Usefulness**. Topically-relevant pages were then given to in-house annotators to be labeled for usefulness (i.e., evidentiality) using a two-way classification scheme: *evidential* and *not evidential*. Annotators were trained on the task and were instructed to closely look for the source of the evidence in the page. However, they were not asked to explicitly extract the source of evidence as part of the annotated dataset as this will over-complicate the annotation task for them. Each page was labeled by three annotators, and majority voting determines the final label of the page.
3. **Usefulness of Passages**. In addition to identifying evidential pages, we are also interested in finding out which passages in these pages contain the evidence. We manually split the *evidential* pages into passages, as we found that automatic splitting techniques not accurate enough. Finally, I labelled each passage as evidential or not.

### *6.1.3. Verifying Annotations Quality*

After constructing the dataset, evaluating its quality was essential to ensure it is representative of the task of evidential pages retrieval. I took two approaches to verify the quality of the dataset relying on the collected annotations. First, I verify that the task definition itself is accurate (Section 6.1.3.1). Then, inter-rater agreement is used to estimate consensus in annotations across annotators for the different annotation stages (Section 6.1.3.2).

#### *6.1.3.1. Validating Usefulness Definition*

When annotating the dataset, we hypothesized that an evidential page must be topically-relevant to the claim, and thus a non-relevant page cannot be useful. I examine the validity of this assumption in CT19–T2 by relabelling a set of non-relevant pages for usefulness. For each claim, I relabelled the 5 highest-ranked pages that received unanimous "non-relevant" judgment from the original annotators. I stress full agreement to maximize the chance that selected pages are indeed non-relevant. Only 260 non-relevant pages for 56 out of the 59 claims matched that condition. Results show that *none* of the relabelled non-relevant pages were eventually found evidential, reassuring the validity of the earlier assumption for CT19–T2 dataset. This outcome is also inline with observations found in literature where relevance is shown to be a condition for usefulness [70]. This suggests that it is very unlikely that a page which

is not topically-relevant to a claim will be useful in verifying it. *That encourages search-based fact checking systems to focus on retrieving topically-relevant pages first to downsize the pool of potentially-evidential pages.*

#### *6.1.3.2. Inter-annotator Agreement (IAA)*

I next evaluate the labels quality by computing inter-annotator agreement among the three judges of topical relevance and also usefulness of pages using Fleiss Kappa [162]. Inter-annotator agreement describes the degree of consensus in judgments among annotators and has been found to be a reasonable measure of judgment quality [188]. I first compute Fleiss $\kappa$ for the topical relevance labels over 2,641 Web pages. I found $\kappa$ = 0.7, which is considered substantial agreement according to a widely-adopted interpretation of Kappa values [163]. Next, $\kappa$ over usefulness labels on *topically-relevant documents* is computed. Agreement on usefulness was moderate ($\kappa$ = 0.49), which is lower than agreement on topical relevance. A possible justification is that usefulness is more complex to judge and requires good understanding of the fact checking process and what characterizes evidence in a Web page. Furthermore, explaining the concept of "evidence" to crowd-workers is quite difficult. *Overall, agreement level for both tasks is comparable to or higher than those achieved in literature for relevance judgments, e.g. [189].*

## 6.2. Evaluating Evidential Retrieval

Before studying the quality of evidential pages retrieval, I need first to establish a suitable evaluation approach for the task. I argue that using typical precision-oriented ranked retrieval evaluation measures might not be enough for this task, due to the different objective I envision the fact checking user has.

I assume the evaluation approach simulates an *artificial user* interacting with the search engine [190] to retrieve web pages useful in fact-checking a given claim. This allows us to put together a user model capturing the user interaction with the system and a corresponding evaluation measure. In designing this proposed user model, I benefit from existing models proposed for two related tasks: (1) focused retrieval tasks (e.g., passage retrieval or question answering), where the system is expected not only to retrieve a ranked list of relevant documents, but also to identify relevant text snippets from these pages given an initial query [191], and (2) the argument retrieval task, such as that described by [192], where the system should return a list of pages ranked by their potential of containing claims supporting or denying an argument. Differently from the latter task, my work only focuses on *factual evidence*, while in [192]'s work, retrieved claims can be opinionated and/or factual.

Let us assume that our user is trying to verify a claim by searching through a Web search engine using the claim as the input query. I hypothesize that the user's objective is to find as many *relevant* evidence from as few Web pages. More specifically, I hypothesise that:

- In order to verify the given claim, the user seeks to find as many evidential pages as possible, to help support or refute the claim. This means that the task is more *recall-oriented.*

- Due to the scale of claims a user might face each day, she has very limited time to verify claims; therefore, she is willing to spend some time looking for evidence for the given claim, but not so much. This calls for a cut-off point, $k$, at which the user stops looking at the retrieved pages. In this work, I set this cut-off point to a small value ($= 10$), as in typical Web search (with 10 results per results page), users are not likely to switch to the next results page. For professional fact-checkers, that cut-off point can be set to a larger value; this is left for future work.

- Focusing more on the task of fact checking, the user is more lenient about the rank of the retrieved evidential pages within the ranked list, before reaching the cutoff point.

Based on the above user model, I adopt a recall-based evaluation measure [191], [192]. The proposed measure is $Recall@k$, or $R@k$ for short. I specifically chose this measure since it was shown through fidelity testing that it is able to model and evaluate focused retrieval tasks [191]. The measure captures the percentage of retrieved evidential pages, for a given input claim, within the top $k$ retrieved pages. In my experiments, I set $k$ to 10.

## 6.3. Comparison of Topical and Evidential Relevance

In this section, I answer **RQ6.1**: *To what extent topical pages are evidential, and how correlated is the effectiveness of retrieving these two types of pages?* I conduct two studies addressing the following sub-questions:

1. How much does topical-relevance imply usefulness for claim verification?
2. How effective is the search engine in evidential pages retrieval?
3. How correlated is retrieval of evidential pages to retrieval of topically-relevant pages?

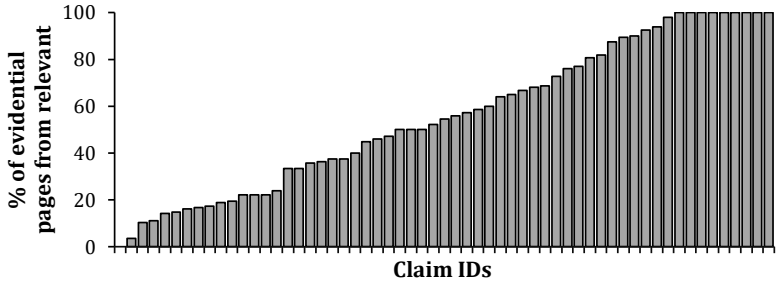



Figure 6.2. Percentage of evidential Web pages out of relevant per claim

### *6.3.1. How much does topical-relevance imply usefulness for claim verification? (RQ6.1.a)*

In this section, I test the hypothesis that usefulness is different from topical relevance by first examining the percentage of evidential pages from those topically-relevant per claim.

Figure 6.2 shows that for about two thirds of the claims, the percentage of evidential pages out of the relevant ones is less than 75%. Moreover, for third of the claims, this percentage is lower than 30%. The average percentage of evidential pages out of the relevant ones is 55.7% per claim. This indicates that, for many claims, a small percentage of topically-relevant pages are indeed useful for verification.

But is it the case that we can observe more evidential pages by getting more topically-relevant pages? To answer this question, I also look at the correlation of number of topically-relevant and evidential pages over all claims (Figure 6.3). I found that Pearson's Correlation $r$ [193] is 0.78 (significant with $p < 0.05$, and two-tailed paired t-test). High correlation is somewhat expected, since evidential pages are a subset of the relevant ones; however, the two sets are far from being equal or even close. As the figure shows, more topically-relevant pages among search results does *not* always imply more evidential pages. In fact, I observe that only few claims have equal number of relevant and evidential pages. I also observe some extreme cases. At one end, claims 20 and 2 for example had many topically-relevant pages *and* almost all were found evidential. At the other end, claims 43 and 53 had many topically-relevant pages too *but* very few were found evidential. Finally, the topical category of the claim does not generally influence correlation between topical relevance and evidentiality. In conclusion, topical-relevance is indeed not equivalent to usefulness for verification.

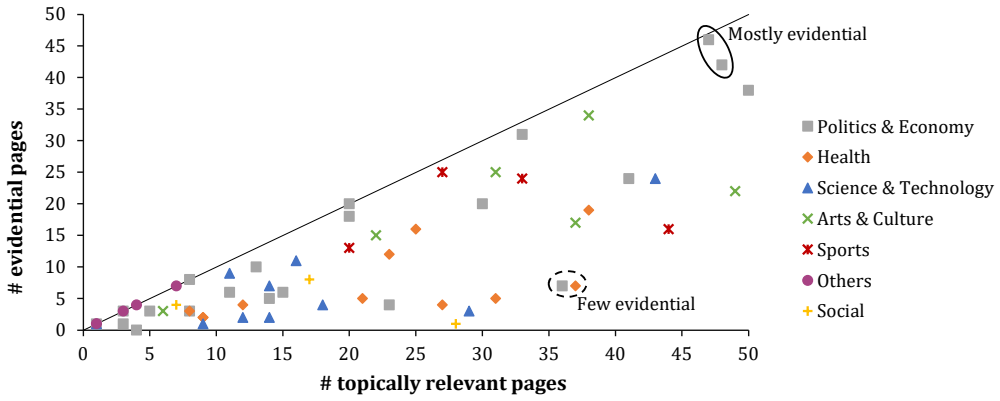



Figure 6.3. Correlation between number of evidential and topically-relevant pages per claim. Line shows ideal case where all relevant pages are evidential.

### *6.3.2. How effective is the search engine in evidential pages retrieval? (RQ6.1.b)*

Search engines have been proven to have great influence on users' opinions and reshaping their perceptions [120]. To assist human fact-checkers who use search engines to form an opinion about the veracity of a claim, the engine, given a claim, should optimize retrieval to present evidential pages.

To assess how effectively a search engine achieves that goal, I evaluate the commercial search system (i.e., Google) on the task of evidential pages retrieval using the aforementioned recall measure ($R@10$). For that task, evidential pages have a label of 1, while the remaining non-relevant and non-evidential pages get a label of 0. I find that, using our dataset, the engine achieves $R@10 = 0.54$, which is very far from the maximum possible value of 1. I believe this is the case because Web search engines are usually more optimized for precision than recall. In terms of topical relevance, I

evaluate the engine using measures typically used for ranked retrieval, namely, precision and average precision @ rank $k$ ($P@k$ and $AP@k$ respectively) setting $k = 10$. The same engine achieves $AP@10 = 0.77$ and $P@10 = 0.72$ in topical relevance retrieval, which is expected from a powerful commercial search engine typically optimized for the task of topical relevance retrieval. This shows a big gap in how the engine is optimized across the two tasks, reflected in the estimated performance, hence the estimated user's satisfaction, in each. I also note that, in terms of $P@10$, the system performs much better in the relevance retrieval task compared to earlier work on the Arabic Web [194]. This can be due to the fact that the existing work is a decade old and commercial search engines are expectedly better now.

### *6.3.3. How correlated is retrieval of evidential pages to retrieval of topically-relevant pages? (RQ6.1.c)*

Since the two tasks are assumed to have different user models, I find that direct comparison of system performance in the two retrieval tasks using the same evaluation measure is not meaningful. Instead, I opt to characterize the difference between the two tasks using correlation between the system performance per claim for both tasks. If the search engine sees the two tasks very similar, I expect perfect correlation. Correlation is a standard approach used in the information retrieval field to compare pairs of systems [195], [196]. I first compute Kendall's $\tau$ correlation coefficient [197] between the *rankings* of claims by the effectiveness of relevance retrieval (measured by $P@10$ or $AP@10$) and by the effectiveness of usefulness retrieval (measured by $R@10$). A linear correlation measure, Pearson's $r$ [193], between scores of both tasks is also reported.

Table 6.2. Correlation between retrieval performance of evidential pages (measured by $R@10$) and topically relevant pages (measured by $P@10$ or $AP@10$).

| **Usefulness Measure** | **Relevance Measure** | **Kendall $\tau$** | **Pearson $r$** |
|---|---|---|---|
| ***R@10*** | ***P@10*** | -0.24 | -0.42 |
| ***R@10*** | ***AP@10*** | 0.13 | 0.15 |

Results in Table 6.2 demonstrate that search system performance across the two retrieval tasks is consistently different. Moreover, the correlation is generally low in 3 out of the 4 scores from the table. Interestingly, recall of evidential pages and precision of relevance ranking has a negative correlation. I further investigate this observation by plotting this correlation in Figure 6.4. The figure shows that for many claims with perfect or near perfect precision in retrieving relevant pages in the first results page, recall of evidential pages varies a lot. In fact, out of the 14 claims for which the system got perfect $P@10$ (the right most vertical line), 8 claims of them featured less than 50% of total evidential pages among the top 10 ranks. Similarly, for claims with perfect recall in retrieving evidential pages in the first results page (the top horizontal line), precision of relevant pages greatly varies across the board. The figure correlating $R@10$ and $AP@10$ is omitted since I observe a similar pattern to Figure 6.4.

Overall, experiments presented so far demonstrated that (1) topically-relevant pages are *not* all evidential, i.e., they are not all useful for fact checking, and (2) performance of retrieval of evidential pages is not correlated with that of retrieval

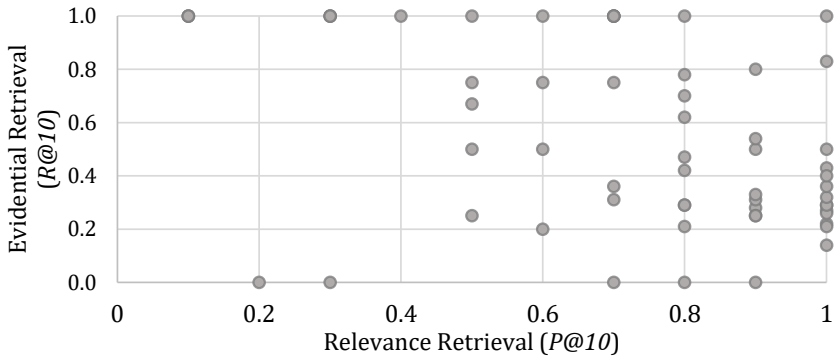



Figure 6.4. Correlation between evidential retrieval performance ($R@10$) and topical-relevance retrieval ($P@10$) per claim.

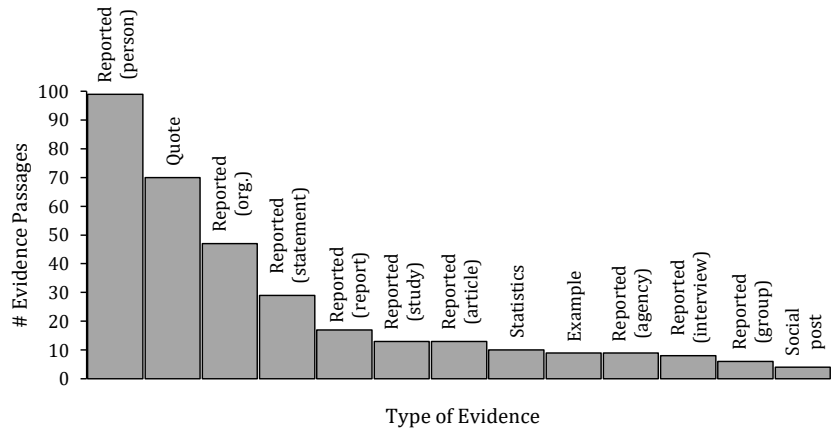



Figure 6.5. Distribution of evidence passages by type of evidence.

of topically-relevant pages. *This suggests that a Web search system used in a fact-checking setting need to be optimized to retrieve evidential pages to better support claim verification.*

## 6.4. Content Analysis

Establishing that retrieval of evidential and relevant pages is different through a statistical analysis has been insightful. However, understanding distinctive linguistic features in evidential pages can have more direct contribution to improving search system design for claim verification. I first develop a categorization of evidence types by manual inspection of evidential pages. Then, I characterize differences between evidential and non-evidential pages through a study of their lexical features.

### *6.4.1. What types of evidence can be found in evidential pages? (RQ6.2)*

Studies from the argumentation theory domain identify three types of evidence that one can use to support a claim [198]: (1) anecdotal (giving examples), (2) statistical (providing statistics and results), and (3) testimonial (quoting testimony from a source, e.g., person or a group). Starting from this high-level taxonomy, I identify a set of

common patterns or types of evidence in the dataset. For each claim, the 5 highest-ranked evidential pages that got unanimous annotation were inspected, resulting in 163 evidential pages with 334 evidential passages covering 44 claims.

I annotated each passage by the category of the main evidence in a passage, adding new categories while annotating. More precisely, I read each evidence, answered a question on whether the evidence type is anecdotal, statistical, or testimonial. Then, a sub-category was given whenever possible, for example, for testimonial (reported) evidence, I mentioned the source (e.g., person, organization, etc.) of the testimony. I then re-visited all the labelled sub-categories, grouped or re-defined some of them, and then reached a final taxonomy of thirteen types. Figure 6.5 shows the distribution of passages across types of evidence. Most evidential passages reported evidence as stated in or from a source, e.g., a person, an organization, an official statement, etc. Moreover, directly quoting from a source was the second commonly-used type of evidence. To clarify how SRE evidence was present in annotated paragraphs, consider this claim "*CT19–T2-050: Steve Jobs was the son of a Syrian immigrant*". An example evidential paragraph reporting evidence from a person is: "*Banksy, who is a famous artist, said: We are often tempted to believe that immigration is a drain of the country's resources, but Steve Jobs was the son of a Syrian immigrant.*" This example paragraph reported both the claim and that Banksy is a source of evidence.

*In conclusion, the study under RQ6.2 suggests that features capturing reported and quoted speech can help retrieve evidential pages.* I further study this observation next.

### *6.4.2. What textual features distinguish evidential pages? (RQ6.3)*

I examined several lexical and stylistic features in evidential and non-evidential (but topically relevant) pages. I run the analysis on a sample of the labelled pages since I had to manually extract content from the live version of some pages due to poor HTML structure. The random sample covers *all claims* from both evidential and non-evidential (but topically relevant) pages. For each category of pages, each claim was covered by the minimum of 5 pages (if any). Eventually, I analyzed 234 evidential and 220 non-evidential pages (representing 35% and 39% of pages per category respectively). Pages were tokenized, and stop words and URLs were removed. Features computed per page are listed below.

- **Length**: Tokens counts
- **Quotes**: Number of quoted statements, since many evidential pages contained quotes as evidence.
- **Statements**: Number of (Arabic) reported speech words, as addressing RQ6.2 showed that most evidence passages had reported statements. I compiled a list of 50 words frequently used to convey or report a statement made by others, such as (translated to English) "said", "reported", and "announced".
- **Unique tokens**: Number of unique tokens. This feature is assumed to represent lexical diversity in text.
- **Claim frequency**: Total frequency of claim words. The prevalence of claim words in a page is considered as an indication of its topical focus.

- **Entities**: Number of named entities, extracted using a multilingual named-entity recognition tool [199]. This aims at capturing the frequency of mentioning names of sources.
- **Numbers**: Count of numbers. It captures usages of statistics as evidence.
- **Sentiment frequency**: Count of words with positive polarity (e.g., "holy") and negative polarity (e.g., "corruption") identified using a large-scale multilingual sentiment lexicon [200]. I hypothesize that evidential pages, by definition, contain *objective* evidence and thus, will show less sentiment.
- **Exclamation frequency**: Number of characters associated with conveying emotions such as '!' and '?'. I hypothesize that evidential pages will show less emotions.
- **POS**: Counts of Part-of-Speech tags using Stanford tagger [201].

Overall, 9 features and 31 POS tags features are computed. For each feature, I compute per-page count, normalize it by page length, and compute average per class. Table 6.3 reports ratio of averages between evidential and non-evidential pages. Ratios $> 1$ indicate features more prevalent in evidential pages, while ratios $< 1$ denote features more prevalent in non-evidential pages. *Note that the table only shows features with values that are significantly different across the two classes* ($p < 0.05$ using two-tailed t-test). Additionally, the table shows *translated* examples from the Arabic pages.

I gauge power of each feature in discriminating among evidential and non-evidential classes by computing Kendall's $\tau$ correlation. $\tau$ is computed between two lists: feature value and page label for all pages. Rank of pages in the search result list per claim was used to break ties in both lists of scores.

Results in Table 6.3 strengthen the conclusion in RQ6.2. Features capturing reported speech, named entities, and quotes were most indicative of page usefulness. POS tags also exhibited similar trend. Participles describing actions (e.g., "saying"), and past-tense verbs (e.g., "stated") were more prevalent in evidential pages showing a tendency to refer to and report information. Furthermore, the language in non-evidential pages was more subjective/opinionated with more use of comparative adjectives, which has been shown to be a strong indicator of opinions [202]. However, this feature is not correlated to page label. Stronger correlation was found between noun quantifiers and page evidentiality. Interestingly, non-evidential pages are longer on average, also showing more lexical redundancy with less unique words. Closer inspection of non-evidential pages showed that several of them were actually directory pages that list a summary of many pages including one that is relevant to the claim. Some pages were forum pages with long discussion threads. Other pages were long articles covering a very general topic, in which the claim's topic is a sub-topic.

## 6.5. A Proof-of-Concept: Evidential Pages Retrieval Model

Following the prior identification of features that can characterize evidential pages, I now study their effectiveness by implementing a ranking model, as a proof-of-concept, for evidential pages retrieval that employs those features. The model re-ranks the same documents returned by the search engine allowing for comparison between the two scenarios.

Table 6.3. Relationship of page usefulness and linguistic features. Ratios indicate how frequently a feature appears in evidential pages compared to non-evidential. Examples show translated text from the pages matching features.

| Feature | Ratio | $\tau$ | Example |
|---|---|---|---|
| POS-Active & passive Participles | 1.49 | 0.14 | Prince Mohammad pointed out ... **saying**, "But I have learnt from previous experience ..." |
| Quotes | 1.36 | 0.15 | Described as **"a new Hitler in Middle East"**... |
| Statements | 1.34 | 0.12 | He **announced** the news on Twitter **saying**: ... |
| Entities | 1.19 | 0.12 | **HFPA** has announced ... **Golden Globe** Awards |
| POS-Verb (past) | 1.10 | 0.03 | Macron **pledged** economic reforms ... |
| Unique tokens | 1.06 | 0.09 | - |
| POS-Verb (present) | 0.90 | -0.07 | ... **enables** the immune system to ... |
| POS-Adjective (comparative) | 0.69 | -0.03 | **Most common** question is why... |
| Length | 0.66 | -0.13 | - |
| POS-Noun quantifier | 0.66 | 0.13 | **Most** operating systems uses a GUI. . . |
| POS-Verb (command) | 0.26 | 0.41 | **Put** half a teaspoon of olive oil and mix. |

### *6.5.1. Features and Classifiers*

The supervised model integrates features from Table 6.3 using traditional machine learning models. Additionally, the rank returned by the search engine is tested as a feature that somewhat indicates the relative page relevance.

The experiments are based on three models: random forest (RF), logistic regression (LR), and multilayer perceptron (MLP). For all classifiers, prediction probability of the positive class (i.e., probability that the page is evidential) is used to rank pages per claim.

### *6.5.2. Dataset*

The models are trained over the CT19–T2 dataset extended with 10 claims (and corresponding annotated pages) of the train set from the CheckThat! lab [37], [186] to increase the number of training examples. I only consider the topically-relevant Web pages from the ground truth, since the previously carried analysis was done over relevant pages which resulted in proposing features that differentiate between evidential and non-evidential *but topically-relevant* pages. The experiments are run using 69 claims and 1314 relevant pages (half of which are evidential).

### *6.5.3. Experimental Setup*

The classifiers are set using the default parameters provided by scikit-learn Python library.[2] I follow a leave-one-claim-out cross validation setup due to the relatively small dataset size. The average performance over folds (i.e., claims) reflect the model's performance. Runs were evaluated using the aforementioned recall measure. I also report precision and average precision at the top 10 ranks to support other potential user models.

### *6.5.4. Results*

Table 6.4 shows performance results of seven models:

- The search engine (**SE**) (i.e., Google) baseline.
- The three traditional learning models using the proposed 11 features.
- The three traditional learning models using the 11 features and the original rank feature (given by the search engine). I hypothesize this feature can capture relative topical relevance of the page following Google's scoring model.

Table 6.4. Evidential retrieval model performance. Results for best model by $R@10$ are boldfaced.

| **Model** | **Features** | ***R@10*** | ***P@10*** | ***MAP@10*** |
|---|---|---|---|---|
| SE | SE rank | 0.599 | 0.537 | 0.539 |
| LR | 11 features | 0.597 | 0.543 | 0.561 |
| MLP | 11 features | 0.600 | 0.544 | 0.559 |
| RF | 11 features | 0.614 | 0.548 | 0.571 |
| LR | 11 features + SE rank | 0.589 | 0.548 | 0.573 |
| MLP | 11 features + SE rank | **0.631** | **0.560** | **0.588** |
| RF | 11 features + SE rank | 0.618 | 0.562 | 0.590 |

Table 6.4 shows all the models only employing the proposed features had comparable or higher performance scores compared to the search engine, with an increase up to 2.5% and 6% by RF model measured by $R@10$ and $MAP@10$ respectively. Interestingly, adding the SE rank feature to the proposed features shows further increase in scores over using the features alone, with an increase up to 5.3% and 9.5% over the search engine performance measured by $R@10$ and $MAP@10$ respectively. This suggests that the proposed features are better coupled with features capturing Web page topical relevance. While these results demonstrate potential effectiveness of the proposed features, the difference in performance over the search engine model was not statistically significant ($p<0.05$, two tailed paired t-test), indicating that more effective features are needed to attain higher improvements.

[2] `https://scikit-learn.org/`

## 6.6. Benchmarking

Previous sections provided insights on some features to consider when designing systems for retrieving evidential pages. I wonder if current systems, designed for reranking Web pages for usefulness, are good enough (**RQ6.4**). Although the problem of re-ranking pages by usefulness for claim verification is relatively new to the automated fact-checking domain [184], there is already some effort in literature to design such systems, in particular in Subtask A of Task 2 of the CheckThat! lab at CLEF2019. These systems were also evaluated using the dataset we are proposing in this work: CT19–T2 [37].[3] It is worth noting here that in Subtask A, the task required systems to rank *potentially-relevant* Web pages by usefulness which is slightly different than the task as defined in the previous section where the system ranked *relevant* Web pages (given ground truth). Table 6.5 compares performance of the best run submitted to the lab (**CLEF-Best**) against two other runs. The first is the original ranking returned by Google (**SE**), representing the performance of existing search engines for the task. The other is an **Oracle** run that perfectly re-ranks pages retrieved by **SE** by placing the evidential pages at the top of the list. This Oracle run is indeed a "cheating" run that knows the labels of the pages and orders pages using these labels. The goal of that run is to establish an upper bound for usefulness-oriented retrieval systems on *this dataset*.

Systems were evaluated using recall, precision and average precision at 10. I also report statistical significance of performance difference between the **Oracle** and the other runs ($p<0.05$, two tailed paired t-test).

Table 6.5. Performance of retrieving evidential pages. Oracle scores marked with $^*$ and $^\dagger$ indicate statistically-significant difference from SE and CLEF-Best respectively.

| **Run** | ***R@10*** | ***P@10*** | ***MAP@10*** |
|---|---|---|---|
| CLEF-Best | 0.48 | 0.40 | 0.45 |
| SE | 0.54 | 0.42 | 0.49 |
| Oracle | $0.80^{*\dagger}$ | $0.63^{*\dagger}$ | $1.00^{*\dagger}$ |

Results demonstrate that a significantly large performance improvement can potentially be attained by search engines or existing fact-checking systems. In the modern age of rapid spread of fake news, efficient fact-checking is a primary goal [203]. More emphasis should be given to designing systems that provide the users with a short but highly-effective list of evidential pages. Such list will help the user (and a fact-checking system) reach a fact-checking decision faster since she only need to look at few documents to make a decision as opposed to having a longer list of topically-relevant but not fully evidential documents.

## 6.7. Conclusions and Future Work

In this chapter, I carried an analytical study where several features were employed to characterize differences between evidential and non-evidential Web pages in the context of fact checking. Furthermore, I showed that those features have some potential

[3] Summary on these systems is presented in the related work section.

by leveraging them in a learning model for evidential pages retrieval. I also examined the performance of existing search systems in retrieving such pages. The main aim was to provide insights on how to better design usefulness-oriented search systems for claim verification. My study has showed that: (1) topically-relevant pages retrieved by a search engine do not always contain evidence needed to verify the given claim, (2) performance of an effective commercial search engine is different in usefulness retrieval compared to topical relevance retrieval and the system performance is weakly correlated, (3) most evidential pages include reported statements from sources, quotes, and entities; these linguistic cues are strong predictors of page usefulness, and (4) Significantly large performance improvements can be attained to better support evidential page retrieval.

There are several potential directions for future work. A more thorough textual analysis using more sophisticated features such as subjectivity (e.g. [204]) can be conducted. Investigating other aspects of evidence, such as reliability [24], is another interesting direction. Another factor to consider is studying the effect of the source of evidence on the evidence quality and availability in a Web page. Moreover, it is necessary to quantify the accuracy of the user model proposed in this chapter to capture the task of retrieving evidential pages for claim verification. Finally, given that this dissertation provided solutions for two components in claim verification pipeline, it is necessary to investigate the performance of these components in a case study over real from-the-wild data.

# CHAPTER 7: CONCLUSIONS AND FUTURE WORK

Searching the Web and accessing social media have become a crucial part of our daily activities. Online content, especially that posted on social media, offers a great opportunity for both normal and professional users to follow news as it emerges, and publish and read updates and opinions on ongoing events. In the Arab world, and among Arabic-speaking users, this online content is mostly written and consumed in Arabic. While navigating this deep sea of online content, Arabic-speaking users are flooded with overwhelming and noisy information; automatic IR systems are more needed than ever. Not only typical information retrieval tasks like ad hoc search are currently important, new challenges (e.g., spread of misinformation) have emerged, requiring new types of IR systems.

In this dissertation, I argue that the progress in developing IR systems for Arabic content is very slow. Although many IR systems for a variety of problems have been proposed in literature for over three decades, we observe that these solutions were mainly implemented over English datasets, leaving an answered question on the effectiveness of state-of-the-art models over content in a language as unique as Arabic. Moreover, building of IR systems that were specifically designed by modelling Arab users, addressing their specific needs, and tackling the challenges of their language, is still in its infancy. This dissertation attempts to push research in that direction by solving one of the main obstacles hindering this research, which is identified as the lack of training and evaluation test collections, and annotated datasets. This quest is very complex, thus, I chose to go the breadth-first route rather than a depth-first approach. I focused on two important IR problems that are independent in some tasks, but can also interplay in others. Specifically, I proposed and implemented approaches to create evaluation datasets, and provided benchmarking experiments of state-of-the-art models, for ad hoc retrieval and misinformation detection, focusing on two domains: the Web and Twitter. Furthermore, I proposed alternative approaches to training and evaluation of IR systems to two sub-problems of misinformation detection, namely, check-worthy claim identification and evidence retrieval.

**Ad hoc Retrieval over the Web.** In Chapter 3, I presented ArTest, which is the first large-scale Arabic Web test collection. ArTest was constructed with the help of in-house annotators, and depended on query variations to eliminate the need for a shared-task evaluation campaign. The test collection is composed of 50 topics (and the queries used to develop them), and an associated set of 10,529 judged document-topic pairs; all are made publicly available. After establishing the test collection, I tested existing effective neural retrieval models, that use BERT transformer, over ArTest. The preliminary results showed that these models did not beat the typical BM25 model, conflicting with the relative systems ranking over English datasets. Further experiments are needed to understand the reasons behind this suboptimal performance, however, initial inspections indicated that this might be due to the types of documents in ArTest.

**Ad hoc Retrieval over Social Media.** Chapter 4 described a language-independent approach to creating a high-quality multi-task tweet test collection without running a shared-task campaign. The proposed Arabic tweet collection, EveTAR, supports four retrieval tasks, namely event detection, ad hoc search, timeline generation, and real-time summarization. It includes 355M Arabic tweets, 50 topics, 62K relevance judgments, and novelty annotations. For future research interested in using this collection, I provided baseline performance by testing multiple effective systems per task.

**Misinformation Detection: Check-worthy Claim Detection over Social Media.** In Chapter 5, the aim was to study whether we can build effective check-worthiness de-

tection systems over Arabic tweets, without the need for an Arabic training dataset. To meet this goal, I proposed and ran the first extensive comparative study of techniques of cross-lingual learning transfer for the task. The in-depth experiments showed that cross-lingual transfer models result in comparable performance to the monolingual models fully trained on many Arabic examples. Moreover, multilinguality during fine-tuning negatively affected the model transfer performance. I also showed that the proposed models are not far behind the state-of-the-art models on the same Arabic test set. Finally, as part of this work, I proposed and constructed the first Arabic tweets dataset that is manually annotated by claim check-worthiness.

**Misinformation Detection: Evidence Retrieval over the Web.** In Chapter 6, I proposed and created the first manually annotated dataset for the task of evidence retrieval for claim verification over the Arabic Web. I then carried an analytical study where several features were employed to characterize differences between evidential and non-evidential Web pages in the context of fact checking. Furthermore, I showed that those features have some potential by leveraging them in a learning model for evidential pages retrieval. I also examined the performance of existing search systems in retrieving such pages. The main aim was to provide insights on how to better design usefulness-oriented search systems for claim verification.

## 7.1. Future Work

For at least three IR problems: ad hoc retrieval, check-worthy claim detection and evidence retrieval, my work paved the way for several advancement opportunities in the area of constructing and evaluating Arabic IR systems. Few ideas for future directions are presented next.

- **Arabic-specific IR systems:** With the datasets that this work made available, and baseline performance established, research on Arabic IR can now focus on the actual development phase of IR systems over Arabic content.
- **Arab user understanding and modelling:** For at least the task of ad hoc retrieval, the datasets are constructed over very large samples of the Arabic Web and Twitter. These samples, coupled with relevance annotations, can enable many interesting user studies such as the investigation of how Arab users interact with results returned by a search system. Another example study can be to investigate the prevalence of dialectal content in the annotated datasets, and how it correlates with relevance.
- **Cross-lingual transfer learning for IR:** My work showed that at least for one IR problem, transfer from other languages to Arabic is possible. This offers great progress opportunity in the field, since it allows us to benefit from the wealth of existing non-Arabic datasets to train effective systems over Arabic data for other IR tasks. Evaluation of such systems can then be carried using the datasets released with this dissertation. Not only these results are promising for Arabic IR, the same idea of cross-lingual transfer can be tested for other lower-resource languages, at least on the task of claim check-worthiness detection.
- **Extended evaluation resources:** The proposed approaches for constructing the test collections and annotated datasets in this work can be replicated for other

languages, domains and tasks such as evidence passage retrieval. Moreover, the same datasets of this work can be extended to cover other tasks. For example, if the claims in the check-worthiness detection dataset were to be labelled by veracity, then the dataset can now serve the first and last steps in the fact-checking pipeline.

## 7.2. Publications

In this section, I list my publications that are directly related to each of the core four chapters in this dissertation.

- **[Chapter 3] Ad hoc Retrieval over the Web**:
  - **M. Hasanain**, Y. Barkallah, R. Suwaileh, M. Kutlu, and T. Elsayed, "ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2017–2020
- **[Chapter 4] Ad hoc Retrieval over Social Media**:
  - H. Almerekhi, **M. Hasanain**, and T. Elsayed, "EveTAR: A new test collection for event detection in Arabic tweets," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 689–692
  - **M. Hasanain**, R. Suwaileh, T. Elsayed, M. Kutlu, and H. Almerekhi, "EveTAR: building a large-scale multi-task test collection over Arabic tweets," *Information Retrieval Journal*, vol. 21, no. 4, pp. 307–336, 2018
- **[Chapter 5] Misinformation Detection: Check-worthy Claim Detection over Social Media**:
  - **M. Hasanain**, F. Haouari, R. Suwaileh, Z. Ali, B. Hamdan, T. Elsayed, A. Barrón-Cedeño, G. Da San Martino, and P. Nakov, "Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020
  - **M. Hasanain** and T. Elsayed, "bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020
  - S. Shaar, **M. Hasanain**, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, J. Beltrán, T. Elsayed, and P. Nakov, "Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021
- **[Chapter 6] Misinformation Detection: Evidence Retrieval over the Web**:

- **M. Hasanain** and T. Elsayed, "Studying effectiveness of web search for fact checking," *Journal of the Association for Information Science and Technology*, Oct. 2021. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24577
- **M. Hasanain**, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, and P. Nakov, "Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality," L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds., ser. CEUR Workshop Proceedings, 2019

### 7.3. Additional Publications

In this section, I list publications that are extended/summarized versions of the previously mentioned publications, and additional publications serving the main goals of this work including improving Arabic IR, and evaluation of IR systems.

- **Misinformation Detection**:
  - T. Elsayed, P. Nakov, A. Barrón-Cedeño, **M. Hasanain**, R. Suwaileh, G. Da San Martino, and P. Atanasova, "Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2019, pp. 301–321
  - A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, **M. Hasanain**, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, and Z. S. Ali, "Overview of CheckThat! 2020: Automatic identification and verification of claims in social media," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, and N. Ferro, Eds., 2020, pp. 215–236
  - F. Haouari, **M. Hasanain**, R. Suwaileh, and T. Elsayed, "ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 72–81
  - P. Nakov, D. Corney, **M. Hasanain**, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. Da San Martino, "Automated fact-checking for assisting human fact-checkers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Survey Track, International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4551–4558
  - P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, **M. Hasanain**, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, and Y. S. Kartal, "Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news," in *Experimental IR Meets Multilinguality, Multimodality, and*

*Interaction*, K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, and N. Ferro, Eds., Springer International Publishing, 2021, pp. 264–291

- **IR over Tweets**:
  - **M. Hasanain**, T. Elsayed, and W. Magdy, "Improving tweet timeline generation by predicting optimal retrieval depth," in *Proceedings of the 11th Asia Information Retrieval Societies Conference*, ser. AIRS 2015, 2015, pp. 135–146
  - R. Suwaileh, **M. Hasanain**, M. Torki, and T. Elsayed, "QU at TREC-2015: Building real-time systems for tweet filtering and question answering," in *Proceedings of the Twenty-Fourth Text REtrieval Conference*, ser. TREC' 15, 2016
  - R. Suwaileh, **M. Hasanain**, and T. Elsayed, "Light-weight, conservative, yet effective: Scalable real-time tweet summarization," in *Proceedings of the Twenty-Fifth Text REtrieval Conference*, ser. TREC' 16, 2017
  - **M. Hasanain**, M. Bagdouri, T. Elsayed, and D. Oard, "What questions do journalists ask on Twitter?" *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 2, pp. 127–134, Aug. 2021
  - A. Albahem, **M. Hasanain**, M. Torki, and T. Elsayed, "QweetFinder: Real-time finding and filtering of question tweets," in *Proceedings of the 39th European Conference on Information Retrieval Research*, ser. ECIR 2017, 2017, pp. 766–769

REFERENCES


[1] F. Salem, "Social media and the internet of things towards data-driven policymaking in the arab world: Potential, limits and concerns," *The Arab Social Media Report, Dubai: MBR School of Government*, vol. 7, 2017.

[2] K. Darwish and W. Magdy, "Arabic Information Retrieval," *Foundations and Trends in Information Retrieval*, vol. 7, no. 4, pp. 239–342, Feb. 2014.

[3] A. A. T. S. Eldin, "Socio linguistic study of code switching of the arabic language speakers on social networking," *International journal of English linguistics*, vol. 4, no. 6, p. 78, 2014.

[4] N. Al-Qaysi and M. Al-Emran, "Code-switching usage in social media: A case study from oman," *International Journal of Information Technology and Language Studies*, vol. 1, no. 1, pp. 25–38, 2017.

[5] R. Baly, G. Badaro, G. El-Khoury, *et al.*, "A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models," in *Proceedings of the third Arabic natural language processing workshop*, 2017, pp. 110–118.

[6] K. Darwish, W. Magdy, and A. Mourad, "Language processing for arabic microblog retrieval," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 2427–2430.

[7] S. Harrat, K. Meftouh, and K. Smaili, "Machine translation for arabic dialects (survey)," *Information Processing & Management*, vol. 56, no. 2, pp. 262–273, 2019, Advance Arabic Natural Language Processing (ANLP) and its Applications.

[8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. DOI: 10.1017/CBO9780511809071.

[9] K. Darwish, N. Habash, M. Abbas, *et al.*, "A panoramic survey of natural language processing in the arab world," *Communications of the ACM*, vol. 64, no. 4, pp. 72–81, 2021.

[10] C. Moral, A. de Antonio, R. Imbert, and J. Ramírez, "A survey of stemming algorithms in information retrieval," *Information Research: An International Electronic Journal*, vol. 19, no. 1, 2014.

[11] M. El-Haj, U. Kruschwitz, and C. Fox, "Creating language resources for under-resourced languages: Methodologies, and experiments with arabic," *Language Resources and Evaluation*, vol. 49, no. 3, pp. 549–580, 2015.

[12] B. Elayeb and I. Bounhas, "Arabic cross-language information retrieval: A review," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 15, no. 3, 2016, ISSN: 2375-4699.

[13] S. MacAvaney, L. Soldaini, and N. Goharian, "Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning," in *Advances in Information Retrieval*, J. M. Jose, E. Yilmaz, J. Magalhães, *et al.*, Eds., Springer International Publishing, 2020, pp. 246–254.

[14] V. Karpukhin, B. Oguz, S. Min, *et al.*, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[15] C. W. Cleverdon, "The evaluation of systems used in information retrieval," in *Proceedings of the international conference on scientific information*, 1959, pp. 687–698.

[16] M. Sanderson, "Test collection based evaluation of information retrieval systems," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 4, pp. 247–375, 2010.

[17] F. C. Gey and D. W. Oard, "The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries," in *Proceedings of the 10th Text REtrieval Conference*, ser. TREC-2001, 2001.

[18] D. W. Oard and F. C. Gey, "The trec 2002 arabic/english clir track," in *Proceedings of the Eleventh Text REtrieval Conference*, ser. TREC 2002, 2002.

[19] S. Jamil, "10 years after the arab spring: The dark and the bright sides of social media," *Digital Journalism*, pp. 1–5, 2022.

[20] J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen, *Types, sources, and claims of covid-19 misinformation*, Reuters Institute for the Study of Journalism, University of Oxford, 2020.

[21] P. Nakov, D. Corney, M. Hasanain, *et al.*, "Automated fact-checking for assisting human fact-checkers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Survey Track, International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4551–4558.

[22] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov, "Integrating stance detection and fact checking in a unified corpus," in *Proceedings of NAACL-HLT'18*, 2018.

[23] W. Mansour, T. Elsayed, and A. Al-Ali, "Did i see it before? detecting previously-checked claims over twitter," in *Proceedings of the 44th European Conference on Information Retrieval*, ser. ECIR '22, 2022.

[24] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, and X. Tannier, "A content management perspective on fact-checking," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 565–574.

[25] J. Unkel and A. Haas, "The effects of credibility cues on the selection of search engine results," *Journal of the Association for Information Science and Technology*, vol. 68, no. 8, pp. 1850–1862, 2017.

[26] M. Abualsaud and M. D. Smucker, "Exposure and order effects of misinformation on health search decisions," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '19, 2019.

[27] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, 2021.

[28] M. Pikuliak, M. Šimko, and M. Bieliková, "Cross-lingual learning for text processing: A survey," *Expert Systems with Applications*, vol. 165, 2021.

[29] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: A large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, ser. NAACl-HLT'18, 2018, pp. 809–819.

[30] R. Suwaileh, M. Kutlu, N. Fathima, T. Elsayed, and M. Lease, "ArabicWeb16: A new crawl for today's Arabic web," in *Proceedings of the 39th International ACM Conference on Research and Development in Information Retrieval*, ser. SIGIR' 16, 2016, pp. 673–676.

[31] M. Hasanain, F. Haouari, R. Suwaileh, *et al.*, "Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020.

[32] "Clef 2020 working notes," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020.

[33] P. Nakov, G. Da San Martino, T. Elsayed, *et al.*, "Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, K. S. Candan, B. Ionescu, L. Goeuriot, *et al.*, Eds., Springer International Publishing, 2021, pp. 264–291.

[34] S. Shaar, M. Hasanain, B. Hamdan, *et al.*, "Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[35] M. Hasanain, R. Suwaileh, T. Elsayed, M. Kutlu, and H. Almerekhi, "EveTAR: building a large-scale multi-task test collection over Arabic tweets," *Information Retrieval Journal*, vol. 21, no. 4, pp. 307–336, 2018.

[36] M. Hasanain and T. Elsayed, "Studying effectiveness of web search for fact checking," *Journal of the Association for Information Science and Technology*, Oct. 2021. eprint: `https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24577`.

[37] M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeño, and P. Nakov, "Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 2: Evidence and factuality," L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds., ser. CEUR Workshop Proceedings, 2019.

[38] M. Hasanain, Y. Barkallah, R. Suwaileh, M. Kutlu, and T. Elsayed, "ArTest: The First Test Collection for Arabic Web Search with Relevance Rationales," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2017–2020.

[39] M. Hasanain and T. Elsayed, "bigIR at CheckThat! 2020: Multilingual BERT for ranking Arabic tweets by check-worthiness," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020.

[40] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06, 2006, pp. 619–620.

[41] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling for large collections," *Information Retrieval*, vol. 10, no. 6, pp. 491–508, 2007.

[42] K. Roitero, J. S. Culpepper, M. Sanderson, F. Scholer, and S. Mizzaro, "Fewer topics? a million topics? both?! on topics subsets in test collections," *Information Retrieval Journal*, pp. 1–37, 2019.

[43] M. Kutlu, T. Elsayed, and M. Lease, "Intelligent topic selection for low-cost information retrieval evaluation: A new perspective on deep vs. shallow judging," *Information Processing & Management*, vol. 54, no. 1, pp. 37–59, 2018.

[44] M. Hosseini, I. J. Cox, N. Milic-Frayling, M. Shokouhi, and E. Yilmaz, "An uncertainty-aware query selection model for evaluation of ir systems," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR'12, 2012, pp. 901–910.

[45] D. E. Losada, J. Parapar, and A. Barreiro, "Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems," *Information Processing & Management*, vol. 53, no. 5, pp. 1005–1025, 2017.

[46] T. McDonnell, M. Lease, M. Kutlu, and T. Elsayed, "Why is that relevant? collecting annotator rationales for relevance judgments," in *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[47] H. Almerekhi, M. Hasanain, and T. Elsayed, "EveTAR: A new test collection for event detection in Arabic tweets," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 689–692.

[48] R. Malhas and T. Elsayed, "Ayatec: Building a reusable verse-based test collection for arabic question answering on the holy qur'an," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 6, Oct. 2020, ISSN: 2375-4699.

[49] I. Bounhas and S. Ben Guirat, "Kunuz: A multi-purpose reusable test collection for classical arabic document engineering," in *Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 2019, pp. 1–8.

[50] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2515–2519.

[51] M. El-Masri, N. Altrabsheh, and H. Mansour, "Successes and challenges of arabic sentiment analysis research: A literature review," *Social Network Analysis and Mining*, vol. 7, no. 1, pp. 1–22, 2017.

[52] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Information Processing & Management*, vol. 56, no. 2, pp. 320–342, 2019.

[53] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," *Applied Computational Intelligence and Soft Computing*, vol. 2020, 2020.

[54] I. A. Farha and W. Magdy, "A comparative study of effective approaches for arabic sentiment analysis," *Information Processing & Management*, vol. 58, no. 2, p. 102 438, 2021.

[55] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 82–91.

[56] T. Alhindi, A. Alabdulkarim, A. Alshehri, M. Abdul-Mageed, and P. Nakov, "Arastance: A multi-country and multi-domain dataset of arabic stance detection for fact checking," in *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2021, pp. 57–65.

[57] M. S. H. Ameur and H. Aliane, "Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, 2021.

[58] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Workshop of the cross-language evaluation forum for european languages*, Springer, 2001, pp. 355–370.

[59] D. Harman, "Overview of the first text retrieval conference (TREC-1)," in *Proceedings of TREC 1992*, 1992.

[60] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance," *Journal of the American Society for information Science and Technology*, vol. 58, no. 13, pp. 2126–2144, 2007.

[61] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: Nature and manifestations of relevance," *Journal of the American society for information science and technology*, vol. 58, no. 13, pp. 1915–1933, 2007.

[62] J. Jiang, D. He, and J. Allan, "Comparing in situ and multidimensional relevance judgments," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 405–414.

[63] D. Harman *et al.*, "Information retrieval: The early years," *Foundations and Trends® in Information Retrieval*, vol. 13, no. 5, pp. 425–577, 2019.

[64] S. Yigit-Sert, I. S. Altingovde, C. Macdonald, I. Ounis, and O. Ulusoy, "Explicit diversification of search results across multiple dimensions for educational search," *Journal of the Association for Information Science and Technology*, 2020.

[65] C. da Costa Pereira, M. Dragoni, and G. Pasi, "Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting," *Information Processing & Management*, vol. 48, no. 2, pp. 340–357, 2012.

[66] 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*, 2016, pp. 965–968.

[67] F. Johnson, J. Rowley, and L. Sbaffi, "Exploring information interactions in the context of google," *Journal of the Association for Information Science and Technology*, vol. 67, no. 4, pp. 824–840, 2016.

[68] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer, "Web credibility: Features exploration and credibility prediction," in *Proceedings of the 35th European Conference on Information Retrieval Research*, ser. ECIR 2013, 2013, pp. 557–568.

[69] P. Vakkari, "The usefulness of search results: A systematization of types and predictors," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020, pp. 243–252.

[70] J. Mao, Y. Liu, K. Zhou, *et al.*, "When does relevance mean usefulness and user satisfaction in web search?" In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 463–472.

[71] P. Vakkari, M. Völske, M. Potthast, M. Hagen, and B. Stein, "Modeling the usefulness of search results as measured by information use," *Information Processing & Management*, vol. 56, no. 3, 2019.

[72] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," in *Proceedings of the 3rd Text REtrieval Conference*, ser. TREC-3, 1994.

[73] S. E. Robertson, S. Walker, M. Beaulieu, and P. Willett, "Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track," *Nist Special Publication SP*, no. 500, pp. 253–264, 1999.

[74] A. Almazrua, M. Almazrua, and H. Alkhalifa, "Comparative analysis of nine arabic stemmers on microblog information retrieval," in *2020 International Conference on Asian Language Processing (IALP)*, 2020, pp. 60–65.

[75] J. Lin, R. Nogueira, and A. Yates, "Pretrained transformers for text ranking: Bert and beyond," *Synthesis Lectures on Human Language Technologies*, vol. 14, no. 4, pp. 1–325, 2021.

[76] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[77] P. Shi and J. Lin, "Cross-lingual relevance transfer for document retrieval," *arXiv preprint arXiv:1911.02989*, 2019.

[78] P. Shi, R. Zhang, H. Bai, and J. Lin, "Cross-lingual training with dense retrieval for document retrieval," *arXiv preprint arXiv:2109.01628*, 2021.

[79] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, "Fake news types and detection models on social media a state-of-the-art survey," in *Intelligent Information and Database Systems*, P. Sitek, M. Pietranik, M. Krótkiewicz, and C. Srinilta, Eds., 2020.

[80] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *arXiv preprint arXiv:2108.11896*, 2021.

[81] I. Barrutia-Barreto, R. Seminario-Córdova, and B. Chero-Arana, "Fake news detection in internet using deep learning: A review," in *Combating Fake News with Computational Intelligence Techniques*, M. Lahby, A.-S. K. Pathan, Y. Maleh, and W. M. S. Yafooz, Eds. Springer International Publishing, 2022, pp. 55–67.

[82] M. Lahby, S. Aqil, W. M. S. Yafooz, and Y. Abakarim, "Online fake news detection using machine learning techniques: A systematic mapping study," in *Combating Fake News with Computational Intelligence Techniques*, M. Lahby, A.-S. K. Pathan, Y. Maleh, and W. M. S. Yafooz, Eds. Springer International Publishing, 2022, pp. 3–37.

[83] M. A. Al-Asadi and S. Tasdemir, "Using artificial intelligence against the phenomenon of fake news: A systematic literature review," in *Combating Fake News with Computational Intelligence Techniques*, M. Lahby, A.-S. K. Pathan, Y. Maleh, and W. M. S. Yafooz, Eds. Springer International Publishing, 2022, pp. 39–54.

[84] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Association for Computing Machinery, 2017, pp. 1803–1812, ISBN: 9781450348874.

[85] J. M.-R. Juan R. Martinez-Rico and L. Araujo, "NLP&IR@UNED at CheckThat! 2021: Check-worthiness estimation and fake news detection using transformer models," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[86] R. Sepúlveda-Torres and E. Saquete, "GPLSI team at CLEF CheckThat! 2021: Fine-tuning BETO and RoBERTa," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[87] E. Williams, P. Rodrigues, and S. Tran, "Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[88] J. Beltrán, R. Míguez, and I. Larraz, "Claimhunter: An unattended tool for automated claim detection on twitter," in *Proceedings of the 1st International Workshop on Knowledge Graphs for Online Discourse Analysis*, ser. KnOD 2021, 2021.

[89] T. Alhindi, B. McManus, and S. Muresan, "What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 380–391.

[90] D. Wright and I. Augenstein, "Claim check-worthiness detection as positive unlabelled learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 476–488.

[91] P. Atanasova, L. Marquez, A. Barrón-Cedeño, *et al.*, "Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness," L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, Eds., ser. CEUR Workshop Proceedings, 2018.

[92] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, and G. Da San Martino, "Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness," L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds., ser. CEUR Workshop Proceedings, 2019.

[93] S. Shaar, A. Nikolov, N. Babulkov, *et al.*, "Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media," L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., ser. CEUR Workshop Proceedings, 2020.

[94] F. Arslan, N. Hassan, C. Li, and M. Tremayne, "A benchmark dataset of check-worthy factual claims," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 821–829.

[95] K. Meng, D. Jimenez, F. Arslan, J. D. Devasier, D. Obembe, and C. Li, "Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims," *arXiv preprint arXiv:2002.07725*, 2020.

[96] I. Baris Schlicht, A. Magnossão de Paula, and P. Rosso, "UPV at CheckThat! 2021: Mitigating cultural differences for identifying multilingual check-worthy claims," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[97] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525.

[98] F. Alam, F. Dalvi, S. Shaar, *et al.*, "Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, no. 1, pp. 913–922, 2021.

[99] F. Alam, S. Shaar, F. Dalvi, *et al.*, "Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 611–649.

[100] L. Uyangodage, T. Ranasinghe, and H. Hettiarachchi, "Can multilingual transformers fight the covid-19 infodemic," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 1432–1437.

[101] S. Shaar, F. Alam, G. Da San Martino, *et al.*, "Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection," in *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Online: Association for Computational Linguistics, Jun. 2021, pp. 82–92.

[102] M. S. Zengin, Y. S. Kartal, and M. Kutlu, "TOBB ETU at CheckThat! 2021: Data engineering for detecting check-worthy claims," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.

[103] Y. S. Kartal and M. Kutlu, "Re-think before you share: A comprehensive study on prioritizing check-worthy claims," *IEEE Transactions on Computational Social Systems*, pp. 1–14, 2022.

[104] D. Varshney and D. K. Vishwakarma, "Hoax news-inspector: A real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 9, pp. 8961–8974, 2021.

[105] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Systems with Applications*, vol. 169, p. 114 171, 2021, ISSN: 0957-4174.

[106] L. Wu, Y. Rao, L. Sun, and W. He, "Evidence inference networks for interpretable claim verification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 058–14 066.

[107] A. T. Nguyen, A. Kharosekar, M. Lease, and B. Wallace, "An interpretable joint graphical model for fact-checking from crowds," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[108] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," 1, vol. 52, 2015, pp. 1–4.

[109] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2561–2571.

[110] A. T. Nguyen, A. Kharosekar, S. Krishnan, *et al.*, "Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 189–199.

[111] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 22–32.

[112] A. Hanselowski, H. Zhang, Z. Li, *et al.*, "Ukp-athene: Multi-sentence textual entailment for claim verification," in *Proceedings of FEVER Workshop 2018*, 2018.

[113] C. Malon, "Team papelo: Transformer networks at fever," in *Proceedings of FEVER Workshop 2018*, 2018.

[114] L. Favano, M. Carman, and P. Lanzi, "Theearthisflat's submission to clef'19checkthat! challenge," in *CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, 2019.

[115] F. Haouari, Z. Ali, and T. Elsayed, "Bigir at clef 2019: Automatic verification of arabic claims over the web," in *CLEF 2019 Working Notes*, 2019.

[116] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim, "Show me your evidence-an automatic method for context dependent evidence detection," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, ser. EMNLP'15, 2015, pp. 440–450.

[117] X. Wang, C. Yu, S. Baumgartner, and F. Korn, "Relevant document discovery for fact-checking articles," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 525–533.

[118] S. Jiang, S. Baumgartner, A. Ittycheriah, and C. Yu, "Factoring fact-checks: Structured information extraction from fact-checking articles," in *Proceedings of The Web Conference 2020*, 2020.

[119] S. Jiang and C. Wilson, "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–23, 2018.

[120] D. Trielli and N. Diakopoulos, "Search as news curator: The role of google in shaping attention to news information," in *Proceedings of the 2019 CHI Conference on human factors in computing systems*, 2019, pp. 1–15.

[121] 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*, 2017, pp. 2931–2937.

[122] B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, 2017, pp. 759–766.

[123] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2017, pp. 647–653.

[124] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, "User variability and ir system evaluation," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'15, 2015, pp. 625–634.

[125] S. Robertson, H. Zaragoza, *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[126] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.

[127] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-stage document ranking with bert," *arXiv preprint arXiv:1910.14424*, 2019.

[128] C. Li, A. Yates, S. MacAvaney, B. He, and Y. Sun, "Parade: Passage representation aggregation for document reranking," *arXiv preprint arXiv:2008.09093*, 2020.

[129] Z. A. Yilmaz, S. Wang, W. Yang, H. Zhang, and J. Lin, "Applying bert to document retrieval with birch," 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): System Demonstrations*, 2019, pp. 19–24.

[130] C. L. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2004 terabyte track," in *Proceedings of the 13th Text REtrieval Conference*, ser. TREC-2004, 2004.

[131] E. M. Voorhees, "Overview of the trec 2004 robust retrieval track," in *Proceedings of the 13th Text REtrieval Conference*, ser. TREC-2004, 2004, pp. 52–69.

[132] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *ACM SIGIR Forum*, vol. 51, 2017, pp. 235–242.

[133] K. S. Jones and C. J. Van Rijsbergen, "Information retrieval test collections," *Journal of documentation*, vol. 32, no. 1, pp. 59–75, 1976.

[134] C. L. Clarke, N. Craswell, and E. M. Voorhees, "Overview of the trec 2012 web track," Tech. Rep., 2012.

[135] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber, "The effect of threshold priming and need for cognition on relevance calibration and assessment," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR'13, 2013, pp. 623–632.

[136] K. Sparck-Jones and C. Van Rijsbergen, *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*, ser. British Library Research and Development reports. University Computer Laboratory, 1975.

[137] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98, 1998, pp. 282–289.

[138] A. Moffat, F. Scholer, P. Thomas, and P. Bailey, "Pooled evaluation over query variations: Users are as diverse as systems," in *proceedings of the 24th ACM international on conference on information and knowledge management*, ser. CIKM'15, 2015, pp. 1759–1762.

[139] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *LREC 2020 Workshop Language Resources and Evaluation Conference*, 2020, p. 9.

[140] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Aug. 2021, pp. 7088–7105.

[141] A. Bruns, T. Highfield, and J. Burgess, "The Arab Spring and Social Media Audiences English and Arabic Twitter Users and Their Networks," *American Behavioral Scientist*, vol. 57, no. 7, pp. 871–898, Jul. 2013.

[142] B. S. Wasike, "Framing news in 140 characters: How social media editors frame the news and interact with audiences via Twitter," *Global Media Journal*, vol. 6, no. 1, 2013.

[143] P. Grover, A. K. Kar, Y. K. Dwivedi, and M. Janssen, "Polarization and acculturation in us election 2016 outcomes – can twitter analytics predict changes in voting preferences," *Technological Forecasting and Social Change*, vol. 145, pp. 438–460, 2019.

[144] N. Alsaedi, P. Burnap, and O. Rana, "Sensing Real-World Events using Arabic Twitter Posts," in *Proceedings of the tenth International AAAI Conference on Web and Social Media*, ser. ICWSM '16, 2016, pp. 515–518.

[145] W. Magdy and T. Elsayed, "Unsupervised adaptive microblog filtering for broad dynamic topics," *Information Processing & Management*, vol. 52, no. 4, pp. 513–528, 2016.

[146] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff, "Overview of the TREC-2011 Microblog Track," in *Proceedings of the 20th Text REtrieval Conference*, ser. TREC '11, 2011.

[147] I. Soboroff, I. Ounis, C. Macdonald, and J. Lin, "Overview of the TREC-2012 Microblog Track," in *Proceedings of the 21st Text REtrieval Conference*, ser. TREC '12, 2012.

[148] J. Lin and M. Efron, "Overview of the TREC-2013 Microblog Track," in *Proceedings of the 22nd Text REtrieval Conference*, ser. TREC '13, 2013.

[149] J. Lin, M. Efron, Y. Wang, and G. Sherman, "Overview of the TREC-2014 Microblog Track," in *Proceedings of the 23rd Text REtrieval Conference*, ser. TREC '14, 2014.

[150] J. Lin, M. Efron, Y. Wang, G. Sherman, and E. Voorhees, "Overview of the TREC-2015 Microblog Track," in *Proceedings of the 24th Text REtrieval Conference*, ser. TREC '15, 2015.

[151] J. Lin, A. Roegiest, L. Tan, R. McCreadie, E. Voorhees, and F. Diaz, "Overview of the TREC 2016 Real-Time Summarization Track," in *Proceedings of the 25th Text REtrieval Conference*, ser. TREC '16, 2016.

[152] 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*, ACM, 2011, pp. 35–44.

[153] J. Lin and G. Mishne, "A study of "churn" in tweets and real-time search queries," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ser. ICWSM '12, 2012.

[154] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, ACM, 2013, pp. 1147–1158.

[155] Z. Zhao and Q. Mei, "Questions about questions: An empirical analysis of information needs on twitter," in *Proceedings of the 22nd international conference on World Wide Web*, ACM, 2013, pp. 1545–1556.

[156] D. Elsweiler and M. Harvey, "Engaging and maintaining a sense of being informed: Understanding the tasks motivating twitter search," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 264–281, 2015.

[157] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a Large-scale Corpus for Evaluating Event Detection on Twitter," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13, 2013, pp. 409–418.

[158] N. Alsaedi and P. Burnap, "Arabic event detection in social media," in *Computational Linguistics and Intelligent Text Processing*, 2015, pp. 384–401.

[159] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," in *Technical Report cucs-012-11*, Columbia University, 2011.

[160] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL HLT'10, Association for Computational Linguistics, 2010, pp. 181–189.

[161] V. Pavlu and J. Aslam, "A practical sampling strategy for efficient retrieval evaluation," *College of Computer and Information Science, Northeastern University*, 2007.

[162] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[163] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[164] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ser. CSCW '11, 2011, pp. 355–358.

[165] J. Benhardus and J. Kalita, "Streaming trend detection in twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, 2013.

[166] S. Petrović, "Real-time event detection in massive streams," Ph.D. dissertation, School of Informatics, University of Edinburgh, 2013.

[167] M. Hasanain and T. Elsayed, "QU at TREC-2014: Online Clustering with Temporal and Topical Expansion for Tweet Timeline Generation," in *Proceedings of the 23rd Text REtrieval Conference*, ser. TREC '14, 2014.

[168] W. Magdy, T. Elsayed, and M. Hasanain, "On the evaluation of tweet timeline generation task," in *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings*, N. Ferro, F. Crestani, M.-F. Moens, *et al.*, Eds. Springer International Publishing, 2016, pp. 648–653.

[169] R. F. Sear, N. Velásquez, R. Leahy, *et al.*, "Quantifying COVID-19 content in the online health opinion war using machine learning," *IEEE Access*, vol. 8, pp. 91 886–91 893, 2020.

[170] "Working notes of clef 2018–conference and labs of the evaluation forum," L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, Eds., ser. CEUR Workshop Proceedings, 2018.

[171] "Working notes of CLEF 2019 conference and labs of the evaluation forum," L. Cappellato, N. Ferro, D. Losada, and H. Müller, Eds., ser. CEUR Workshop Proceedings, 2019.

[172] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," 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)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844.

[173] C. Samarinas, W. Hsu, and M. L. Lee, "Improving evidence retrieval for automated explainable fact-checking," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Association for Computational Linguistics, Jun. 2021, pp. 84–91.

[174] M. Zhao, Y. Zhu, E. Shareghi, *et al.*, "A closer look at few-shot crosslingual transfer: The choice of shots matters," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5751–5767.

[175] M. A. Hedderich, D. Adelani, D. Zhu, J. Alabi, U. Markus, and D. Klakow, "Transfer learning and distant supervision for multilingual transformer models: A study on African languages," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 2580–2591.

[176] H. Schwenk and X. Li, "A corpus for multilingual document classification in eight languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. ( chair), K. Choukri, C. Cieri, *et al.*, Eds., European Language Resources Association (ELRA), May 2018, pp. 3548–3551.

[177] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4996–5001.

[178] S. Wu and M. Dredze, "Are all languages created equal in multilingual BERT?" In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 2020, pp. 120–130.

[179] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[180] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, ser. AAAI'19, 2019, pp. 6859–6866.

[181] Y. Zhang, Z. Ives, and D. Roth, "Evidence-based trustworthiness," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 413–423.

[182] I. Augenstein, C. Lioma, D. Wang, *et al.*, "Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims," 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)*, 2019, pp. 4685–4697.

[183] A. Sathe, S. Ather, T. M. Le, N. Perry, and J. Park, "Automated fact-checking of claims from wikipedia," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6874–6882.

[184] K. Yasser, M. Kutlu, and T. Elsayed, "Re-ranking web search results for better fact-checking: A preliminary study," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM'18, 2018, pp. 1783–1786.

[185] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *NAACL-HLT'16*, 2016, pp. 1163–1168.

[186] T. Elsayed, P. Nakov, A. Barrón-Cedeño, *et al.*, "Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2019, pp. 301–321.

[187] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees, "Trec 2014 web track overview," Tech. Rep., 2015.

[188] T. T. Damessie, T. P. Nghiem, F. Scholer, and J. S. Culpepper, "Gauging the quality of relevance assessments using inter-rater agreement," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'17, 2017, pp. 1089–1092.

[189] O. Alonso and S. Mizzaro, "Using crowdsourcing for trec relevance assessment," *Information processing & management*, vol. 48, no. 6, pp. 1053–1066, 2012.

[190] A. F. Wicaksono and A. Moffat, "Metrics, user models, and satisfaction," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 654–662.

[191] J. Pehcevski and J. A. Thom, "Evaluating focused retrieval tasks," in *SIGIR 2007 Workshop on Focused Retrieval*, 2007.

[192] H. Roitman, S. Hummel, E. Rabinovich, B. Sznajder, N. Slonim, and E. Aharoni, "On the retrieval of wikipedia articles containing claims on controversial topics," in *Proceedings of the 25th international conference companion on world wide Web*, 2016, pp. 991–996.

[193] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[194] W. Tawileh, T. Mandl, J. Griesbaum, *et al.*, "Evaluation of five web search engines in arabic language," in *Proceedings of LWA2010*, 2010, pp. 221–228.

[195] B. Carterette, "On rank correlation and the distance between rankings," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09, 2009, pp. 436–443.

[196] E. Yilmaz, J. A. Aslam, and S. Robertson, "A new rank correlation coefficient for information retrieval," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR'08, 2008, pp. 587–594.

[197] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[198] H. Hoeken and L. Hustinx, "When is statistical evidence superior to anecdotal evidence in supporting probability claims? the role of argument type," *Human Communication Research*, vol. 35, no. 4, pp. 491–510, 2009.

[199] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, "Polyglot-NER: Massive multilingual named entity recognition," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, SIAM, 2015, pp. 586–594.

[200] Y. Chen and S. Skiena, "Building sentiment lexicons for all major languages," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ser. ACL'14, 2014, pp. 383–389.

[201] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 252–259.

[202] B. Liu *et al.*, "Sentiment analysis and subjectivity," *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[203] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, 21:1–21:42, 2019, ISSN: 2157-6904.

[204] M. Abdul-Mageed, M. Diab, and S. Kübler, "Samar: Subjectivity and sentiment analysis for arabic social media," *Computer Speech & Language*, vol. 28, no. 1, pp. 20–37, 2014.

[205] A. Barrón-Cedeño, T. Elsayed, P. Nakov, *et al.*, "Overview of CheckThat! 2020: Automatic identification and verification of claims in social media," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Arampatzis, E. Kanoulas, T. Tsikrika, *et al.*, Eds., 2020, pp. 215–236.

[206] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV19-Rumors: Arabic COVID-19 Twitter dataset for misinformation detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 72–81.

[207] M. Hasanain, T. Elsayed, and W. Magdy, "Improving tweet timeline generation by predicting optimal retrieval depth," in *Proceedings of the 11th Asia Information Retrieval Societies Conference*, ser. AIRS 2015, 2015, pp. 135–146.

[208] R. Suwaileh, M. Hasanain, M. Torki, and T. Elsayed, "QU at TREC-2015: Building real-time systems for tweet filtering and question answering," in *Proceedings of the Twenty-Fourth Text REtrieval Conference*, ser. TREC' 15, 2016.

[209] R. Suwaileh, M. Hasanain, and T. Elsayed, "Light-weight, conservative, yet effective: Scalable real-time tweet summarization," in *Proceedings of the Twenty-Fifth Text REtrieval Conference*, ser. TREC' 16, 2017.

[210] M. Hasanain, M. Bagdouri, T. Elsayed, and D. Oard, "What questions do journalists ask on Twitter?" *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 2, pp. 127–134, Aug. 2021.

[211] A. Albahem, M. Hasanain, M. Torki, and T. Elsayed, "QweetFinder: Real-time finding and filtering of question tweets," in *Proceedings of the 39th European Conference on Information Retrieval Research*, ser. ECIR 2017, 2017, pp. 766–769.

[212] "CLEF 2021 working notes. Working notes of CLEF 2021–conference and labs of the evaluation forum," G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, Eds., ser. CEUR Workshop Proceedings, 2021.