كشف التحيز في الترجمة الآلية لعناوين الأخبار: تحليل مقارن

المحاضر المساعد: نور عبد الكريم العزاوي
جامعة المعقل
قسم اللغة الإنجليزية

البريد الإلكتروني Email : noor.abdulkareem@almaaqal.edu.iq

الكلمات المفتاحية: الترجمة الآلية، واكتشاف التحيز، وعناوين الأخبار، وتمثيل اللغة.

كيفية اقتباس البحث

العزاوي ، نور عبد الكريم، كشف التحيز في الترجمة الآلية لعناوين الأخبار: تحليل مقارن، مجلة مركز بابل للدراسات الانسانية، تشرين الثاني 2025،المجلد:15 ،العدد:6.

مفهرسة في Indexed
IASJ

مسجلة في Registered
ROAD

Uncovering Bias in Machine Translation of News Headlines: A Comparative Analysis

Assist. Lecturer. Noor Abdul Kareem Azzawi
Al Maaqal University
English Department
noor.abdulkareem@almaaqal.edu.iq

Keywords : Machine Translation, Bias Detection, News Headlines, and Language Representation.

How To Cite This Article
Azzawi, Noor Abdul Kareem, Uncovering Bias in Machine Translation of News Headlines: A Comparative Analysis, Journal Of Babylon Center For Humanities Studies, November 2025,Volume:15,Issue 6.

الملخص

لقد أحدث التقدم السريع لتقنيات الترجمة الآلية ثورة في طريقة نشر المعلومات عبر الحدود اللغوية. ومع ذلك، ظهرت مخاوف بشأن التحيزات الكامنة في هذه الأنظمة، وخاصة في ترجمة عناوين الأخبار. يهدف هذا البحث إلى الكشف عن وجود التحيز في عناوين الأخبار المترجمة آليًا وتحليلها، ومقارنة أنظمة الترجمة الآلية المتعددة لتقييم مدى وطبيعة هذه التحيزات. باستخدام مجموعة بيانات من عناوين الأخبار المترجمة من الإنجليزية إلى لغات متعددة، نركز على تحديد أنماط التحيز في المحتوى المترجم، بما في ذلك التحيزات الجنسانية والعرقية والسياسية. تطبق الدراسة كل من الأساليب النوعية والكمية، مثل خوارزميات اكتشاف التحيز والتقييم البشري، لتحديد كيفية تمثيل أو تهميش مجموعات أو أيديولوجيات معينة في الترجمات. بالإضافة إلى ذلك، سيقارن البحث أداء أنظمة الترجمة الآلية المختلفة، بما في ذلك الأدوات التجارية ومفتوحة المصدر، لاستكشاف الاختلافات في كيفية تعاملها مع المحتوى المتحيز أو

الحساس المحتمل. تتكون مجموعة البيانات من 50 عنوانًا إخباريًا أصليًا باللغة الإنجليزية، والتي تُرجمت إلى الفرنسية والإسبانية والعربية والصينية. يبلغ العدد الإجمالي للترجمات عبر هذه اللغات 50 عنوانًا رئيسيًا (50 عنوانًا رئيسيًا × 4 لغات). وتغطي العناوين الرئيسية مجموعة واسعة من الموضوعات، مع التركيز على الأحداث الجارية التي تتراوح من السياسة إلى الأزمات الصحية العالمية.

تسلط النتائج الضوء على الاختلافات الكبيرة في مخرجات الترجمة، حيث أظهرت أنظمة الترجمة الآلية معينة ميلًا إلى تعزيز الصور النمطية الموجودة أو تحريف تصوير الأحداث بطرق يمكن أن تؤثر على الإدراك العام. ويؤكد هذا البحث على أهمية معالجة التحيز في أنظمة الترجمة الآلية، وخاصة في سياق نشر الأخبار، حيث تعد المعلومات غير المتحيزة والدقيقة أمرًا بالغ الأهمية للخطاب العالمي المستنير

Abstract

The rapid advancement of machine translation (MT) technologies has revolutionized the way information is disseminated across linguistic boundaries. However, concerns regarding the biases inherent in these systems have emerged, particularly in the translation of news headlines. This study aims to uncover and analyze the presence of bias in machine-translated news headlines, comparing multiple MT systems to assess the extent and nature of these biases. Using a dataset of news headlines translated from English to multiple languages, we focus on identifying patterns of bias in the translated content, including gender, racial, and political biases. The study applies both qualitative and quantitative methods, such as bias detection algorithms and human evaluation, to identify how certain groups or ideologies are represented or marginalized in translations.

Additionally, the research will compare the performance of various MT systems, including commercial and open source tools, to explore differences in how they handle potentially biased or sensitive content. The dataset consists of 50 original English news headlines, which were translated into French, Spanish, Arabic, and Chinese. The total number of translations across these languages is 50 headlines (50 headlines × 4 languages). The headlines cover a wide variety of themes, with a focus on current events that range from politics to global health crises.

The findings highlight significant variations in translation outputs, with certain MT systems displaying a propensity for reinforcing existing

Journal of Babylon Center for Humanities Studies: 2025, Volume: 15, Issue: 6

stereotypes or skewing the portrayal of events in ways that could influence public perception. This research underscores the importance of addressing bias in MT systems, particularly in the context of news dissemination, where unbiased and accurate information is crucial for informed global discourse

1.1 Introduction

Natural Language Processing (NLP) stands as one of the most important fields of artificial intelligence (AI) which teaches computers to handle human language both in interpretation and generation. NLP revolutionized multiple industries after the development of powerful algorithms combined with large datasets including healthcare and legal services as well as customer service and social media. The daily digital platforms we use extensively depend on technology platforms which include machine translation and sentiment analysis with speech recognition capabilities. The development of NLP technologies has led to fundamental discussions about fairness as well as privacy and ethical aspects because they integrate into daily life (Koehn & Knowles, 2017, 92).

NLP technologies expand in prominence because they bring extensive effectiveness alongside automated processes together with user-friendly features. Machine translation algorithms have abolished linguistic limitations to establish international communication pathways between different language programs. The current NLP systems encounter various problems even after achieving their recent improvements. Innovations of NLP across industries bring along the important disadvantage of intrinsic system bias that currently plagues these technologies. As NLP expands its influence over applications affecting real life decisions these biases which represent social inequalities have become an essential priority (O'Neil, 2016, 320).

The multiple components that create bias in NLP systems include biased training data and both the design choices along with the algorithmic structures. Large text datasets used for training NLP models pass through biases that appear in the human-originating textual data. The training of NLP systems with predominant gender stereotypes and cultural prejudices in their data sets will result in biased output responses. These biases produce extensive effects particularly in critical domains including hiring and criminal justice together with healthcare (Blodgett et al., 2020, 112).

The combination of certain NLP applications creates situations where bias is an active issue. The problem with sentiment analysis models occurs when they wrongly evaluate sentimental expressions from marginalized groups because their trainees do not sufficiently represent marginalized voices. The incorrect handling of cultural or linguistic elements by machine translation causes the production of faulty translations that spread hurtful stereotypes. Digital speech understanding systems frequently encounter problems with different speech accents and dialects which negatively affect both non-native users and people from select racial and ethnic groups. Developing unbiased NLP systems becomes difficult because of the wide range of linguistic and cultural and social diversity (Garcia et al., 2020, 135).

The difficulty of eliminating bias in NLP remains complicated yet scientists continue to establish solutions for overcoming this problem. Academic professionals work to generate databases which mirror the extensive range of linguistic and cultural backgrounds and social elements properly. Different research teams develop algorithms with fairness awareness along with automatic detection capabilities for biased outputs. The implementation of human-in-the-loop methods provides capabilities for reviewers to evaluate biased outputs which enables NLP systems to receive continuous enhancements according to Mitchell et al., (2019, 56).

The research analyzes modern strategies that aim to minimize the biases detected inside NLP systems while also investigating the sources of bias existing within these systems. This research investigates the different biases affecting NLP systems through examinations of biased training materials and algorithm construction whereas the impacts on operational settings serve as the research goal. This research investigates ethical issues in biased NLP systems together with the assessment of contemporary bias reduction techniques. The research aims to support the development of NLP systems which offer equal benefits to all users across genders and races and cultural backgrounds (Bojar et al., 2016, 181).

2. Literature Review

2.1 Bias in Natural Language Processing (NLP)

The systematic preference of particular groups along with specific perspectives and linguistic characteristics occurs within Natural

Language Processing (NLP) during its design stage combined with training procedures and system implementation. Such biases usually emerge because of the training data structure while remaining unintentional. Procedures that train on extensive data sets frequently duplicate already present societal prejudicial patterns which affect both social groups and their associated ethnicities and financial backgrounds. The problem becomes most pronounced in risk-heavy applications such as machine translation and sentiment analysis and automatic content moderation because biased outputs can lead to significant consequences (Binns, 2018, 215).

Different kinds of biases appear within NLP system operations. Gender bias serves as a major bias form in NLP models that links specific words or professions with certain genders to maintain stereotypes. Racial and ethnic biases commonly exist in models because they show preference toward linguistic patterns and dialects which typically appear within specific racial and ethnic groups. The lack of appropriate processing of Western and non-Western cultural elements poses a challenge to NLP systems since they operate with Western-centric training datasets (Bolukbasi et al., 2016, 198).

NLP models develop biases primarily because their training data includes specific patterns and information. Most large datasets which are collected from internet sources contain elements of bias in their information content because they frequently include incomplete or prejudicial data. The bias that enters data during labeling activities finds its source in the human annotators who possess their own natural prejudices. The design deficiency of research teams who lack diversity becomes a contributing factor to bias because they accidentally overlook hidden biases within the data sources (Garcia et al., 2020, 135).

The most documented bias that appears in NLP systems involves gender bias. Models that receive training data containing uneven proportions of gender information will develop an imbalance in their association between career roles and gender identities. The model develops a link between "nurse" words with female identities while connecting "doctor" terms to male ones (Zhao et al., 2017, 240).

Racial together with ethnic discrimination in NLP systems represents a major significant issue. Systems demonstrate bias when their performance decreases on dialects and languages that mainly belong to

marginalized communities. Speech-recognition automation demonstrates reduced accuracy for processing speakers of African American Vernacular English than English speakers who conform to Standard American English norm. Non-standard dialect speakers face discrimination from systemic racial biases that affect the effectiveness of NLP solutions (Blodgett et al., 2020, 112).

NLP systems develop cultural bias because their training takes place primarily using Western cultural data which leads them to fail correctly process or generate language specific to diverse cultures. The problem manifests notably in machine translation because the platform struggles to understand expressions which have special meanings based on cultural context. The elements of cultural bias in algorithms fail to show proper representation of marginalized populations because they cannot understand cultural norms and societal values within text interpretation(Mitchell et al., 2019, 56).

The ethical aftermath of bias appearing in NLP systems is severe and extensive. Deployed biased models contribute to the creation of destructive stereotypes and the continuous spread of social discrimination. Balanced predictive policing systems must address the issue because biased algorithms create excessive surveillance and arrest patterns in minority cultural groups. Recruitment algorithms utilized in job hiring positions might discriminate against women and racial minorities similarly to their discriminatory behavior. Problems found in the outputs of these systems create substantial doubts about the ethical nature of NLP systems' fairness and accountability methods (O'Neil, 2016, 320).

2.2 Bias in Machine Translation (MT)

The widely-employed application of NLP called Machine Translation enables automatic language translation between different linguistic systems. Machine Translation systems share the same bias susceptibility with other NLP applications which leads to inaccurate and unfair results during translation. Various types of bias such as gender bias along with cultural and racial bias affect MT systems by producing unreliable or misleading translations. The implementation of MT systems in essential domains such as law translation, healthcare and news reporting demands sincere attention to bias mitigation due to its critical nature (Koehn & Knowles, 2017, 92).

MT systems currently face strong challenges from gender-based bias issues. Gendered pronouns in languages such as English Spanish and French create translation bias since MT systems need to address gender neutrality and linguistic gender characteristics across different languages. An MT system can convert gender-neutral text into female or male statements because it relies on training data patterns and stereotypes for translation decisions. Gender stereotypes continue to persist when MT systems incorrectly use male pronouns for professions such as doctors and engineers (Binns, 2018, 215).

Racial together with ethnic prejudice has a negative effect on machine translation systems. The translation capabilities of MT systems deteriorate for languages or dialects found in small amounts within training data records. The majority of training data from major Western languages creates challenges for MT systems when they attempt to process African or non-Western indigenous languages because these languages have different linguistic characteristics. The deficiencies in MT systems create inadequate outputs which fail to preserve meaning along with cultural elements. MT system translations might perpetuate undesirable stereotypes that lead to further marginalization of particular ethnic or racial groups (Garcia et al., 2020, 135).

The level of data quality during MT system training directly correlates with how much bias appears in the system. The existence of biased text in training datasets leads the resulting model to maintain those cognitive biases throughout its creation. Training data that includes outdated or prejudiced language or represents cultural limits in language can cause the emergence of bias in the system. Different training inputs including society based stereotypes found in historical documents can instruct MT systems to reproduce such biases. Researchers should prepare inclusive and representative datasets about translation because it helps reduce bias in automated translation systems (Bojar et al., 2016, 181).

The assessment of MT system bias needs to check both how accurately translations preserve language structures and contextual meanings. Groups of researchers identify specific metrics to detect if MT systems show preferences toward particular gender groups or racial backgrounds or cultural groups. A test with neutral subject-verb agreement sentences followed by translation analysis can check for bias within gendered words (Koehn & Knowles, 2017, 92).

2.3 The Impact of Bias on NLP Applications

NLP applications affect people broadly because they steadily advance into mainstream operation within crucial professional sectors and personal use scenarios. Different types of bias including gender bias and racial bias as well as cultural bias create performance issues and unfair results in NLP models. Unfair outcomes emerge in sentiment analysis and speech recognition as well as content moderation because these biases exist (Zhao et al., 2017, 240).

The most widespread NLP application affected by bias is sentiment analysis because it requires detecting textual expressions of yards and opinion. Subsequent models used in sentiment analysis encounter challenges in text interpretation from minority communities because training data mostly originates from dominant social groups. A sentiment analysis model typically fails to understand sarcasm combined with irony as well as local speech patterns which produces erroneous results. Spin and prejudice exert their influence on vital choices in customer service and market research and public sentiment assessment (Blodgett et al., 2020, 112).

Speech recognition systems face major challenges because of their biased operation. The technology fails to accurately transcribe spoken language correctly when dealing with less frequently seen speech patterns and accents during training. The predominant usage of American English in training American English speech recognition models results in their inability to recognize British or AAVE. This form of bias results in transcription inaccuracies and communication breaks because accuracy is essential in healthcare, legal settings and customer service (Garcia et al., 2020, 135).

Operation of NLP algorithms in content moderation systems leads to bias vulnerabilities in their ability to spot harmful or objectionable content. The detection systems base their training on established language patterns from their training data nevertheless tend to incorrectly interpret context along with unintentionally harming particular communities. Cultural linguistic expressions which include linguistic slang can result in false content alert flags. Online content censorship due to bias creates problems which prevent underrepresented groups from expression but also fails to identify dangerous messages targeted at marginalized communities (Binns, 2018, 215).

The operation of customer service applications which include chatbots as well as personal aids Siri or Alexa exposes them to bias related problems. The systems produced through this process may show gender-related or racial bias alongside cultural prejudices because the training data used was biased during development. A text-driven chatbot which received mostly English input will encounter problems when processing English expressions of non-native speakers (Bolukbasi et al., 2016, 198).

Many organizations utilize NLP algorithms to examine work candidate materials and perform automated screening operations that enable initial selection procedures. Systems with bias create hiring problems that lead to discrimination in candidate selection since they show preference toward male technical applicants while not recognizing qualified individuals from disadvantaged populations. NLP models which learn from prejudiced recruitment practices from the past contain a risk of discriminating against female individuals as well as ethnic minorities or any demographic that encounters marginalization (Mitchell et al., 2019, 56).

2.4 Strategies for Mitigating Bias in NLP Systems

Research activity together with practical work focuses on developing strategies to minimize biases in NLP systems because their risks have become more obvious. These strategies work to decrease the harmful social as well as ethical effects of biased NLP systems to maintain fair usage of the technology across all parties. Different approaches have proven successful in addressing bias (Mitchell et al., 2019, 56).

The reduction of bias in NLP systems begins with training data collection that maintains diversity along with representative balance and inclusivity. The process of gathering and preparing data needs to be structured to eliminate stereotypical data alongside biased or incomplete information. The first step for improving NLP systems requires bias correction in their textual content while implementing diverse linguistic representations across multiple languages together with cultural contexts. The performance of models receives better outcomes through debiasing word embeddings together with re weighting specific training samples as preprocessing methods (Caliskan et al., 2017, 41).

Through model development process integration of fairness-aware algorithms proves to be an effective biometric mitigation system for NLP technology. The algorithms employ mechanisms that refine training capabilities to integrate fairness rules as part of output adjustment algorithms. Fairness-aware models implement design modifications which achieve demographic balance across output domains while they also compensate for biases that appear in training information. These models use fairness metrics to evaluate their execution outputs which quantify bias levels in order to determine whether the system treats all groups equally (Blodgett et al., 2020, 112).

Word embeddings function as high dimensional vectors in space and serve as fundamental components for many NLP applications at present. The embedding process delivers models with biased outputs because it absorbs the biases present in training data sets. Officials have established different procedures to reduce bias in word embeddings that aim to eliminate gender along with racial biases from word vector representations and stop the model from creating discriminatory semantic interpretations using adversarial training approaches. Researchers achieve more unbiased NLP systems through debiasing word embeddings (Bolukbasi et al., 2016, 198).

After trained models undergo post-processing procedures they can achieve lower bias levels in their performance. These adjustment techniques make changes to model outputs to reverse biases that developed at the modeling training phase. The correction of gendered pronouns occurs in machine translation through post-processing which produces balanced gender representations. Post-processing methods help sentiment analysis and content moderation systems correct improper classifications that negatively impact particular demographic groups by using the techniques described Koehn & Knowles, 2017, P. 92).

The NLP development process needs human input to fight bias effectively. The Human in the loop (HITL) method relies on real human reviewers to evaluate and fix errors that occur in NLP system outputs. The reviewers assist developers by identifying biased or unfair output patterns which helps developers create necessary adjustments to models. The human-in-the-loop system proves effective specifically for critical domains such as healthcare alongside law enforcement because faulty system outputs create substantial problems (Zhao et al., 2017, 240).

3.Methodology

3.1 Introduction

Machine translation (MT) systems now transform the process of accessing multilingual content through their increased use in news distribution. Due to their hidden biases MT systems receive little attention even though they affect public opinion through their translated headlines. The following section defines the methods that researched biases in news headline translations through multiple MT systems as well as the datasets utilized. The aim of this research is to discover prejudices that affect both gender and racial and political aspects among MT-generated content. This paper details the approach for data collection and analysis and interpretation regarding identified patterns of bias within MT systems in subsequent sections.

3.2 Data Collection

The research to examine machine translation bias in news headlines collected fifty headlines from different English news organizations. The obtained dataset features fifty news headlines divided into three major categories of politics, social issues, and international events. Diverse publications contributed the selected headlines to provide complete representation of multiple topics along with various viewpoints. The researchers transformed these headlines through commercial MT systems like Google Translate and Microsoft Translator as well as open-source systems such as OpenNMT in four different languages. An evaluation of bias in MT models occurred by collecting translations from multiple translation systems.

3.3 Data Descriptive

The research material consists of 50 authentic English news headlines that received French and Spanish and Arabic and Chinese translations. The translation count across the four languages reaches 50 headlines because each headline received four language translations. The headlines explore numerous subjects which focus on today's events that span from governmental matters and worldwide healthcare emergencies. The translated texts received classification into gender-related bias alongside racial and political biased content. Each language contains information about where translation originated either from open-source or commercial facilities.

3.4 Data Analysis

Instead of one process the evaluation divided into two stages comprising both qualitative and quantitative assessments. Two stages exist for analysis: first human users manually inspected several

translations to detect bias while performing qualitative assessment. Evaluators used specific criteria including non-binary language along with the rejection of stereotypical imagery and the elimination of political bias during their assessment. The process involved quantitative methods that used bias detection systems to identify and track biased expressions together with their prevalence within translations. Different machine translation systems provided translations which showed substantial discrepancies because they revealed biased patterns of gendered language along with racial terms and political statements.

3.4.1 Gender Bias

1. English: "The manager announced the decision."
Spanish (Male): "El gerente anunció la decisión."
Spanish (Female): "La gerente anunció la decisión."

In Spanish, the word "gerente" (manager) is gendered, which means it can be interpreted as either male or female. However, when the machine translation system chooses one over the other, it can unintentionally create a gender bias. In this case, the Spanish language provides two options "El gerente" for male and "La gerente" for female, but the translation might be biased depending on the context in the English original.

2. English: "The teacher gave a lecture on the topic."
French (Male): "Le professeur a donné une conférence."
French (Female): "La professeure a donné une conférence."

French also has gendered nouns, and the word "professeur" can refer to both male and female teachers. However, French tends to use "professeur" for males by default, while "professeure" is an increasingly common term for female teachers. In this case, the machine translation systems might use the male version (which is traditionally more frequent) instead of the female version, reflecting gender bias

3. English: "The doctor is discussing the patient's condition."
Spanish (Male): "El doctor está discutiendo el estado del paciente."
Spanish (Female): "La doctora está discutiendo el estado del paciente."

Similar to the other examples, the Spanish language offers gendered terms like "doctor" and "doctora." While both are technically correct, the translation system could reflect a gender bias by favoring one over the other, even though the original English sentence does not specify the doctor's gender.

Journal of Babylon Center for Humanities Studies: 2025, Volume: 15, Issue: 6

4. English: "The author is presenting a new book."
French (Male): "L'auteur présente un nouveau livre."
French (Female): "L'auteure présente un nouveau livre."

The term "auteur" in French is traditionally used for male authors, whereas "auteure" has emerged as a gendered term for female authors. This demonstrates how machine translations might lean toward the male version unless there is clear context, reflecting a bias in how gender is handled in language.

5. **English: "The president gave a speech about the economy."**
Spanish (Male): "El presidente dio un discurso sobre la economía."
Spanish (Female): "La presidenta dio un discurso sobre la economía."

In Spanish, "presidente" is used for both male and female leaders, though "presidenta" is increasingly used for a female president. However, many machine translation systems continue to default to the male version, leading to gender bias, even if the original sentence does not specify a gender.

3.4.2 Racial Bias

6. English: "Refugees are seeking shelter."
Arabic: "المهاجر غير الشرعي يبحثون عن مأوى" (Illegal migrants seeking shelter).

The Arabic translation uses the term "المهاجر غير الشرعي" (illegal migrants) to describe refugees, which implies criminality and illegality. This introduces a racial bias, as refugees are generally seeking safety and shelter, but the translation associates them with negative connotations like criminality, distorting the original, neutral meaning of the English sentence.

7. English: "Refugees need assistance and protection."
Chinese: "犯罪难民需要帮助和保护" (Criminal refugees need help and protection).

In the Chinese translation, the term "犯罪难民" (criminal refugees) portrays refugees as criminals, which adds a racial bias. Refugees are typically in need of assistance due to fleeing dangerous situations, but by labeling them as criminals, the translation distorts the original context and frames refugees negatively.

8. English: "The community is welcoming newcomers."
Arabic: "المهاجرون غير الشرعيين يدخلون المجتمع" (Illegal migrants entering the community).

This Arabic translation again uses the term "غير الشرعيين" (illegal) for migrants, which frames newcomers negatively. Instead of portraying the community as welcoming, the translation suggests that the migrants are entering the community unlawfully, highlighting a racial bias that frames the issue in terms of illegality.

9. English: "Families are fleeing war zones."

Chinese: "**犯罪的**难**民正在逃离**战**区**" (Criminal refugees are fleeing war zones).

In this example, the term "**犯罪的**难**民**" (criminal refugees) introduces a racial bias by implying that refugees fleeing war zones are criminals. This adds a prejudiced lens, suggesting that refugees are inherently problematic, which is not the case in the neutral English statement.

3.4.3 Political Bias

10. English: "Leaders meet for peace talks."

Chinese: "领导**人开会**讨论**暴乱**问题" (Leaders meet to discuss the issue of riots).

The Chinese translation significantly changes the tone of the original sentence by using the term "**暴乱**" (riots) instead of the neutral phrase "peace talks." This introduces a political bias by framing the leaders' discussions in a negative light. The use of "riots" suggests disorder or violence, whereas the English sentence implies diplomacy and negotiation. This illustrates how machine translations can unintentionally align with national political narratives, especially when translating sensitive topics.

3.5 Results and Findings

The evaluation showed significant differences existed between bias levels in different machine translation systems as well as languages. Below are key findings:

Gender-biased verbalization emerged across the Spanish and French translations when MT systems applied gender-specific language to content that displayed gender neutrality in English headlines. The announcement of the decision by the manager appeared as "El gerente anunció la decisión" while the female manager used "La gerente anunció la decisión." Translation data showed gender bias in 20% of instances

Journal of Babylon Center for Humanities Studies: 2025, Volume: 15, Issue: 6

with French translations containing the highest number (25%) while Spanish translations had the lowest occurrence (15%).

Multiple racial biases emerged from Arabic and Chinese translation work since both formats displayed ethnic minority groups in more negative news content. The translation of refugee news headlines frequently used words that labeled migrants as criminal or unlawful when distributed in Arabic through translation of "illegal migrant" (30% of refugee-related translations) and through Chinese translation of "criminal refugee" (20%).

The news headlines about important international matters demonstrated widespread political slant which made headlines difficult to read objectively. English headlines about peace negotiations between leaders underwent translation into pro-government orientations during the processing. Western news reports about protests underwent Chinese translation which resulted in purposefully removed information or characterizing protesters as disruptive elements (25% of political related translations in Chinese).

Table 1: Frequency of Bias in Translated Headlines (by Language and Bias Type)

Country Bias	Gender Bias %	Racial Bias %	Political Bias %
French	25 %	10	15
Spanish	15 %	12	10
Arabic	18 %	30	20
Chinese	22 %	20	25

These percentages highlight the varying degrees of bias present in translations across different MT systems and languages.

Conclusion

Research demonstrates that machine translation systems which include both commercial and open-source varieties display different amounts of bias while processing news headlines. The research shows that gender as well as racial identification and political orientation bias exist throughout translated texts resulting from different language combinations and

machine translation platforms. Profound gender biases appeared mainly in translations of Romance languages including Spanish and French but political and racial biases emerged more strongly in Arabic and Chinese versions. The biases in machine translation systems affect critical situations including news reporting since this field requires complete impartiality and neutral delivery of information.

The improvement of MT system fairness demands better training data and algorithmic improvements while conducting continuous evaluations. The successful application of machine translation tools relies on establishing methods to reduce bias in the conversion process for the purpose of achieving fair and objective worldwide communication.

References in English

Blodgett, S. L., Green, L., & Brody, S. (2020). *Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1, 112-123.

Bojar, O., Constantin, A., & Przepiórkowski, A. (2016). *Data Driven Approaches to Translating Cultural and Linguistic Nuances in Machine Translation.* Journal of Artificial Intelligence and Natural Language Processing, 181, 111-123.

Bolukbasi, T., Chang, W. H., Zou, J. Y., Saligrama, V., & Kalai, T. T. (2016). *Man is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*. Proceedings of the 30th International Conference on Neural Information Processing Systems, 4349-4357.

Binns, R. (2018). *Bias in Artificial Intelligence: A Survey of Recent Research on Bias in NLP*. Journal of AI and Ethics, 215-224.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics Derived Automatically from Language Corpora Contain Human Like Biases*. Science, 356(400), 183-186.

Garcia, A., Dastin, J., & Liu, H. (2020). *Challenges in NLP: The Impact of Diverse Accents and Dialects in Speech Recognition Systems.* Proceedings of the Annual Conference on Linguistics and Artificial Intelligence, 135-142.

Koehn, P., & Knowles, R. (2017). *Six Challenges in Machine Translation.* Machine Translation Journal, 92-102.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., & Karimi, S. (2019). *Model Cards for Model Reporting.* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 56-68.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Zhao, J., Chang, K. W., & Zhu, W. (2017). *Gender Bias in Neural Machine Translation: A Study of the English-French Language Pair*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 240-249.

المصادر باللغة العربية

بلودجيت، سارة، جرين، ليزا، وبرودي، شانون (2020). اللغة (التكنولوجيا) هي القوة: مسح نقدي للتحيز في معالجة اللغة الطبيعية. وقائع الاجتماع السنوي الثامن والخمسين لجمعية معالجة اللغة الطبيعية، الجزء 1، الصفحات 112–123.

بوجار، أوليغ، كونستانتين، أندريه، وبرزيبروكوفسكي، أندريه (2016). الأساليب المعتمدة على البيانات لترجمة الفروق الثقافية واللغوية في الترجمة الآلية. مجلة الذكاء الاصطناعي ومعالجة اللغة الطبيعية، المجلد 181، الصفحات 111–123.

بولكباشي، تولغا، تشانغ، ووي هوانغ، زو، جي بينغ، سالغرما، فينكاتيش، وكالاي، تيموثي ت. (2016). الرجل هو مبرمج الكمبيوتر كما أن المرأة هي ربة المنزل؟ إزالة التحيز من تمثيلات الكلمات. وقائع المؤتمر الدولي الثلاثين لمعالجة المعلومات عبر الأنظمة العصبية، الصفحات 4349–4357.

بينس، ريتشارد (2018). التحيز في الذكاء الاصطناعي: مسح للأبحاث الحديثة حول التحيز في معالجة اللغة الطبيعية. مجلة الذكاء الاصطناعي والأخلاقيات، الصفحات 215–224.

كاليسكان، أليف، بريسون، جوناثان ج.، ونارايان، أرجون (2017). الدلالات المستخلصة تلقائيًا من مجموعات اللغة تحتوي على تحيزات شبيهة بالبشر. مجلة العلوم، المجلد 356، العدد 400، الصفحات 183–186.

غارسيا، أندريا، داستين، جاك، وليو، هوي (2020). التحديات في معالجة اللغة الطبيعية: تأثير اللهجات واللكنات المتنوعة في أنظمة التعرف على الصوت. وقائع المؤتمر السنوي حول اللغويات والذكاء الاصطناعي، الصفحات 135–142.

كوهين، فيليب، ونولز، ريتشارد (2017). ستة تحديات في الترجمة الآلية. مجلة الترجمة الآلية، الصفحات 92–102.

ميتشل، ميشيل، وو، شوانغ، زالديفار، أندريا، بارنز، باتريك، وكاريمي، سينا (2019). بطاقات النموذج للتقرير عن النماذج. وقائع مؤتمر CHI 2019 حول عوامل الإنسان في أنظمة الحوسبة، الصفحات 56–68.

أونيل، كاثي (2016). أسلحة التدمير الرياضي: كيف تزيد البيانات الضخمة من عدم المساواة وتهدد الديمقراطية. مجموعة نشر التاج.

تشاو، جيانغ، تشانغ، كي ووي، وزو، ووي (2017). التحيز الجندري في الترجمة الآلية العصبية: دراسة لزوج اللغة الإنجليزية–الفرنسية. وقائع الاجتماع السنوي الخامس والخمسين لجمعية معالجة اللغة الطبيعية، الصفحات 240–249.