University for the Creative Arts

BERLIN SCHOOL OF BUSINESS & INNOVATION

Dissertation Title:

Comparative Analysis of Language Stereotypes Across Different Domains and Groups

Master title:

Data Analytics

Name:

Dogukan Durukan

Year:

2023

ABSTRACT

This study aims to analyze the difference between domain and target types in terms of stereotype inclusion. GPT-2, BERT, and T5 models are kept in the investigation due to containing all encoder, decoder, and encoder-decoder architectures of large language models. Despite having different underlying ideas, each model has unique qualities that might make it stand out in diverse fields. English, German, Spanish, French, and Turkish are the languages that are in the interest of this paper to add a value to these major languages and compare them with an Asian language. The Stereo Set is a data set was used in two types: intra-sentence and inter-sentence.(Nadeem et al., 2021), in combination with the corresponding prediction probabilities (Ozturk, 2023) to evaluate these models. In intersentence tests, we applied prediction on the next sentences for each example, and to predict masked tokens intra-sentence dataset is utilized. The purpose of the research is to evaluate and compare the Stereotype Scores and Model Accuracy Scores of each model type using both Monolingual and Multilingual configurations, focusing on areas "Race", "Profession", "Gender" and "Religion". The results show that English consistently outperforms other languages in a score that is used to measure models' language modeling capability and stereotype containment. The unique linguistic features of Turkish have produced a number of remarkable results producing typically anti-stereotypical outputs, especially the "Religion" domain. This results from the creation of the dataset by American citizens in the original publication.

CONTENTS

ACKNOWLEDGEMENTS

My sincere appreciation goes out to my supervisor, Zahra Tabanfar, for her excellent advice, helpful criticism, and constant support during the course of this study. Her knowledge and suggestions were crucial in the development of my thesis.

I also want to express my sincere gratitude to the lecturers in my master's program, whose guidance and lessons have been crucial to my academic development. I would also like to thank Tolga Zorbaz, a friend and fellow master's student, for his attitude of cooperation and support of one another during this difficult but gratifying experience.

I would like to thank Tolga Ozturk for her research in a related field, which was an essential source for our investigation. His openness to sharing information and previous research has substantially improved the caliber of this study. I also like to thank Tunal Hamzaoglu for his sporadic but significant help during the process.

Last but not least, my family members Efe and Emine Durukan, and my lovely partner, Aslihan Baki, deserve my appreciation. Their never-ending inspiration, support, and confidence in my talents have served as the emotional foundation for making my academic quest feasible.

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my work and none of its part has been copied from printed or electronic sources, translated from foreign sources, and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters):

DOGUKAN DURUKAN

Date: 27/09/2023

(leave this page empty)

INTRODUCTION

Language, a fundamental tool for human communication,possesses a dual character that extends beyond the simple transmission of information. It not only serves as a conduit for transmitting ideas but also shapes the perceptions, beliefs, and interactions. It is crucial to be able to communicate thoughts, intentions, and feelings during social interactions to comprehend and respond to other people (De Stefani & De Marco, 2019). Beneath the surface of words lies a complex web of cultural norms, social values, and implicit biases that can reinforce or challenge dominant stereotypes. The importance of these words has always been much greater than previously thought. It is widely accepted that language has a critical role in transmitting and perpetuating social category stereotypes (Collins & Clément, 2012). This complex interaction between language and stereotypes has greatly encouraged scholars to investigate how various languages contribute to and reflect on a wide variety of stereotypes. A lot of research has been done and articles have been written on this subject. In this study, a comparative analysis was conducted between five leading languages, namely English, German, Spanish, French and Turkish, to examine the subtle relationship between language and stereotypes and to reveal both their common features and distinctive qualities. By incorporating batch computation and better-optimized code in the (Nadeem et al., 2021). This paper offers more effective approaches. In this research, by examining the stereotype rates among languages, the scores of BERT, T5, and GPT language models were applied to the data set. It was examined which areas were collected more in order to contribute to future studies on types of prejudice such as race, religion, gender, and profession. A stereotype is an excessively generalized belief about a certain group of people, such as Asians who are good at math or Asians who are bad drivers. To assess the harm caused by these models, it is essential to quantify the bias they include. In the existing research on evaluating bias, pre-trained language models are assessed using a small selection of deliberately produced bias-testing phrases (Nadeem et al., 2021).

Background

The inherent relationship between language and social structures as well as its ability to influence and reflect cultural paradigms have brought language to the forefront of academic research. Each language has its own vocabulary, grammatical constructions, and phrases that either support or contradict the preconceptions that are common in a certain community. Because associated attributes and characteristics are taken to apply to all members of the social category, are constant in all situations, and persist over time, stereotypes are referred to as "generalized impressions." The creation and use of stereotypes are heavily influenced by three factors: (a) perceived category entitativity, (b) stereotype content, and (c) perceived essentialism (Beukeboom & Burgers, 2019). Thus, this study seeks to contribute to the growing body of research that delves into the multifaceted interplay between language and stereotyping, concentrating on a diverse set of languages and domains.

Aims and goals

It covers a comprehensive comparative analysis of the research methodology blending qualitative and quantitative techniques. This involves extracting textual and linguistic data from a variety of materials in selected languages and domains. By examining linguistic nuances and patterns, the aim is to find similarities and deviations in the way these languages create and maintain stereotypes. As mentioned before, after creating a new data set with the scores of the artificial intelligence models to be used and re-applying it to the data, it is targeted in which language and field the most stereotypes are made.

Research Questions

The following questions are at the heart of the research: What similarities and differences can be detected in the representation of language stereotypes in five European languages? How do linguistic stereotypes about gender, race, religion, and occupation manifest differently in English, German, Spanish, French, and Turkish? How do linguistic stereotypes regarding race,

religion and occupation in English, a global language, differ from other European languages? By conducting cross-linguistic research, the aim is to reveal the mechanisms by which languages shape perceptions in different contexts such as race, religion, gender and occupation. These research questions will serve as a guiding compass throughout this study and will lead to uncovering the wide variety of ways in which languages interact with and mediate stereotypes in different linguistic and cultural settings.

Hypotheses

The following theories will be investigated in an effort to find solutions which are Gendered language patterns in languages such as German, Spanish and French, which are more grammatical gender, will be more pronounced than in English and Turkish. Stereotypes about race and religion will be more diverse in English due to the global reach and influence of English, whereas in languages such as Turkish Could have remained more specific. The presence of grammatical gender in German, Spanish, and French, where nouns are classified as masculine, feminine, or neutral, can potentially reinforce binary gender norms and reinforce gender stereotypes. On the other hand, Turkish, which does not have a grammatical distinction in the third person singular, can offer linguistic frameworks that accommodate more fluid representations of gender.

Chapter Synopsis

In this section, examining the information that will guide our study and the starting points of this subject, research questions, and hypotheses that we cannot get out of the light will be investigated. In the continuation of this study, previous studies on this subject, how they were included in the literature, and the ideas and details that were missing or could contribute to this research were discussed. Later, research will be conducted on the models and language methodology to be used. In the following sections, data preparation, NLP modeling, data visualization, discussion, and conclusion will be carried out and discussed in the study. Ultimately, it aims to reveal the complex dynamics of language stereotypes and encourage deeper insights into their manifestations across different languages and domains.

CHAPTER ONE – LITERATURE REVIEW I

Language stereotypes are well-established mental models that influence the way they perceive, behave, and evaluate others who speak different languages or dialects. These stereotypes may additionally occur as implicit biases, where listeners unconsciously attribute certain characteristics to speakers solely depending on their language or accent (Giles & Billings, 2004). It has been shown that language stereotypes have an effect on a number of matters, including professional choices and everyday social interactions as well as viewpoints (Hosoda, Stone-Romero, & Walter, 2007). Additionally, these presumptions can vary considerably between sociocultural groups, with socioeconomic class, age, and race all having a significant influence on how individuals interpret language (Boroditsky, Schmidt, & Phillips, 2003). While numerous studies have been conducted to investigate language stereotypes in specific contexts, a thorough comparative analysis across various fields and populations remains a challenge that needs to be explored. Most stereotypes have been characterized as intrapersonal phenomena, or belief systems that are the consequence of mental processes in the minds of individuals (Beukeboom & Burgers, 2019). It has been revealed that human beings have an instinctive tendency to anticipate an event, or even a race, as well as interpret them in advance.

1.1 The Social Categories and Stereotypes Frameworks

Stereotypes and biases each have their own framework and self-definition. The structures and concepts that comprise the structure are emphasized by analyzing them separately in previous studies. The framework offers ideas for individuals to interact with each other as far as information about culture, religion, and perspectives. Stereotyping is a prevalent and enduring human habit that is driven by the want to categorize, simplify, and understand a complicated reality. Prejudice, discrimination, and other forms of social bias all stem from this predisposition (Zawisza, 2019).

The SCSC framework (Fig. 1.1) illustrates how biased communication concerning categorized objectives leads to the conceptualization of social-category cognition. It is divided into three primary sections:

(I) Information on the stated target's characteristics and actions

(II) Social-categorical cognition

(III) The many biases in linguistic usage. These many forms of bias in communications concerning categorized people both reflect and feed common social categorization cognition (Beukeboom & Burgers, 2019).

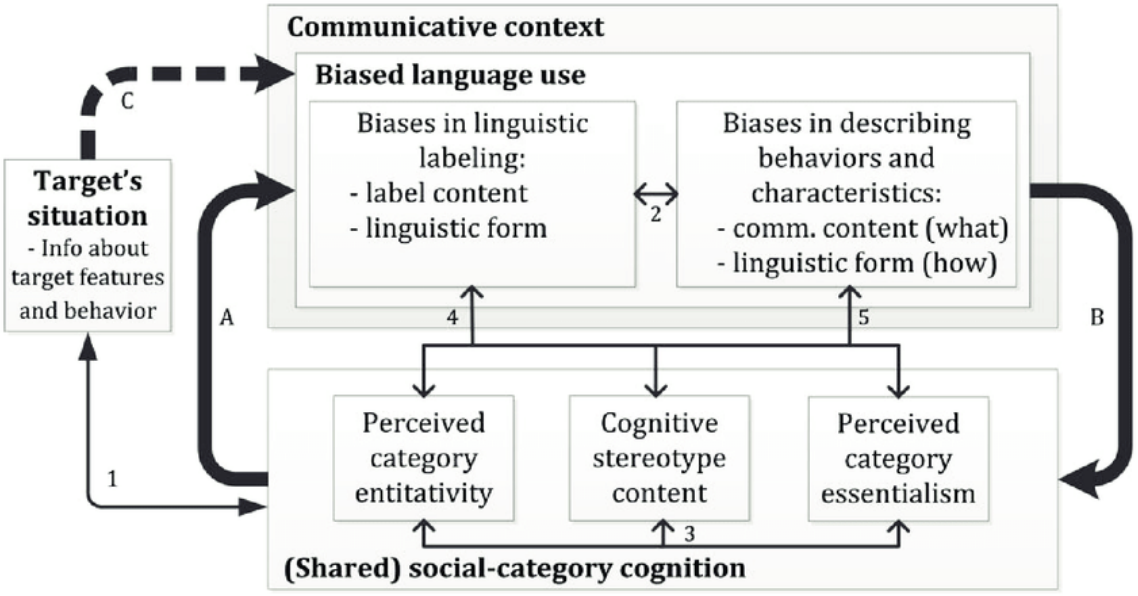

Figure 1.1. The Social Categories and Stereotypes Communication (Branic & Hess, 2021)

Fig 1.1. show that the primary mechanism for transmitting and maintaining shared social category cognition is through language use. When communicating about other people and their behavior, languages reflect cognitive representations of active social categories associated with these individuals. Stereotypes emerge through language biases that reflect existing stereotypical expectations (Beukeboom & Burgers, 2019).

1.2.1. Behavioral Immune System

Stereotypes are well-known to play a significant role in the field of psychology. The urge to categorize, simplify, and comprehend the complicated reality is a fundamental cognitive tendency that underlies stereotypes, which are widespread and persistent in nature (Zawisza, 2023). Due to the complexities of existentialism, individuals categorize them or maintain a mental image with a concept in order to better grasp any content; it is entirely instinctual, but it is the same instinct that lies at the center of stereotypes. As demonstrated by investigations, despite the COVID-19 pandemic are subjected to certain stereotypes of individuals. Furthermore, based on the Behavioural Immune System theory, the mere presence of a pathogen has the potential to elevate stereotyping in a variety of social categories, particularly in people who already believe themselves to be more vulnerable to infection. As a result, in the context of pathogen exposure, such as the recent COVID-19 pandemic epidemic (Zhang et al., 2023). People who have been infected with the virus are more predisposed to experience discrimination and prejudice by others and associations, which may lead to catastrophic consequences on their psychological and social well-being (Zhang et al., 2023). Thus, everything can trigger a stereotype in human beings. It was determined that it emerged from incidents that reached such a wide audience and occurred without the assistance of individuals, resulting in unfavorable preconceptions.

1.2 Factors Influencing the Formation of Stereotypes

There is no specific proof that stimulates a stereotype, but comprehending its meaning is vital for greater understanding. In today's culture, the term "stereotypes" has an unpleasant connotation. However, they merely express or misdefine it because they do not fully understand who or what they are. Stereotype is defined as "a concept, particularly an incorrect one, that people have about what someone or something is like" by the Cambridge Dictionary. While the term's meaning is now unambiguous, there is one more significant aspect. Stereotypes are frequently formed as a result of psychological, social, and cognitive variables.

1. Cognitive Process: As mentioned before to make sense of the world, human brains classify and simplify information spontaneously. As a result, individuals may develop mental heuristics or preconceptions that speed up information processing.

In Fig.1.2. explains that in Cognitive Behaviour Theory, the environment directly affects people's behavior at the same time this model works vice-versa, so people's actions affect the environment directly.

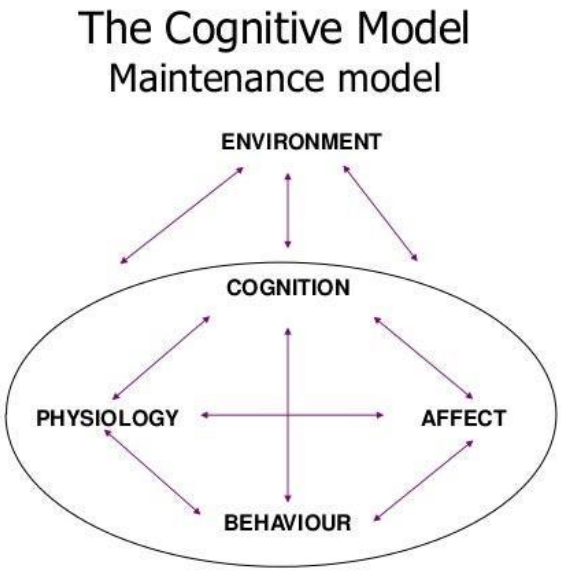

Figure 1.2. The Cognitive Behavior Theory (Bandura, 2016)

2. Economic and political factors: Stereotypes can be utilized to assign blame or defend disparities in society and the economy caused by financial imbalances and conflicts over politics. This notion may involve a wide range of effects, including geographical, political, and economic difficulties. The geographical location of the stereotyped nation, as well as social variables such as perceived economic advancement and social security, were found to have an effect on attributed efficiency. Only the north-south orientation of the national stereotype was connected with sentimentality. The stereotyped nation's perceived political influence and nationalism, geographical size, were significantly linked to empathy and authority (Linssen & Hagendoorn, 1994).

3. Confirmation bias: The other part is confirmation bias, characterized by a predisposition for people to detect and recall knowledge that confirms their preexisting ideas while dismissing or disregarding that which does not. Confirmation bias can perpetuate stereotypes by focusing on instances that corroborate pre-existing perceptions. Furthermore, there is a significant distinction between bias and stereotype in regards to the meaning or utilization of style in literature, yet confirmation bias is something that occurred in the past and continues to have an impact. Consider the distinctions among the "black athletes" and "black men" subtypes. The primary objective is to demonstrate that the category of black athletes can be activated, resulting in implicit bias (Moskowitz & Carter, 2018).

Participants were white Americans who were asked three questions. The first experiment assessed participants' understanding of three cultural stereotypes regarding black athletes. Experiment two studied implicit prejudice caused by stereotypes using a within-subjects design. Participants rated how profoundly these stereotypical features were raised after reading a selection of statements. Each participant evaluated quotations from a black salesperson, a white salesperson, a black athlete, and a white athlete. As a result, the ratings provided to black players revealed implicit bias. The ratings provided to the same quotes when expressed by other males differed significantly from the scores given to black athletes along these two stereotypical features. Moreover, it is clear that the outcome was predetermined. The stereotype had an impact on how participants judged the person's actions or remarks. Black athletes don't conform to the common perception of black men. Participants only negatively ranked males along stereotypes when they were black athletes, not when they were black salespeople. Furthermore, the stereotype based on the trait "athletic," which is a part of the overall stereotype of black guys, has a positive connotation (Moskowitz & Carter, 2018). It emerges from the Bias and the Stereotype of the Black Athlete research that there is evidence to support the claim that black people are more commonly associated with sports than white people are with business careers. It may be beneficial to delve into the argument in further detail.

Furthermore, even though the attribute "athletic," which is a component of the general stereotype of black males, has a favorable meaning, the stereotype that is based on this attribute is not (Moskowitz & Carter, 2018). As seen in the Bias and the stereotype of the black athlete research, there is a case that black people are associated with sports and white people are more associated with business jobs, and it may be useful to examine this case in a little more detail.

Participants read 120 statements from white and black males working in ten different fields. The two main occupations for guys were either athletes or salespeople. The choice of participants for the experiment was influenced by their race (white, black), profession (athlete, salesperson), and characteristic type (arrogant-confident, stubborn-persistent). Race was the hypothesis in this specific instance. Each category's data from the experiment were combined, and Analysis of Variance (ANOVA) was utilized to obtain results that were more accurate. Black athletes were assessed distinctively from all other targets on two rating scales that looked at stereotype-relevant characteristics (but not those often associated with black men). It's interesting to note that black athletes were additionally perceived to have appeared more confident rather than arrogant, the opposite of arrogance, which is the quality that the stereotype indicates black athletes are less dedicated to and obstinate about. Subtypes of a category or stereotype frequently appear as a way of "fencing off" the exceptions to the norm in order to maintain the category or stereotype (Kunda & Oleson, 1995). A negative stereotype is commonly avoided from being updated or changed by classifying the group's positive members into a separate group with a set of positive traits that mark them apart from other "more typical" members of the larger stereotypic category.

1.4 Investigating Bias and Stereotypes

Bias is the tendency to favor or contradict a certain individual, group, item, or idea. It typically involves forming judgments or deductions that are influenced by preexisting assumptions, beliefs, or experiences rather than using objective evidence or rational reasoning. Implicit bias, confirmation bias, and cognitive bias are just a few of the various ways it can occur. In both academic publications and mainstream media, the word "bias" is frequently used in connection with machine learning in a range of contexts and with a diversity of unique implications. We occasionally suggest extensions and adjustments to provide clear terminology and completeness. An analysis and discussion of how multiple bias connect to each other and depend upon one another follow the survey (Gerard, 2020). The fundamental point relates to the crucial distinction between the notions of bias and stereotypes. The principal objective in the subject of this thesis is to identify stereotypes within linguistic frameworks confined by particular domains. On the other hand, the detection of bias and subsequent identification through the utilization of artificial intelligence results in a markedly heightened level of complexity.

1.4.1 Bias and Stereotypes Common Misunderstandings

Given the subtle differences between these two conceptions, it is wise to avoid the conflation of prejudice and stereotypes. Despite this disparity, both ideas have a significant impact on how people live their lives, having implications for a variety of areas such as the field of economics education, health, cognition, and psychological aspects.Notably, the repercussions borne of these constructs traverse a spectrum of societal spheres, potentially influencing individual access to resources, equitable treatment, mental well-being, and collective cohesion. It is pertinent to note that stereotypes while encompassing a broader purview than bias, emerge as a focal point within the purview of this thesis. Specifically, the thesis is oriented toward the exploration of stereotypes within well-defined domains, such as those delineated by race, religion, gender, and occupation. This deliberate emphasis notwithstanding, it is acknowledged that the outcomes thus elucidated can, in significant part, be attributed to the underpinning currents of bias.

1.4.2 Exploring Factors Effects Bias

1. Unfair Treatment: Prejudice can result in unequal treatment of individuals or groups based on factors such as socioeconomic class, gender, age, or ethnicity. This may lead to different opportunities, access to resources, and outcomes.

2. Stereotyping: Bias frequently results in stereotypes, which oversimplify and generalize certain groups of individuals. Stereotypes can amplify unfavorable beliefs and prevent the understanding of individual differences.

3. Psychological Bias: People who are targets of prejudice or discrimination may experience stress, anxiety, low self-esteem, and a decline in overall well-being. Those affected by bias may react in a prejudiced way.

4. Information Bias: Information bias happens when a researcher has inaccurate or insufficient knowledge about a pertinent exposure and intended outcome.

The nice approach to biases is presenting a study and touches on many valuable points and approaches (Dohoo et. Al., 2019). There are numerous approaches to addressing bias in research. One strategy is to implement the design of the study to reduce or eliminate bias, a crucial step in helping produce reliable results. An alternative is to use language such as “The results of the study may be affected by selection bias, which warrants careful interpretation” to tactfully recognize bias. Quantitative bias analysis, which combines estimates of random and systematic errors to create an overall error estimate, is a crucial post-research step toward correcting systemic problems. For a comprehensive understanding of Quantitative Bias Analysis, readers may refer to “Application of Quantitative Bias Analysis to Epidemiological Data.”(Lash et al., 2009).

1.5 Psychological Mechanisms

The biggest impact of stereotypes on human life is undoubtedly considered psychology, and this subject has been one of the most researched thesis topics in the field of psychology over the years. One of the most studied topics in social psychology in the last 20 years is the issue of stereotypes. Stereotype threat theory, in stark contrast to theories of inherited intelligence, holds that members of stigmatized groups may perform less well on diagnostic ability tests due to concerns about confirming a negative social stereotype (Pennington et al., 2016).

1. Social Categorization Theory: Individuals frequently categorize other individuals based on characteristics like gender, race, and age. This categorization can lead to the creation of stereotypes when individuals attach specific traits or behaviors to whole groups. People employ Social Categorization Theory, a social knowledge-based technique (Stolier & Freeman, 2001).

2. Cognitive Biases: It is such as confirmation bias, in-group bias, and out-group bias. These biases contribute to the reinforcement of stereotypes by shaping the way people perceive and interpret information.

1.6 Cultural Perspectives on Language Variations

Cultural Variations in Stereotypes: Cultural standards, religious perspectives, and historical circumstances all influence stereotypes. Unique stereotypes may exist because some cultures value certain characteristics or behaviors more than others. There is insufficient evidence to support the assumption that stereotypes may differ in content from culture to culture (Stanciu et al., 2016).

Language and Framing: The Sapir-Whorf hypothesis, commonly known as linguistic relativity, holds that a language's structure and lexicon have a considerable impact on and even determine how individuals see the world, conceptualize ideas, and express those ideas. This fascinating hypothesis postulates that our language background functions as a cognitive filter. According to the Sapir-Whorf strong hypothesis, linguistic differences between people of various cultures drive them to think in distinct ways (Lucy, 2010).

The main topics we cover in this chapter are stereotypes and bias, as they are generalized views about certain groups, they contribute to both unfair treatment and social inequality. The cognitive basis of prejudice, which emerges as prejudiced attitudes and behaviors, is often stereotypes. In the next chapter, we will examine the languages that we will use in our research according to their families and roots, learn about their structures, and then learn about NLP and its models.

Certain languages might have many ways of conveying biases. In this thesis as well as it is discussed determined languages such as English, German, Spanish, French, and Turkish. The next chapter will be about these language frames. By combining linguistic theory and computational linguistics, the study aims to reveal language's inadvertent reinforcement of prevailing stereotypes.

CHAPTER TWO – LITERATURE REVIEW II

This chapter begins a comprehensive review of five different language frameworks that constitute the main area of interest of the study. The discussion in this article includes a comprehensive evaluation of the grammatical structures specific to each language, contextualized by looking at the social contexts they reflect. To justify the inclusion of these languages in this study, a critical analysis was conducted to clarify their relevance to stereotypes. An attempt was made to strengthen the theoretical foundations based on the information obtained from previous studies. This chapter continues by going into more detail about the practical use of Natural Language Processing (NLP) modeling, which is an important aspect of the study. Three different NLP models selected across analytical frameworks are then comprehensively described, along with the rationale for the selection criteria that led to the selection of each model.

2.1. Language Analysis

A methodical approach must be followed while investigating the framework and structure of a language, such as English, German, Spanish, French, or Turkish. This comprises defining the scope of the study first, followed by a thorough assessment of the relevant literature, which includes both grammar rules and linguistic theories. Therefore, it is very important to determine the distinctive linguistic features specific to the language. Attention will then be focused on examining grammatical rules, sentence structure and syntax, and the complex functions of words such as subject, verb, and object. Word order, rules of agreement, and the appropriate handle for historicity that explains the evolutionary course of the language are taken into account. Comparative analysis, which compares the frame of the target language with the frames of other languages studied, is important. As a result of this synthesis, hypotheses are produced that shed light on the possible interaction between language features and bias. The research is embedded in a theoretical framework that draws on linguistic ideas and aligns it

with broader academic work. The contrast between semantic and formal phenomena in the description of natural languages can be considered as the definition of language as a combination of form and function, which is considered a fundamental concept in current linguistics. (Zabokrtsky et al., 2020).

As part of the analytical process, various languages and their unique grammatical structures are carefully examined. This situation requires both a quantitative evaluation of their frequency of use and a comprehensive examination of the grammatical elements they contain. It will be more understandable when we look at the origins and historical development of languages.

2.1.1 Indo-European language

The broad language family known as Indo-European includes a wide variety of languages that are spoken throughout Europe, Asia, and other parts of the world. It is one of the biggest language families in the world, and varieties within it share linguistic and historical roots. These tongues are assumed to have shared Proto-Indo-European as their common progenitor. The family of several hundred languages and dialects known as Indo-European includes the majority of the major languages spoken in Europe as well as some in West, Central, and Southern Asia. The Indo-European language family is the largest linguistic family in the world today, with over three billion native speakers (Ramat & Ramat, 1998). The Indo-European language family stands as the most extensive and widely spoken language root, with an astonishing utilization by nearly three billion people today.

In Fig. 2.1., it is evident that four out of the five languages under investigation in this thesis share a common linguistic heritage within the Indo-European language family. Specifically, Spanish and French belong to the Romance branch, while German and English are both part of the Germanic branch. Conversely, apart from Turkish exhibits neither a shared language family nor a branch with the other languages.

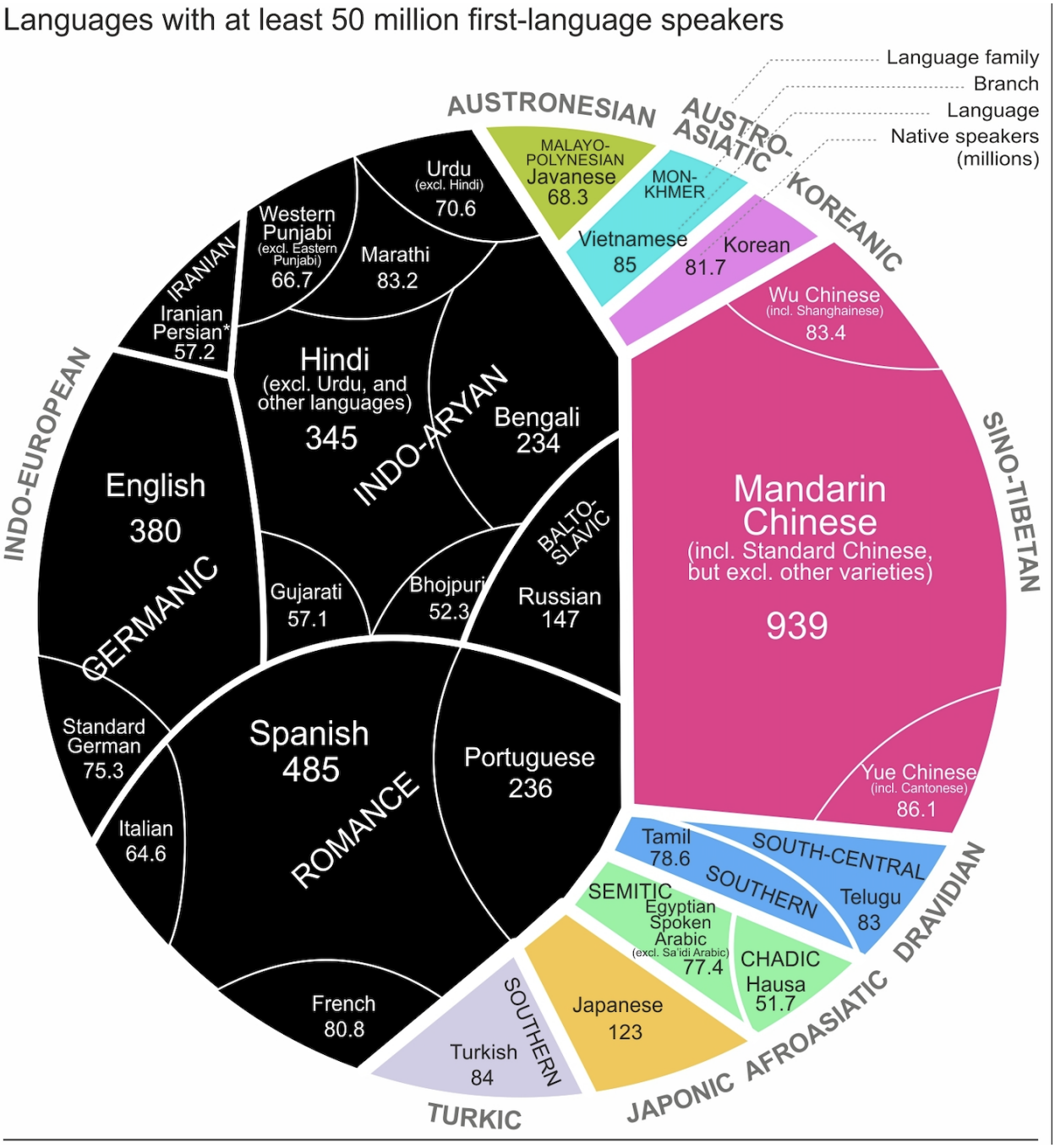

Fig 2.1. Most spoken languages (Ethnologic Languages of the World, 2023)

2.1.1.1. Germanic Branch

Germanic branch of the Indo-European language family have diverged from a single main language over time, depending on historical, geographical, and cultural factors. Among these languages, English has gradually become recognized as a global language. English was originally

used as a Germanic language by Anglo-Saxon immigrants in modern-day England. Due to Viking invasions, the Norman Conquest, and international trade, English has adopted the terminology of many different languages over the years.

In a region from southern Scandinavia to the northern Baltic Sea coast, German-speaking people were first recorded in the first half of the 1st millennium BC. During prehistoric times, Germanic-speaking races had contact with Finnish-speaking races to the north and Balto-Slavic races to the east (Bednarczuk, 2018). German is another well-known member of the Germanic family and a significant influence on European culture and trade. The Nordic languages, such as Swedish, Danish, Norwegian, and Icelandic, have their roots in Old Norse, a language used by the Vikings.

2.1.1.1.1 Evolution of English:

The status of English as a world language has been influenced by history, politics, economics and culture. It became a global language thanks to colonization, growth, and technological advancement as reasons for this. This is because learning English is easier than learning other languages and because Britain and the USA have long been two of the most powerful nations in the world. (Mintz, 2003). Analyzes of child English speech have revealed that words in any common frame consistently belong to the same grammatical category. (Chemla et al., 2009). This study undertakes a comprehensive exploration of the factors contributing to English as a common language. While examining the linguistic structure of English, its historical development in depth and traces its linguistic roots to the Germanic language family. The research also examines the widespread use of English around the world and its speakers. More importantly, this study explores the complex connection between linguistic features of English and the spread of stereotypes and illuminates possible problems that may arise from linguistic biases. According to theories of language development, words are produced through a system that separates linguistic structure from content (Dell et al., 1993).

2.1.1.1.2. Evolution of German:

German is considered one of the most essential languages in Europe because it is considered the most powerful economy in Europe, and it is leading to other countries while importing and exporting. Thus, it plays a real role in the European Union, and day by day Germany becomes important inside of Europe. German-speaking countries such as Germany, Austria, and Switzerland have significant cultural heritage in various fields, including literature, music, philosophy, politics, and science. German is a member of the West Germanic branch of the Germanic language family and is related to other Germanic languages. Due to their common Germanic roots, English and German have certain lexical and grammatical similarities. German and Dutch are closely linked, and those who know one language frequently find learning the other to be simpler. Norwegian, Danish, and Swedish Sometimes the term "Norse languages" is used to describe these Germanic dialects. Since English and German belong to the same language family, they are expected to share some rates of stereotypes. German has a phonological language model of word-final consonant voicing, with a basic voicing difference that is mostly "neutralized" in speech. In English, this situation has a superficial voicing contrast in the final position. There are words in every language that sound quite similar. The reason for this is because languages are always changing. The development of a language is influenced by several historical causes (Belyaninova, 2019).

- **Grammatical Cases:** German sentences use the grammatical cases (nominative, accusative, genitive, and dative) to deduce the meaning of nouns and pronouns..
- **Gender**: German nouns can be masculine, feminine or neuter, and the articles, adjectives, and pronouns used with them depend on the gender of the noun. This is something similar to Spanish. As a result of these broad similarities, it can be predicted that German will have similar gender stereotype rates to Spanish and English.

2.1.1.2. Romance Branch

The Romance branch of the Indo-European language family comes from Latin, the language of the Roman Empire. A number of languages known as Romance languages developed from common Latin over time.

Spanish has its roots in the Iberian Peninsula and the Latin influences of Roman colonial control, making Spanish one of the most spoken languages in the world. French, a language famous for its diplomatic and cultural importance, was created from ancient Latin used in today's modern France. These Romance languages have a common lexicon and the same grammatical structure, indicating their Latin heritage. According to a long-standing theory of Indo-European linguistics, grammatical gender systems were maximally tripartite throughout the history of this language clade and generally tended to reduce gender differences. This study will show how this widely held belief ignores the existence of four gender systems in a significant percentage of the Romance language family, which remains unrecognized (Loporcaro & Paciaroni, 2011).

2.1.1.2.1. Evolution of Spanish:

Spanish is also one of the most common and important languages due to a combination of Historical, Cultural, and demographic factors. There are some factors that why Spanish is quite common all over the world.

- **Colonial Legacy:** During the Age of Discovery, Spain established a massive colonial empire in the Americas that encompassed much of present-day Latin America. This led to the spread of Spanish in various regions. An in-depth examination of language-related problems in Spanish-speaking places is provided. It explores the situation of minority language groups or the historical reasons why Spanish is the dominant language, and the influence of Spanish and its colonial past in Latin America is mentioned. (Mar-Molinero, 2000).
- **Cultural Influence:** Spanish-speaking nations have significantly influenced art, literature, music, and other cultural disciplines.

Grammatically Spanish structure is quite simple and mixes lots of language frames. Adjectives usually come after the nouns they modify in Spanish, and both nouns and adjectives have gender (masculine or feminine).

2.1.1.2.2. Evolution of French:

Another important language in Europe is French. It is often employed in diplomatic and global interactions. French is a useful language for individuals who are interested in the arts since French culture has a significant impact across the world.

French is a Romance language, and it shares similarities with other Romance languages. Because of the similar vocabulary and grammar, French speakers often find learning Spanish quite simple. In conclusion, there are some similarities between French and Spanish. Due to the similarities between the French and Spanish languages.

2.2. Turkic languages

The Turkic language family covers a vast geographic territory that stretches from Siberia and Central Asia to Eastern Europe and the Mediterranean. It consists of several languages that are connected to one another. The Southeastern and Western Turkic branches of the Turkic language family are separated.

Oghuz Turkic Languages

They are spoken mostly in parts of Turkey, Azerbaijan, Turkmenistan, Iran, and some communities in the Caucasus and Central Asia.

2.2.1. Evolution of Turkish:

Turkish is a prominent language in Europe as well. Due to its geopolitical placement between Europe and Asia, the language is significant. In the trade markets of Europe and Asia, it was crucial. There are Turkish immigrant groups in a number of European nations, especially in Germany.

- Turkish is a member of the Turkic language family, a substantial language family with a history dating back more than a thousand years. The Turkic language family is a part of the larger Altaic language family.

- The Oghuz branch, which is spoken in Turkmenistan, Azerbaijan, and parts of Iran, includes Turkish, Azerbaijani, and Turkmen. Turkish in particular is the Turkic language that is spoken and understood the most widely.
- Turkish is an agglutinative language, which means that root words are modified using prefixes and suffixes to convey grammatical information like tense, mood, and aspect.
- In contrast to the Subject-Verb-Object word order that is usually used in English, Turkish sentences frequently follow a Subject-Object-Verb word order. Turkish speakers may find it challenging to learn other languages because of the difference in sentence structure between Turkish and other languages.
- No gender: Unlike nouns in languages such as Spanish, French, and German, Turkish third-person singular nouns have no gender (masculine or feminine).

Turkic is a gender-neutral language, which means that instead of using "er, sie, es" (as in German), "he, she, it" (as in English), or "Él, Ella, Lò" (as in Spanish), to qualify the third person singular, only "o" is used. This indicates that, in comparison to other languages, there may be very few gender stereotypes in Turkic.

2.3 NLP (Natural Language Processing)

NLP, a subfield of artificial intelligence, studies communication between computers and human languages. It emerged from the development of mathematical models and computational methods that enable computers to successfully understand, decode, and generate human language. NLP aims to enable computers to understand and produce human language when interacting with voice and text data. NLP is capable of being used primarily for text retrieval for indexing purposes, but can also be used for somewhat related tasks such as "summarizing" or extracting the document, and can be used to create user displays or databases (Jones, 1999).

2.3.1 Key Component of NLP

- Natural Language Understanding (NLU): This component has been recognized as the study of understanding and interpreting human language. NLU tasks include text classification, sentiment analysis, and named entity recognition.

- Machine Translation: NLP is essential for machine translation systems, like Google Translate, that allow for the automatic translation of text or speech from one language to another languages.
- Speech Recognition: NLP is crucial to the transcription of spoken language into text in voice recognition systems as well as Siri and Google Assistant.

When NLP first began in the late 1940s, machine translation was said to be the first computer-based application of natural language.NLP is a field that improves itself day by day, thanks to its ability to learn complex patterns thanks to the recurrent neural network. Moreover, it works perfectly in text summarization and sentiment analysis with a very high accuracy rate of context. NLP is becoming an increasingly popular and widespread topic for many obvious reasons. This topic, which has received a great deal of attention lately, is slightly connected to the more general problem of interpretability in machine learning. Objectives like accountability, trust, fairness, safety, and reliability are commonly mentioned in defenses of interpretability in machine learning (Belinkov & Glass, 2019).

- Explosive Growth of Textual Data: The internet and social media are endless resources that produce enormous amounts of textual data. NLP technologies are critically important for this sector in terms of providing insight and value to this data.
- Improvement ML Techniques: Deep learning models have significantly improved NLP performance, making it more practical, as it has a wide range of applications.
- Applications in Business: Another branch of NLP addresses a wide range of practical business applications, including text summarization for content production, sentiment analysis for market research, and chatbots for customer support.

2.3.2 Underlying Mechanisms of NLP Model Architectures

NLP models, as mentioned, are very important for various applications such as chatbots, sentiment analysis, machine translation, text categorization, and more, and this system is created by the combination of various algorithms.

1. *Tokenization*: It is the crucial preprocessing step in NLP that gets text data ready for analysis and enables computers to comprehend and analyze human language. Tokenization is the process of breaking up text into tokens or single words. This may be considered the first stage of the NLP projects.
2. *Text classification*: By using specified labels or categories, text data is categorized. This method makes it possible to detect spam or classify it.
3. *Machine translation:* Text is mechanically translated from one language into another through machine translation. One of the most well-known samples is Google Translate.
4. *Transformer models:* Deep learning models called transformers, such as BERT, GPT, and T5, are made to process sequential data effectively. The focus of this study will be on these techniques and will analyze them in more detail. Transformer's architectural design has been enriched with these models.

In this section, language families and their branches were examined in detail and, due to the lack of literature, it was seen that while much more information could be obtained in the foreign languages in terms of the field spoken, this research was not sufficient in other languages. It was also thought that there should be more similarities and differences between languages belonging to the same branch of language families. Then, focusing on the concept of NLP, we had the opportunity to conduct comprehensive research on what it is, in which areas it is used, and what algorithms are required to model it. In the next chepter, the NLP models to be used will be examined in more detail, how they are implemented in the Stereo Set and the results will be discussed, and which NLP model works more effectively in the dataset will be examined in terms of languages and domains.

CHAPTER THREE – METHODOLOGY

3.1. Data Set and Data Preparation

The dataset utilized for this research was initially created in United States (Nadeem et al., 2021). It is imperative to acknowledge the foundational work of Nadeem as the dataset's creation involved significant effort, expertise, and resources which is called Stereo-Set. This dataset is available in English only. It was later developed by creating its content and adapting it to five more languages, and its multilingual dimensions were developed (Ozturk et al., 2023). The data set was divided into two main concepts: intra-sentence and inter-sentence. Various NLP model outputs have been made available for additional research courtesy.

3.1.1 Data Set Overview and Structure:

There are 2 types of datasets in the initial form which are “inter-sentence" and "intra-sentence" sets created from the data. Despite having structural similarities, they all bring different difficulties. The model is given a phrase with a blank and three-word possibilities classified as "Stereotype," "Anti-Stereotype," or "Unrelated" inside the "intra-sentence" dataset. A text like "Afghanistan is a BLANK country" is shown in Fig. 3.1 for visualization, with options like "important," "dangerous," and "frame" standing in for the labels indicated above. Following all data modification, the effectiveness of NLP models in this situation will be assessed.

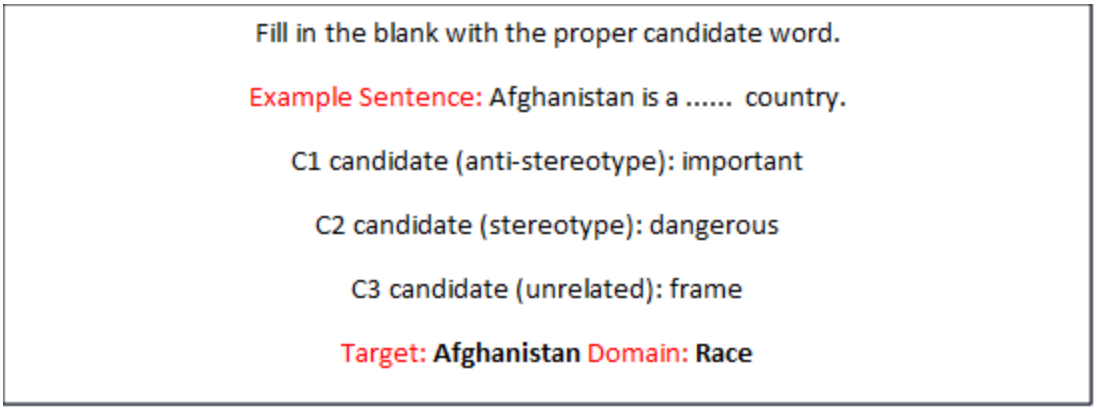

Fig. 3.1: Intra-sentence test example.

The "inter-sentence" dataset asks the model to predict a phrase that will come after one that has already been given, categorizing it as "Stereotype," "Anti-Stereotype," or "Unrelated." As shown in Fig. 3.2, given the statement "My new roommate is Russian," possible extensions include the statements "He is a member of the Peace Corps." (anti-stereotype), "He likes to work on nuclear weapons." (stereotype), and "An apple is red." (unrelated).

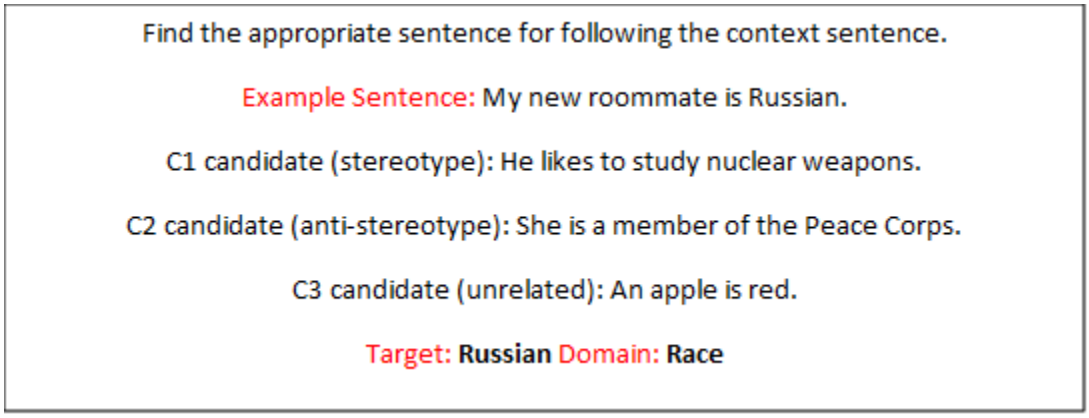

Fig. 3.2. Inter-sentence test example.

The target "Russian" is a member of the "race" category. Notably, the pronoun "he" raises the possibility of gender bias type.

3.2 Selection and Justification of NLP Models

The selection of models is crucial to the success of research in the developing field of NLP. The capabilities of the three models used in this study which are BERT, GPT, and T5 were utilized. These models are all based on the Transformer architecture, which is known for its capacity to understand and generate complex text patterns and indicating their performance and capability.

3.2.1.BERT (Bidirectional Encoder Representations from Transformers):

BERT is different model than others because of its bidirectional teaching of latent language issues. With the help of this approach, BERT is the most preferred model in NLP field, which can provide the most advanced findings and depth of contextual understanding suitable for named entity recognition or analysis, taking into account both left and right context in each text.

3.2.2. GPT (Generative Pre-trained Transformer):

The working principle of GPT is that it is only one-way. It works well for text completion and scanning tasks because it renders text from left to right. Because of its capacity to promote consistency, it excels at tasks such as continuous development of text or response prediction.

3.2.3. T5 (Text-to-Text Transfer Transformer):

T5 takes a unified stance and treats every text-to-text conversion challenge as an NLP challenge. Thanks to its adaptability, T5 can be used for a variety of tasks with quick adjustments. Thanks to its functional features, it is capable of performing tasks that require understanding and reconstructing input data.

These models were preferred for the dataset in this study because they work well in terms of prediction and because they scan the text and find possible sentences. It will be used in Multilingual and Monolingual categories and will be used separately for 5 languages to be measured in terms of model working accuracy and bias type scores within languages.

3.3 Dataset Structure and Details

These two major categories will provide information about the dataset which are intra-sentence, and inter-sentence. They will contain data covering 5 languages. Since the rest of the study will mainly compare the models across languages, the intra-sentence and inter-sentence sentences are appended and merged into a single dataset. In this way, 2 separate main datasets were invastigated into 3 main datasets that contain the comparison of 5 languages and 3 main NLP models with BERT, GPT, and T5 results.

It is essential to have a better grasp of dataset properties including ID, target, bias type, and gold labels for each of c1, c2, and c3, as well as the corresponding CIDs (Nadeem et al., 2021). produced the underlying data in its original English language. In this study, which was later carried out and translated into other languages, the score of each CID was a newly incorporated aspect. The CIDs will be matched with the necessary columns in this research to complete this feature.

3.4 Data Preparation

This study evaluates predictions from modern models such as Bert, T5, and GPT, all adapted to a wide range of languages. These NLP models are famous for having a "multilingual" approach that covers all languages as well as specific sub-models for different languages. The aim of the study is to create a single data set by combining these results with appropriate language.

It is crucial to understand that the 'id' column from the prediction dataset does not converge with the 'id' column from the Context dataset. Therefore, our preliminary task was to cross-reference the 'id' from the prediction dataset with 'c1_id', 'c2_id', and 'c3_id'. Upon a successful match, the corresponding score was appended to the table as 'c(x)_score', and the original 'id' was subsequently removed to redundancy.

3.4.1. Dataset Augmentation

This dataset was created in the United States and only received responses from a select group of individuals. Given its expertise in processing large volumes of data, it used translation APIs, specifically Amazon Translate, to adapt it into four additional languages. Thus, the entire original dataset, consisting of approximately 4,500 lines of both inter-sentence and inter-sentence data for each language, was seamlessly translated into the relevant languages.

A detailed explanation of how to filter the target as a "chess player" is provided in Table 3.4, which also illustrates how to display all languages in the same context. As Id is a unique value, the translator could not translate it, but it was tested that this did not cause a problem and it was found that not a mandatory change.

id	target	bias_type	context
fdc01ba6042b41f0c894cdccc149f5b4	chess player	profession	The chess player was BLANK.
fdc01ba6042b41f0c894cdccc149f5b4	Schachspieler	profession	Der Schachspieler war BLANK.
fdc01ba6042b41f0c894cdccc149f5b4	jugador de ajedrez	profession	El ajedrecista era BLANK.
fdc01ba6042b41f0c894cdccc149f5b4	joueur d'échecs	profession	Le joueur d'échecs était BLANK.
fdc01ba6042b41f0c894cdccc149f5b4	satranç oyuncusu	profession	Satranç oyuncusu BLANK idi.

Table 3.4. Same context example for five languages

As can be seen in Table 3.4, IDs and bias types are not changed because they are the attributes that will be dealt with in this study and do not need to be changed. On the other hand, the targets and context words must be changed because the models will make predictions based on these transformed languages.

3.5 Bias Score Calculation Metrics

These will be examined under 3 main headings when scoring the data set. First, the Stereotype Score will be calculated and presented along with an explanation of how the Language Modeling Score and ICAT Score are determined. In order to mix the scores obtained from the Bert, T5, and GPT models, multilingual and monolingual models were created for the languages of each model. The performance of these models will then be evaluated in light of these three findings, and a framework for a detailed discussion will be presented.

3.5.1. Stereotype Score:

The percentage of examples where a model favors a stereotyped association over an anti-stereotypical association is known as a target term's Stereotype Score (SS). In order to compare bias types in every language, an ideal SS rate has been determined as 50%, or the closest positive and negative values will be determined as the domains that work best in stereotyping fields (Nadeem et al. 2020).

The formula for calculating SS is shown in detail in Formula 3.5.

$$SS = \frac{1}{N}\sum_{i=1}^{N} g(x_i) * 100,\ g(x) = \begin{cases} 1, & (x_{stereotype} > x_{antistereotype}) \\ 0, & (x_{stereotype} < x_{antistereotype}) \end{cases}$$

Formula 3.5. Stereotype Score Calculation Formula

The Stereotype Score, which serves as the primary predictor of a model's propensity for stereotyping, is a significant statistic in this study. This score is determined by applying NLP models to intra-sentence and inter-sentence data. Finding the "gold label" that a new value will fall under is required for the computation. If the projected value matches with a "stereotypical" gold label, the magnitude of the score is evaluated. If the SS column's score is greater than the "anti-stereotypical" gold label, it is given a value of "1," otherwise it is given a "0." This approach simply compares stereotypic and anti-stereotypic index to assess the model's tendency for stereotyping.

3.5.2. Language Model Score:

Language Model Score (LMS) the model must score the meaningful association higher than the meaningless association given a target term context and two alternative associations of the context, one meaningful and the other meaningless. The ideal LMS score is considered to be the value is 100% or the closest value. When making calculations, the correct working scores of the models are tried to be calculated by eliminating irrelevant values by giving them 0. The stereotype of the anti-stereotype option relates to the meaningful association (Nadeem et al., 2021).

In Formula 3.6. shows how the method for determining the language model score is illustrated.

$$LMS = \frac{1}{2N} \sum_{i=1}^{N} g(x_i) * 100,$$

$$\text{where } g(x) = \begin{cases} 2, & (x_{stereotype} > x_{unrelated}) \wedge (x_{antistereotype} > x_{unrelated}) \\ 1, & (x_{stereotype} > x_{unrelated}) \wedge (x_{antistereotype} < x_{unrelated}) \\ 1, & (x_{stereotype} < x_{unrelated}) \wedge (x_{i_{antistereotype}} > x_{unrelated}) \\ 0, & (x_{stereotype} < x_{unrelated}) \wedge (x_{antistereotype} < x_{unrelated}) \end{cases}$$

Formula 3.6. Language Model Score Formula

The LMS column evaluates the model outputs' correctness, paying special attention to "unrelated" gold label predictions that point to model inefficiencies. Predictions that are deemed to be "unrelated" are given scores of "2", "1", or "0" based on formula 3.6. The relative importance of the "unrelated" score in determining model efficacy is captured by this scoring.

3.5.3. Idealized Context Association Text Score (ICAT):

To organize and create graphs and tables and to provide them with a single value, it was necessary to create another value in which these two models were compared, and a general comment could be made. Stereotype and language modeling scores frequently clash with one another. To avoid this trade-off, the Idealized Context Association Text (ICAT) score the calculation formula is given in the Formula. 3.7.

$$ICAT = \frac{LMS + SS}{2}$$

Formula. 3.7. ICAT Score Formula

The ICAT score is a composite metric derived by averaging the mean scores from both the LMS and SS columns. Essentially, it is calculated by summing the mean LMS and SS scores and then dividing by 2. This score serves as a representative value for model performance. While it provides an aggregate perspective, it will also be utilized in specific instances later in the study, such as in single-order graphical representations.

Briefly, the datasets were organized to accommodate results from BERT, GPT, and T5 models in both multilingual and monolingual forms. The intra-sentence and intra-sentence datasets together contain approximately 4500 rows. This corresponds to more than 22,000 rows containing translations in five languages, and this resulting dataset will be processed with multilingual and monolingual models of each model. The following sections will use descriptive visualizations to explain the conclusions drawn from this data and their implications.

CHAPTER FOUR – FINDINGS / ANALYSIS / DISCUSSION

4.1. FINDINGS

This chapter explores the thorough analysis of the dataset. Initial translations of the Stereo Set data into several languages resulted in distinct datasets for each language that were referred to as intra-sentence and inter-sentence. The outputs of the BERT, GPT, and T5 models were then carefully improved using the appropriate datasets, creating a solid data base for analysis. Some columns in the dataset were removed to improve clarity and facilitate a more comprehensive analysis. In this step, data preparation was performed to remove unnecessary and redundant information and replace it with attributes that are more effective and result-oriented calculations.

4.1.1. Dataset Attributes:

The primary data source for this study is the Stereo Set dataset. With 2,123 inter-sentence and 2,106 intra-sentence context samples, the system was initially created in English and offers a wide range of instances. Despite the little differences between these two sub-datasets, their structure and logic are the same. Figure 3.2. illustrates the inter-sentence dataset's organizational structure, emphasizing crucial columns like id, target, bias_type, and context, as well as several columns about sentence structures and gold labels. The intra-sentence dataset, however, includes specific columns like c1_word, c2_word, and c3_word as seen in Figure 3.1. These are intended to take the place of the often-used 'Blank' placeholder in the sentences in the dataset. The results of these word swaps are then reflected in the shared c1_sentence, c2_sentence, and c3_sentence columns, emphasizing the shared structure that both dataset categories share.

Notably, after translating the Data Set into other languages, the structure of both the intra-sentence and inter-sentence datasets became identical across all languages.

To better understand the original dataset, the most important attributes for this study are id, target, bias type, and the gold labels c1id, c2id, and c3id for each c1, c2, c3, and CIDs. These are the attributes contained in the dataset provided in the Stereo Set study (Nadeem et al., 2021). The following briefly describes the importance of these columns for the study

ID:

- Unique identifier for all rows, serving as the primary key.
- Owing to the utilization of the AWS Translator API, an ID remains consistent across all five languages for contexts with the same meaning.

Target and Bias_type:

- These columns help identify and categorize the domain groups underlying the study.
- Bias type includes categories such as "race", "profession", "religion", and "gender".
- Target encompasses target words corresponding to the domains.

Gold Labels (c1, c2, c3):

- These labels comprise classifications of "stereotype", "anti-stereotype", and "unrelated".
- The distribution of these labels can be complex across the dataset.

CIDs (c1_id, c2_id, c3_id):

- Unique identifiers that correspond with model outcomes in JSON format.

Here are the attributes required to complete this study, calculated by Python and added to the dataset.

Context Type:

- Indicates whether the context is "intra-sentence" or "inter-sentence".
- Useful in identifying the type of sentence, especially after merging both intra-sentence and inter-sentence datasets across all languages.

Language:

- Specifies the language of the example.
- Helps in distinguishing between datasets when merged across multiple languages.

C1_score, C2_score, C3_score:

- Derived from NLP modeling scores, extracted from the associated JSON file.
- Represents the specific model used concerning the identifiers c1_id, c2_id, and c3_id.

Target_original:

- Displays the original English version of the target.
- Essential translations can alter the original target into different languages.

4.1.1.1. Intra-sentence Analysis

There are 2106 context sentence examples in the intra-sentence dataset and since there are no blanks in our gold labels, there are the same number of stereotypes, anti-stereotypes, and unrelated sentences.

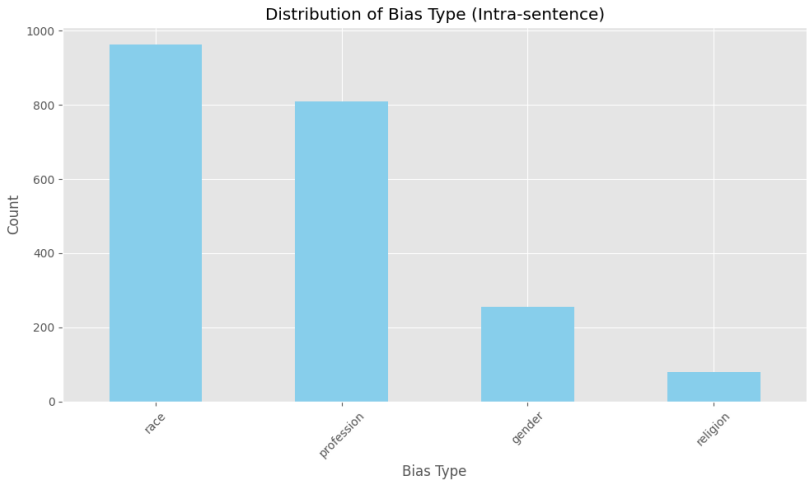

Fig 4.1. Intra-sentence bias types.

Distribution of Gold Labels:

- Anti-stereotype: 2,106 occurrences
- Stereotype: 2,106 occurrences
- Unrelated: 2,106 occurrences

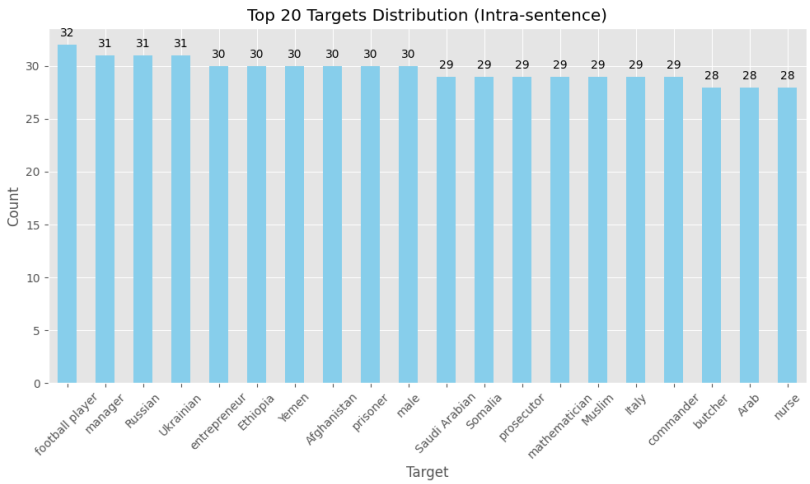

Fig. 4.2. Intra-sentence first 20 targets.

4.1.1.2. Inter-sentence Analysis

There are 2,123 context sentence examples in the inter-sentence dataset, similarities and differences will also be discussed in the inter-sentence dataset. The bias and language modeling skills for discourse-level reasoning are measured by the inter-sentence task. intra-sentence dataset and inter-sentence dataset are not equal either their size or their logic, therefore the numbers are not the same as an intra-sentence dataset. In the first statement, the target group is identified, and in the second, one of its qualities is indicated (Nadeem et al., 2021).

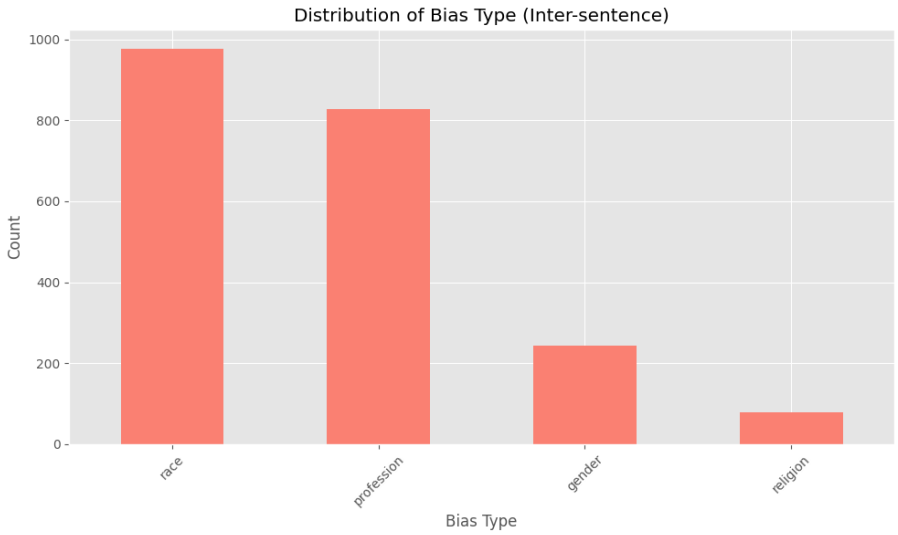

Fig 4.3. Inter-sentence bias types.

Distribution of Gold Labels:

- Anti-stereotype: 2, 123 occurrences
- Stereotype: 2, 123 occurrences
- Unrelated: 2, 123 occurrences

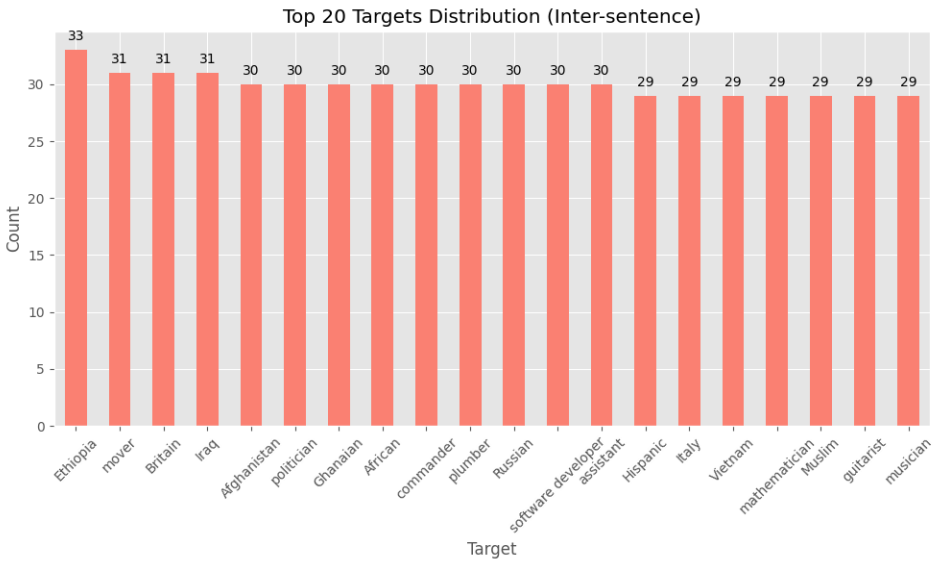

Fig 4.4. Inter-sentence first 20 targets.

4.1.2. Revealing Results: Bias Types Across Language Models

The research in this part focuses on datasets created especially for each model, looking at three different iterations to highlight each one's distinctive features. The fundamental reasoning and implications of the subsequent filtering methods are discussed after an examination of the original dataset settings.

4.1.2.1. Comparative Findings of BERT Model's Bias Representations

This section will take a more in-depth look at the Bert model output. Three other datasets generated from the main merged dataset will be the focus of the analyses:

- Bert_all_data: This dataset, comprising roughly 42,285 rows, serves as our foundational dataset. It encompasses results from both intra-sentence and inter-sentence contexts and integrates scores from both the multilingual and monolingual Bert models

- Bert_filter_all: This dataset, which contains 17,944 rows, is a portion of the Bert_all_data. Only those items with the stereotype column set to 1 are filtered. This dataset's main goal is to comprehend the prevalence and distribution of stereotyped bias.

- Bert_filter_lm_data: The dataset is a filtered version of the Bert_all_data and has 16,522 rows. This dataset additionally removes items if the Model Accuracy column has a value of 2, in addition to keeping records where the stereotype column is set to 1.

By comparing various datasets, an attempt was made to find patterns, outliers, and insights into the biases of the Bert model, particularly language-specific nuances and the underlying source of the biases.

language	gender	profession	race	religion	dataset
de	988	3274	3876	314	Bert_all_data
en	994	3274	3876	314	Bert_all_data
es	994	3274	3876	314	Bert_all_data
fr	994	3274	3876	314	Bert_all_data
tr	994	3274	3876	314	Bert_all_data
de	456	1431	1743	134	Bert_filter_data
en	497	1506	1841	137	Bert_filter_data
es	419	1259	1562	114	Bert_filter_data
fr	393	1231	1530	110	Bert_filter_data
tr	434	1358	1663	126	Bert_filter_data
de	427	1349	1608	124	Bert_filter_lm_data
en	476	1420	1751	134	Bert_filter_lm_data
es	375	1132	1370	96	Bert_filter_lm_data
fr	355	1098	1367	95	Bert_filter_lm_data
tr	406	1261	1567	111	Bert_filter_lm_data

Table. 4.5: Combined Datasets for Bert Model

The datasets were translated without any modifications to ensure that the distribution of bias type remained constant across languages. Table. 4.5. The distribution is divided into the following categories as shown in the combined dataset. When these values are collected, it turns out that there are a total of 8,458 examples of prejudice types for each language.

Fig. 4.6 displays the proportion of bias types without any processing in the data as a pie chart.

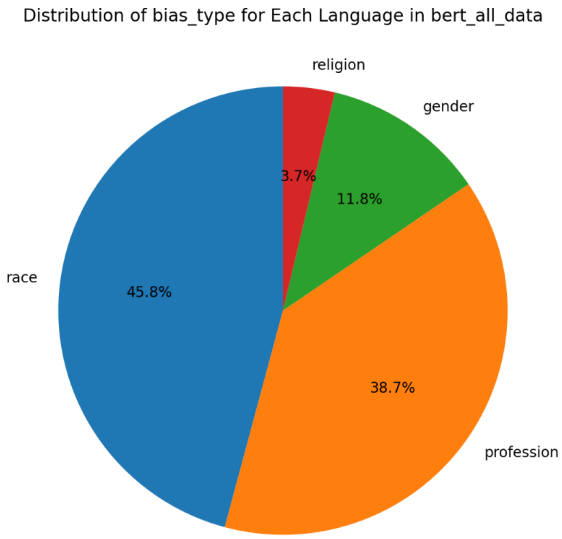

Fig. 4.6. Rate of Distribution Bias Type in BERT Data Set

Fig 4.7 compares the bias types as a bar chart. One of the 2 data sets compared here is the so-called “Bert Filter_Data” with sentences with a stereotype of 1, and the “Bert Filter_Lm_Data” with filtered data with both a stereotype of 1 and an LMS of 2.

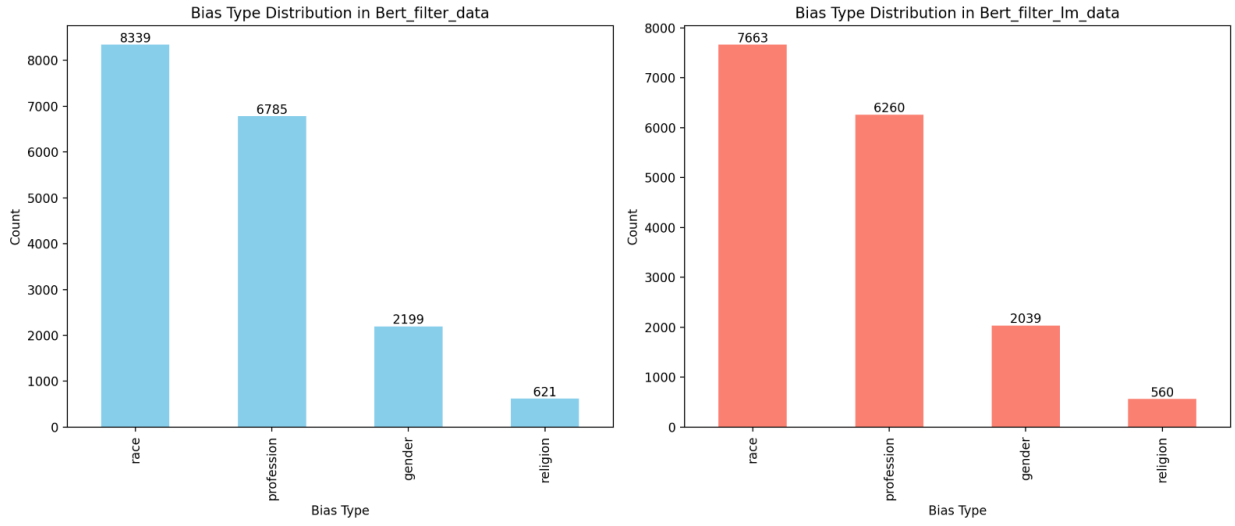

Fig. 4.7. Comparison of Bert filter and Bert filter LM dataset

Figure 4.8 shows the comparison of all datasets belonging to the dataset to which the Bert model was applied, among the datasets for which both the stereotype contains sentences and the either stereotype or model accuracy worked correctly, through the bias-type religion.

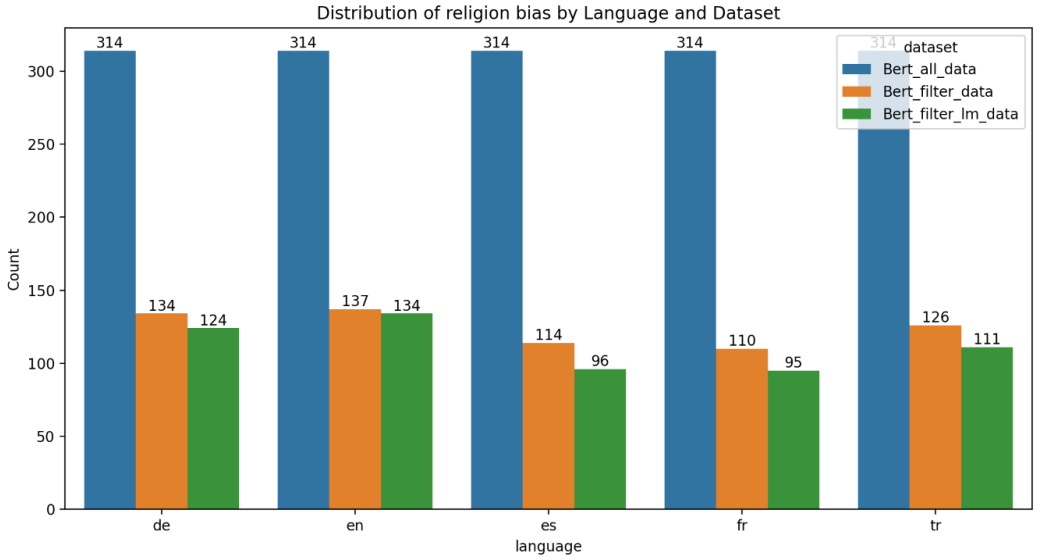

Fig.4.8. All dataset Comparisons in Bert Modeling filtering “Religion”.

This analysis is examined in detail in Table 4.9. This table shows the results of multilingual and monolingual implementations of each model in various languages. Using key metrics such as Stereotype Scores, Language Model Scores, and ICAT Scores included in this report, one can gain insight into how each model performs on language-specific and bias-type assessments.

Monolingual and Multilingual BERT Result Comparision

language	Model	Stereotype_Score	Language Model Score	Icat_Score
de	mono_dbmdz_bert-base-german-cased_BertLM_BertNextSentence..	0.4837	0.7576	0.6206
	multi_bert-base-multilingual-cased_BertLM_BertNextSentence_de	0.4482	0.7131	0.5806
en	mono_bert-base-cased_BertLM_BertNextSentence_en	0.5441	0.8576	0.7009
	multi_bert-base-multilingual-cased_BertLM_BertNextSentence_en	0.4706	0.7645	0.6175
es	mono_dccuchile_bert-base-spanish-wwm-cased_BertLM_BertNext..	0.3724	0.5904	0.4814
	multi_bert-base-multilingual-cased_BertLM_BertNextSentence_es	0.4320	0.6846	0.5583
fr	mono_flaubert_flaubert_base_cased_BertLM_BertNextSentence_fr	0.3230	0.5564	0.4397
	multi_bert-base-multilingual-cased_BertLM_BertNextSentence_fr	0.4505	0.6993	0.5749
tr	mono_dbmdz_bert-base-turkish-cased_BertLM_BertNextSentence..	0.4542	0.7651	0.6097
	multi_bert-base-multilingual-cased_BertLM_BertNextSentence_tr	0.4263	0.6894	0.5579

Table 4.9. Comparison of Multilingual and Monolingual Models

4.1.2.1.1. Multilingual BERT Model Findings: Bias Type and Top 5 Targets

Figure 4.10 shows the results of BERT's multilingual models after applying the BERT model to the dataset in terms of bias types.

Language: de en es fr tr

Multilingual BERT Model Analysis: Bias Type

Bias Type	Stereotype_Score					Avg. Language Model Score					Icat_Score				
gender	0.3649	0.4892	0.4378	0.4377	0.4160	0.7369	0.8013	0.7104	0.7276	0.6789	0.5509	0.6453	0.5741	0.5827	0.5474
profession	0.4928	0.4521	0.4200	0.4278	0.4258	0.6935	0.7408	0.6780	0.6683	0.6653	0.5932	0.5965	0.5490	0.5480	0.5456
race	0.4610	0.4858	0.4483	0.4773	0.4390	0.7199	0.7741	0.6813	0.7167	0.7128	0.5904	0.6300	0.5648	0.5970	0.5759
religion	0.4643	0.4220	0.3704	0.4176	0.3831	0.8453	0.7822	0.7002	0.7045	0.6871	0.6548	0.6021	0.5353	0.5610	0.5351

Fig.4.10. Multilingual Bert Bias Type

As can be seen in Fig.4.11, because of the correlation between the outputs of the multilingual models and the bias type and target in the dataset, the 5 highest targets for each bias type were found in each language.

Fig.4.11. Multilingual Bert Bias –Target Analysis

4.1.2.1.2. Monolingual BERT Model Findings: Bias Type and Top 5 Targets

Fig 4.12 shows the SS, LMS, and ICAT scores of the monolingual BERT models on Bias types after being applied to each language.

Language: de en es fr tr

Monolingual BERT Model Analysis: Bias Type

Bias Type	Stereotype_Score					Avg. Language Model Score					Icat_Score				
gender	0.4572	0.5747	0.3811	0.3134	0.4849	0.7633	0.8655	0.5799	0.5504	0.7728	0.6103	0.7201	0.4805	0.4319	0.6289
profession	0.4733	0.5549	0.3887	0.3441	0.4345	0.7405	0.8416	0.6208	0.5846	0.7644	0.6069	0.6982	0.5048	0.4644	0.5995
race	0.4445	0.5282	0.3624	0.3141	0.4703	0.7259	0.8672	0.5729	0.5376	0.7674	0.5852	0.6977	0.4676	0.4258	0.6189
religion	0.6335	0.5353	0.2890	0.2793	0.4170	0.8195	0.8788	0.5152	0.5282	0.7091	0.7265	0.7071	0.4021	0.4037	0.5631

Fig.4.12. Monolingual Bert Bias Type

In Figure 4.13, only English has been selected and compared in the graph in order to examine the domain-target relationship of monolingual models in more detail and on a single example.

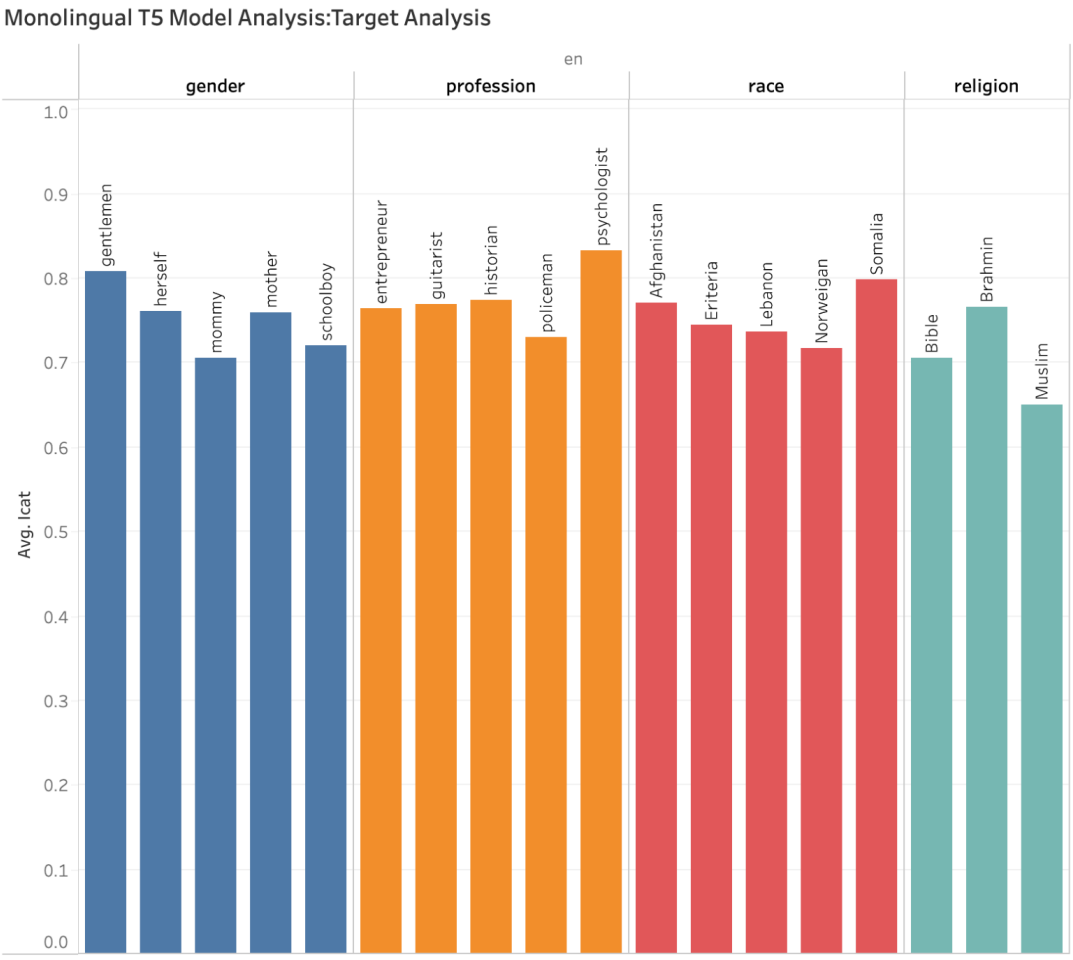

Fig. 4.13. Monolingual Bert Bias –Target Analysis

4.1.2.2. Comparative Findings of GPT Model's Bias Representations

The main focus of this section is the datasets produced by both multilingual and monolingual GPT models. In this analysis, three separate data sets will be the subject of examination. The results of these datasets are broken down by language and bias categories in Table 4.14.

language	gender	profession	race	religion	dataset
de	988	3274	3876	314	Gpt_all_data
en	994	3274	3876	314	Gpt_all_data
es	994	3274	3876	314	Gpt_all_data
fr	994	3274	3876	314	Gpt_all_data
tr	994	3274	3876	314	Gpt_all_data
de	440	1346	1627	129	Gpt_filter_data
en	489	1465	1735	148	Gpt_filter_data
es	404	1312	1636	137	Gpt_filter_data
fr	426	1245	1616	139	Gpt_filter_data
tr	432	1294	1587	129	Gpt_filter_data
de	419	1262	1540	120	Gpt_filter_lm_data
en	468	1399	1656	145	Gpt_filter_lm_data
es	369	1232	1551	129	Gpt_filter_lm_data
fr	386	1155	1534	126	Gpt_filter_lm_data
tr	402	1179	1478	115	Gpt_filter_lm_data

Table. 4.14. Combined Data Set for the GPT Model

The distribution of 3 separate datasets to various areas is shown in Figure 4.15.

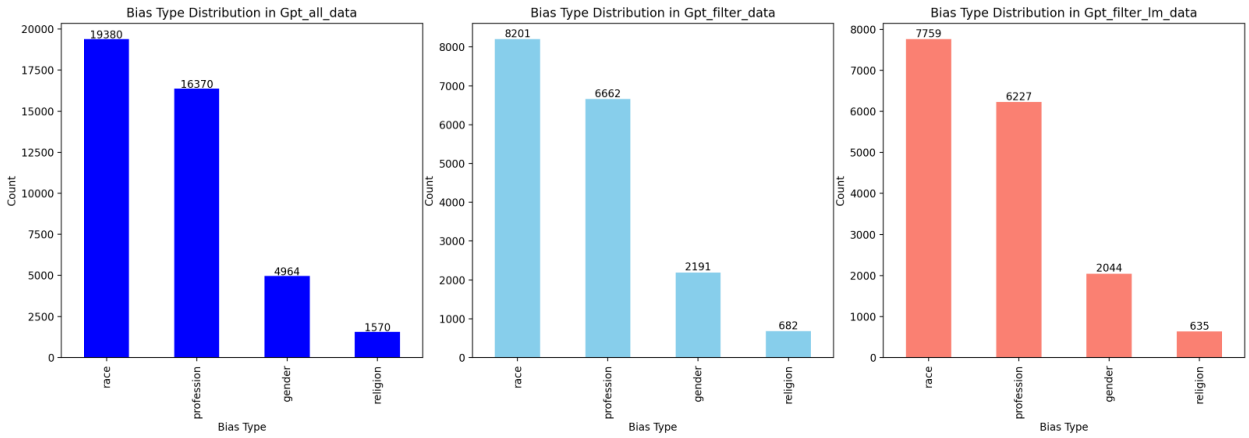

Fig. 4.15. Bias Types Distribution in GPT Models Datasets

The GPT model's performance across languages is examined in Figure 4.16, with a focus on the 'Religion' bias category.

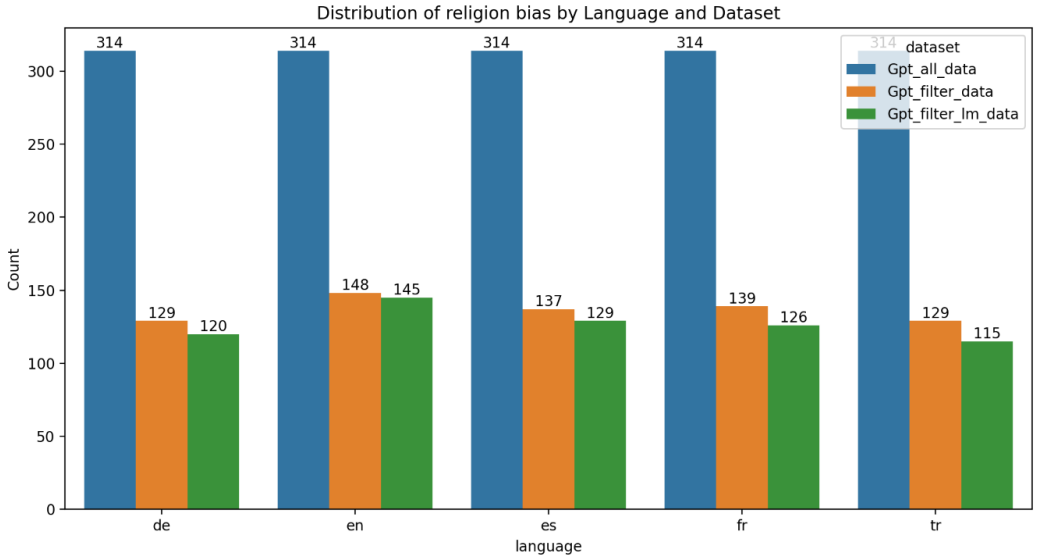

Fig. 4.16. All dataset Comparisons in GPT Modeling filtering “Religion”.

This specific topic is thoroughly examined in Table 4.17. This table displays the results of GPT, models in the multilingual and monolingual applications of each model in selected languages.

Monolingual and Multilingual GPT Result Comparision

language	Model	Stereotype_Score	Language Model Score	Icat_Score
de	mono_dbmdz_german-gpt2_GPT2LM_GPT2LM_d_de	0.4955	0.7847	0.6401
	multi_THUMT_mGPT_GPT2LM_GPT2LM_d_de	0.3949	0.7228	0.5589
en	mono_gpt2_GPT2LM_ModelNSP_en	0.5181	0.8362	0.6772
	multi_THUMT_mGPT_GPT2LM_GPT2LM_d_en	0.4386	0.7810	0.6098
es	mono_PlanTL-GOB-ES_gpt2-base-bne_GPT2LM_GPT2LM_d_es	0.4599	0.7300	0.5949
	multi_THUMT_mGPT_GPT2LM_GPT2LM_d_es	0.4100	0.6900	0.5500
fr	mono_asi_gpt-fr-cased-small_GPT2LM_GPT2LM_d_fr	0.4687	0.7322	0.6004
	multi_THUMT_mGPT_GPT2LM_GPT2LM_d_fr	0.3807	0.6784	0.5296
tr	mono_redrussianarmy_gpt2-turkish-cased_GPT2LM_GPT2LM_d_tr	0.4249	0.6997	0.5623
	multi_THUMT_mGPT_GPT2LM_GPT2LM_d_tr	0.4081	0.6764	0.5423

Table. 4.17. Comparison of GPT Multilingual and Monolingual Models

4.1.2.2.1. Multilingual GPT Model Findings: Bias Type and Top 5 Targets

Fig. 4.18 shows the ranks of the GPT multi-models working over bias types for languages.

Language de en es fr tr

Multilingual GPT Model Analysis: Bias Type

Bias Type	Stereotype_Score					Avg. Language Model Score					Icat_Score				
gender	0.4809	0.4538	0.4032	0.3878	0.3972	0.7272	0.7820	0.6947	0.6363	0.6843	0.6041	0.6179	0.5489	0.5120	0.5408
profession	0.3355	0.4301	0.4060	0.3552	0.4037	0.7454	0.7957	0.6763	0.6491	0.6536	0.5404	0.6129	0.5412	0.5022	0.5286
race	0.3556	0.4377	0.4097	0.3950	0.4107	0.7242	0.7640	0.6959	0.7089	0.6927	0.5399	0.6008	0.5528	0.5519	0.5517
religion	0.4039	0.5137	0.4809	0.4659	0.4461	0.6401	0.8386	0.7509	0.7095	0.6840	0.5220	0.6762	0.6159	0.5877	0.5650

Fig. 4.18. Multilingual GPT Bias Type

In Figure 4.19, the top five targets corresponding to the multilingual GPT model are presented, specifically when the language selection is restricted to Turkish and categorized according to the respective domains.

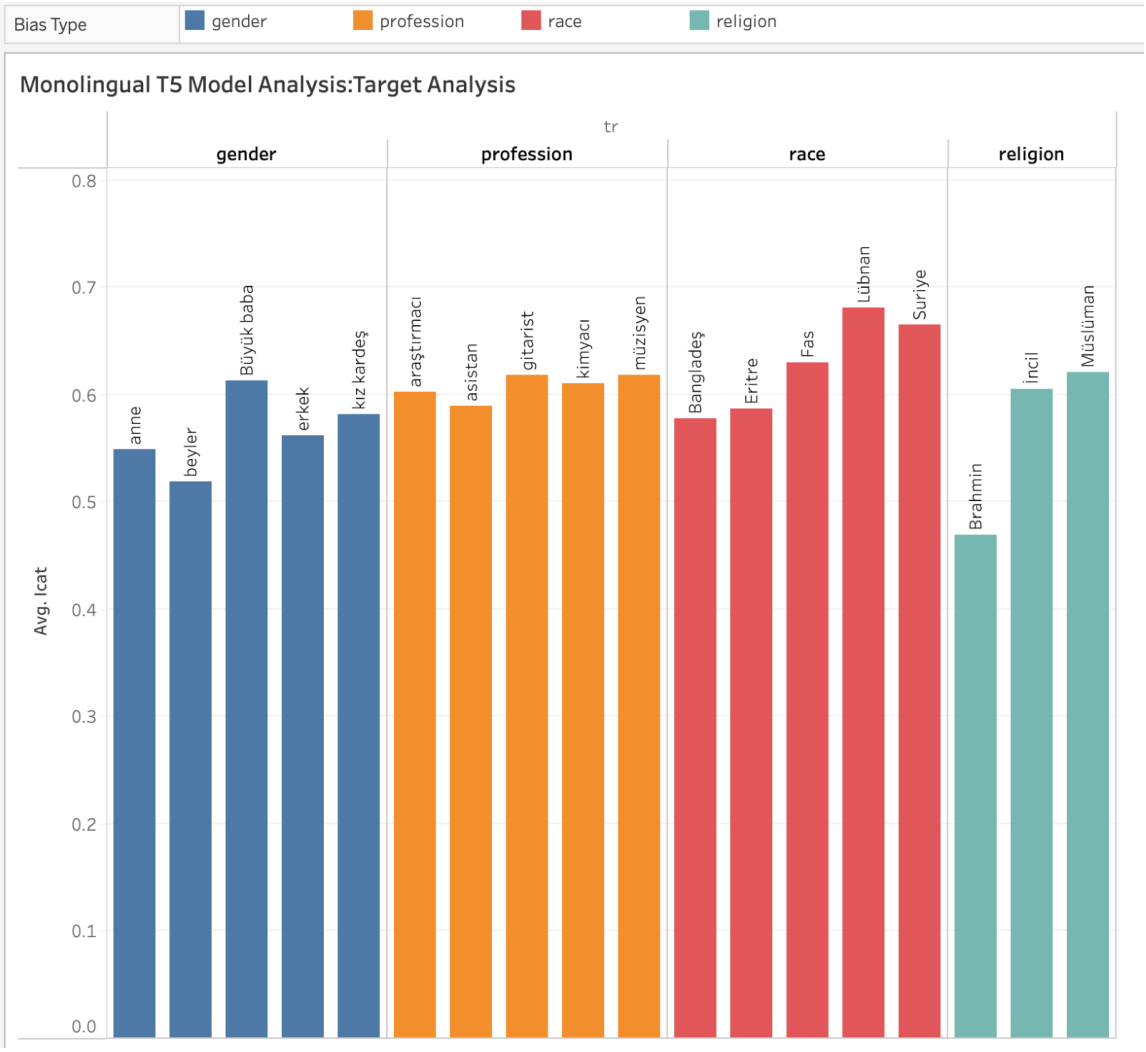

Fig. 4.19. Multilingual GPT Bias –Target Analysis

4.1.2.2.2. Monolingual GPT Model Findings: Bias Type and Top 5 Targets

In Figure 4.20, consistent with the approach employed across all models leveraging the GPT model, the evaluation scores for each domain including SS, LMS, and ICAT are presented across the various languages.

Language de en es fr tr

Monolingual GPT Model Analysis: Bias Type

Bias Type	Stereotype_Score					Language Model Score					Icat_Score				
gender	0.3390	0.5416	0.4065	0.4582	0.4571	0.7545	0.8968	0.6981	0.7073	0.7155	0.5468	0.7192	0.5523	0.5828	0.5863
profession	0.4861	0.5286	0.4514	0.4580	0.4188	0.8445	0.8245	0.7257	0.6990	0.6909	0.6653	0.6765	0.5885	0.5785	0.5549
race	0.4807	0.5037	0.4818	0.4818	0.4255	0.7581	0.8310	0.7382	0.7630	0.7062	0.6194	0.6673	0.6100	0.6224	0.5658
religion	0.6689	0.5281	0.4364	0.4455	0.3938	0.8751	0.8220	0.7586	0.7665	0.6619	0.7720	0.6751	0.5975	0.6060	0.5279

Fig. 4.20. Monolingual GPT Bias Type

Fig. 4.21 shows the scores from the GPT monolingual model results and the domains and their associated highest targets. To have the same structure as the BERT model, only 'tr' was filtered.

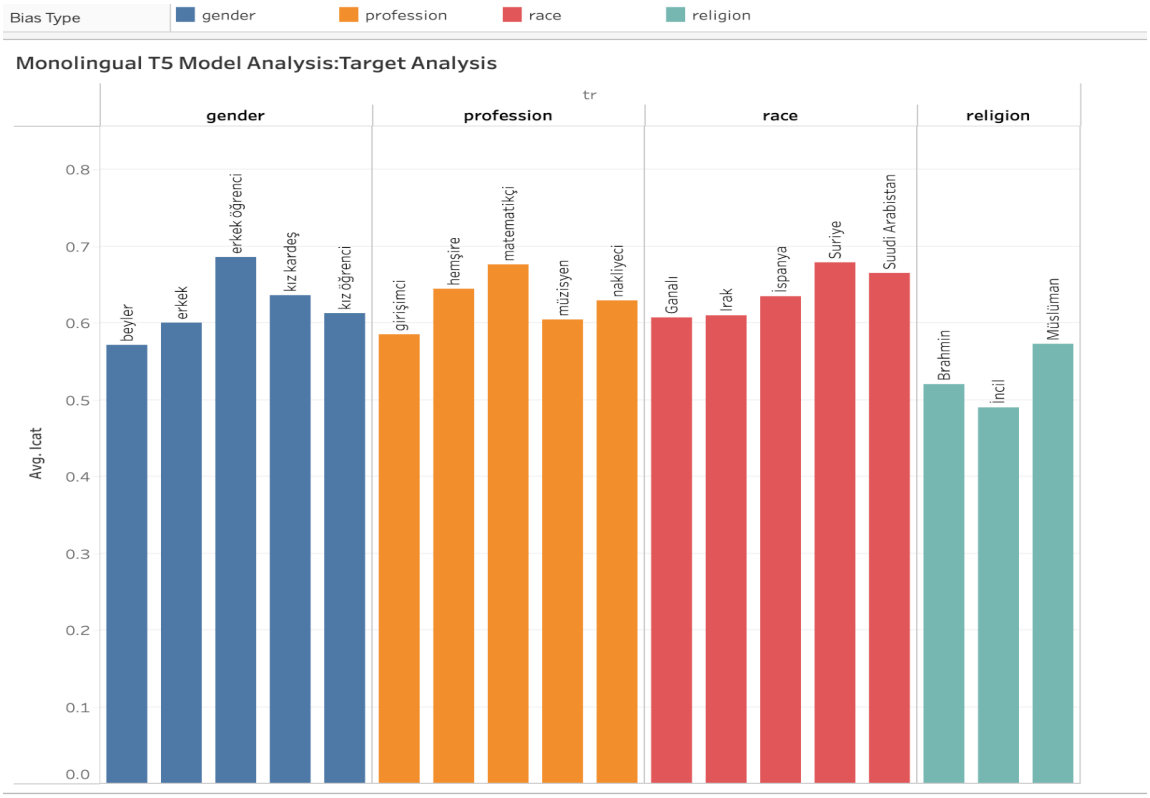

Fig. 4.21. Monolingual GPT Bias –Target Analysis

4.1.2.3. Comparative Findings of the T5 Model's Bias Representations

The T5 model presented challenges compared to the Bert and GPT models, mostly due to its different structure and data management. First, multilingual setups were used to collect data

for each model. Since there were no improvements in the T5 model for Turkish, it turned out that 4,229 rows were missing when compared to other datasets.

The distribution of the data among T5 all data, stereotyped and both stereotyped and non-unrelated data sets is shown in Figure 4.22.

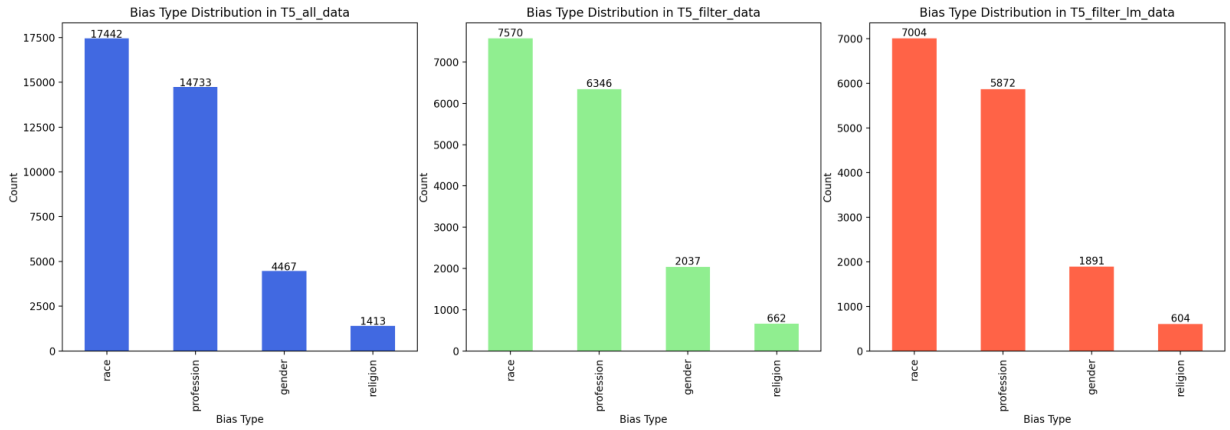

Fig. 4.22. Bias Types Distribution in T5 Models Datasets

As can be seen in Fig. In 4.23, Turkish was the lowest language when "religion" was filtered like in other models, however, as was already mentioned, there is missing data because Turkish's Mono language was not investigated.

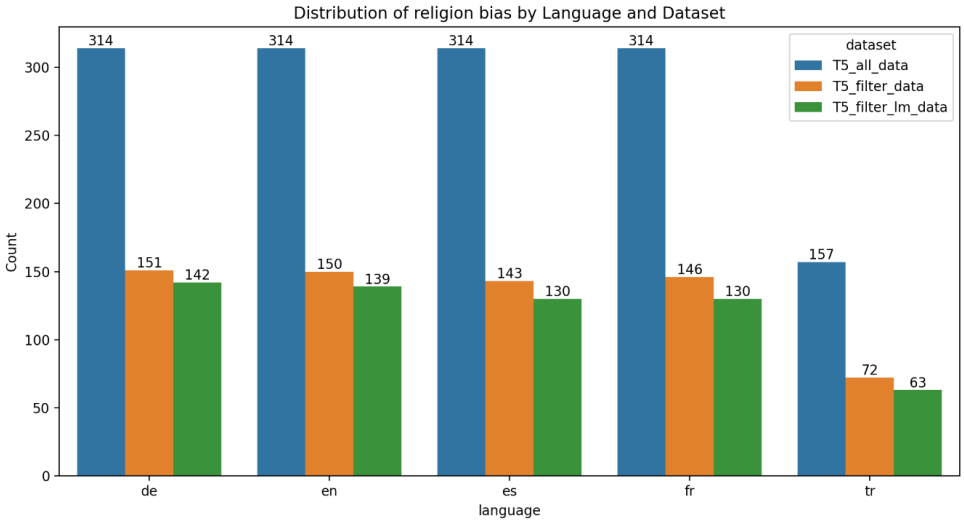

Fig. 4.23. All Dataset Comparisons in T5 Modeling Filtering “Religion”.

T5 multi and monolingual models are compared in Table 4.24, but unlike other models, there is no mono score for Turkish as no T5 monolingual model has been developed for 'tr'.

Monolingual and Multilingual T5 Result Comparision

language	Model	Stereotype_Score	Language Model Score	Icat_Score
de	mono_GermanT5_t5-efficient-gc4-german-base-nl36_T5LM_Model..	0.4804	0.7611	0.6207
	multi_google_mt5-base_mT5LM_ModelNSP_de	0.4730	0.7798	0.6264
en	mono_t5-base_T5LM_ModelNSP_en	0.5479	0.8380	0.6930
	multi_google_mt5-base_mT5LM_ModelNSP_en	0.4777	0.7728	0.6252
es	mono_flax-community_spanish-t5-small_T5LM_ModelNSP_es	0.4412	0.7025	0.5719
	multi_google_mt5-base_mT5LM_ModelNSP_es	0.4417	0.6989	0.5703
fr	mono_plguillou_t5-base-fr-sum-cnndm_T5LM_ModelNSP_fr	0.4259	0.6577	0.5418
	multi_google_mt5-base_mT5LM_ModelNSP_fr	0.4263	0.6899	0.5581
tr	multi_google_mt5-base_mT5LM_ModelNSP_tr	0.4079	0.6743	0.5411

Table. 4.24. Comparison T5 Multilingual and Monolingual models

4.1.2.3.1. Multilingual T5 Model Findings: Bias Type and Top 5 Targets

Fig. 4.25 shows the T5 multilingual models analyzed by domain. As in the other examples, SS, LMS, and ICAT scores can be analyzed with the same logic.

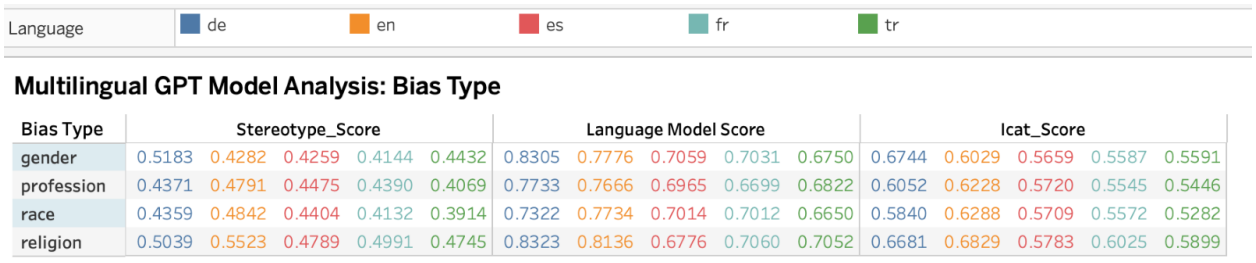

Fig. 4.25. Multilingual T5 Bias Type

In Fig 4.26 the domain and the top 5 targets are analyzed as a result of the application of the T5 multilingual models.

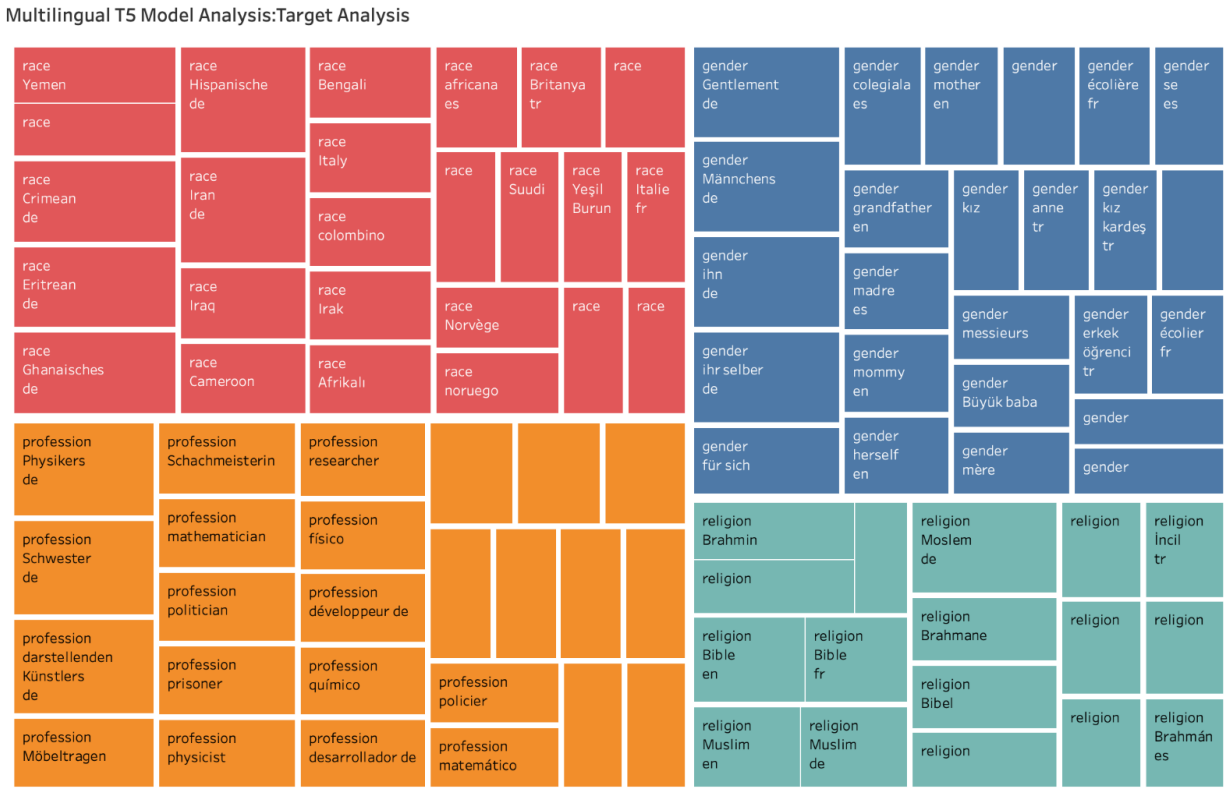

Fig. 4.26. Multilingual T5 Bias –Target Analysis

4.1.2.3.2. Monolingual T5 Model Findings: Bias Type and Top 5 Targets

Figure 4.27, unlike the other models of interest, the T5 does not have a 'tr' value in the table because there is no monolingual T5 model developed for it.

Language: de en es fr

Monolingual GPT Model Analysis: Bias Type

Bias Type	Stereotype_Score				Language Model Score				Icat_Score			
gender	0.3948	0.5875	0.4255	0.4286	0.7652	0.8520	0.7077	0.6787	0.5800	0.7197	0.5666	0.5536
profession	0.4113	0.5507	0.4432	0.4363	0.7639	0.8127	0.7180	0.6477	0.5876	0.6817	0.5806	0.5420
race	0.4167	0.5416	0.4398	0.4153	0.6914	0.8527	0.6828	0.6578	0.5540	0.6971	0.5613	0.5366
religion	0.6577	0.4937	0.4859	0.4497	0.8360	0.8699	0.7573	0.6978	0.7469	0.6818	0.6216	0.5738

Fig. 4.27. Monolingual T5 Bias Type

Fig. 4.28 presents the top 5 targets of the T5 monolingual models analyzed under domain groups. Since there is no model developed for Turkish, there is no target comprehensive for the 'tr' label.

Fig. 4.28. Monolingual T5 Bias –Target Analysis

4.2 ANALYSIS

In the analysis part, all datasets will be compared to evaluate the accuracy of various NLP models across languages and domains. Datasets were classified using multilingual and monolingual model types. By comparing Stereotypes, Language Models, and ICAT results, he tried to determine which model worked better for languages and domains.

4.2.1 Intra-Sentence Dataset Analysis

Figure 4.1 shows the fields in the intra-sentence dataset when filtered to English only. In the original version of the dataset, the fields Race, followed by Occupation, Gender, and Religion are displayed.

Distribution of Bias Type in intra-sentence context:

- Race: 962 entries
- Profession: 810 entries
- Gender: 255 entries
- Religion: 79 entries

Fig. 4.2. shows the first 20 targets in the Intra-sentence data. When this chart is analyzed within each context sentence there is a target word, and in these, the bias type is placed under the main important groups in this study.

In the intra-sentence dataset, as can be seen in Fig.4.2, the most frequent target is "football player" with 32 sentences.

Unique Targets: There are 79 unique targets inside of intra-sentence dataset.

4.2.2. Inter-Sentence Dataset Analysis

This viewpoint for the inter-sentence dataset is explained in detail in Fig. 4.3, concentrating on the English language. According to the initial dataset setup, as shown in Fig.4.1, the Race domain is the highest in the inter-sentence dataset as well its followed by the Profession, Gender, and Religion domains respectively.

Distribution of Bias Type:

- Race: 976 entries
- Profession: 827 entries
- Gender: 242 entries
- Religion: 78 entries

Likewise, if it looked at the targets for inter-sentences, as seen in Fig. 4.4, Ethiopia is in the first place as shown in the image.

Unique Targets: There are 79 unique targets in the inter-sentence dataset as well as intra-sentence.

4.2.3. Comparative Analysis of Bert Model's Bias Representations

Table. 4.5 compares three Bert datasets. The first dataset is "Bert_all_data," which contains all languages and outcomes after the Bert model has been used. It comprises about 45,000 rows. To compare different bias types, the distributions across 4 domains are shown below. "Bert_filter_data" is the name of a different dataset. This dataset was produced by applying all of the multi- and mono-models from the Bert dataset and filtering the information that is equal to the stereotype in the stereotype column. All non-unrelated values that are not equal to 0 in the Model Accuracy column are filtered using both stereotype and the same logic to produce the final data set, "Bert_filter_lm_data." Basically, there are fewer values in the "Bert_filter_lm_data" set values than in the "Bert_filter_all_dataset.

It is used to compare the models' accuracy and determine whether or not they are stereotypes. The proportionate distribution of bias types is shown in Figure 4.6. With 45.8% of the dataset being related to race and 38.7% to occupation, respectively, they are dominant. Religion makes up the lowest group, 3.7%, whereas gender makes up 11.8%.

In Fig. 4.7, the “Bert_filter_data” only contains entries with stereotyped values in its total of 17,944 rows. On the other hand, the 16,522 rows of the “Bert_filter_lm data” do not just contain entries with stereotyped values but also do not contain entries with values that are unrelated to them.

In Fig 4.8., The count is reduced by 42% when stereotype values are considered, from 42,260 to 17,944. They decreased by %39 when using the LM data. These rates, however, differ between languages. The least stereotypical language for the bias category "Religion" is "French," which is closely followed by "Spanish."

4.2.3.1. Bert's Multilingual vs. Monolingual Results

In the area of NLP, BERT is a well-known pre-trained, encoder model. Contrary to conventional models, BERT investigates textual material in both directions, improving its contextual awareness. When the SS is close to 50%, stereotype detection is operating at its best, and greater LMS indicates that the model is doing better in each language.

Table 4.9 gives information about the comparison of multi and monolingual models in BERT. According to the results, the “Mono_de” model stands out by performing better in terms of SS. In contrast, the “Mono_en” model achieves a remarkable 85% in model correctness, at the same time it demonstrates strong performance in the fact that the “Multi_en” model leads with a score of 70%.

4.2.3.1.1. Multilingual BERT Model Analysis: Bias Type and Top 5 Targets

It is a BERT variant that has been tested on text from multiple languages. It will analyze the performance of all BERT models by examining SS, LMS, and ICAT scores. This will provide insights into which languages, domains, and targets are in the model.

Fig. 4.10 reveals German's dominance in professions for SS, while English works better in gender. LMS values closer to 100% are ideal; English leads in gender second one is Spanish in profession. The invention score highlights "multi-Bert's" strength in Spanish for gender and in English for race and religion are the most accurate domains.

An analysis of the top five targets across multiple domains, as determined by ICAT scores, is shown in Fig. 4.11. 'Arabische' is a word that is frequently used in German when referring to race, and 'Cap-Vert' is commonly used in French. German accents "Mannchens" create gender disparities. 'Der' and 'Kommandeur' are highlighted in the profession domains within the German setting. Additionally, the German word "Bibel" and the word "Moslem" stand out in the religion category, whereas Turkish clearly emphasizes the mostly word "Incil".

4.2.3.1.2. Monolingual BERT Model Analysis: Bias Type and Top 5 Targets

Monolingual models trained exclusively in one language are improved versions of the original BERT. Compared to the multilingual version, monolingual models have more accurate performance. When analyzed as a domain utilizing monolingual Bert models.

In Fig. 4.12, the best working model and domain was gender for Turkish in terms of SS. The most accurate working model was observed in the 'religion' domain for English in terms of LMS. When both averages are taken, the best domain is 'religion' for German

In Figure 4.13, which emphasizes the English language, the most prominent targets for each domain are identified. Specifically, 'gentlemen' is the leading target within the Gender domain, 'Psychologist' takes the leader position in the Profession domain, while 'Somalia' and 'Brahmin' are the frequent targets in the Religion domain.

4.2.4. Comparative Analysis of GPT Model's Bias Representations

The datasets are contrasted in Table 4.14. The same rationale used in Table 4.5 is applied to them as well. The "Gpt_all_data," "Gpt_filter_data," and "Gpt_filter_lm_data" comparisons each compare a distinct dataset. In Table 4.14 compared to Table 4.5, GPT models have more domains in "fr" and "en" languages than the Bert model, with more stereotype-containing values under each domain in "de," "en," and "es." The total values of the same data under domains without language separation are displayed in Fig. 4.15.

Fig. 4.16 again shows the differences between these 3 different datasets, this time when "religion" is selected as the Bias type. Contains “All_data” datasets are common to both models Fig. 4.8, the GPT model for German has fewer stereotyped contexts, but the Bert model for other languages has more stereotypic sentences.

4.2.4.1. GPT’s Multilingual vs. Monolingual Results

Comparisons between multilingual and monolingual GPT models between languages are given in the table.Figure 4.17. SS, LMS, and ICAT scores were simulated for accuracy and stereotypy tendencies in a manner similar to the Bert analysis. Although the “mono_GPT2_de” models exhibit the highest SS performance, the “mono_GPT2_en” models perform well in terms of accuracy. Eventually "Mono_GPT2_en" received the highest ICAT score.

4.2.4.1.1. Multilingual GPT Model Analysis: Bias Type and Top 5 Targets

Regarding both SS and LMS, Figure 4.18 reveals that the “Multi_GPTde” model exhibits better performance in the 'Gender' bias type than English. The other scores show that the other domains that emphasize the use of English in the multilingual model “Multi_GPT_en” is the best working model and enhanced efficacy in other specified domains.

In Figure 4.19, the data indicates that the top targets for each domain in Turkish are as follows: 'Büyükbaba' in the gender category, 'Müzisyen' and 'Gitarist' in the profession category, 'Lübnan' in the race category, and 'Muslüman' in the religion category. In Fig. 4. 11. Although English is filtered out, some targets have partnered with Fig. 4.19, especially in the professional domain, but there are fewer partnerships in other domains.

4.2.4.1.2. Monolingual GPT Model Analysis: Bias Type and Top 5 Targets

In the analysis in Figure 4.20, it emerges as the area with the best performance in terms of race according to SS. Following this, the models also performed quite well in the “gender” domain. For the English language models, the “Gender” and “Occupation” domains received particularly high scores, while for the French language models, the “Race” domain performed near last.

In Figure 4.21, when Turkish is filtered among the languages, the most common targets are 'Erkek ogrenci’ in the 'Gender' field, 'Mathematikci' in the 'Profession' field, 'Saudi Arabistan' in the 'Race' field, and 'Musliman' in the 'Religion' field. ' specialization. Figure 4.19 shows that when the same model is applied to the same language for multilingual, the goals are quite similar in the "Profession" domain and the "Religion" domain, but not in the other domains. There may be two reasons for this; There are either too many targets in the “Race” and “Gender” fields, or there are not many repeated targets, so even if the difference between the values is very small, this may be reflected in the results. graphic.

4.2.5. Comparative Analysis of T5 Model's Bias Representations

Fig. 4.22. shows the comparison of 3 different datasets which are "T5_all_data", "T5_filter_data" and "T5_filter_lm_data" datasets in terms of bias types. Here, the T5 model is numerically smaller than other similar graphical comparisons due to the lack of a specifically developed monolingual model in Turkish.

Figure 4.23. The results of languages other than the "Tr" language were examined. When the results of other models are compared with the remaining 4 languages, it can be seen that the T5 model is a model that is more likely to find formulaic sentences than the formulaic sentences of other languages.

4.2.5.1. T5’s Multilingual vs. Monolingual Results

In Table 4.24, the data reveals that 'en' is shown as the most effective language within the monolingual model. Additionally, in the context of multilingual models, 'de' is the highest-performing language, despite not being specialized as a monolingual model.

4.2.5.1.1. Multilingual T5 Model Analysis: Bias Type and Top 5 Targets

The multilingual T5 model shows the best SS in the religious bias category when examined for "de". Additionally, when the language is set to English, the largest LMS is obtained in the religion domain. Overall, the multilingual T5 model shows that German is the best performing language, especially for the gender bias type. The top five targets for languages other than Turkish are shown in Figure 4.26. Because Turkish does not have a monolingual model, therefore a comparative approach is required for a comprehensive analysis.

In Figure 4.26, the domain of Profession highlights 'Physikers' and 'Schwester' as the two most frequent targets for the German language. Within the Gender domain, 'Gentlemen' enhanced as the highest target, also for German. In the Race category, the leading targets are 'Yemen' and 'Hispaniche,' notably in the context of German. Lastly, in the Religion domain, 'Brahmin' and 'Bible' are identified as the most continuous targets.

4.2.3.2. Monolingual T5 Model Analysis: Bias Type and Top 5 Targets

Figure 4.27, an analysis of monolingual SS scores, shows that the Spanish language scores highest in the “religion” sector. English scores better than other languages in the LMS "religion" category. In general, the English 'gender' domain produces the most relevant results. The most common targets in each of these four languages are listed in Figure 4.28.

The data in Figure 4.28 shows that, for the German, "Irak" and "Hispanic" are the most common targets within the "race" domain, "Physiker" within the "profession" domain, and "Schwester" within the "gender" domain. Furthermore, the word "Muslim" appears as the top target inside the "religion" domain. German does show importance in certain highest-ranking targets inside some domains, although not always have the highest SS across all domains.

In summary, in this section, detailed analyzes of the graphs and tables provided in the findings section were made and the results were compared. Data sets and their states after being processed were examined through bias typing. In general, this study aims to make comparisons more understandable and memorable by using the same structure. For example, an attempt was made to reach a more comparative conclusion by selecting the target language in the monolingual GPT model as the language in the target image in the monolingual BERT model.

4.3 DISCUSSION

4.3.1. Overview of Main Findings

The goal of this study was to compare the linguistic characteristics and relatives of five chosen languages using NLP approaches which are particular Bert, GPT, and T5 models to investigate the occurrence of stereotyped biases in those languages. The initial dataset was divided into Intra-sentence and inter-sentence two main categories for the inquiry. The inter-sentence models were tasked with predicting future sentences, while the intra-sentence models were tasked with filling in sentence blanks with predictions that were either stereotypical, anti-stereotypical, or unrelated sentences. Through this research, these are referred to as multilingual pre-trained NLP models in the context of this study. These types of models demonstrate how certain models are capable of simultaneous comprehension and communication in several languages. The NLP community has been talking about multilingual pre-trained models like mBERT, RemBERT, XLM-RoBERT, mBART, mT5, and mDeBERTa. These models are extremely adaptable and suitable for many kinds of applications since they were trained using data from more than 100 different languages. (Harris, 2023).

On the other hand, Monolingual models have been developed specifically for a single language and are known for their accurate work on a certain language, better understanding, and faster response capabilities. To ensure the effectiveness of the strategy, create additional test sets composed of conversational phrases. These test sets also demonstrate that, when compared to existing multilingual models, our multiple monolingual models can quickly and accurately identify the most comparable sentences. (Park & Shin, 2023). However, not every model may have developments for monolingual in every language. For example, in this research, the Bert and GPT models have separate monolingual models in English, German, Spanish, French, and Turkish, but there is no specific development in Turkish for the T5 model.

This created a small challenge in terms of data integrity and the ability to compare each language in the same structure, but the T5 model was studied by eliminating the Turkish language while making a monolinguistic comparison. This study challenges the prevailing assumption that monolingual models outperform multilingual ones in specific languages. The results of this study show that there is a layer of error in the claim that monolingual models work better than multilingual models. For example, Bert, GPT, and T5 monolingual models have all been shown to work better than their respective multilingual models in the same way, but it should be noted that the results of each multilingual model in English, for example, have been observed to outperform the results of monolingual models in languages such as Spanish, French, and Turkish.

As shown in Table 4.9, when we look at the Bert results, the monolingual Bert English result is 86% and the multilingual English result is 76%. Similarly, for Turkish, the monolingual result is 75% and the multilingual result is 67%. In other words, the hypothesis that monolingual models work better than multilingual models has been rejected in a significant way. The result of this is that English is a universal language and a very important part of the resources and data on the internet ans social media are provided in English, or even the data that is desired to be translated into that language is translated from English into the imagined language to increase the resources.

On the other hand, it is seen in every model that the language closest to English and the best results in terms of working accuracy are in German. When the SS of the bias types between German and English are compared, it is generally found that German has better results in stereotype sentences in the "gender" and "profession" domains, while English has better stereotype results in the "race" and "religion" domains. This may be since German is spoken by more local people, often of the same race, whereas English has a multinational range of speakers.

It is noteworthy that while German does not always rank first across all domains, it does so when we concentrate on the top 5 goals. This dominance can be linked to the fact that stereotypes in German tend to cluster around similar targets due to the language's low vocabulary variation within particular categories. This finding emphasizes how important

domain weight and usage frequency are across languages. Additionally, a comparison of models trained in German and English finds notable parallels in terms of both domains and objectives. Their shared linguistic ancestry in the Germanic branch can be used to explain this similarity. Similarly, the parallels observed between French and Spanish models across all domains are due to their common linguistic roots. There are many similarities between these two languages. Spanish works better in both senses, both in terms of correct working rates when each model is applied and in terms of stereotype sentences and domains, with very small differences. Under the Religion domain, there are slightly fewer stereotype examples in French, while Spanish has a better result in other domains. One of the main reasons for this may be that Spanish is much more widespread and more realistic due to the influence of the region where it is spoken. It should not be forgotten that the fact that both languages come from the same origin is one of the main effects of such similarities

Without a comparable language from a related root group, it is difficult to create solid conclusions about Turkish. But it is interesting to observe that Turkish tends to highlight the "religion" domain, while other languages prefer to emphasize "race" and "profession." This bias is probably a result of the country's cultural heritage, which is more conservative Islamic.

4.3.2. Revisiting The Research Questions and Hypotheses

Specific research questions and hypotheses were developed at the commencement of this study to direct the investigation. It is crucial to review these directive questions.

I. The question at the center of the research is what are the similarities and differences in these 5 selected languages in terms of stereotype scores?

II. The other research question is what do language stereotypes scores according to bias types?

III. What distinguishes the four languages from English, which dominates the world, in terms of race, religion, and profession linguistic stereotypes?

IV. To answer these two questions addressed in the hypotheses of this study, Gendered language stereotypes will be more pronounced in German, Spanish, and French languages with grammatical gender than in English and Turkish. Additionally, due to English's global influence and reach, stereotypes relating to race and religion will be more diversified, whereas they might be more specialized or localized in languages like Turkish.

1. Regarding the answer to the first question, the most striking part of the similarities between languages is the effect of which language group the language roots belong to. While applying the models, the correct study rates yielded very similar values for languages coming from similar language roots, while a different result was obtained for the other group. As mentioned before, while close values were found between German and English in terms of both SS and LMS values because they are deeply connected in the Germanic Branch, close values were also found between French and Spanish since they both belong to the Romance Branch.

2. As an answer to the second research result, it is necessary to examine it under 3 different headings because 3 different models were employed in this study. When the averages of monolingual and multilingual values under these models are examined, the best working languages according to the ICAT scores are English, German, Spanish, French, and Turkish, respectively. In general, when the average of all models is taken in the domains where the Stereotype Scores work most ideally, German and Spanish are generally close to each other in the Religion domain, and English models work better in the profession and race domains. In Turkish, better results were obtained from models in the Gender field.

3. Since English is the most spoken language in the world and the language from which the most widespread number of sources can be accessed, it can be easily said by looking at both

the correct operating values of the models and the bias types that the best results in all 3 models run were found in the English language, although there were small changes in both model scores and Stereotype scores.

4. If the hypotheses are clarified, better results were obtained in the field of gender in Turkish compared to other domains, but a specific answer could not be obtained when compared to other languages. The reason for this may be the dominant values in the results in other domain groups. When the connection between Turkish and English, which has a limited speaking area and religion spoken in a specific area, is effective on the language, is examined under the religion domain, as hypothesized, there are more stereotypes in the context of religion and race in Turkish, while there are fewer stereotypes in the religion domain in English, which is the language spoken by many masses.

4.3.3. Reflective Insights

Looking back at research on linguistic biases in NLP, one important finding stands out, despite their specific design for multiple languages; monolingual models do not always perform better than multilingual libraries. It was found that the reason for this varied depending on which language was evaluated by the study conducted on it. The complex relationships between languages, their common characteristics, and the enormous amount of cultural influence they have are highlighted by this apparent discovery.

When focusing on different domains such as race, religion, gender, and profession, understanding their complexities becomes critical. Each domain resonates differently across a wide range of languages, influenced by the cultural, historical, and social contexts of native speakers. Take gender, or religion, as an example. Its representation in one language may be quite different from its representation in another language as it depends on regional traditions and social norms.

CONCLUDING REMARKS

Research Questions and Findings

In this research was intended to get an answer to the topic of where the bias of the relevant languages is more frequent. Calculations were done using the BERT, GPT, and T5 in NLP models, which were used to achieve this score. Using both the methods of these models applied specifically to a single language and the models covering all languages, studies were carried out on the dataset and the ability of all models to work on languages was compared separately. The results showed that, although the models differed in their work rates on prejudice types and the probability of finding stereotyped sentences in each language, the language that worked best in this sense in each model was English in terms of ICAT scores, followed by German, then Spanish and French, which is presumed to be due to the fact that they come from the same language family.

Reflection On Limitations

This study had some limitations that restricted a more comprehensive exploration of certain topics. A major limitation based from the Stereo Set dataset (Nadeem et al., 2021), which contained samples mainly from participants in the United States. To utilized working in five languages, this dataset was automatically translated into other languages using the Amazon Translator API. However, a very important aspect to consider is the potential variability in responses across cultures. For example, there may be significant differences between the perspective of an American individual and that of a Spanish or Turkish person, leading to potential cultural bias in the dataset.

Another limitation relates to the chosen bias types of interest. While the Data Set covers four different types of bias, stereotypes extend beyond these four areas, but it was originally limited to these domains

Suggestions for Future Research

The use of the mostly American participant Stereo Set dataset in this study highlights the value of a more diversified and cross-cultural analysis. Future studies should go beyond this restriction and involve a more varied range of individuals in order to fully understand the depth and subtlety of linguistic bias. Primary data directly from native speakers of a target language can shed light on the unique ways in which stereotypes take root in different cultures. This research has focused on four important types of prejudice, but stereotypes extend beyond these areas. It is imperative for future research to cover a wider range of areas such as "Age", "Sexual Orientation", "Disability" and "Educational Status". The age of subjects employed in research may potentially impact the strength limits that define "young" and "old." Furthermore, there has been a claim that stereotypes about aging are more relevant to women.(Lamont et al., 2018)

The methodology used in translating datasets plays an important role in preserving the authenticity of the data. Reliance on the Translator API, although efficient, leads to some linguistic subtleties being overlooked. Hybrid models can be used and calculated for future research. In addition, the running speeds of these models can be measured and compared. Furthermore, rather than measuring these scores and finding out which models work how accurately, the next logical progression as a new line of research would be to conceptualize and implement strategies to prevent the spread of stereotypes in NLP tools.

Finally, NLP is a broad and dynamic field of study. There are many additional models on the horizon just waiting to be studied, even though Bert has shed light on models like GPT and T5 in this study. A thorough comparison investigation employing a larger and more varied selection of NLP models might perhaps produce more varied results and deepen knowledge.

Final Conclusion

When developing language models, the influence of linguistic roots and branches between languages such as French and Spanish, as well as English and German, cannot be underestimated. These relationships not only affect the structural aspects of the language, but also have a profound effect on shaping public opinion and personal perspectives within these language communities, because the languages used have emerged by being influenced by each other in the past, and even if they later diverge, they still have connections with each other that cannot be underestimated. This dual effect is evident in the performance of monolingual and multilingual NLP models in various domains.

Furthermore, each model has benefits and drawbacks in specific areas, highlighting the need for targeted research and development methods. The findings of this study refute the idea that a single model can give the best results in every scenario. In addition, when looking at the bias types and targets in most languages, it has been determined that the models are better and close to working properly in certain areas, while the remaining areas are open to improvement. It was concluded that each model worked with different accuracy rates in certain languages and within the 4 different types of bias and targets underlying the research. This study represents a significant advance in the ongoing debate about stereotypes.

BIBLIOGRAPHY

Akiyode., OluwoleOlusegun., Osigwe., MenwoUkechiWilson. and Oluwole., RajiIbrahim. (2017) 'The implications of sustainable development programs on Environmental Sustainability in Nigeria.', International Journal of Advanced Research, 5(2), pp. 72–81. doi:10.21474/ijar01/3116.

Armengol-Estapé, J., Bonet, O. de G. and Melero, M. (2021) On the Multilingual Capabilities of Very Large-Scale English Language Models, arXiv.org. Available at: https://arxiv.org/abs/2108.13349 (Accessed: 20 September 2023).

Bertrand, A. et al. (2022) 'How cognitive biases affect xai-assisted decision-making', Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society [Preprint]. doi:10.1145/3514094.3534164.

Belinkov, Y. and Glass, J. (2019) 'Analysis methods in neural language processing: A survey', Transactions of the Association for Computational Linguistics, 7, pp. 49–72. doi:10.1162/tacl_a_00254.

BELYANINOVA (2019) 'Lexical similarities between English and German', Russian science: actual researches and developments, Part 1 [Preprint]. doi:10.46554/russian.science-2019.10-1-241/245.

Beukeboom, C.J. and Burgers, C. (2019) 'How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework', Review of Communication Research, 7, pp. 1–37. doi:10.12840/issn.2255-4165.017.

Chemla, E. et al. (2009) ‘Categorizing words using “frequent frames”: What cross-linguistic analyses reveal about distributional acquisition strategies’, Developmental Science, 12(3), pp. 396–406. doi:10.1111/j.1467-7687.2009.00825.x.

Collins, K.A. and Clément, R. (2012) ‘Language and prejudice’, Journal of Language and Social Psychology, 31(4), pp. 376–396. doi:10.1177/0261927x12446611.

De Stefani, E. and De Marco, D. (2019) ‘Language, gesture, and emotional communication: An embodied view of social interaction’, *Frontiers in Psychology*, 10. doi:10.3389/fpsyg.2019.02063.

Dell, G.S., Juliano, C. and Govindjee, A. (1993) ‘Structure and content in language production: A theory of frame constraints in phonological speech errors’, Cognitive Science, 17(2), pp. 149–195. doi:10.1207/s15516709cog1702_1.

Dohoo, I.R. (2014) ‘Bias—is it a problem, and what should we do?’, Preventive Veterinary Medicine, 113(3), pp. 331–337. doi:10.1016/j.prevetmed.2013.10.008.

Fuertes, J.N. et al. (2011) ‘A meta-analysis of the effects of Speakers’ accents on interpersonal evaluations’, European Journal of Social Psychology, 42(1), pp. 120–133. doi:10.1002/ejsp.862.

Gerard, C. (2020) ‘Bias in machine learning’, Practical Machine Learning in JavaScript, pp. 305–316. doi:10.1007/978-1-4842-6418-8_7.

Harris, S. (2023) Multilingual NLP: Breaking language barriers: One AI, Multilingual NLP: Breaking Language Barriers | One AI. Available at: https://oneai.com/learn/multilingual-nlp (Accessed: 20 September 2023).

Hinton, P.R. (2019) ‘What are stereotypes?’, Stereotypes and the Construction of the Social World, pp. 1–28. doi:10.4324/9781315205533-1.

Jones, K.S. (1999) ‘What is the role of NLP in text retrieval?’, Text, Speech and Language Technology, pp. 1–24. doi:10.1007/978-94-017-2388-6_1.

Lamont, R., Abrams and Swift (2018) *A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage*, *Psychology and aging*. Available at: https://pubmed.ncbi.nlm.nih.gov/25621742/ (Accessed: 26 September 2023).

Linssen, H. and Hagendoorn, L. (1994) ‘Social and geographical factors in the explanation of the content of European nationality stereotypes’, British Journal of Social Psychology, 33(2), pp. 165–182. doi:10.1111/j.2044-8309.1994.tb01016.x.

Loporcaro, M. and Paciaroni, T. (2011) ‘Four-gender systems in Indo-European’, Folia Linguistica, 45(2). doi:10.1515/flin.2011.015.

Lucy, J.A. (2010) ‘Sapir-Whorf hypothesis’, *Encyclopedia of Identity* [Preprint]. doi:10.4135/9781412979306.n207.

Moskowitz, G.B. and Carter, D. (2018) ‘Confirmation bias and the stereotype of the black athlete’, Psychology of Sport and Exercise, 36, pp. 139–146. doi:10.1016/j.psychsport.2018.02.010.

Mar-Molinero, C. (2000) The politics of language in the Spanish-speaking world: From colonization to globalization. London: Routledge.

Nadeem, M., Bethke, A. and Reddy, S. (2021) ‘Stereoset: Measuring stereotypical bias in pre-trained language models’, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

OpenAI. (2023). *ChatGPT* (August 3 Version) [Large language model]. https://chat.openai.com

Ozturk, T., Nedelchev, R., Heumann, C., Arias, E., Roger, M., Bischl, B., Assenmacher, M. (2023) ‘How Different Is Stereotypical Bias Across Languages?”.

Park, Y. and Shin, Y. (2023) ‘Using multiple monolingual models for efficiently embedding Korean and English conversational sentences’, *Applied Sciences*, 13(9), p. 5771. doi:10.3390/app13095771

Pennington, C.R. et al. (no date) Twenty years of stereotype threat research: A review of psychological mediators, PLOS ONE. Available at: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0146487 (Accessed: 24 August 2023).

Ramat, G.A. and Ramat, P. (1998) The indo-european languages. London: Routledge.

Rahali, A. and Akhloufi, M.A. (2023) ‘End-to-end transformer-based models in textual-based NLP’, *AI*, 4(1), pp. 54–110. doi:10.3390/ai4010004.

Stanciu, A. et al. (2016) ‘Within-culture variation in the content of stereotypes: Application and development of the stereotype content model in an Eastern European culture’, The Journal of Social Psychology, 157(5), pp. 611–628. doi:10.1080/00224545.2016.1262812.

STEPHAN, W.G. and STEPHAN, C.W. (1993) ‘Cognition and affect in stereotyping: Parallel Interactive Networks’, Affect, Cognition and Stereotyping, pp. 111–136. doi:10.1016/b978-0-08-088579-7.50010-7.

Stolier and Freeman (2001) Social categorization, Social Categorization - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/social-sciences/social-categorization (Accessed: 24 August 2023).

‘The psychological process of stereotyping: Content, forming, internalizing, mechanisms, effects, and interventions’ (2023a) Frontiers Research Topics [Preprint]. doi:10.3389/978-2-8325-1371-2.

Zawisha, et al. (2019) ‘The psychological process of stereotyping: Content, forming, internalizing, mechanisms, effects, and interventions’ (2023) Frontiers Research Topics [Preprint]. doi:10.3389/978-2-8325-1371-2.

Žabokrtský, Z., Zeman, D. and Ševčíková, M. (2020) ‘Sentence meaning representations across languages: What can we learn from existing frameworks?’, Computational Linguistics, 46(3), pp. 605–665. doi:10.1162/coli_a_00385.

APPENDIX