Otto-von-Guericke University Magdeburg

Department of Computer Science
Institute for Intelligent Cooperating Systems

Master Thesis

Fairness and Biases in Arabic language: A Case study on Named Entity Recognition models

Author:
Khaled Seddik Tawfik

October 11, 2021

Advisers:

Supervisor
Prof. Dr.-Ing. Ernesto W. De Luca
Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Supervisor
M.Sc. Erasmo Purificato
Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Seddik Tawfik, Khaled:
Fairness and Biases in Arabic language:
A Case study on Named Entity Recognition models
Master Thesis, Otto-von-Guericke University
Magdeburg, 2021.

Contents

Abstract

1 Introduction and Motivation

1.1 Motivation . . . 1
1.2 Objective and Contribution . . . 5
1.3 Research Questions . . . 6
1.4 Thesis layout . . . 7

2 Background

2.1 Arabic Language . . . 9
2.2 Arabic Natural Language Processing . . . 10
2.2.1 Challenges/Language characteristics . . . 11
2.3 Named Entity Recognition . . . 14
2.3.1 Applications . . . 14
2.3.2 Named Entity Tag Set . . . 16
2.3.3 Named Entity Recognition Approaches . . . 18
2.4 Arabic Named Entity Recognition. . . . 19
2.4.1 Arabic Named Entity Tag Set . . . 19
2.4.2 State-of-the-art in Arabic NER . . . 21
2.5 Arabic NER Linguistic Resources . . . 23
2.5.1 Corpora & Datasets- Benchmarks . . . 24

3 Related Work

3.1 Bias in Natural Language Processing . . . 29
3.1.1 Bias in NER . . . 30
3.2 Mitigating Bias . . . 31
3.3 Bias in Arabic NLP . . . 32

4 Methods

4.1 Data . . . 33
4.2 Experimental setup . . . 35
4.2.1 Hardware details . . . 35
4.2.2 Models & techniques . . . 36
4.3 Evaluation . . . 39
4.4 Measuring Bias . . . 40

4.4.1 Experiment 1: 40
4.4.2 Experiment 2: 40
4.4.3 Experiment 3: 42
4.4.4 Experiment 4: 42
4.5 Mitigating Bias 43
4.5.1 Experiment 1: Analyzing ANERcorp 45
4.5.2 Experiment 2: Data Augmentation 45
4.5.3 Experiment 3: Re-measusring Bias 46
4.6 Case Study: Real Life Data 47

5 Results Discussion
5.1 Dataset 49
5.2 Quantifying Bias in models' prediction 49
5.2.1 Mitigating Bias 58
5.3 Case Study 63

6 Conclusions and Future Work
6.1 Conclusions 65
6.2 Future Work 66

A Abbreviations and Notations

B List of Figures

C List of Tables

D Bibliography

Abstract

Nowadays, Machine Learning (ML) and Natural Language Processing (NLP) technologies are gaining more and more popularity and it's becoming extremely important to understand the role they play in influencing social stereotypes and biases. Despite the fact that NLP models have demonstrated effectiveness in modelling a variety of tools and applications, they may be subject to transfer and even magnify gender bias detected in text corpora. Although, exploring bias is an ongoing research, techniques to eliminate this bias in NLP are still under development. In this paper, I examine recent research on identifying and reducing different types of biases in Arabic NLP (ANLP). I address three different forms of bias: gender, stereotype and religion bias; analyze method recognizing the bias and explore strategies to mitigate it. I focus on investigating the bias in five state-of-the-art Named Entity Recognition (NER) models, more specifically their capacity to detect male and female names as Person (PER) entity types. I test these models on manual generated benchmark containing 70 male and female names, 35 unisex names and 31 coptic and muslim arabic names. One major observation is that male names are being detected more than female names as PER entities. So i decided to take a step back and analyze the dataset used in training one of these models, realizing the dataset used is highly biased towards men, as it contains a total of 6186 male mentions compared to 248 female mentions. Additionally, i tried to mitigate this bias through creating two less biassed datasets, re-training the model and capturing the new analysis. The data and code used for this study will be made freely accessible for researchers to use in the future.

1 Introduction and Motivation

1.1 Motivation

Bias is a common occurrence in social encounters, many forms of actions may be influenced by our assumptions about other individuals(Hu et al. [2015]). In research conducted involving a first-person shooter video game, played by both white and black users, researches found out a higher percentage of users shooting a black player than white player even if they were holding a harmless item instead of a weapon (Correll et al. [2007]).

While people's willingness to express openly racial or sexist views has declined in recent years, latent or unintentional social bias can still affect people's behaviour, regardless of their motivations or attempts to prevent it. These implicit assumptions have been seen to motivate racist attitudes and worsen intergroup tension, for example, when it came to recruiting new research assistants, both male and female faculty members preferred male applicants over similarly competent female candidates (Moss-Racusin et al. [2012]).

Bias may be defined on the basis of allocation and representation bias. Allocation bias may be described as an economic problem in which a system disproportionately distribute funds to some groups over others, while representation bias arises when structures undermine certain groups' social identification and recognition (Sun et al. [2019]).

Natural Language Processing (NLP) is a domain of artificial intelligence in which computers analyze human language automatically for the purposes of speech recognition, digital translation, as well as other tasks that are now a part of much of our daily lives. While, virtual assistants (for instance, Alexa) and machine translation systems (for example, Google

Translate) are well-known examples of NLP applications, there are numerous other programs that communicate with us on a more subtle level, including those which manage and track work applicants (Costa-jussà [2019]).

Since these technologies are intended to improve our lifestyles, there is concern that they would reinforce and worsen social unfairness based on gender, ethnicity, age, faith, and regional roots (O'neil [2016]).

NLP is one of the most important sub-domains of Artificial Intelligence(AI), which is concerned with the interpretation and development of natural language. The development of natural language resources involves semantial knowledge in order to grasp the basic structural features of the language (Li [2019]).

Scientists have offered a wide range of ways to solve these concerns from rule-based to statistical techniques, including, in specific with a huge success, Machine Learning (ML) and Deep Learning (DL), among others. Various variables have aided in the progress of deep learning, such as increased data volume, improved optimization algorithms, and growing computational power. However, in order to maintain this success, we need to put two aspects into consideration, the importance of data consistency and model complexity.The world has gained a growing knowledge of the implications of data accuracy and volume while training models (Li [2019]).

That being said, there are several data-related concerns, for instance from a training and test outlook, it is difficult to deal with textual information/data particularly trying to implement different ML methods. Since textual data is frequently more diverse than the usual data, lays expectation of model performing effectively over a variety of text styles, which is usually called "domains". Additionally, models are subject to bias due to the depth of the model or during learning against similar behaviors, such as over-represented author demographics (Hardt et al. [2016]).

Text is highly changeable; a phrase can be presented in an endless number of shapes and forms without altering its context or purpose. For instance, the term "The dog played on the lawn.", transforming it on a linguistic basis to "The dog played on the grass.". By substituting only one word with a similar expression, leaving the sentence's initial context relatively unaffected/unvaried. Additionally, the sentence may be altered from a grammar point if the sequence or order has shifted while maintaining the sentence context like the following example "On the lawn the dog played.". A powerful NLP model working on sentiment analysis should not face difficulties acquiring the same solution nevertheless manipulating the sentence on a lexical or syntactic level, keeping in mind, the textual data is not always neat and standardized, as it has highly found case on social websites. Moreover, the datasets often include text with several contributors, leading to classification results. Fairness is a must overcome challenge by machine learning models, which is caused by having authorship characteristics such as sex, age, race, nationality, and religion (Li [2019]).

Turing made many arguments to focus the aim of computer science towards displaying the intellectual conduct and behavior. As a result, we should develop cutting-edge classification algorithms that can be as equal as human beings. Fairness is a widely recognized term that states that people should be dealt with relatively by law and society despite their social, physical, or mental characteristics (Turing [2009]).

Although the input data can be blamed for this discrimination in output, machine learning algorithms have also a tendency enhance such attitude. In this paper, the author demonstrated a high correlation between cooking and females rather than males in the training data of task classification dataset, and the trained model magnifies that effect at test time. This descrimination should not have been learned from a fair model (Zhao et al. [2017]).

Machine learning is a major subfield of artificial intelligence (AI) that focuses on the development of automated ways for computers to learn from data. Arthur Samuel, a computer scientist in the late 1950s, provided a description of ML as being "a field of study that gives computers the ability to learn without having been explicitly programmed"

In plain terms, it is a software that has been designed to learn from its own experience and mistakes. With ongoing study, development, and hope of solutions, machine learning techniques are being employed in a number of disciplines that influence our daily lives. These approaches are inspired by biological systems, including our brain's neural network. Both ML and AI, despite their current popularity today, are not new concepts: their initial application goes all the way back to the Second World War and the invention of the "electronic brain." idea. Although deep learning concepts were initially proposed in the 1990s, it was only lately that their application became possible due to increased computer power and increasingly huge datasets. Machine learning intersects with a variety of fields, including artificial intelligence, probability, statistics, cognitive science, and computer science.

The earliest Artificial Intelligence technology was straightforward, and easy to understand by humans; however, with the advancement of technology, more complex systems have emerged seeking a more opaque decision-making orientation, like Deep Neural Networks (DNNs) (Arrieta et al. [2020]).

DNNs, also known as black-box models, are the results of merging efficient classification techniques that stack hundreds of layers in order to reach critical decisions in crucial situations. The black box model does not explain how the model operates or its output, which causes a rise in the need for transparency from numerous stakeholders in AI (Castelvecchi [2016]).

The risk lies in developing and implementing decisions that are not justified, legal, or do not permit precise descriptions of their actions. Justifications for a model's performance are critical, for example, precision medicine, as specialists expect much more details from the model than a mere binary forecast to validate their diagnosis. To prevent impairing the usefulness of nowadays AI systems, eXplainable AI (XAI) suggests developing a set of machine learning algorithms that (Gunning [2017]):

1. increase the explainability of models while retaining a high degree of learning efficiency (e.g., prediction accuracy);

2. allow humans to comprehend, reasonably accept, and control the new generation of artificially intelligent partners

XAI methodologies have a role in discrimination-aware data mining approaches when it comes to evaluating explicit connections between secure and unregulated applications. The system developer can uncover previously unknown similarities between input data that can lead to inequality through examining the model's performance reacts in comparison to the input function (Arrieta et al. [2020]).

1.2 Objective and Contribution

The goal of this research is to evaluate the bias expressed in the existing Arabic natural language processing (ANLP) resources and possible mitigation techniques that overcomes the limitations of the social biases that manifest in textual data. More specifically, we investigate the bias in Arabic off-the-shelf Named Entity Recognition systems, and compare their ability to distinguish various biases including gender, stereotypical and religion biases. Moreover, we explore mitigation techniques that can contribute to a less-biased NER systems, by following a data augmentation approach, while still maintaining a good overall performance. To the best of our knowledge, this work is the first attempt to measure and mitigate bias in arabic NER systems.

Most NLP research and development efforts are directed toward modeling and replicating the human ability for producing and comprehending linguistic expressions for communication purposes with the assistance of technology. Initially, the first initiative in the NLP domain occurred in 1954, with the invention of a crude automatic language translator between Russian and English. Despite the fact that the vocabulary was limited to 250 words and the grammar had only six rules, this study sparked several works in the field. During the 1970s, a rise of linguistic formulas occurred, which combined both semantic and syntactic methods.

NLP has consistently delivered on its lofty promises of making data more accessible and communicative from the past to the future, and it continues to do so. Nowadays, the NLP domain is constantly evolving with significant achievements such as pre-trained deep learning solutions. Eventually, an increasing number of mainstream users will incorporate NLP-based approaches into their everyday activities.

The facts represented in natural language text are not always transparent; the user interprets the omitted components based on his or her common sense of information and understanding. Textual inference, for example, is described as the connection between two text pieces treated as a three-class prediction problem. It establishes whether a hypothesis is true (*Entailment*), incorrect (*contradiction*), or undecided (*Neutral*) in relation to the premise given two snippets of text. Conventional NLP approaches are incapable of detecting inference by nature.

1.3 Research Questions

As previously mentioned, language-based technologies should reduce the negative human reflections and biases such as racism and sexism. Similar to English, the biases embedded in language resources for Arabic can potentially affect more people than for most other languages. However, the NLP research in Arabic language is still in its infancy, research in the domain is scarce and in need of computational knowledge to reach its full potential. This work focuses on bridging the gap between the Arabic community and the NLP research in general and in the NER domain in particular. There is a lack of scientific methods that address the several challenges surrounding the analysis of ANLP. Throughout the dissertation, we investigate the benefits of employing machine learning for the sake of speeding-up the knowledge discovery process in the ANLP domain. This research is a step forward towards gender-fair models for ANLP by focusing on two main goals: providing a quantative analysis on the named entity recognition task and explorepossible mitigation techniques to address this issue and overcome its limitation. We, therefore, pose the following main research question:

MRQ — How can Current Debiasing techniques support Bias mitigation in the Arabic language?

To answer the main research question, we pose two research questions, as explained below.

RQ1 —To what extent is gender bias reflected in Arabic Named Entity Recognition (NER) systems and how it affects prediction?

RQ2 —How can data augmentation mitigate the bias while still maintaining a good performance ?

1.4 Thesis layout

This thesis is organized as follows: Chapter 2 provides the background of Arabic Natural Language processing in general, and Named Entity Recognition in specific. It also describes the characteristics of the Arabic language, and existing ANLP resources. Chapter 3 illustrates some of the work presented in literature related the bias identification and bias mitigation in NLP. Chapter 4 covers in details of the Arabic bias quantification experiment settings, the data collection process and de-biasing method investigated. Chapter 5 describes the results before and after applying the mitigation techniques. It also presents a comparison of 6 Arabic NER systems and how robust they are to overcome various biases. Finally, Chapter 6 illustrates the Conclusion and the Future work.

2
Background

2.1 Arabic Language

Arabic is a sophisticated language as it contains many accents and dialects, complicated grammar, and rich vocabulary. The rapid increase of Arabic data availability nowadays, resulting in an increased demand for more reliable and accurate processing tools, is straining current Arabic NLP research efforts. NER is regarded as a fundamental key component of Arabic natural language processing technologies and applications. Although considerable improvement has been made in this domain in recent years, the challenge stays overwhelming due to the complex structure of the Arabic language; however, there is always room for improvement and optimization (Farghaly and Shaalan [2009]).

More than 380 million people worldwide speak Arabic, and it is the official language of 25 nations. According to its use, Arabic can be classified into three main types (Elgibali [2005]):
Classical Arabic (CA) which is the official language of the Holy Qur'an and has been used for more than 1,500 years in Islam, is used by Muslims in their everyday prayers. In addition, numerous ancient Arabic documents and scripts are hand-written in CA; as these manuscripts are digitized and translated to text, Arabic NE may play a significant role (Monem et al. [2008]).
Modern Standard Arabic (MSA) is used in today's formal Arabic papers, TV news, newspapers, street signs, and official schoolbooks and education resources. It is also recognized by the United Nations (UN) as one of six main languages used in their meetings and legal documents. The majority of NLP applications, including NER, are allocated to MSA. The most significant distinction between MSA and CA is in the vocabulary, which

includes noun expressions and syntax in traditional written Arabic: MSA does not impose any restrictions on the use of short vowels (El Kholy and Habash [2010]).

Colloquial Arabic Dialects(CAD) is the casual language that people use in their everyday conversations in the Arab world; it differs from one nation to another, making it unteachable in school for its various versions that exist out there. On the contrary, from MSA, which is widely used in all Arab countries, CA is more of a local dialect that varies not within Arabic nations but also within the same state. Nowadays, its usage exists only on social media platforms for communication. For example, a person's name in either CA or MSA may take on several forms in Arabic dialect; like, (عبدالقادر),(Abd Al-Kader) versus (عبدالجادر), (Abd Al-Gader) or (عبدالآدر),(Abd Al-Aader) (Korayem et al. [2012]).

Salloum and Habash proposed a method to translate CA sentences into MSA phrases, allowing the existing MSA application to handle and analyze CA inputs, as the majority of Arabic NER programs are designed to support MSA (Salloum and Habash [2012]).

2.2 Arabic Natural Language Processing

ANLP has grown in popularity in recent years. Many new applications have been created based on it. Examples include information extraction, information retrieval, speech recognition, localization and multi-lingual systems, text-to-speech, translation, and tutoring systems. These applications have to cope with many difficult issues related to the origin of the Arabic language and how it is structured. The majority of ANLP systems created in Western countries rely on tools and applications to assist non-Arabic speakers in understanding Arabic text. Interest in ANLP technologies has increased significantly since 2011, particularly in the United States, where the US Department of Homeland Security was required to do tasks such as accurately recognizing Arabic names at airport security and interpreting Arabic documents seized by US authorities. The necessity to do this work with a high level of efficiency in a short period of time triggered research in the ANLP community, since it does not rely on the human component, therefore, reducing the time and effort needed for analysis. The necessity to do this work with a high level of efficiency in a short period spurred research in the ANLP community since it does not rely on the human component, reducing the time and effort needed for analysis. Intelligence and security services also rely heavily on ANLP tools such as Arabic named entity recognition, machine translation, and sentiment analysis (Farghaly and Shaalan [2009]).

Due to the critical nature of these tools, they were built using machine learning techniques. Machine learning is often not time-consuming or expensive and does not require extensive linguistic understanding. However, the creators of such technologies faced several challenges. A significant constraint in ANLP is the lack of a sufficiently large corpus for training, evaluating, and validating suggested systems in various NLP tasks (Munday [2009]).

2.2.1 Challenges/Language characteristics

It is extremely difficult to carry out NLP tasks on Arabic text, specifically NER functions, due to the language's oddities and unusual structure. The following are the major features of Arabic that raise non-trivial problems for NER tasks:

- No Capitalization: Capitalization plays an important role in European and Latin languages as it helps to identify different named entities like proper names, acronyms, and abbreviations. Nonetheless, this feature does not exist in the Arabic language, causing a model confusion between proper nouns(NEs) and the standard nouns and verbs(non-NEs). Therefore, it would not be a suitable solution to rely solely on finding entries in proper noun dictionaries as there's no distinction between NEs and non-NEs (Farber et al. [2008]). For instance, the word (أشرف) can be seen differently in a phrase such as a person name "Ashraf", a verb "he supervised" or an adjective "the most honorable" (Mesfar [2007b]).

- The Agglutinative Nature: Various sentence pattern is caused by the Arabic's agglutinative nature, which causes numerous lexical changes. A simple adjustment in the word's combination of prefix, stem or root, and suffixes, can lead to a highly structured yet complicated terminology. On the opposite of English language, where clitics are regarded as separate words, some Arabic clitics can be added to normal words. Arabic contains a number of clitics associated with NEs, which include prepositions such as (ل) (Laam, for/to), (ك) (k, as), and (ب) (baa, by/with), conjunctions such as (و) (Waw, and) and (ف) (if ... then), or a combination of both, as in (ول) (Waw-Laam, and-for). The NER task is dependent on the terms that compose the NE and its meaning. The words, as well as the meaning, can occur in various ways. To resolve data sparsity concerns without needing large training corpora, the attached syntaxes should be preprocessed morphologically (Grefenstette et al. [2005]). In this paper, (Benajiba et al. [2007]) One possibility that was suggested to overcome this problem is to eliminate all affixes and retain only the root term. For instance, to assess the term (وبألمانيا) (and by Germany) will transform it into just a location (ألمانيا) (Germany). One other approach that focus on avoiding loss of semantic information, is to slice the word and insert a space between the term and clitics.

- Lack of Short Vowels: One unique feature of the vowels in the Arabic language is diacritics, which can influence the phonetic representation of a word giving an utterly different meaning to the same word without changing the lexical structure. However, the current Arabic version is composed without diacritics, causing a confusion both unspoken and verbal to many, which provides analysis contradiction for the same word. For example, the majority of the Arabic language used in press and media, handwritten or digitized, is undiacritical (Alkharashi [2009]). For native Arabic speakers, this is understandable, but not for a computer. Since different diacritics indicate multiple interpretations, the simplification achieved by avoiding certain diacritics resulted in structural and lexical ambiguity (Benajiba et al. [2007]). Context-awareness and a thorough understanding of the vocabulary are the best ways to overcome these ambiguities. For example (قطر) can be assiosiated with the country Qatar (LOC) if pronounced as qatar, or can refer to word radius (a trigger for measuring expressions NE) if pronounced as qutr.

- Multiple Named Entities: The uncertainty between two or more named entities is a common mistake in many languages. Consider the following sentence as an example, (محمد بدوي أكل الت)

(Mohamed Badawi ate the apple). Since (محمد بدوي) Mohamed Badawi can be identified as both person's name and a location, confusion arises when the same word is labeled as two separate NE forms. This challenge can be overcome by using heuristic methods such as cross-recognizing NEs. Shaalan and Raza [2009] suggested a heuristic strategy using heuristic rules in order to choose one type of NE over another. Benajiba and Rosso [2008] suggested a strategy that supports the type for which the classifier is most precise.

- Inconsistency in Writing Styles: Arabic has a high degree of transcriptive ambiguity: a NE can be transcribed in several forms (Cavalli-Sforza and Zitouni [2007]). This diversity is due to Arabic authors' variations and inconsistent translation schemes. This inconsistency is crucial, as it results in multiple variations/versions of the same term that are pronounced out differently but still match the corresponding term with the same context (Halpern et al. [2009]). For instance, various versions are created when Arabizing (which is translating foreign language into Arabic handwritten words) a NE such as the city of Washington, which can be written as (واشنطن، وشنجتون، وشنجطن، وشنغتن، واشنجتن). One explanation for this is that there are more speaking patterns in the Arabic language than in a foreign language like Europe, which can vaguely or incorrectly result in a named entity with more than one variation. One suggestion is to maintain the possibility of connecting each variation of the name variants and link them all together (Cavalli-Sforza and Zitouni [2007]).

- Insufficient Resources: When trying to implement and evaluate the output of Arabic NER systems, some great resources include corpora (large sets of labeled records) and gazetteers (predefined lists of typed NEs). Fair distribution and the usage of non-sparse representative NEs are crucial in order for these linguistic resources to be efficient. However, sad to say, there is always insufficient potential or coverage of available Arabic tools for NER study. Furthermore, creating or licensing these valuable Arabic NER data has a very high cost. This is why scientists often depend on their own Corpus, which involves annotating and verifying manually by specialists. Unfortunately, only a few of these corpora can be found online for free and public use for academic educational purposes (Abouenour et al. [2010]).

2.3 Named Entity Recognition

The role of Named Entity Recognition (NER) is to recognize and define real names found within unstructured texts into fixed categories like Person, Location and Organization. The term "Named Entity" (NE) refers as well to some time and numerical entities as it was initially implemented as a data extraction task and was considered significantly valuable by the scientific world (Shaalan [2014]). To design a NER model, three techniques are presented: rule-based approach which is based on custom-made grammatical guidelines, machine learning (ML) based approach that uses a collection of variables obtained via NEs annotated datasets, and finally hybrid approach that utilize the previous two approaches to enhance the system efficiency (Oudah and Shaalan [2012]).

2.3.1 Applications

In this section, we will be discussing five main applications where NER is highly used in NLP domain, with pointing out it's various role that varies from one task to another (Oudah and Shaalan [2012]).

Information Retrieval (IR) The process of finding and retrieving related records from a database based on the input query. IR utilizes NER in two main methods:

1. identifying the NEs included inside the input query
2. identifying the NEs inside the searched documents in order to retrieve the related files while keeping NEs confidentiality and their relation to the query in mind.

For example, "Apple" can be identified as a type of fruit or an organization name; figuring out the proper classification may aid in the extraction of the related records (Benajiba et al. [2009]).

Machine Translation (MT) is converting textual data from one language into another chosen language. High-quality NE translation plays a major role in optimizing MT system efficiency. Hence it requires additional consideration and attention in order to be properly converted, especially with the applications that support multi-language. In case of translating from Arabic language into English, some words might get interpreted incorrectly due to it's double meaning, for example the word (كريم) can be translated into the adjective "generous" or into a name "Karim", distinguishing between these two words is crucial in the MT performance (Babych and Hartley [2003]).

Question Answering (QA) QA role is quite relevant to IR tasks; however, the findings are more advanced and complex. QA application provides detailed and accurate responses given some questions by the user. NER starts with a query analysis process to classify the NEs contained inside it, which will assist in finding related records and developing the answers from these acquired documents. Usually, the answer to most questions can be found in a NER system, in case of "who?" an answer is a Person or Organization, "where?" entail a location, and "when" can be answered by expressions of time (Hamadene et al. [2011]).

Text Clustering (TC) Clustering the output from a query will take advantage of the NER system by ordering the results into a group of clusters depending on the number of NE contained in each group. This strengthens the clustering methodology of defined properties and the method of evaluating the structure of every cluster. One example of this application is using time and location NEs to help find the related document or information of a particular event (Benajiba et al. [2009]).

Navigation Systems This feature have become extremely important in the last few years, as it allows a better, more easy navigation experience on digital maps. This is achieved by offering constant updates of traffic alerts, local location details, and related online services. The idea is simple: an extensive database containing all NEs referred here to points of interest and their geographical coordinates. Which act as suggestion points to tourists and travelers on vacation that can help them the nearest or a specific place or services like hospitals, parking lots, stores, food, monuments, and many other things (Kim et al. [2012]).

2.3.2 Named Entity Tag Set

The process of classifying each named entity to it's corresponding label is often referred to as tagging. In the case of a continuous string of words having the same label, it is called a single multiword Named Entity, and the label given to each NE can vary from model to model depending on the user's need (Mohit et al. [2012]).

There are three most common tagging techniques for text annotation will be discussed and explained. These techniques can be used as a standard to annotate textual information and system outputs (Shaalan [2014]).

The 6th Message Understanding Conference (MUC-6) This method can be regraded as the starting point for all NER tasks. By simply classifying named entities into three main tag groups and assigning a type to corresponding tag element (Shaalan [2014]):

1. ENAMEX, contains persons names, location and organization.
2. NUMEX, defines all numerical entities like money and percentage.
3. TIMEX, specifically for temporal expressions like time and date.

For example, a NER model using the MUC style on the sentence
"Thomas bought 150 shares of Apple Corp in 2020" is shown in table 2.3.2

MUC tagging		
TERM	TAG ELEMENT	TYPE
Thomas	</ENAMEX>	⟨ENAMEX TYPE=PERSON⟩
150	</NUMEX>	⟨NUMEX TYPE=CARDINAL⟩
Apple Corp	</ENAMEX>	⟨ENAMEX TYPE=ORGANIZATION⟩
2020	</TIMEX>	⟨TIMEX TYPE=DATE⟩

The Conference on Computational Natural Language Learning (CoNLL) In this method, CoNLL uses an inside-outside-beginning (IOB) format to label fragments of text expressing the named entities in a dataset, giving out an output of four categories: person, location, organization, and miscellaneous. Following a word-based classification problem, a tag is given to each term, defining either the word is at the (B) beginning of a particular NE,(I) inside a certain NE, or (O) outside any NE. This annotation technique is applied when NEs are not nested hence do not overlap (Benajiba et al. [2007]). As such, each term appearing in the text should be marked with one of the following tags:

- B-PERS : The Beginning of the name of a PERSon.
- I-PERS : The continuation (Inside) of the name of a PERSon.
- B-LOC : The Beginning of the name of a LOCation.
- I-LOC : The Inside of the name of a LOCation.
- B-ORG : The Beginning of the name of an ORGanization.
- I-ORG : The Inside of the name of an ORGanization.
- B-MISC : The Beginning of the name of an entity which does not belong to any of the previous classes (MISCellaneous).
- I-MISC : The Inside of the name of an entity which does not belong to any of the previous classes.
- O : The word is not a named entity (Other).

This is the most common method used by researches. Table 2.3.2 shows a NER model using the CoNLL stagging scheme for the following sentence:
"Steve jobs, Apple Corp's CEO, has died in 2011"

CoNLL tagging	
Token	Tag
Steve	B-PERS
Jobs	I-PERS
Apple	B-ORG
CORP	I-ORG
CEO	O
has	O
died	O
in	O
2011	B-LOC

2.3.3 Named Entity Recognition Approaches

NER is often used to fulfill two primary objectives: identifying NEs and categorizing those NEs into predefined types. To achieve those two objectives, three techniques are used: the rule-based, the ML-based and the hybrid approaches.

Rule-based NER

NER systems that are rule-based rely on local handcrafted linguistic rules to detect NEs inside texts via linguistic and contextual information, as well as indications. These systems make use of gazetteers/dictionaries to provide further information about the rules and guidelines. Typically, the rules are implemented using regular expressions or finite-state transducers. The major benefit of rule-based NER systems is that they are built around a strong foundation of linguistic knowledge. However, maintaining rule-based systems is not a simple task, since skilled linguists must be accessible to make necessary modifications. As a result, any modification to such systems would be difficult and time-consuming (Shaalan and Raza [2007]).

Machine learning-based NER

Machine learning-based NER systems make use of machine learning techniques to learn NE tagging recommendations from labeled texts. Supervised Learning is the most popular technique associated with this approach, which portrays the NER challenge as a classification task that require large labeled datasets for training and testing. Conditional Random Fields (CRF) is considered one of the most frequently used SL methods for NER and used in several state-of-the-art applications, alongside Hidden Markov Models (HMM), Maximum Entropy (ME), Support Vector Machines (SVM), and Decision Trees as well. The advantage of the Machine learning-based approach is that they are easy to adapt and update without wasting time and energy, as long as enough data is available. Additionally, if we are dealing with open fields, it is preferable to use ML methods since acquiring or deriving linguistic rules would be prohibitively costly in terms of both cost and time. The operational workflow can be directed from the rule-based system to the ML-based system or vise - versa (Nadeau and Sekine [2007]).

Hybrid NER

The Hybrid NER approach is a combination of previous mentioned methods, the Rule-based and ML-based. By using the NEs' rule-based preferences as variables in the ML classifier, the hybrid approach has resulted in significant overall performance (Petasis et al. [2001]).

2.4 Arabic Named Entity Recognition.

While the NER challenge is non-trivial in general, but it is specifically challenging the Arabic language than in the English language. This is mainly due to the inherent linguistic variations between the two languages, precisely the absence of a straightforward indication such as capitalization in Arabic to identify a named object (Benajiba and Rosso [2008]).

2.4.1 Arabic Named Entity Tag Set

As mentioned in section 2.3.2, NER systems was initially introduced and received a lot of interest from the research community. Three major components were identified in the 6th MUC:
ENAMEX: consists of Person, Location and Organisation.
TIMEX: consists of temporal expressions.
NUMEX: consists of numerical expressions.

Additionally, A customised NER system may need additional sub-divisions within one or more of the NER components in order to achieve the system's aims and objectives.

The following section gives an example of previous tag methods and how Arabic can be integrated and highlighting of a new method specifically for the Arabic language

The 6th Message Understanding Conference (MUC-6)

Implementing the following sentence
(خالد إشترى ٠٥١ سهم من شركة سامسونج في ٠٢٠٢)
(Khaled bought 150 shares of Samsung Corp. in 2020)
will give the following output tags in the table below:

MUC tagging		
TERM	TAG ELEMENT	TYPE
خالد	</ENAMEX>	⟨ENAMEX TYPE=PERSON⟩)
150	</NUMEX>	⟨NUMEX TYPE=CARDINAL⟩
شركة سامسونج	</ENAMEX>	⟨ENAMEX TYPE=ORGANIZATION⟩
2020	</TIMEX>	⟨TIMEX TYPE=DATE⟩

The Conference on Computational Natural Language Learning (CoNLL)

For example an Arabic NER model implementing the CoNLL tagging on the sentence
(وقال حسني مبارك خلال مؤتمر صحفي في القاهرة أن مجلس الامن)
(Hosny Mobarak said during a press conference in Cairo that the Security Council)
will give the following output tags in the table below:

CoNLL tagging		
Arabic	English Trans.	Tag
وقال	said	O
حسني	Hosny	B-PERS
مبارك	Mobarak	I-PERS
خلال	during	O
مؤتمر	conference	O
صحفي	press	O
في	in	O
القاهرة	Cairo	B-LOC
أن	that	O
مجلس	Council	B-ORG
الامن	Security	I-ORG

The Automatic Content Extraction (ACE) program

In the context of the ACE project, Arabic tools for IR has been created. Initially, four tag types were defined as a benchmark in the ACE 2003: person, facility(FAC), organization and geographical and political entities(GPE). Two types were added later on in ACE ACE 2004 and 2005: vehicles and weapons (El Kholy and Habash [2010]).

For example, a Arabic NER model using the ACE-style tagging on the sentence (زار الملك حسين لبنان في العام الماضي) (King Hussein visited Lebanon last year) will give the following output tags in the table below:

ACE tagging		
Arabic	English Trans.	Tag
الملك	King	<PER>
حسين	Hussein	</PER>
لبنان	Lebanon	</GPE>

2.4.2 State-of-the-art in Arabic NER

Meanwhile, several techniques have been developed, especially for solving the NER problem in Arabic. The majority of traditional Arabic NER models are rule-based (Shaalan [2014]). Lately, researchers have begun associating this activity with machine learning techniques (Elrazzaz et al. [2017]). To further enhance performance, efforts have been conducted to merge rule-based and learning-based methods into one single frame(Pasha et al.,2014; Abdelali et al., 2016).

Rule-based

Maloney and Niv [1998] were the first to address the Arabic NER challenge, developing the TAGARAB system. A rule-based framework uses a pattern recognition algorithm in conjunction with a morphological tokenizer (MT) to recognize the following entities: Person, Location, Organisation, Time, and Number. Their findings demonstrate that when implementing TAGARAB to a random data source from AL-HAYAT, integrating NE scanner with a MT surpasses the standalone NE detector in terms of precision.

Mesfar [2007a] created a rule-based Arabic module for the NooJ linguistic system to facilitate Arabic text and NER processing. The module comprises three components: a tokenizer, a morphological analyzer, and a NE detector. The NE locator uses a collection of datasets and identifier lists to assist in the development of rules. The system recognizes five types of entities: Person, Location, Organization, Currency, and Time expressions. This method makes use of morphological features to remove unidentified proper nouns, therefore improving the system's overall outcome.

Another study that utilizes a rule-based method for NER is (Shaalan and Raza [2007]). PERA is a grammar-based method that was developed to recognize person names in Arabic scripts accurately. PERA consists of three parts: gazetteers, grammars, and a filtration system. First, a list of full names is given to ensure that matched names are extracted independently of the grammars in the gazetteers part. Following that, the input words is submitted through grammar, which is composed of regular idioms, to locate the remaining Person NEs. Lastly, the detected NEs are filtered using specific linguistic criteria to exclude any invalid NEs. When implemented to the ACE and Treebank Arabic datasets, PERA demonstrated promising results.

NERA system (Shaalan and Raza [2009]) extends (Shaalan and Raza [2007]) previous study. NERA is a rule-based framework that is able to identify up to 10 distinct sort of entities: Person, Location, Organization, Date, Time, ISBN, Price, Measurement, Phone Numbers, and Filenames. The system consists of the three main components found in the PERA system, implemented using the FAST ESP framework, and provides the same functionality for all ten NE kinds. In addition, the authors created their own corpora from various sources to ensure that each NE type has a representative amount of examples.

ML based

Benajiba et al. [2007] were the first to design an Arabic NER system, ANERsys 1.0, based on Maximum Entropy (ME). The researchers created their own datasets: ANERcorp (a corpus of annotated texts) and ANERgazet (i.e. gazetteers). The system exploits lexical, contextual and gazetteers characteristics. The framework is capable of distinguishing between four different categories of entities: Person, Location, Organization, and Miscellaneous. However, ANERsys 1.0 experienced difficulty identifying NEs made of several tokens. As a result, Benajiba et al. [2007] created ANERsys 2.0, which utilises a two-step NER process: 1) identifying the beginning and ending points of each NE, and 2) categorising the identified NEs. Benajiba and Rosso [2008] attempted to enhance performance by using CRF rather than ME and found out that CRF based method produces more accurate findings.

Abdul-Hamid and Darwish [2010] presented a simpler feature set for Arabic NER. They used CRF to classify NEs into three categories: Person, Location, and Organization. The system examines just surface structure (i.e., starting and following character n-grams, term location, word length, word unigram likelihood, preceding and subsequent word n-grams, and character n-gram probability) and ignores all additional characteristics. ANERcorp and ACE2005 datasets were used to assess the system. The findings indicate that the system outperforms (Shaalan and Raza [2007]) CRF-based NER system.

2.5 Arabic NER Linguistic Resources

Generally, Arabic NLP faces a tremendous challenge due to the unavailability and shortage of digital linguistics resources, specifically Arabic NER. Hence, it is necessary to invest in developing these resources, as it will result in several advantages, including renewability, broad coverage, and frequency and disreputable data, along with a method for analyzing and comparing models. In NER, corpora and lexical assets are two of the most frequently used forms of linguistic resources (Shaalan [2014]).

To build a NER corpora, a relatively enormous labeled corpus is needed in which each NE has its own type assigned to it. A well structured corpus with a good NE type allocations is an essential trait in a reliable corpus. A corpus may be genre-independent/domain-specific and include texts written in a single natural language (monolingual corpus), two natural languages (bilingual, parallel, or comparable corpus), or multiple natural languages (multilingual corpus) (a multilingual or cross-lingual corpus). A popular method suggested in (Hegab et al. [2007]), is to obtain NE translating combination using both comparable and parallel corpora.

Parallel corpora, which are sentence-level compatible, are used to identify one corpus based on the tagged knowledge in the other corpus, allowing them to balance and enhance one another. For instance, Samy et al. [2005] approach generates a NE-aligned bilingual corpus based on the simple hypothesis that the same NE can either be translated in a single sentence, provided the two sentences where one is the translation of the other. The solution is very successful, as it includes Arabic, a case-insensitive script, and Spanish, where names and non-names are cast differently.

When NLP studies use freely accessible data sets or corpora, their experimental findings are more effectively comparable.Due to their widespread use in the research community, these resources have become regular datasets or corpora, acting as a standard benchmark data for monitoring performance and evaluating systems based on their annotation capability.

2.5.1 Corpora & Datasets- Benchmarks

Automatic Content Extraction (ACE) Program

The ACE program aims to create extraction technologies that can help the processing of source language data automatically (in the form of raw text and as text derived from Automatic Speech Recognition(ASR) and Optical Character Recognition(OCR)). "Automatic processing, defined at that time, included classification, filtering, and selection based on the language content of the source data, i.e., based on the meaning conveyed by the data.". Therefore, the ACE program necessitated the creation of technology capable of recognizing and characterizing its meaning automatically. These were the program's objectives as they aimed to discover and characterize Entities, Relations, and Events.

The Linguistic Data Consortium(LDC) provided annotation guidelines, corpora, and other language resources to assist the ACE program. In collaboration with the Translingual Information Detection, Extraction, and Summarization(TIDES) program, several of these materials were produced to aid in the assessment of TIDES Extraction.

ACE annotators labeled broadcast transcripts, newswire, and newspaper data in three languages: Arabic, Chinese, and English, in order to generate both training and test data for the evaluation of proper research tasks.

Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), and Event Detection and Characterization (EDC) were the three fundamentals annotations tasks. Entity Linking (LNK) was a fourth annotation activity that consolidated all references to a single entity and all of its properties into a Composite Entity.

However, this developed linguistic resource is not for free; ACE is only available to organization researchers with a paid license agreements. Hence, Small research organizations have difficulty accessing them (Muhammad [2017]).

Example of ACE Arabic corpora that commonly used in general for NLP, specifically classification are:

- ACE 2003 corpus: This includes Broadcast News (BN) and Newswire (NW) genres. The total size is 55.29 KB and the number of NEs is 5,505.

- ACE 2004 corpus: This includes BN and NW from Arabic Tree Bank (ATB) genres. The total size is 154.12 KB and the number of NEs is 11,520.

- ACE 2005 corpus: This includes BN, NW, and Weblogs (WL) genres. The total size is 104.65 KB and the number of NEs is 10,218.

ANERcorp

Aner corpus is one of the publicly available corpora that can be found online. Anercorp was manually annotated only by a single individual to ensure the annotation's consistency. It was developed for Arabic NER tasks and is divided into two sections; training and testing. The corpus contains more than 300 that were collected from news wire and other forms of online sources. Prior to labeling the corpus, the data were subjected to prepossessing. In Arabic, a single word can be written in a variety of ways. To balance these disparities and reduce data sparsity, the data were normalized by merging several word forms into a single form. The corpus follows a standard CONLL format containing more than 150K tokens in the dataset, and 11% of them are NEs (Benajiba et al. [2007]). Each token in the corpus has one of the following annotations: person, location, organization, miscellaneous, or other. The following table summarises the distribution of the Named Entities:

- Person : 39%
- Location : 30.4%
- Organization : 20.6%
- Miscellaneous : 10%

American and Qatari Modeling of Arabic(AQMAR)

The AQMAR corpus is a small corpus that consists of 28 Arabic Wikipedia articles for Arabic NEs that has been manually annotated (Mohit et al. [2012]). AQMAR dataset has a total of 74000 tokens and 2687 sentence and consists of four tag elements:

- PERSON
- LOCATION
- ORGANIZATION
- MISCELLANEOUS

and nine entity class: O, B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC

	Documents	Words	Sentences	Entities
Test	20	52,650	1,976	3,781
DEV	8	21,203	711	2,073

Fine-grained Arabic Named Entity Corpora (FANE)

Alotaibi [2015] developed 4 sets of fine-grained Arabic NE corpora.

1. Automatically- developed:

   These corpora were generated automatically using the Arabic Wikipedia and are published under the Creative Commons Attribution-ShareAlike 3.0 Unported License. These datases shares the same annotation methodology as ACE, however it includes a new tag, PRODUCT, which covers Books, Movies, Sound, Hardware, Software, Food, Drugs and Other. Additionally, the corpora split the PERSON tag in 10 feature labels to offer a broader coverage, (i.e. Politician, Athlete, Businessperson, Artist, Scientist, Police, Religious,Engineer and Group). This new structure of labelling can be easily integrated with the CONLL and ACE format.

- WikiFANE_Whole: The corpus was created by extracting all sentences from Arabic Wikipedia articles.

- WikiFANE_Selective: The corpus was created using sentences with at least one NE sentence.

2. Manually gold-standard:

The purpose of this corpora was to coduct an in-depth experiment for fine-grained Arabic named entities. Hence they manually compiled Arabic gold-standard fine-grained NE corpora focusing using two distinct genres. This establishes a vital standard or benchmark for assessment and comparison with the corpus generated automatically.

- WikiFANE_Gold: This corpus is created from Arabic Wikipedia. The articles were chosen using a random heuristic, which involved picking articles that discussed a certain topic while keeping a reasonable amount of dispersion throughout the classes.

- NewsFANE_Gold: This corpus is created based on newswire and uses the same textual data as ANERcorp. The whole corpus was re-annotated to the fine-grained level.

Dataset	Data Used	Size	Named Entities	Publicly Available
ACE 2003	Broadcast News (BN) and Newswire (NW)	55.29 KB	5,505	No
ACE 2004	BN and NW from Arabic Tree Bank (ATB)	154.12 KB	11,520	No
ACE 2005	BN, NW, and Weblogs (WL)	104.65 KB	10,218	No
ANERcorp	Newswire (NW)	174.76 KB	12,989	Yes
AQMAR	Arabic Wikipedia Articles	912 KB	5,854	Yes
FANE	Arabic Wikipedia Articles	25,8 MB	2M	Yes

3 Related Work

3.1 Bias in Natural Language Processing

Bolukbasi et al. [2016] were the first to set the ground for most of the upcoming research. The use of Machine Learning in terms of word embedding is a danger that can lead to the accidental amplification of biases in data. In this paper, they trained Word Embeddings on Google News articles, resulting in gender stereotypes to a distressing extent. The primary contribution was to demonstrate that embeddings were able to accurately capture the connection between words in order to resolve similarities such as man king, woman queen. However, they found out some similar analogies were biassed, such as relating man to doctor and woman as a nurse, whereas assigning the term doctor to woman would be more appropriate. Following the same approach, they acquired a list of stereotyped terms for each gender, demonstrating that this was not a unique example.

Caliskan et al. [2017] suggested a new technique to measure bias using the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). WEAT uses the cosine between the vectors of two words to calculate their similarity. To achieve this, they reproduced a set of common human biases as identified by the Implicit Association Test and other well-known psychological research using the GloVe model (Global Vectors for word representation) - trained on a text dataset extracted from the web. The findings show consequences, not for artificial intelligence only, but psychology, sociology, and human ethics, because they suggest that simply being exposed to daily language might account for the biases duplicated here.

3.1.1 Bias in NER

Mehrabi et al. [2020] investigated bias in multiple state-of-the-art NER models, particularly gender bias and the capability of a model to detect men and women as PERSON entities. First, they collected historical data containing baby names per year from 1880 to 2018(139 years) from the US census data website. Additionally, they introduced a benchmark containing nine templates, with the first template having only the collected names to gain information on the distribution of the training data, following in the eight templates, the name, and a human-like activity. The NER models used in this experiment were: Flair, CoreNLP, and Spacy with small, medium, and large models. Afterward, they tested the NER models against their newly developed dataset and evaluated the performance, through six sets of error calculations, of every template for males and females by comparing yearly models' outcomes. One of their findings, that female names were being wrongly classified more than male names, for instance the name "Charlotte", which was one of the popular female names in 2018 was mainly classified as a Location more than Person. In order to try and debias their findings, two methods were proposed, first by updating the model versions and recalculating their outputs; however, this showed that upgrading to the model's latest version might result in an increase of model unfairness. Secondly, by analyzing the dataset implemented in each NER model, they found out that the bias exists in the data itself, which can directly impact the model performance.

Using Artificial corpora consisting of 123 names, Mishra et al. [2020] analyze the bias in several NER models in the English language throughout different demographic groups, precisely the ethnic stereotype. Their demographic structure falls under four groups; Black, White, Hispanic, and Muslim. Then it has categorized either male or female. The NER algorithms used in this research were; spaCy, Stanford CoreNLP, and BiLSTM-CRF architecture, with applied GloVe, CNET, and ELMo techniques. Their experiments indicate that models are more accurate in detecting White names than the other categories such as Black names. Additionally, they illustrate that debiased embeddings do not assist in solving the fairness issue in name detection. Furthermore, they demonstrate that character-based techniques like ELMo, have the least biased outcome among the other NER models; however, they are still unable to completely eliminate unfairness.

3.2 Mitigating Bias

It has been established that usually, mitigating bias in NER applications and similar NLP tasks may be accomplished during three stages (Garrido-Muñoz et al. [2021]):

- **Before:** Mainly this involves altering or enriching the data source in order to eliminate the bias or balance the data used in the training model.
- **During/Train:** Modifying the model training procedure or fine-tuning it.
- **After:** Typically, altering the vector space model following the training phase

This section will discuss previous work on mitigating bias in NLP using which technique and stage.

Following on Bolukbasi et al. [2016] work,to eliminate the bias that they have found, they recommended determining the direction of the gender vector subspace and adjusting the vector to make occupational words gender-neutral. Their algorithms can reduce gender bias in embeddings significantly while preserving their useful properties. These findings can be used in applications that are designed to minimize gender bias.

Zhao et al. [2018] research was conducted on gender bias in English language, using GloVe model. The data was collected from OntoNotes 5.0 and Occupation Data (BLS) and managed to create their own new benchmark WinoBias. They used prediction accuracy to evaluate their model performance, additionally they tried Data Augmentation and Vector Space Manipulation techiniques for debiasing in the after stage.

Another research was conducted by Manzini et al. [2019], where they used online text across different domains from reddit platform to tackle ethnicty, and religion bias in English language. They used Word2Vec to run their experiment and PCA, WEAT, MAC, Clustering as evaluation methods. They debiased their model through Vector Space Manipulation in After stage, and they focused on POS tagging, POS chunking, NER tasks.

3.3 Bias in Arabic NLP

To the best of our knowledge, there has been minimal research conducted directly to measure Bias in the Arabic language. Below, we summarize scattered attempts published in this field:

Habash et al. [2019] introdcued a method for enhancing biased single-output gender-neutral NLP algorithm with gender-specific alternative re-inflections. The focus of the paper was machine translation(MT) from English, a neutral gendered language, to Arabic, a highly gendered language and only the first-person expression. An example for this bias is translating "I am a doctor/ I am a nurse", which results into " Ana Doktor(masculine)" and " Ana Momareda(feminine)" which is unacceptable for female doctors and male nurses. Thus, they develop an independent system wrapper that is aware of the first-person gender, giving options to the user of which gender they would like to select. For example, when the user translates " I am a doctor" it gives out two selections " / Ana Doktor / ana Doktora". This is being achieved by firstly identify the gender, then reflect it. Additionally, they produced the first parallel gender corpora for training and evaluating first-person singular gender identification and re-inflection in Arabic.

4 Methods

In this thesis, we propose a methodology to quantify, analyze and de-bias existing textual biases included in Arabic NLP resources. Our proposed evaluation approach relies on synthetically generated data to quantify the bias in existing off-the-shelf and popular NLP libraries with embedded Arabic NER models based on various architectures trained on standard datasets. This section describes the included experiments and gives further details on the models used in our analysis and the evaluation procedures for each task.

4.1 Data

In order to evaluate the bias in NER among different categories, a corpus of phrases is required, in which the identified entity is similar to each other or belongs to the same category. This can be obtained by creating sentence templates that provide placeholders that can be replaced with different names. The following sections describe the procedure for creating named entity corpora that consists of templated-sentences that begins with an Arabic names followed by a noun that represents a human-like activity. The dataset created is inspired by Bolukbasi et al. [2016], however the details of the names collection is different as there were no available name censuses for Arabic names, and the pattern are also different to reflected biases more apparent in the middle-eastern culture and to take into consideration the inherent distinctions in the Arabic language. Our data set is concerned with the *person* entity only and does not include any other entities. More specifically, the dataset relies on the first person named entities. The ambiguous impact produced by the sentence's grammatical syntax is removed by applying a different name to the exact sentence.

Names The names collection includes names representing five distinct demographic classes based on gender and religion. These classes are divided into: Female, Male, Unisex, Coptic and Muslim. Table 1 has a complete list of names along with their demographic class.

All the names included in our evaluations are compiled through a two-phase collection scheme. The first step follows the guidelines of any dataset collection, it involves two native-speaking Arabic Annotators that constructed a list of popular Egyptian names through this website: `"https://www.behindthename.com"`. The compiled list include all name categories that this research focused on. The original list after website curation included a total of 300 first name. The second step involved classifying the names collected into specific categories, namely; male, female, unisex, Coptic and Muslim. To achieve the most accurate classification as possible, a questionnaire was created where users were to choose one class y only for each name displayed where they think the name best belongs to. As the original list of names was too long, and since we aimed at making the questionnaire time limited to 5 minutes, two version of the questionnaire were created (each with 80 name). Next both were shared with 40 Egyptians, responses were gathered and analyzed and finally names with low users agreement were discarded from our evaluations. The final list is shown in table 6.2 that can be found in the appendix, it contains a total of 175 names divided into 3 subsets:

- 70 female and male first names
- 35 unisex first name
- 31 Muslim and Coptic first names

I recognize that my research is constrained by the availability of names from various classes, and I understand that individuals will not always identify their names with the associated demographics group. Additionally, I do not recommend using this name list to estimate or predict any demographic characteristics about an individual, as demographic characteristics are private identifiers, and this approach is subject to mistake when used at the individual level.

Templates The goal was to compile a list of template that covers a variety of existing biases. For that purpose we chose 4 main template categories.

1. **BASIC** we start with a basic template as this provides no context to the NER models, it simply contains a name. With this template, it should show the ability of NER architecturs to detect different name classes.

2. **OCCUPATION** A total of 6 distinct professions. Each phrase is a factual statement and lacks the ability to express either positive or negative impression by itself.My selection of 6 occupations was motivated by the desire to represent a range of gender distribution features and career kinds. I selected 3 professions that are slightly male dominated *doctor, student, mechanic,* and also add slightly female dominated *nurse, teacher, secretary.*This dataset is genderly equal in size containing 70 phrase for each gender and a total of 210 per gender.

3. **NEUTRAL** This category mainly focuses on the religious bias where names from Muslim/Coptic names are followed by a statement that does not reveal any religious identity such as (*in Egypt)* . This set constructed of 31 phrases for each religious name.

4. **STATEMENT** This category mainly focuses on the unisex names that can be used to reference both males and females in Arabic. The list is duplicated to include the same verbs but conjugated twice to match males and females mentions. The list contains 70 phrases for each gender.

4.2 Experimental setup

4.2.1 Hardware details

All the following experiments were run on my laptop DELL XPS with Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz. Google Colab was employed to run the proposed experiments and validate the results using NVidia Tesla. K80 GPU.

4.2.2 Models & techniques

In order to asses the presence of bias in NER systems, We investigated five off-the-shelf named entity models used in industry and academia. We provide next details of their architecture, sources and usage scenarios of each.

STANZA

STANZA is a publicly available python NLP toolkit developed by Stanford NLP Group in 2020 to process more than 66 human languages. Stanza is a complete neural network workflow specializing in comprehensive text analysis covering the following features; tokenization, lemmatization, multi-word token (MWT) expansion, part-of-speech (POS), morphological features tagging, dependency parsing, and named entity recognition. For training the model, 112 datasets were used, including the Universal Dependencies treebanks and additional multilingual corpora, giving a consistent performance across all languages. Furthermore, more capabilities, such as coreference resolution and relation extraction, are achieved using the python API against the commonly known Java Stanford CoreNLP package.

For the Arabic language, STANZA uses the AQMAR corpus giving out four types of entities (PER, LOC, ORG, and MISC). A contextualized string representation-based sequence tagger is used. They start with a single language model trained to simulate how a forward and a backward character level long short term memory (LSTM) would communicate. Next, they include word embeddings at word positions to imitate how characters and words connect at each position, then inject the results into a standard Bi-directional long short-term memory (Bi-LSTM) sequence tagger with a conditional random field (CRF)-based decoder (Qi et al. [2020]).

The model uses python 3.6 version and documentation can be found in the following link :
`https://stanfordnlp.github.io/stanza/index.html`

POLYGLOT

Al-Rfou et al. [2015] developed the POLYGLOT tool using a different approach by exploiting word embeddings techniques [i.e., Semantic and grammatical aspects are coded using vectorial representations that were created by pretraining from a substantial quantity of text] as the main functions in a primary neural network-based word-level classifier, for example, a single hidden layer model that uses word embeddings to provide a unique classification for each word within a given range of text-centered around each word. Polyglot covers 40 languages, training its models using datasets automatically collected from Wikipedia and outputting three types of entities (PER, LOC, and ORG). The authors analyzed phrases from Wikipedia articles in several languages for creating the training datasets, taking into account the relevant hyperlink structure. For example, if a link within a Wikipedia phrase is connected to an article classified as an entity by Freebase, the anchor text was used as a positive training example for that entity type. Additionally, due to Wikipedia's style constraints that prevent all entity references from being connected, the authors used oversampling and surface word matching to boost model training.

The following link contains implementation examples and documentation on Named entity recognition Polyglot library:
`https://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html`

Computational Approaches to Modeling Language Lab (CAMeL Tools)

In 2020, New York University's research in Abu Dhabi developed the first library that offers a toolkit for Universal Arabic NLP in terms of features. The library is a collection of open-source tools developed in Python for processing and analyzing Arabic text that supports both MSA and Arabic dialects. It now offers APIs and command-line interfaces(CLIs) for preprocessing, Disambiguator, Tagger, Tokenizers, morphological modeling, dialect identification, named entity Recognization (NER), and sentiment analysis (Obeid et al. [2020]).

For the NER application, CAMeL Tools uses ANERcorp as their training set, using HuggingFace's Transformers to fine-tune AraBERT for labeling named entities in the commonly used IOB (inside, outside, beginning) NER tagging format. They output four entity types (PER, LOC, ORG, and MISC)

CAMeL Tools installation and setup can be found in:
`https://github.com/CAMeL-Lab/camel_tools`

FLAIR

FLAIR is an open-source framework developed using BiLSTM-CRF sequence labeling referred to as contextual string embeddings. FLAIR encodes phrases as a series of letters and performs auto-regressive training on them. By combining forward and backward character language models, FLAIR can represent sequenced bidirectionally. An important property is that by representing context as characters, FLAIR models may generate many embeddings for the same word depending on its surroundings and deal with uncommon and misspelled words (Akbik et al. [2018]).

According to author, their model outperformed earlier models on English and German NER task, enabling them to publish new state-of-art F1 scores for the CONNLL shared challenge.

For Arabic NER task, ANERcorp was used as well for training the model, outputing four entity types (PER, LOC, ORG, MISC) A list of all word embeddings techniques can be found here:
`https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_4_ELMO_BERT_FLAIR_EMBEDDING.md`

Flair Embeddings *Contextual string embeddings* are extremely powerful embeddings that collect additional syntactic-semantic information than ordinary word embeddings. The important difference is that they are taught with no explicit concept of words and hence basically model words as sequences of letters. Furthermore, they are contextualised by the information that surrounds them, which means that the same word may have many embeddings depending on its context.

Flair Embeddings github repository:
`https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md`

FastText Embeddings *FastText Embeddings* may provide vectors for out of vocabulary (oov) terms, through exploiting sub-word information.

Flair FastText Embeddings github repository:
`https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FASTTEXT_EMBEDDINGS.md`

Model	Dataset	Entities	Technique	Year
POLYGLOT	Wikipedia, Freebase	PERS, lOC, ORG	language agnostic techniques	2014
FLAIR	ANERcorp	PERS, LOC, ORG, MISC	BiLSTM-CRF character-based model	2018
STANZA	AQMAR corpus	PERS, LOC, ORG, MISC	Rule based model	2020
CAMEL	ANERcorp	PERS, LOC, ORG, MISC	Rule based model	2020

4.3 Evaluation

A major dilemma is determining how to compare different models that use word embeddings, rule-based appraoch, or machine learning approach. While normally NER task is evaluated through the traditional metrics such as *Precision, Recall and Accuracy*. The variations in performance across models might be ascribed to a variety of factors, including improved pre-trained word embeddings, a different architecture, a different target or goal, or the normalization layer.

To address the bias issue in the existing NER models, we resort to different evaluation metrics. For each of the three experiments, I ran each NER model on my generated benchmark dataset and assessed the performance of each template across female and male genders, as well as comparing the outcomes of these models between genders. We rely on two main factors when comparing results across models:

1. Person Count: First evaluation technique is basically just detecting the person name. If the name is tagged as a PERSON, the system simply add it to the count

2. Non-tagged Error" The second evaluation metric is error based technique, similar to recall calculations. This sort of error is used to determine if a name is tagged as non-PERSON or not tagged at all. To put it another way, any name that isn't labelled as a PERSON is deemed as an error (Mehrabi et al. [2020]).

4.4 Measuring Bias

We conduct 4 main experiments to measure and quantify the bias in Arabic textual resources. Our aim is not only detect biases in text but to also compare which model and architectures are more robust against others in capturing such biases and by-passing them. Each experiment is concerned with a certain type of bias.

4.4.1 Experiment 1:

In the first experiment, we investigate the capabilities of the NER models to detect masculine and feminine proper nouns. For this experiment we use the first template , the BASIC one **<NAME>** which contains only the name and is considered context-independent. As mentioned earlier, the name list contains popular and commonly used Arabic names for both genders. Assuming fairness among genders, we expect that for the curated name lists, the detection of male and female names should be the same or slightly comparable.

4.4.2 Experiment 2:

The second experiment explores the stereotypes and gender roles in our society, more specifically in occupations. It involves the six templates of the second category listed in the table 4.4.2 below. Each template begins with a name from the collected male/female lists and ending with a human-life occupation.

	Template	Sentence
1	<NAME>	is a doctor
2	<NAME>	is a nurse
3	<NAME>	is a student
4	<NAME>	is a teacher
5	<NAME>	is a mechanic
6	<NAME>	is a secretary

For each template, the models were run on female names only first and then on male names.

1. Opposite to the English language, Arabic is a gendered language.

2. With each gendered name adaptation in the template, the occupation word ending changes, defining whether it is suitable for masculine or feminine form.

3. A masculine profession can be switched to feminine by adding (ة) or *T̃eh Marbouta* to the end of word, as shown in the following table.

English	Arabic Translation
Ahmed is a doctor	أحمد دكتور
Mariam is a doctor	مريم دكتورة
Ahmed is a nurse	أحمد ممرض
Mariam is a nurse	مريم ممرضة
Ahmed is a student	أحمد طالب
Mariam is a student	مريم طالبة
Ahmed is a teacher	أحمد مدرس
Mariam is a teacher	مريم مدرسة
Ahmed is a mechanic	أحمد مكانيكي
Mariam is a mechanic	مريم ميكانيكية
Ahmed is a secretary	أحمد سكرتير
Mariam is a secretary	مريم سكرتيرة

4.4.3 Experiment 3:

In this experiment, once again we aim to evaluate the models' performances against gender bias through the unisex name list. This is a two-fold evaluation achieved by evaluating the models on: (1) Only on the first name (35 unisex names) and (2) the name followed by a human-like activity through a verb. There are two reasons why using an activity verb in this case will be the most suitable :

1. In the Arabic language, the verb can define a person gender by only changing the first letter of the verb while maintaining the same word base.

2. Masculine verbs usually starts with (ي) or *Ỹeh* and feminine verbs starts with (ت) or *T̃eh*

3. To mitigate any gender bias that might occur, for instance like occupation.

Sample examples of the data used in this experiments are shown in table 4.4.3

Form	English	Arabic Translation
Masculine	Nour plays	نور يلعب
Feminine	Nour plays	نور تلعب
Masculine	Esmat eats	نور يأكل
Feminine	Esmat eats	نور تأكل

4.4.4 Experiment 4:

In the final experiment, we measure the religion bias by investigating the final template category that includes Muslim and Coptic first names. The models were tested on:

1. First name only (31 Muslim and Coptic names)

2. The existing name following with a gender-less, non biased statement.

As we only want to evaluate the religion biases in this experiment, we chose that the pattern following the name is gender independent so we minimize the gender bias effect and not accumulate it on top of the religious bias.

Religion	English	Arabic Translation
Coptic	Michael is in Egypt	مايكل في مصر
Muslim	Mohammed is in Egypt	محمد في مصر

4.5 Mitigating Bias

The next phase of our study involves eliminating and reducing biases effects found in any of the previous experiments. As explained in chapter 2.5.1, various techniques for mitigating bias in NLP have been suggested and developed. They mainly focus on two main approaches: debasing the text corpora or debasing the model algorithms. They are also referred to as retraining and inference. Retraining techniques involve retraining the model, whereas inference approaches decrease bias without the need for changing the original training set. Retraining approaches are frequently used to tackle gender bias in its early phases or even at its origin. Nevertheless, retraining the model from the beginning on a new source of data might be time and resource consuming. On the other hand, inference approaches do not necessitate retraining of models; rather, they adjust preexisting models to alter their outcome, therefore giving a debiasing effect during testing (Sun et al. [2019]).

A popular debiasing technique is **Vector space manipulation**. This approach is also known as "word embedding debiasing" or "hard debiasing." Word embeddings are vector representations of words. This technique was first mentioned by Bolukbasi et al. [2016], where he suggests finding the gender's vector representation to correct for its divergence and equalize some terms with regard to the neutral gender. This suggestion has been improved significantly in order to properly capture the bias while avoiding harm to the model.

On the other hand **Data augmentation** handles biases that exist in data.

In many tasks, a data collection has an abnormally high number of references to a single attribute. To address this,Zhao et al. [2020] recommended creating an augmented data set that was similar to the original data set but biased towards the other lower class. Then training on the combination of the original and new biased data. Hence, try to represent that specific attribute in a less biased manner. The overall goal would be to balance the data on the bias-level. In our work, we followed the data augmentation technique as we find the Word embeddings debiasing is not the best methods when it comes to the Arabic data and models:

1. Fairness is a tremendously complicated topic that no algorithm can define on its own. According to research, training an algorithm to perform similarly well on all population subsets does not assure fairness and actually cripples the model's learning ability.

2. Applying more objective functions can reduce the accuracy of the model, resulting in a trade-off. Rather than that, it is preferable to simplify the method and guarantee that the data is balanced, therefore increasing model performance and eliminating the trade-off.

3. It is unrealistic to assume the model will perform effectively in situations when it has not seen many examples. Improving the diversity of the data is the greatest approach to assure good results.

4. Using engineering approaches to de-bias a model is both complicated and tricky. It is far less expensive and time-consuming to train the models on fair data in the first place, allowing the researchers to focus on development.

While some argue that the data is merely a subset of the bias issue. However, it is fundamental, impacting all that follows. That is why we believe it is crucial and would pave the way to better unbiased performance in the Arabic NER model and is plays a key role in the problem solution. This was also motivated by the fact that the data have a significant impact on fairness limitations if it includes any biases. The road to debiasing start by measuring how biased is the data.So first we analyse the dataset used in the training of the NER model to evaluate whether they could have a bias towards a particular group leading to the discriminatory behavior observed from the results of the previous sections.

4.5.1 Experiment 1: Analyzing ANERcorp

We mainly focus on the ANERcorp, full details of the dataset are given in chapter , blow we explore in depth the class distribution of the dataset to unveil the existing biases. In the this experiment, we manually annotated the ANERcorp dataset in a straightforward method, by identifying the person's names and adding a label to each name, "F" for female and "M" for male. Annotation was only concerned with the person entity that exists in the dataset and covered both B-PER and I-PER which refers to the first name and the second name respectively. Over all, redadd a number of the total number of persons tag in anercorp were re-annotated and a gender label was added. As shown in figure redinsert the graph here ,the corpus shows extreme bias towards the male gender resulting in 96% names of men and less than 4% for females. The corpus contains a total of 6186 male mentions and only 248 female mentions. That means for every female mention; there is 24 male mention. Although gender-specific bias was unavoidable, male dominance is greater than predicted. Previously a similar analysis was conducted on the English NER dataset (Lennon [2020]), it was found that for every mention, there exists almost 5 male mention. This entails that bias found in Arabic resources is more apparent than its counter English resources.

4.5.2 Experiment 2: Data Augmentation

After the annotation step was completed and realizing that the training dataset has a severe gender imbalance, a data augmentation step is required. In order to overcome this bias through the data augmentation approach, more sentences with female mentions were needed. While many publications suggest generation of synthetic data to counter-balance the included sentence, in other word including the same sentence with changing the male names to female names and applying corresponding changes. However, we beleive that this does not reflect real-world scenarios so we researched other options. We resorted to downloading two publicly available corpora **AQMAR** and **FANE**. Again both data set were re-annotated and for each person's name, an extra label(F or M) was appended. The next step involved extracting all the sentences that contain any female mentions, leaving male mentions sentences only out. The end result of this step was two extra additional training corpses to mitigate the male dominating biased ANERcorp. It is worth mentioning that Female names are only available as first person name only. They are mostly accompanied by a male name as the last name which again contribute to the original male bias but it does balance it nonetheless. To better evalaute the NER model we divide our analysis according to the dataset imbalance :

1. **ANER ORIGINAL:** The original corpus with 3990 training size. It contains a total of 6434 total mentions divided into 3601 first names (B-PERS tag) and 2833 secondary name (I-PERS tag).
2. **Extracted sub-dataset:** The extracted dataset contains 8456 mentions, broken down to 2152 female and 6304 male mentions. The male count still exceeds the female count even in this dataset due to the existence of the last name as mentioned.
3. **ANER BALANCED:** Due to the apparent gender inequality and the low proportion of female sentences in the original corpus, I proposed to try and balance the corpus by adding the extracted sub-dataset to the original dataset. This union does not entirely remove the bias; however, it improves the male to female ratio. Hence, giving a chance for the model to predict a better output.
4. **BALANCED 35:** Given the shortage in the balance between females and males in the previously mentioned dataset, we tried to extract a maller corpus with a better ratio. This dataset consists of all the sentences that contain female mentions from ANER BALANCED, resulting in a smaller corpus than the original but with the most minor ratio between the two genders.

4.5.3 Experiment 3: Re-measusring Bias

After evaluating the severity of gender bias in each Off-the-shelf models, we choose one tool to re-evaluate the bias after the mitigation step. In our case it was the FLAIR NER model to evaluate whether they could have a bias towards a particular group leading to the discriminatory behavior observed from the results of the previous sections. Our choice for the FLAIR model to conduct our experiments was mainly because of the availability of the documentation, the straight forwardness of the model re-training, as is it an important part of the mitigation, both in terms of the implementation and the available resources. Moreover, their NER architecture model is compatible with any corpus, as long it has CoNLL tagging format and offers a various range of different types of embeddings techniques.

Flair Embeddings is based on contextualized string embeddings which are currently considered state-of-the-art models in many NLP tasks, using the ANERcorp for their NER model training. In the previous experiments, Flair Embeddings seems to show a higher performance as well as greater bias than the other models.

Lastly,we repeat the experiments mentioned in section 4.4 after retraining the NER model on the newly compiled datasets. In order to evaluate the model performance after the debias technique, we compare the model's output on the three corpora: **ANER ORIGINAL, ANER BALANCED and BALANCED 35**

4.6 Case Study: Real Life Data

Furthere more, we conduct one extra evaluation experiment to better validate the newly trained models' performance on external data other than the NER datasets. We attempt to test the model ability to detect person name entity in real life data sentences. For that purpose, we mine a popular Egyptian newspaper website `https://www.youm7.com/`. This website contains daily articles about different topics varying from politics and economics to sports and media news. Next we collect short sentences found in various existing articles, we decided to extract the data related to the three major topics, Politics, Media and Sports.

To maintain a robust evaluations, obtain accurate results and try to avoid the noise found in data, some constraints are taken into consideration when collecting the articles:

1. The sentences are kept gender-specific and contains one full name only
2. For each topic, 15 sentences are collected for each gender
3. To ensure fairness in evaluation, each sentence was mitigated to the opposite sex and added to the corpus

This results in a total of 120 sentences, 60 per each topic. Example of sentences collected in each topic. Finally, we preform the NER task using the newly enhanced Flair NER models.

5
Results Discussion

In this chapter we follow the methodology approaches explained previously to quantify the gender bias in existing models and also aim at mitigating the bias through data augmentation. The chapter is divided three main parts where results from the current ANER dataset are reported, results from the augmented datasets are reported and finally the case study to show the efficiency of our proposed mitigation technique.

5.1 Dataset

The final dataset collected was a combination of Arabic names with, optionallym a human-like activity. The final generated dataset is divided into 3 subsets to serve each of the experiments. The first set consists of female/male names which serves experiment 1 and 2 it consist of 210 sentences, 70 names and 8 template for each gender, totalling to 420 sentences and 140 name. The second set is used in the analysis of experiment 2 to detect how gender-less names are perceived by models. This set contains a list of 35 unisex name with one single pattern <dance> resulting in 70 sentences for both gender. The third and last set contains 62 religious names that indicates an Islamic or Coptic religion distributed equally.

5.2 Quantifying Bias in models' prediction

Experiment 1: Gender Bias

In this experiment, we evaluate the ability of all models to detect the *BASIC TEMPLATE* scheme which involves simple a female or a male name. Figure 5.1 shows the name counts for each of the male and female categories.

As illustrated The Flair framework, using flair embeddings, outperformed all model in detecting the PER named entity with 57 and 26 for both male and female names respectively. The STANZA model followed up, with 42 male name captured and 23 female name captures. Both POLYGLOT and FLAIR(using fastext word embeddings) detected less than 20 names for either males or females. Nonetheless, in all models, there is an obvious pattern where male names are detected more than female name. The same results can also be verified in 5.2, that shows the non-tagged error values for all models are lower for the male category. It is also important to note that the models' evaluation should not only be based on the recognition of names in total, i.e counts. But also on the ability of the models to detect both gender names in a balanced way. From this perspective, the FLAIR(using fastext word embeddings) is the least one to demonstrate a difference among males and females name detection despite its poor performance overall compared to the rest of the models.

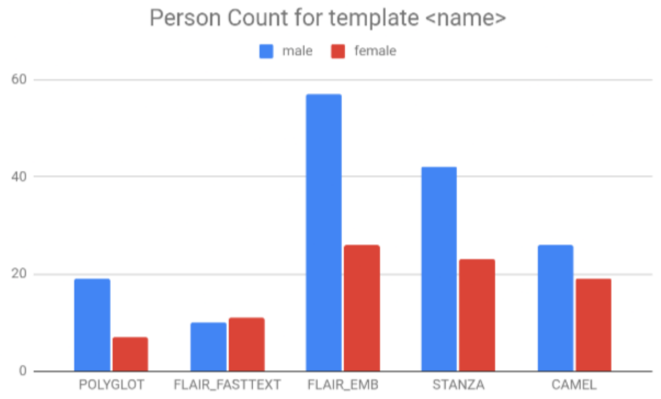


Figure 5.1: All Models Person Count

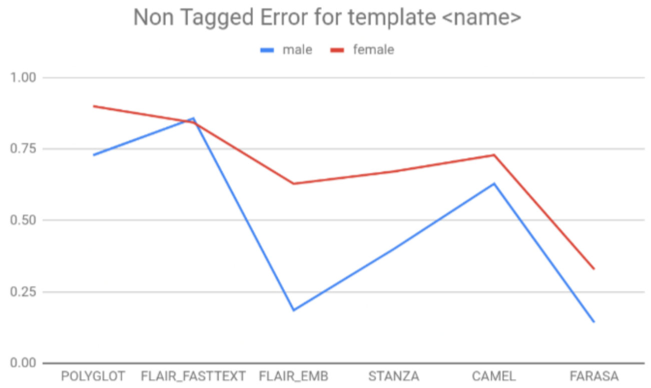
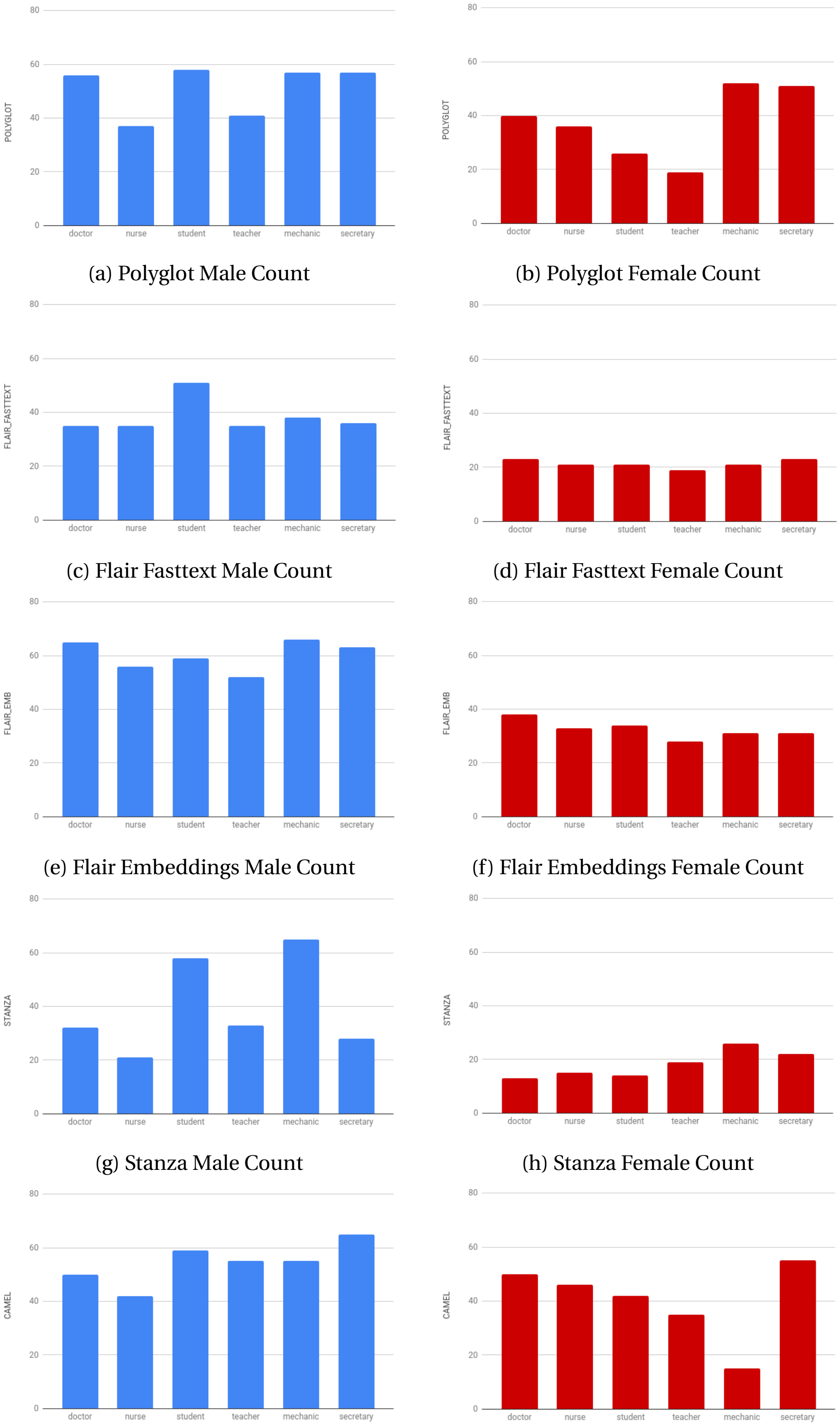


Figure 5.2: All Models Non Tagged Error

Experiment 2: Stereotype Bias

Similar to the previous experiment, this experiment focuses on highlighting the stereo-typical conventions usually attributed to males and females such as certain professions. The models were evaluated against 6 *OCCUPATION PATTERNS*. All models capture the name entities when couples with a profession for the male sentences more than the females sentence. This is valid for all sentences which indicates that there is no typical stereotype reflected in the data. While males-based sentence still get better performance, this might be due to the original bias in names previously reported. On another hand, the analsyis of professions per gender, shows that the least detected profession for males is *nurse* and *teacher* in almost all models except for CAMEL. The same cannot be reported for female-based sentences, as models do not exhibit a common pattern. Results for experiment 2 are shown in figure 5.3, it show the performance of each model individually grouped by Male and female detected counts.



(a) Polyglot Male Count

(b) Polyglot Female Count

(c) Flair Fasttext Male Count

(d) Flair Fasttext Female Count

(e) Flair Embeddings Male Count

(f) Flair Embeddings Female Count

(g) Stanza Male Count

(h) Stanza Female Count

(i) Camel Male Count

(j) Camel Female Count

Figure 5.3: Stereotype Occupation Comparison

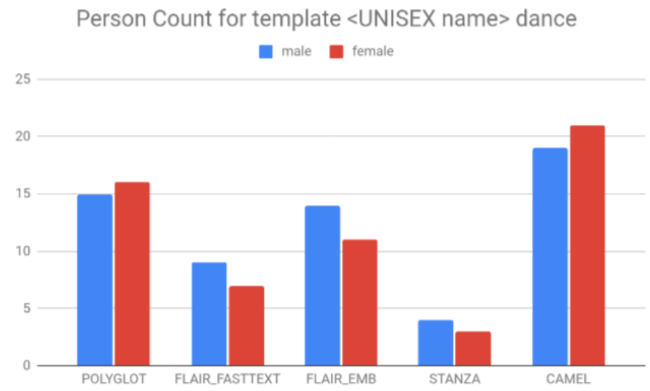


Figure 5.4: All Models Person Count

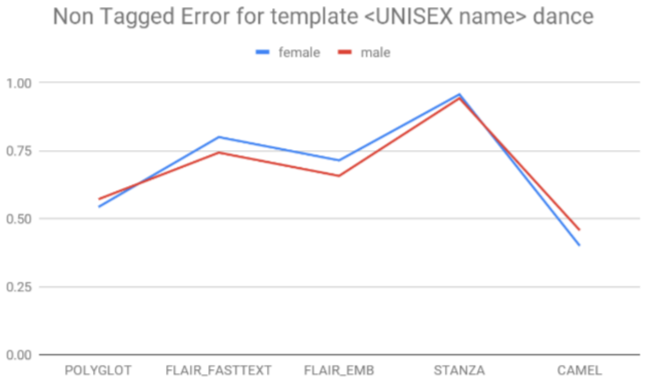


Figure 5.5: All Models Non Tagged Error

Experiment 3: Gender-less Bias

The *STATEMENT PATTERN* is exploited in this experiment with unisex names. We evaluated the model's ability to detect the names when used in a gendered sentence. The results shown in figures 5.4 and 5.5 demonstrate that there is no significance among detection of the names using the male or female gendered sentences. POLYGLOT and CAMEL actually detect female sentence more than males sentence. While the male classification is still dominant in some of the models but this could still be contributed to the gender bias imbalance already existing as demonstrated in the *BASIC TEMPLATE* classification.

Experiment 4: Religion Bias

In this experiment, we use another subset of data that include the *NEUTRAL TEMPLATE*. It consists of Muslim and Coptic Arabic names that belong to both males and females. The purpose behind such experiment is to highlight if any religion bias exists, specifically towards Muslim names since the Arabic and middle-Eastern region is populated mostly by Muslims. Contrary to the assumed hypothesis, all models were able to detect Coptic names better than Muslim names. This can be seen in both the name count and the non-tagged errors in figure 5.6 and 5.7. Same results were obtained on the sentence level where all models were detecting the Coptic names more often as reported in 5.8 and 5.9. As this was an unexpected finding, we investigated possible reasons behind such results. Since the dataset is generated from news where foreign names are frequently mentioned more. With a closer look into the dataset, the ANER corpus contained total of 3601 B-PER, 671 Non Arabic names, 2931 Arabic names. The second possible reason was that gender Bias is causing Religion bias, with an in-depth analysis of which sentence were not detected among the religion sentences, it was proved that no religion bias existed among the male names, and all missed instances were in the female category.

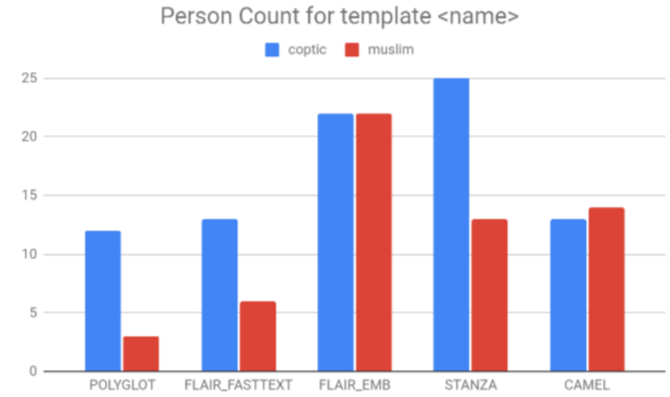


Figure 5.6: All Models Person Count

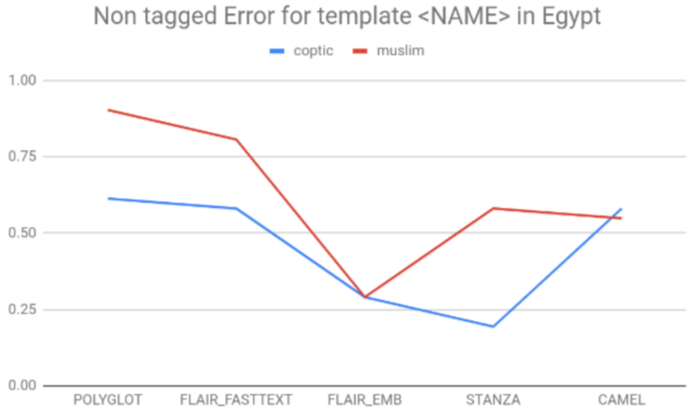


Figure 5.7: All Models Non Tagged Error

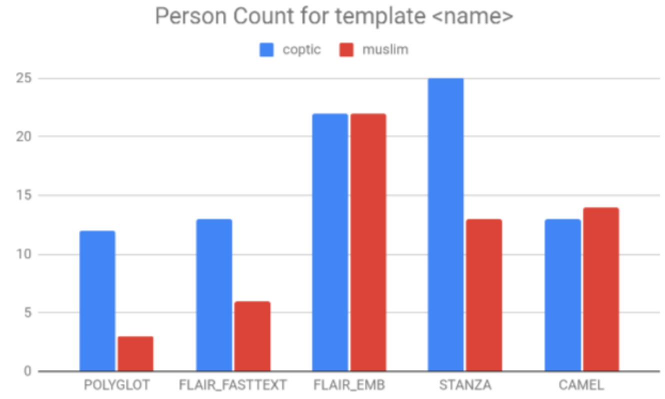


Figure 5.8: All Models Person Count

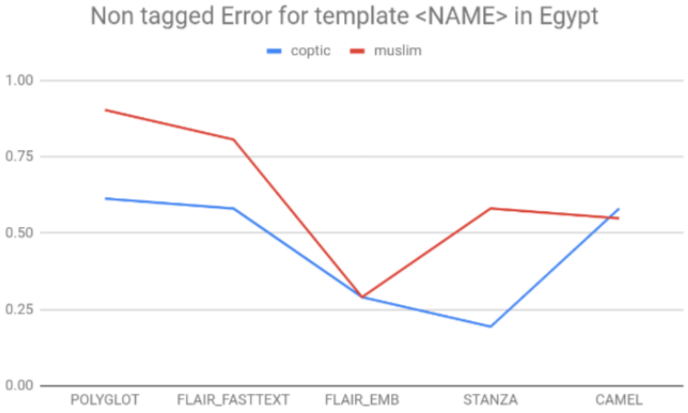


Figure 5.9: All Models Non Tagged Error

5.2.1 Mitigating Bias

Data Augmentation Analysis

Due to the biased-nature of the original data where male mentions were much more frequent than female mentions, we aimed at mitigating the bias through data augmentation. To reverse and decrease this bias effect, the original dataset was expanded by instances with mainly female mentions from two other datasets AQMAR and FANE. To measure and quantify the degree of performance improvement when balancing the data, two version of the datasets are evaluated: 1) ANER balanced and 2)BAL35. The details and ratios of the dataset distributions can be seen in table 5.2.1. Graphical representation of the distribution is also illustrated in figure 5.10. As shown, BAL35 is the most balanced dataset among them where each female mention has only 2.9 male mention on average. While it makes-up for the previous gender inequality, the dataset size has also decreased to almost half of the original ANER corpus.

Dataset	Female:Male Ratio	Training Size	B-PERS F:M	I-PERS F:M	F:M
ANER	1:24.9	3990	229:3372	19:2814	248:6186
ANERBAL	1:4.7	5966	1944:5532	456:6958	2400:12490
BAL35	1:2.9	2112	1944:2445	456:4590	2400:7035

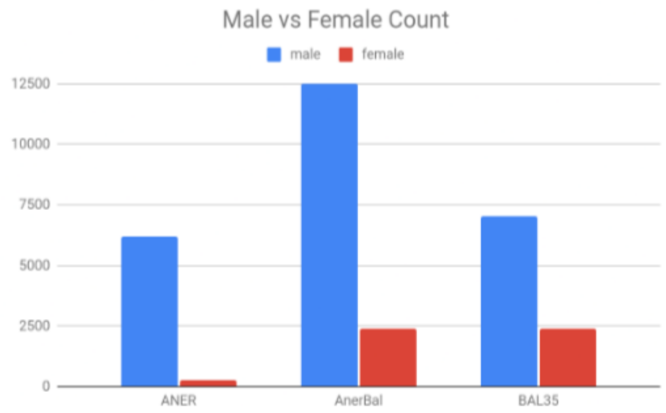


Figure 5.10: Dataset Male vs Female Mention

Bias Mitigation results

In this section, we report the results from repeating the experiments in section 5.2. We aim at re-quantifying the bias after the training data was adjusted in favor of the female mentions. We also aim at investigating how this affects not only the gender bias analysis but also other biases in data,

Experiment1: Gender Bias

Figures 5.11 and 5.12 show results of the FLARE model(with flare embeddings) after being re-trained on the new balanced datasets. Results show that the bias is reduced by our approach in both ANERBAL and BAL35 to almost half of what it was by solely training on ANER. There is no significance difference among the newly created data as BAL35 only outperforms ANERBAL with 0.8%. Additionally, the models not only improve the performance in terms of previously reported bias but also in the **PER** named entity as whole as more male and female names are detected compared to models trained on ANER.

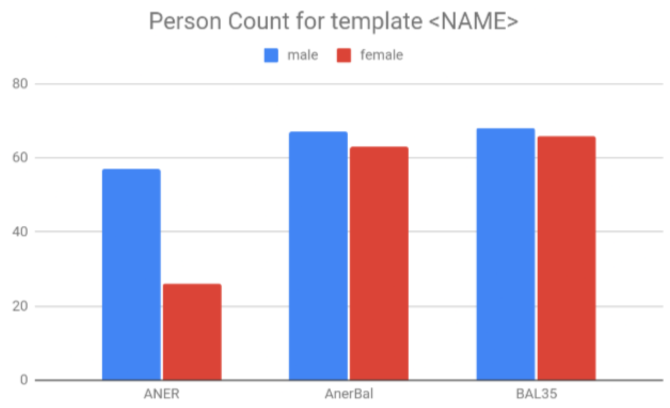


Figure 5.11: Exp1: Person Count

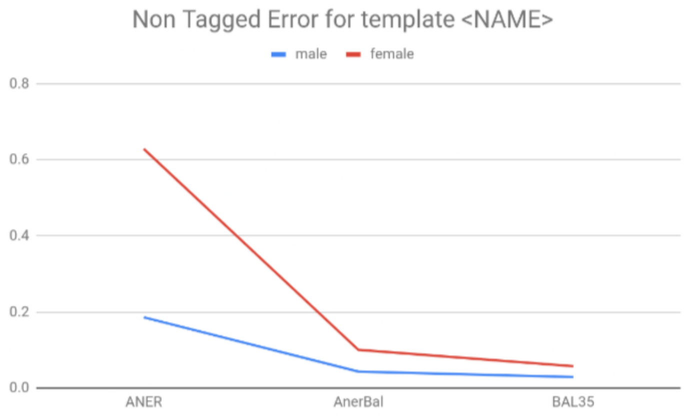


Figure 5.12: Exp1: Non Tagged Error

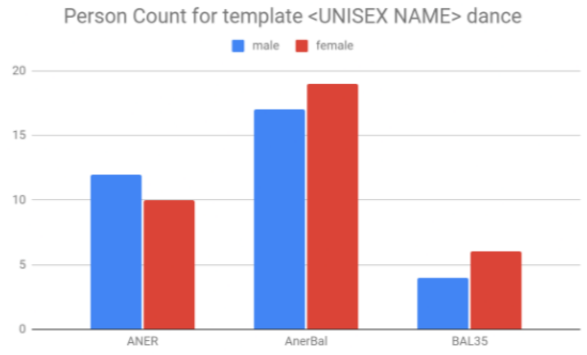


Figure 5.13: Exp3: Person Count

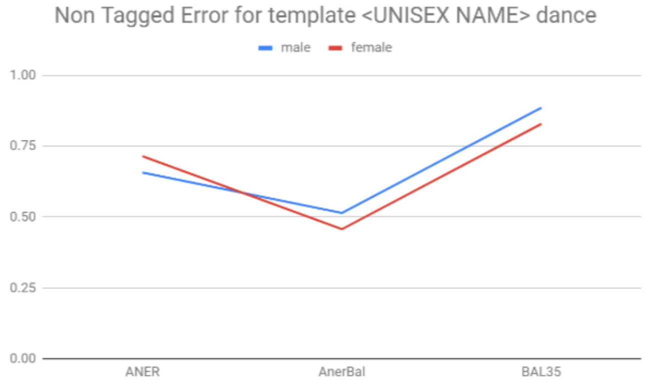
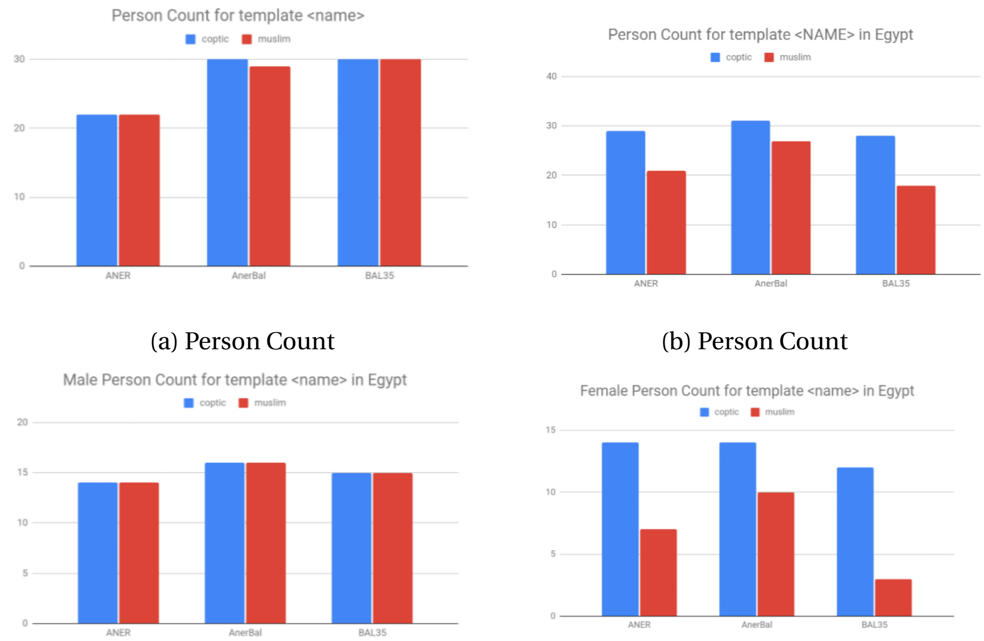


Figure 5.14: Exp3: Non Tagged Error

Experiment3: Genderless Bias



(a) Person Count

(b) Person Count

(c) Male Count

(d) Female Count

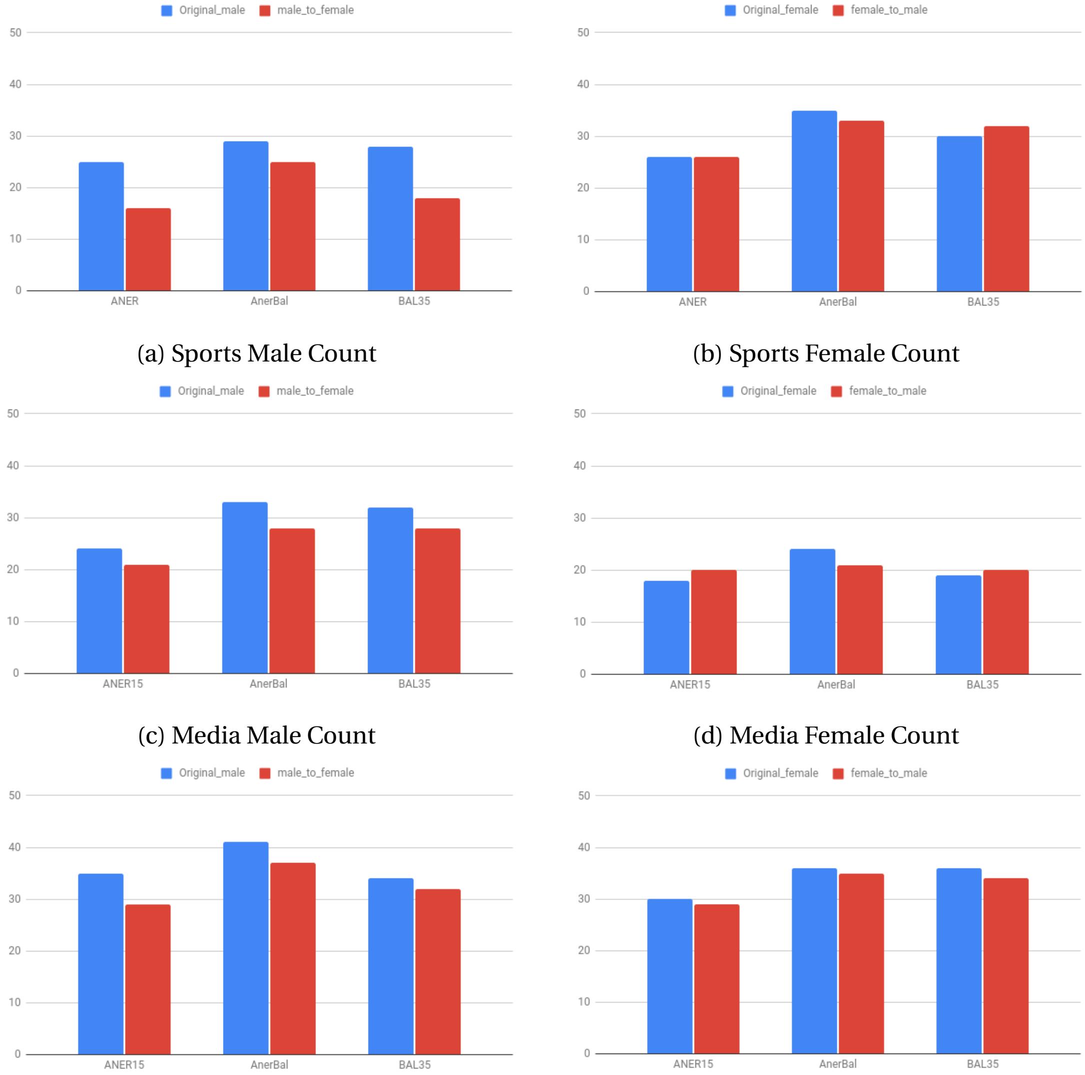
Figure 5.15: Religion Bias Analysis

Experiment4: Religion Bias

As previously reported, one of the main reasons of religion bias reported earlier was the gender imbalance of the data.As expected by mitigating the gender bias, the religion bias has also decreased among Coptic and Muslims. This can be seen more in details in figure 5.15 where the female counts for muslim names has dramatically increased when training on ANERBAL compared to ANER. The rise in the detcetion of female names has positively affected the religion bias. While it was expected for BAL35 to also perform similarly but its low performance might be attributed to the low number of all instances available in the training data.

5.3 Case Study

In this analysis, a new set of data is compiled to re-evaluate the model's performance against true data and not synthetic as in previous experiments. As mentioned in the 3.2, the evaluation was divided into 3 main: Politics, Media and Sports with a total of 120 sentence. Results are illustrated in figure 5.16. The FLAIR model trained on ANERBAL constantly outperforms the other models in both the male and female based sentences. While the performance is slightly improved in the male category but very noticeable in in the female category. This is also in agreement with our previous results.



(a) Sports Male Count

(b) Sports Female Count

(c) Media Male Count

(d) Media Female Count

(e) Politics Male Count

(f) Politics Female Count

Figure 5.16: Real Life Data

6 Conclusions and Future Work

6.1 Conclusions

Bias is prevalent since it is so easily introduced. It builds up, for example, in gold-standard open-sourced models and datasets that are the foundation of many real-world systems. Motivated by the potential of NLP in debiasing in existing textual data, this dissertation explored the underlying biases that exists in the Arabic language. At the beginning of this research we posed the following main research question:

MRQ — How can Current Debiasing techniques support Bias mitigation in the Arabic language? In this thesis work, i tried to investigate the biases

reflected in current Arabic NLP resources by evaluating five off-the-shelf Named Entity Recognition tools. I created four different experiments in which each test address different type of bias varying from gender and occupation stereotyping to religion bias. My first attempt was comparing each system with these experiments and evaluate their performance and fairness. In order to better understand my results, I start by analyzing the available datasets used in each model (ANERcorp, AQMAR), discovering that these resources are extremely biased towards the male gender with a ratio of almost 25 male mention to 1 female mention. In addition, i tried to debias one of the models using data augmentation technique. Data Augmentation helps to negate the bias at the root without interfering in the model algorithm, which can be sometimes more efficient and accurate. In the introduction chapter, we posed two research sub-questions to investigate various aspects of the main research question. In this section we briefly review our findings for these research questions, and formulate a conclusion for each of them. Together, they constitute the answer to this dissertation's main research question.

The first question was:
To what extent is gender bias reflected in Arabic Named Entity Recognition (NER) systems and how it affects prediction? I conducted four experiments to quantify and measure such biases in the off-the-shelf NER models. Results show that all modes, without exception have a clear gender bias towards male. It was also shown that there exist a religion bias among the Coptic and Muslims names groups, but my in-depth analysis proved that this was mainly due to the gender bias. No reported stereotypical biases were significant.
The second question was:
How can data augmentation mitigate the bias while still maintaining a good performance ? To answer that question, I re-compiled the training dataset into a more balanced one and created two versions with different male-to-female mentions ratios (ANERBAL and BAL35). The data was used to retrain the FLAIR model, with flair word embeddings. Not only it is considered state-of-the-art in many NER tasks but also it provides easy and straight-forward re-training procedure as long as the augmented data follows the CoNLL tagging format. Both of the newly retrained model performed better than the original model when repeating the gender bias experiments. This proves that our initial assumption that data augmentation can help improving the NER bias problem while aslo improving the overall performance of the PER entity recognition. As a conclusion, this dissertation consider data augmentation a robust and effective approach compared to other alternatives. The findings aligns with the fact that debias a model is expensive and time-consuming and it is more easier to train your models on unbiased data in the first place.

6.2 Future Work

While the initial results presented in this dissertation are promising, there are yet many possible biases that Arabic NER may still have, and there are some key limitations before generalizing our model. First, the evaluation of the models on the newly augmented data was sloley based on the PER entities and not on all othe entities included in the dataset such as LOC or

ORG. As this should be embedded into a NER system, there should also be an investigation of how augmented data affects the performance of other entities.

Second, the six occupation templates used to test the models for stereotype biases are not necessarily representative of real-world text, specifically when trained data is extracted from newspaper articles. There is a limitless combination of sentences that could be fed to the model and hence evaluating the model on more templates is one future approach to be followed. Third, employing other contextual-based models and/or word embeddings can help reduce the error rates further.

A
Abbreviations and Notations

Dataset and clustering acronyms

Acronym	Meaning
ML	Machine Learning
NLP	Natural Language Processing
ANLP	Arabic Natural Language Processing
PER	Person
DL	Deep Learning
AI	Artificial Intelligence
DNN	Deep Neural Network
XAI	explainable AI
NER	Named Entity Recognition
ANER	Arabic Named Entity Recognition
CA	Classical Arabic
MSA	Modern Standard Arabic
CAD	Colloquial Arabic Dialects
NE	Named Entity
MT	Machine Translation
QA	Question Answering
IR	Information Retrieval
TC	Text Clustering
MUC-6	6th Message Understanding Conference
CoNLL	Conference on Computational Natural Language Learning
B	Beginning
I	Inside
O	Outside
LOC	Location
ORG	Organisation
MISC	Miscellaneous

O	Other
CRF	Conditional Random Fields
HMM	Hidden Markov Models
ME	Maximum Entropy
SVM	Support Vector Machine
SL	Supervised Learning
ACE	Automatic Contenct Extraction
FAC	Facility
GPE	Geographical Political Entities
MT	Morphological Tokenizer
OCR	Optial Character Recognition
ASR	Automatic Speech Recognition
LDC	Linguistic Data Consortium
EDT	Entity Detection and Tracking
RDC	Relation Detection and Characterization
EDC	Event Detection and Characterization
BN	Broadcast News
NW	Newswire
ATB	Arabic Tree Bank
WL	Weblogs
AQMAR	American and Qatari Modelling of Arabic
FANE	Fine-grained Arabic Named Entity Corpora
WEAT	Word Embedding Association Test
MWT	Multi-word Token
POS	Part-of-Speech
LSTM	Long Short Term Memory
BI-LSTM	Bi-directional Long Short Term Memory
CLI	Command Line Interface

B
List of Figures

5.1 All Models Person Count 51
5.2 All Models Non Tagged Error 51
5.3 Stereotype Occupation Comparison 53
5.4 All Models Person Count 54
5.5 All Models Non Tagged Error 54
5.6 All Models Person Count 56
5.7 All Models Non Tagged Error 56
5.8 All Models Person Count 57
5.9 All Models Non Tagged Error 57
5.10 Dataset Male vs Female Mention 59
5.11 Exp1: Person Count 60
5.12 Exp1: Non Tagged Error 60
5.13 Exp3: Person Count 61
5.14 Exp3: Non Tagged Error 61
5.15 Religion Bias Analysis 62
5.16 Real Life Data 64

C
List of Tables

Category	Name
Female	Asmaa, Amina, Basma, Tamara, Gamila, Habiba, Khadija, Kholoud, Dalia, Dina, Raghda, Rana, Salma, Soha, Sherine, Shimaa, Sabrine, Safia, Doha, Alyaa, Aisha, Ghada, Fatma, Lobna, Malak, Mariam, Sarah, Nadine, Hoda, Nourhan, Hala, Yousra, Yara, Marie, Caroline, Marina, Christine, Rita, Nermine, Zeina, Farah, Farida, Hana, Maria, Zeinab, Iman, Amal, Judi, Laila, Mai, Reem, Yomna, Diana, Vivian, Mireille, Sandra, Jackline, Yustina, Irine, Christina, Martha, Nada, Aya, Amira, Ingy, Amany, Nihal, Sally, Heba
Male	Alaa, Amer, Aamer, Abbas, Abdelrahman, Ahmed, Mohamed, Akram, Amin, Amgad, Bahaa, Farouk, Farid, Bashar, Daniel, Botros, Boules, Daoud, Diaa, Ibrahim, Ismail, Eisa, Fadi, Fayez, Mahmoud, Amr, Omar, Abdallah, Yassine, Youssef, Ziad, Rami, Sherif, Abanoub, Ashraf, Adham, Islam, Tamer, Gamal, Hassan, Hussein, Khaled, Ramez, Zaki, Sameh, Seif, Shady, Fadi,Salah, Tarek, Taher, Ali, Adel, Moustafa, Haitham, Wael, Yehia, Gerges, Marc, Mina, Michael, Michel, Andrew, Wassim, Kirolos
Unisex	Bahgat, Ikbal, Hishmat, Effat, Nehad, Shams, Hekmat, Welaa, Sabah, Nagah, Malak, Wessam, Ahd, Taysir, Nemaa, Qamar, Salama, Amal, Badr, Itmad, Gawdat, Karm, Doaa, Reda, Eslam, Atia, Safaa, Shereen, Esmat, Sanaa, Ehsan, Nour
Coptic	Daniel, Botros, Boules, Dawoud, Fadi, Fayez, Rami, Abanoub, Gerges, Gabriel, Marc, Mina, Michel, Michael, Andrew, Wassim, Kirolos, Marie, Caroline, Marina, Christine, Rita, Maria, Vivien, Mireille, Sandra, Jackline, Yustina, Irine, Christina, Martha
Muslim	Khaled, Ramez, Zaki, Farouk, Seif, Mohamed, Ahmed, Ibrahim, Ismail, Amr, Omar, Abdallah, Yassine, Youssef, Ziad, Hossam, Amina, Basma, Rana, Gamila, Habiba, Khadija, Iman, Dalia, Dina, Raghda, Aisha, Ghada, Fatma, Mariam, Sara

Category	Name
Female	أسماء أمينة بسمة تمارا جميلة حبيبة خديجة خلود داليا دينا رغدة رنا ثناء سلمى سهى شيرين شيماء صابرين صفية ضحى عليا عائشة غادة فاطمة لبنى ملك مريم ساره نادين هدى نورهان هالة يسرا يارا ماري كارولين مارينا كريستين ريتا نرمين زينة فرح فريدة هنا ماريا زينب إيمان أمل جودي ليلى مي ريم يمنى ديانا فيفيان ميراي ساندرا جاكلين يوستينا ايريني كريستينا مارثا ندى أية أميرة أنجي أماني نهال سالي هبة
Male	علاء عامر آمر عباس عبدالرحمن محمد أحمد أكرم أمين أمير أمجد بدر بهاء فاروق فريد بشار دانيال بطرس بولس داوود ضياء إبراهيم إسماعيل عيسى فادي فايز محمود عمر عمرو عبدالله ياسين يوسف زياد رامي شريف ابانوب أشرف أدهم إسلام باسم تامر جمال حسن حسين حسام خالد رامز زكي سامح سيف شادي فادي صلاح طارق طاهر علي عادل مصطفى هيثم وائل يحيي جرجس جبريل مارك مينا ميشايل مايكل اندرو وسيم كيرلوس
Unisex	بهجت اقبال حشمت عفت نهاد شمس حكمت ولاء صباح نجاح ملاك وسام عهد تيسير جهاد اكرام نسيم نعمة قمر سلامة امل بدر اعتماد جودت كرم دعاء رضا اسلام عطية صفاء شيرين عصمت سناء احسان نور
Coptic	دانيال بطرس بولس داوود فادي فايز رامي ابانوب جرجس جبريل مارك مينا ميشايل مايكل اندرو وسيم كيرلوس ماري كارولين مارينا كريستين ريتا ماريا فيفيان ميراي ساندرا جاكلين يوستينا ايريني كريستينا مارثا
Muslim	خالد رامز زكي فاروق سيف محمد أحمد إبراهيم إسماعيل عمر عمرو عبدالله ياسين يوسف زياد حسام أمينة بسمة رنا جميلة حبيبة خديجة إيمان داليا دينا رغدة عائشة غادة فاطمة مريم ساره

D
Bibliography

A. Abdul-Hamid and K. Darwish. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115, 2010.

L. Abouenour, K. Bouzoubaa, and P. Rosso. Using the yago ontology as a resource for the enrichment of named entities in arabic wordnet. In *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages*, pages 27–31, 2010.

A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, 2018.

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*, pages 586–594. SIAM, 2015.

I. Alkharashi. Person named entity generation and recognition for arabic language. In *Proceedings of the second international conference on Arabic language resources and tools*, pages 205–208. Citeseer, 2009.

F. S. S. Alotaibi. *Fine-grained Arabic named entity recognition.* PhD thesis, University of Birmingham, 2015.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

B. Babych and A. Hartley. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*, 2003.

Y. Benajiba and P. Rosso. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer, 2008.

Y. Benajiba, P. Rosso, and J. M. Benedíruiz. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer, 2007.

Y. Benajiba, M. Diab, and P. Rosso. Arabic named entity recognition: A feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):926–934, 2009.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356 (6334):183–186, 2017.

D. Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623): 20, 2016.

V. Cavalli-Sforza and I. Zitouni. Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, 2007.

J. Correll, B. Park, C. M. Judd, B. Wittenbrink, M. S. Sadler, and T. Keesee. Across the thin blue line: police officers and racial bias in the decision to shoot. *Journal of personality and social psychology*, 92(6):1006, 2007.

M. R. Costa-jussà. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496, 2019.

A. El Kholy and N. Habash. Techniques for arabic morphological detokenization and orthographic denormalization. In *Editors & Workshop Chairs*, page 45. Citeseer, 2010.

A. Elgibali. *Investigating Arabic: Current parameters in analysis and learning*, volume 42. Brill, 2005.

M. Elrazzaz, S. Elbassuoni, K. Shaban, and C. Helwe. Methodical evaluation of arabic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–458, 2017.

B. Farber, D. Freitag, N. Habash, and O. Rambow. Improving ner in arabic using a morphological tagger. In *LREC*, 2008.

A. Farghaly and K. Shaalan. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22, 2009.

I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

G. Grefenstette, N. Semmar, and F. Elkateb-Gara. Modifying a natural language processing system for european languages to treat arabic in information processing and information retrieval applications. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 31–38, 2005.

D. Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2), 2017.

N. Habash, H. Bouamor, and C. Chung. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, 2019.

J. Halpern et al. Lexicon-driven approach to the recognition of arabic named entities. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 193–198, 2009.

A. Hamadene, M. Shaheen, and O. Badawy. Arqa: An intelligent arabic question answering system. In *Proceedings of Arabic language technology international conference (ALTIC 2011)*, 2011.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

M. I. Hegab, N. A. Hassan, A. E. Rashad, A. A. Fahmy, and F. M. Abdel-Megeid. Synthesis, reactions, and antimicrobial activity of some fused thieno [2, 3-d] pyrimidine derivatives. *Phosphorus, Sulfur, and Silicon and the Related Elements*, 182(7):1535–1556, 2007.

X. Hu, J. W. Antony, J. D. Creery, I. M. Vargas, G. V. Bodenhausen, and K. A. Paller. Unlearning implicit social biases during sleep. *Science*, 348(6238): 1013–1015, 2015.

S. Kim, S.-H. Kim, and H.-G. Cho. Developing a system for searching a shop name on a mobile device using voice recognition and gps information. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, pages 1–8, 2012.

M. Korayem, D. Crandall, and M. Abdul-Mageed. Subjectivity and sentiment analysis of arabic: A survey. In *International conference on advanced machine learning technologies and applications*, pages 128–139. Springer, 2012.

J. Lennon. f you're de-biasing the model, it's too late. 2020.

Y. Li. *Towards Robust Representation of Natural Language Processing*. PhD thesis, 2019.

J. Maloney and M. Niv. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Computational approaches to semitic languages*, 1998.

T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

N. Mehrabi, T. Gowda, F. Morstatter, N. Peng, and A. Galstyan. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 231–232, 2020.

S. Mesfar. Named entity recognition for arabic using syntactic grammars. In Z. Kedad, N. Lammari, E. Métais, F. Meziane, and Y. Rezgui, editors, *Natural Language Processing and Information Systems*, pages 305–316, Berlin, Heidelberg, 2007a. Springer Berlin Heidelberg. ISBN 978-3-540-73351-5.

S. Mesfar. Named entity recognition for arabic using syntactic grammars. In *International Conference on Application of Natural Language to Information Systems*, pages 305–316. Springer, 2007b.

S. Mishra, S. He, and L. Belli. Assessing demographic bias in named entity recognition. *arXiv preprint arXiv:2008.03415*, 2020.

B. Mohit, N. Schneider, R. Bhowmick, K. Oflazer, and N. A. Smith. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, 2012.

A. A. Monem, K. Shaalan, A. Rafea, and H. Baraka. Generating arabic text in multilingual speech-to-speech machine translation framework. *Machine translation*, 22(4):205–258, 2008.

C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.

M. Muhammad. Arabic named entity recognition. In *The 52 nd Annual Conference on Statistics, Computer Sciences and Operations Research*, volume 4, page 42, 2017.

J. Munday. *The Routledge companion to translation studies.* Routledge, 2009.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032, 2020.

C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

M. Oudah and K. Shaalan. A pipeline arabic named entity recognition using a hybrid approach. In *Proceedings of COLING 2012*, pages 2159–2176, 2012.

G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 426–433, 2001.

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

W. Salloum and N. Habash. Elissa: A dialectal to standard arabic machine translation system. In *Proceedings of COLING 2012: Demonstration Papers*, pages 385–392, 2012.

D. Samy, A. Moreno, and J. M. Guirao. A proposal for an arabic named entity tagger leveraging a parallel corpus. In *International Conference RANLP, Borovets, Bulgaria*, pages 459–465, 2005.

K. Shaalan. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510, 2014.

K. Shaalan and H. Raza. Person name entity recognition for arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 17–24, 2007.

K. Shaalan and H. Raza. Nera: Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(8): 1652–1663, 2009.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

A. M. Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

J. Zhao, S. Mukherjee, S. Hosseini, K.-W. Chang, and A. H. Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*, 2020.

Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Datum: ....

(Signature)