Bias in human data : a feedback from social sciences

Reference:
Takan Savas, Ergun Duygu, Yaman Sinem Getir, Kilincceker Onur.- Bias in human data : a feedback from social sciences
Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery - ISSN 1942-4795 - San francisco, Wiley periodicals, inc, 13:4(2023), e1498
Full text (Publisher's DOI): https://doi.org/10.1002/WIDM.1498
To cite this reference: https://hdl.handle.net/10067/1967800151162165141

Conflict of Interest

There is no conflict of interest.

Abstract

The fairness of human-related software has become critical with its widespread use in our daily lives, where life-changing decisions are made. However, with the use of these systems, many erroneous results emerged. Technologies have started to be developed to tackle unexpected results. As for the solution to the issue, companies generally focus on algorithm-oriented errors. The utilized solutions usually only work in some algorithms. Because the cause of the problem is not just the algorithm; it is also the data itself. For instance, deep learning cannot establish the cause-effect relationship quickly. In addition, the boundaries between statistical or heuristic algorithms are unclear. The algorithm's fairness may vary depending on the data related to context. From this point of view, our article focuses on how the data should be, which is not a matter of statistics. In this direction, the picture in question has been revealed through a scenario specific to "vulnerable and disadvantaged" groups, which is one of the most fundamental problems today. With the joint contribution of computer science and social sciences, it aims to predict the possible social dangers that may arise from artificial intelligence algorithms using the clues obtained in this study. To highlight the potential social and mass problems caused by data, Gerbner's "cultivation theory" is reinterpreted. To this end, we conduct an experimental evaluation on popular and commonly used algorithms and data sets such as Word2Vec with the fastText dataset from the Facebook (Meta) company, GloVe algorithm with its dataset from Stanford University, and ELMO (Embeddings from Language Model) algorithm with its dataset from Allen Institute at the University of Washington. The article stresses the importance of a holistic approach combining the algorithm and data as well as an interdisciplinary assessment.

Keywords

Artificial Intelligence, Machine Learning, Fairness, Social Science, Data Bias, Social Computing, Cultivation Theory, New Media

1. INTRODUCTION

The accuracy of modern algorithms is a critical issue, as unfortunate incidents have occurred due to algorithms related to minority and disadvantaged groups. When looking at the reasons for these unfortunate events, a few points stand out. One of them is deep learning algorithms, which are becoming increasingly popular daily. However, these types of techniques have difficulty establishing a retrospective cause-and-effect relationship. Another reason for the unfortunate situations that algorithms can cause is the absence of a standard for statistical or intuitive analysis (Zhang & Harman, 2021). In other words, the boundaries between statistical and intuitive algorithms are unclear. However, this uncertainty is not reflected in the data. Given that most algorithms fall into this category, this is a serious problem.

The accuracy of the algorithm is also highly related to its context. Can a correct algorithm show the correct result on incorrect data? What is error tolerance? How accurately does the correct algorithm convert the correct data into information? Or what kind of results can the wrong algorithm produce with the correct data? How much incorrect data can a wrong algorithm produce? Since the algorithm is tightly coupled to the data, is it accurate to evaluate it separately from the data?

In this regard, the questions are what attributes of data are used or when it can be used, which standards should medical data be used in health care, and how accurate data can be created guiding us to shaping data-focused testing mechanisms that are popular in software engineering (Takan et al., 2023). To answer these questions, we reveal the concept of synthetic data. For such a concept, every data should have metadata that specifies where and in what context the data will be used. An interdisciplinary approach is needed to understand these and to produce accurate results. Statistics is built on an accurate sampling of existing data. Our focus problem is instead to work on the data's *incorrect context*. Common standards should be developed significantly on subjects such as the amount and the features of data, how it is created, and how well tested. From this point of view, our article focuses on how the data should look like, which is not only a matter of statistics.

Moreover, the issue of how the data looks cannot be the research subject of computer science only. Collaboration with different disciplines, such as communication, sociology, law, medicine, psychology, and machinery, is critically required. Producing solutions by cooperating with the discipline to which the data belongs is vital for the accuracy of the algorithms fed from the data. Data can be built on many different and controversial values. For example, the way of thinking in America may connote different things in a different country where context becomes very important. Suppose that any text is used in artificial intelligence (AI) technology that can affect everyone. In that case, this text needs to at least comply with the universal declaration of human rights and, if possible, have values for the culture in which it is used. Sensor data used in autonomous vehicles should be tested with the same meticulousness within the framework of cultural and universal values (Hertweck et al., 2021) (Towsend et al., 2022). This important issue about passing a number through a standardized testing machine must be considered, and methodologies and software tools must be developed to validate cultural and normative concerns. As many numerical quality standards can vary from country to country or globally, basic quality standards must be developed and issued in different contexts.

In the vast majority of today's algorithms, there are only a few tests, and too many programs exist. Cloud-based collaborative version control systems such as GitHub (https://github.com/) and Bitbucket (https://bitbucket.org/) exacerbate this problem. These programs can be added to critical programs and cause severe weaknesses in the program's functionality. Although it seems beneficial that so many programs can be accessed publicly, it can also bring a severe data-oriented security vulnerability. For all these reasons, we consider data-driven standardization essential alongside the algorithm.

For such reasons, it is not enough to focus only on algorithms to avoid such misfortunes where the basic premise of algorithms is data. Focusing on the accuracy of algorithms without researching the data may be the biggest shortcoming of the studies on the subject to date.

In this paper, three different popular algorithms and their datasets were examined in terms of words that could be associated with vulnerable and disadvantaged groups, even though the scope, meaning, and interpretation of these words may vary from culture to culture, geography to geography, and historical perspective. These algorithms are Word2Vec (Mikolov, Tomas; et al., 2013) with the fastText dataset (https://fasttext.cc/) from the Facebook (Meta) company, GloVe algorithm with its dataset from Stanford University (Pennington, J.; et al., 2014), and ELMO (Embeddings from Language Model) algorithm with its dataset from Allen Institute at the University of Washington (Peters, M. E.; et al., 2018). As a result, we discover a similarity between the trends shaped by media messages and the data generated by digital footprints and the artificial intelligence technologies fed by this data. In order to draw attention to the importance of data selection, Gerbner's "cultivation theory" (Gerbner, G. 1998) has been reinterpreted from the perspective of artificial intelligence learning to highlight the possible social and mass problems that artificial intelligence technologies may cause, due to their usage of human data.

In the rest of the paper, section 2 presents the related work on algorithm-centric and data-centric fairness. Section 3 shows our data analysis and results on three different algorithms and corresponding datasets, namely Word2Vec, GloVe, and ELMO. In section 4, we introduce cultural context based on Gerbner's "cultivation theory". Section 5 provides the discussion and conclusion of our work.

2. RELATED WORK

The fairness of machine learning systems can be separated regarding the algorithms employed and the data used. Therefore, we divide the related work into 1) algorithm and 2) data-centric fairness.

2.1 Algorithm-Centric Fairness

The algorithmic bias is mainly caused by software utilized for the machine learning process. Several studies are employed to tackle algorithmic unfairness (Aggarwal et al., 2019; Galhotra et al., 2017; Udeshi et al., 2018). Software testing methods are also employed to test algorithmic discrimination for fairness. Galhotra presents a test case generation tool, Themis, which automatically generates discrimination tests without requiring any test oracle (Galhotra et al., 2017). It calculates causal discrimination scores to measure the algorithm's fairness against race and age.

To test fairness, the machine learning models as software artifacts are considered black-box testing, for which only the inputs and outputs of the systems are checked against discrimination. Aggarwal et al. (2019) present a method to automatically generate test inputs to detect individual discrimination (Aggarwal et al., 2019). They employ a dynamic symbolic execution technique to automatically generate test inputs and collect execution path constraints for different inputs to maximize path coverage. They also provide an experimental evaluation of eight benchmarks concerning age, gender, and race attributes compared to the Themis tool (Galhotra et al., 2017). They achieved a 34.8% average success score, while Themis did just 6.4%.

Moreover, Udeshi et al. (2018) introduced a scalable test generation approach to detect the machine learning model (Udeshi et al., 2018). They employed a global search algorithm for which random test inputs are generated and used to detect sensitive parameters (e.g., race, religion, gender) that become candidate test inputs for local search. They search in the neighborhood of candidate inputs to detect similar test inputs in the local search. They provided an experimental evaluation of six standard machine learning classifiers.

Rather than employing the algorithm's properties, Hertweck focused on the moral aspects of statistical fairness (Hertweck et al., 2021). They utilize statistical fairness as an independent metric related to its moral perspective by extending pure mathematical usage. Their proposal extends Friedler's concern about bias caused by the life of the individual born with developing abilities into realized abilities (Friedler et al., 2016). They also introduced two counterexamples based on the extension and concluded that their proposal is not universally valid.

2.2 Data-Centric Fairness

There are various empirical studies on fairness in datasets in different domains, such as precision medicine, which attempts to find personalized preventative and therapeutic strategies by considering differences in genes, environment, and lifestyle throughout their lives (Ntoutsi et al. 2020; Le Quy et al. 2022). The inequalities in sex and gender are discussed in a review of the domains of biomedicine and healthcare (Cirillo et al., 2020). In this review, the authors highlight the main available biomedical data types and the role of several AI technologies in understanding sex and gender differences in the health domain. The results provide recommendations for improving the global health and disease landscape to reduce inequalities.

An exciting study by Benthall and Haynes (2019) discusses a dilemma for society and designers: first, be blind to racial group disparities and thereby realize racialized social inequality by no longer measuring systemic inequality. Second, be conscious of racial categories in a way that refuses to be racialized. They propose a third option by preceding group fairness interventions with unsupervised learning to detect patterns of segregation dynamically. Machine learning systems can mitigate the root cause of social disparities, social segregation, and stratification without further anchoring status categories of disadvantage. The drawback to this approach might be its applicability.

Kearns proposes an algorithm for rich subgroup segregation (Kearns et al., 2019). They implement an extensive empirical evaluation of the algorithm on four real datasets for which fairness is a concern, and they investigate the fundamental convergence of the algorithm when instantiated with fast heuristics in place of learning oracles, measure the tradeoffs between fairness and accuracy, and compare this approach with recent algorithms.

The separation of data and algorithms in software testing is not new. Almost all studies on this subject distinguish between data and algorithms at the point of error detection. Whether the data or the algorithm is faulty, in our opinion, can cause severe problems because an algorithm can seem completely problem-free when the data is ignored. Keeping the data and the algorithm conceptually separate may cause existing problems to be overlooked. Specifically, we argue that data and algorithms should be considered together to avoid these problems.

On the other hand, none of those mentioned in earlier studies developed a systematic perspective on defining the problem. Moreover, existing studies have not demonstrated the seriousness and breadth of this issue. Our main difference from the studies in the literature is that we aim to define the problem with a systematic approach. Then we emphasize the necessity of an interdisciplinary approach that interacts with the social sciences to how the problem we consider should be solved. Furthermore, we address the problem in this study in a social scientific conceptual framework based on Gerbner's cultivation theory.

To validate the above concepts, we used Turkish and English datasets of three popular algorithms. These are Word2Vec (Mikolov, Tomas; et al., 2013) with the fastText dataset (https://fasttext.cc/), GloVe algorithm with its dataset (Pennington, J.; et al., 2014), and ELMO (Embeddings from Language Model) algorithm with its dataset (Peters, M. E.; et al., 2018). We have identified words that can express specifically vulnerable and disadvantaged groups, and we observe the similarities between the words. Details of the review and results are given below.

3. DATA ANALYSIS AND RESULTS

In our study, three different popular algorithms and datasets were used to convert texts into digital vectors. We chose to use these algorithms and data sets because they have been created by the most popular algorithms in this field, which have been published in many different languages. In our study, current expressions related to vulnerable and disadvantaged groups were identified. The words closest to the words we determined about these groups were obtained from the data sets we used.

Data could be selected from everywhere human beings exist for the scope of the study. However, language is one of humanity's most compressed data transfer methods. For example, the information extracted from an image is incomplete compared to the space it occupies on the computer. Therefore, we chose to use text. However, this choice does not mean that the problem is in the text. The problem is largely related to the traces left by humans. The methodology and results of the data analysis are explained in the subsections below.

3.1 Method

Word2Vec (Mikolov, Tomas; et al., 2013) is an unsupervised (no labels) and prediction-based model, which was developed by Google researcher Thomas Mikolov and his teammates, and tries to express words in vector space that is based on the distributional hypothesis. Say another way: words in the same text tend to have similar meanings, aiming to train words using two different models with an artificial neural network.

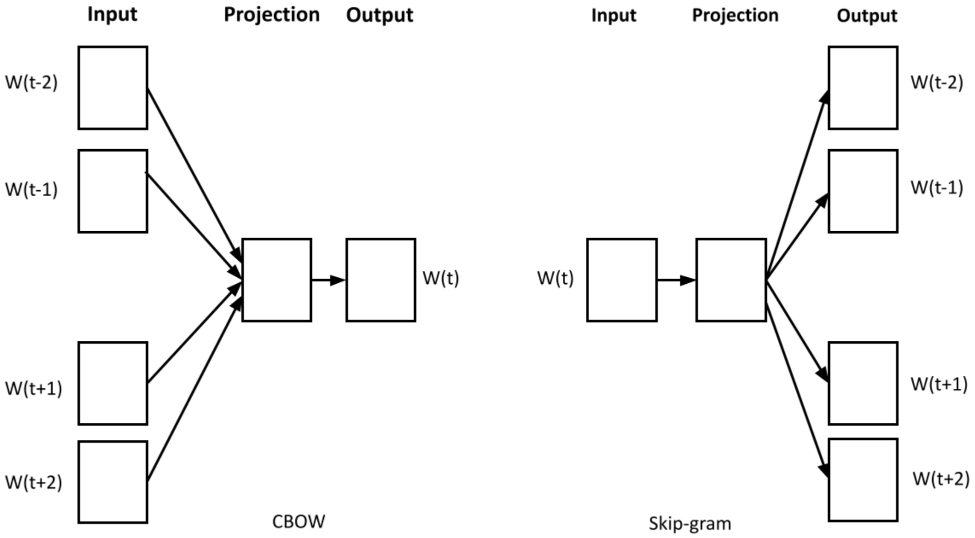

Figure 1: Overview of the method based on Word2Vec

The two models that Word2Vec uses are (The continuous Bag of Words (CBOW) and the Skip-Gram Model. In the CBOW (Continuous Bag of Words) model, words that are not in the window size center are taken as input, and the words in the center are estimated as output. Skip Gram: In the Skip Gram model, words in the window size center are taken as an input, and words not in the center are estimated as an output. One of the hyperparameters is the window size (see Figure 1). The maximum distance between a sentence's current and predicted word is the window size. The other parameter is the number of nodes in the hidden layer that tells us how many-dimensional spaces the words are represented in.

Models like Skip-gram and CBOW capture semantic meanings but do not use co-occurrence statistics. Matrix factorization methods use these statistics but fail to capture semantic relationships. In these models, there is no semantic meaning. The "GloVe" model, proposed by Pennington et al. (2014), aims to solve this problem by utilizing probability statistics and creating a new objective function (1):

$$J \ = \ \sum_{i,j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \ log X_{ij})^2 \tag{1}$$

Here, X_{ij} is the number of times the word pair (i, j) co-occurs in the corpus. The weighting function F(x) has three requirements (Pennington et al., 2014): it should not tend to infinity if f(0) = 0; it should not be attenuating for word pairs that occur infrequently together; and it should be slightly attenuating for large values of X_{ij}.

ELMo (Embeddings from Language Model) is one of the natural language processing models that led to a fundamental change in vectorizing words in the field of natural language processing by contextualizing words within sentences using vectors (Peters et al., 2018). ELMo constructs these word vectors with a neural network equipped with LSTM (Long Short-Term Memory) layers that have a biLM (bidirectional Language Model) architecture. The simplified version of the ELMo model architecture is shown below.

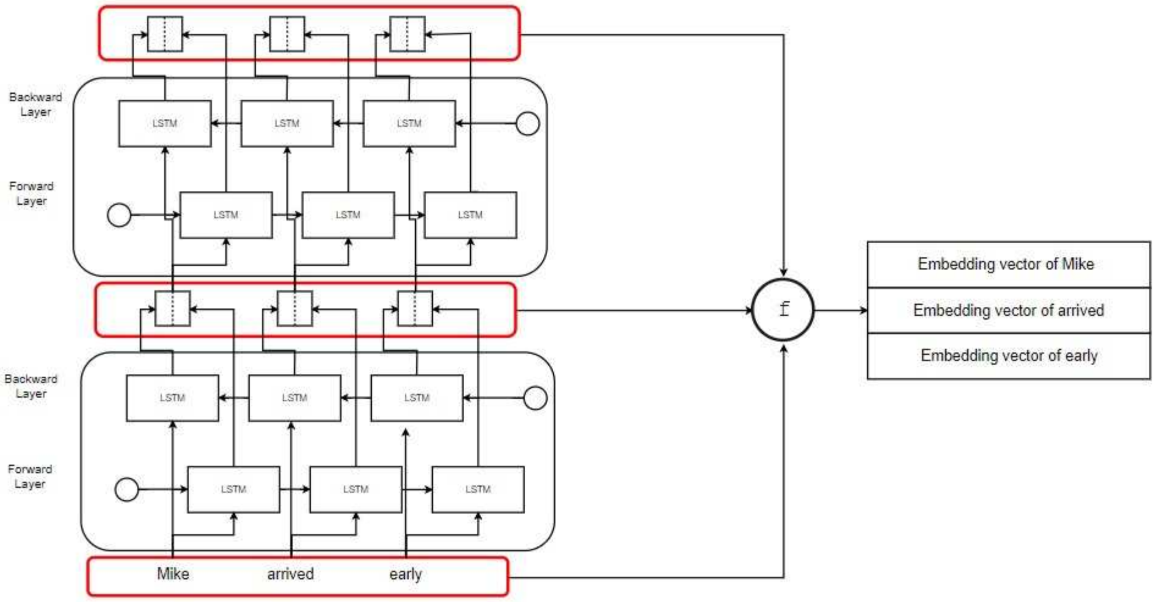

Figure 2. Simplified ELMo architecture

Through these models, specific words describing vulnerable and disadvantaged groups were searched using datasets published by the Facebook (Meta) company, Stanford University, and Allen Institute at the University of Washington. The research findings are detailed below.

3.1 Analysis

Our paper examines Word2Vec (Mikolov, T.; et al., 2013) with the fastText dataset (https://fasttext.cc/), GloVe algorithm with its dataset (Pennington, J.; et al., 2014), and ELMO (Embeddings from Language

Model) algorithm with its dataset (Peters, M. E.; et al., 2018) from the perspective of some words that can be associated with vulnerable and disadvantaged groups.

The scope of the term vulnerable and disadvantaged groups can vary from culture to culture, geography to geography, and historical perspective. However, based on some recent studies (Buolamwini & Gebru, 2018; Hacker, 2018; Ntoutsi et al., 2020), some relatively common and widely accepted words have been identified in this study. These words were analyzed in both Turkish and English datasets to see if a similar problem would arise. The general result that emerged is that the words indicating disadvantaged groups are present in both English and Turkish datasets along with some problematic words. Before briefly discussing the societal problems that this general result may cause, it was thought to be useful to examine the results in detail for each word individually.

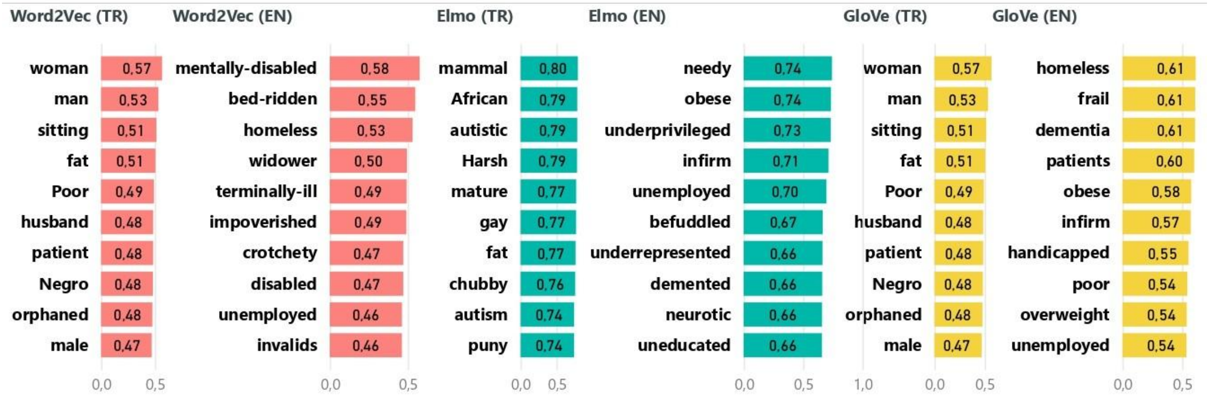

Figure 3: Word2Vec, Elmo, and GloVe results for **Elderly**

In the Turkish dataset, the most commonly used words with the term "yaşlı" (elderly) were first scanned in Word2Vec (Fig. 3). The most common words were found to be extremely negative labels such as "fat," "greasy," "naked," "rusty," and "unskilled." In this regard, similar results were obtained in ELMo and GloVe. Common expressions that stand out in all three techniques with the term "yaşlı" are physically oriented labels focusing on obesity and both sexual insults and gender expressions. In addition, expressions emphasizing illness and pity are also noteworthy. Moreover, it is possible to say that, in comparison to other techniques, GloVe in the Turkish dataset contains slightly less aggressive and divisive expressions.

Looking at the English dataset in Fig. 3, unlike the Turkish dataset, there were fewer sexual and gender-related associations (none in ELMo and GloVe), and generally more associations related to health and specifically mental health. In addition, a notable difference in the English dataset compared to the Turkish dataset is the abundance of expressions related to economic insufficiency in all three techniques, and fewer physically oriented labels such as obesity. Thus, it can be said that in English datasets, aging is more associated with economic and health issues. In contrast, it is associated with physical and gender-based associations in Turkish datasets.

In general, when looking at the overall picture, it is not wrong to say that these references to elderly individuals reveal sociocultural stereotypes about aging. Stereotypes contain clues about negative preconceptions in society about the facts they describe (Yang et al., 2021). In this context, it is clear that the results obtained contain social preconceptions about aging. However, the existence of these preconceptions in artificial intelligence algorithms can lead to extremely unhealthy results for the continuation of social problems.

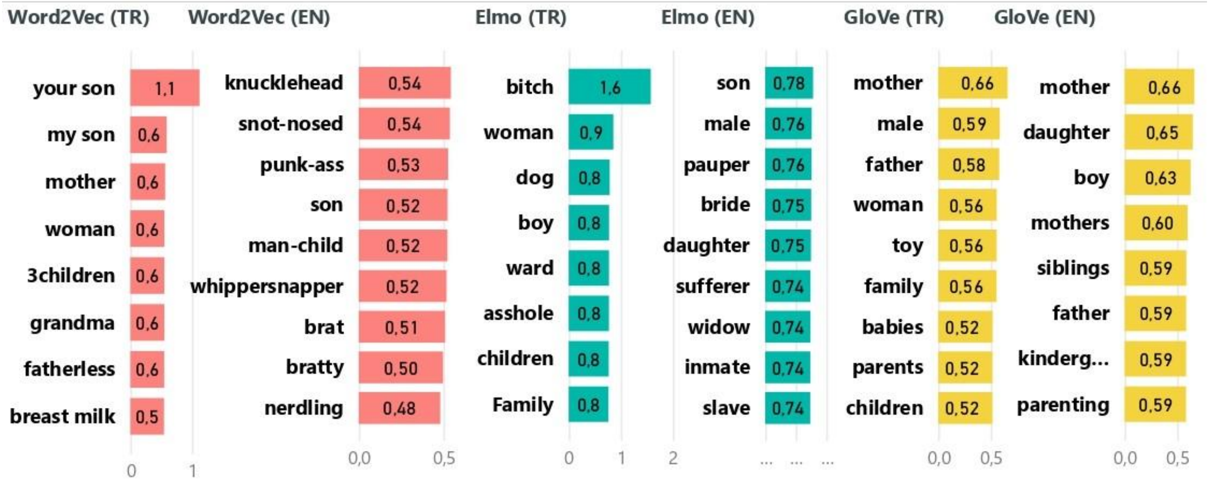

Figure 4: Word2Vec, Elmo, and GloVe results for **Child**

The results obtained for the word "child" in the Turkish dataset of Word2Vec reveal a social pattern, contrary to the negative stereotypes associated with the word "elderly": A focus on son and mother in Fig. 4. Unfortunately, "son" still carries a male-dominated connotation in some cultures. The pattern that emerges here can be interpreted as a concrete example of positive discrimination towards male children due to emphasis on words like "my son," "my dear son," and "my little son" (these words are written as a single word in Turkish as "oğul," "oğlum," and "oğluşum," respectively). The frequency of words like "my daughter," "my little daughter," and "your daughter" (written as a single word in Turkish as "kız," "kızım," and "kızın," respectively) is significantly lower than that of "my son." A similar gender emphasis can also be observed in using the word "mother," which is more associated with children than fathers. The current picture suggests that the relationship between child and mother has become synonymous, with the father's role fading into the background (see Fig. 4).

Unlike other datasets, in the Turkish dataset of ELMo, expressions related to sexual assault are more common. However, GloVe shows fewer expressions related to sexual assault and more balanced results in terms of gender, with expressions like "woman," "man," "girl," "boy," and so on. When looking at the English datasets, Word2Vec has encountered much more sarcastic and belittling expressions compared to other tools when it comes to the word "child." The English dataset of ELMo has produced very similar results to the Turkish dataset.

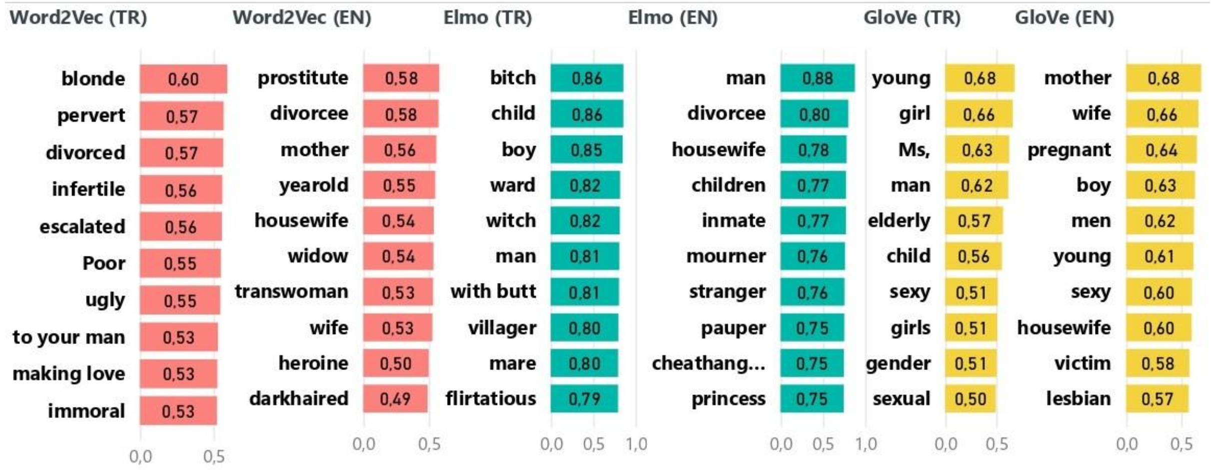

Figure 5: Word2Vec, Elmo, and GloVe results for **Woman**

When looking at the word "woman," both sexist insults and stereotypical gender references are prominently observed in all three techniques and both languages in Fig. 5. Words such as divorced, barren, obsessed, and pervert have negative connotations towards women's marital status and sexuality. Negative descriptions such as blonde, ugly, immoral, and poor give clues to physical attributions towards women and reduce the meaning of womanhood to binary gender roles, emphasizing the negativity of men (see Fig. 5).

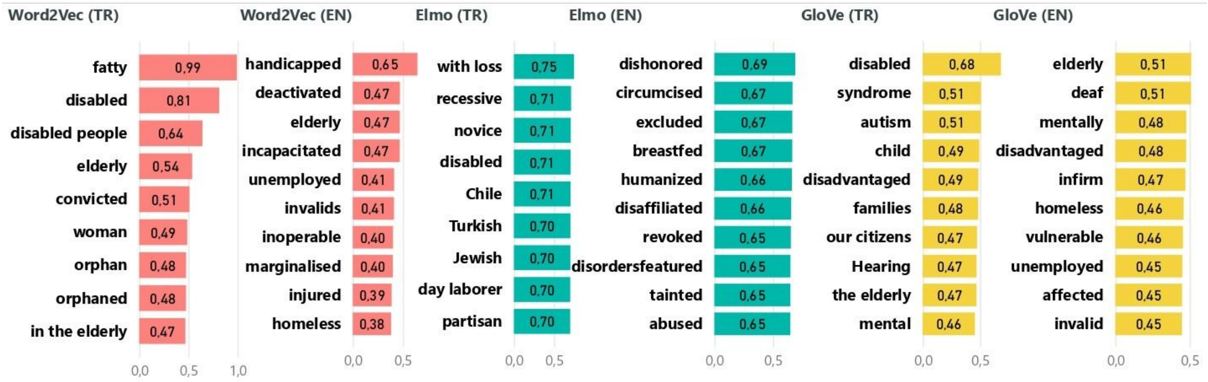

Figure 6: Word2Vec, Elmo, and GloVe results for **Disabled**

In addition to the topics discussed in detail above, when we examine the disabled word in the Turkish dataset, we frequently encounter expressions such as "özürlü" and "kusurlu" that are considered offensive in Fig. 6. It is observed that stereotypical expressions that cannot be associated with being disabled stand out in all three tools and in both languages when looking at the word "disabled." Furthermore, particularly in ELMo's Turkish dataset, racial, religious, and political references are noteworthy. While in the English datasets of Word2Vec and ELMo, offensive expressions that cannot be associated with being disabled are noticeable, expressions of pity stand out in GloVe's English dataset (see Fig. 6).

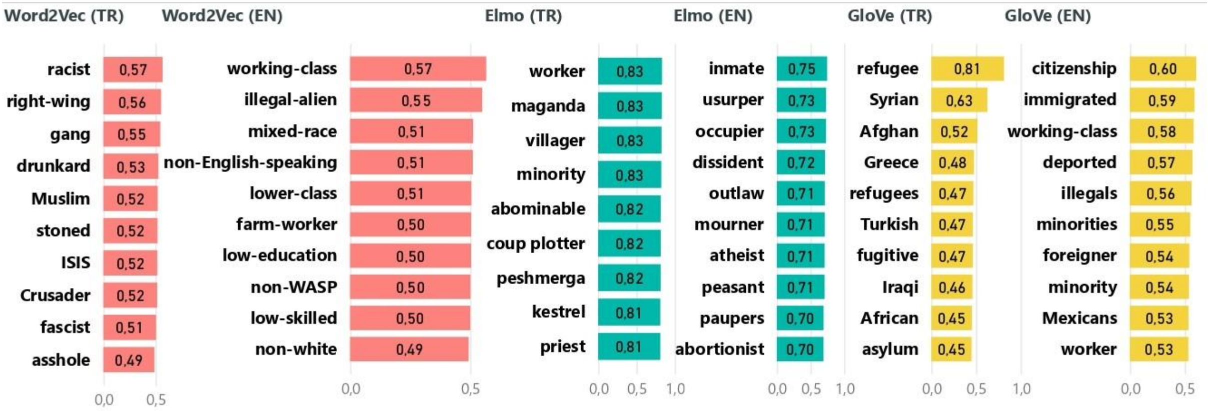

Figure 7: Word2Vec, Elmo, and GloVe results for **Immigrant**

When looking at the word "immigrant," striking results were also found. In all datasets, particularly in GloVe's Turkish dataset, race-based, religious, and political expressions stand out in Fig. 7.

Nevertheless, GloVe's datasets have the least number of offensive expressions. On the other hand, racist, gangster, drunkard, and thug, as well as general economic and religious definitions, are frequently encountered, especially in ELMo's English dataset. This situation contains disturbing clues about society's attitudes towards immigrants. A similar situation applies to the word "refugee" in Fig. 8. When comparing the tools, it can be seen that Word2Vec and ELMo lead to hate speech that cannot be associated with the topic of the words "immigrant" and "refugee." In addition, unlike the emphasis on insults in the Turkish dataset, it can be said that it is frequently used with derogatory descriptions in the English dataset. As a result, all expressions such as homosexual, terrorist, occupier, looter, disabled, fascist, barbarian, and dog are among the most frequently used words with "refugee."

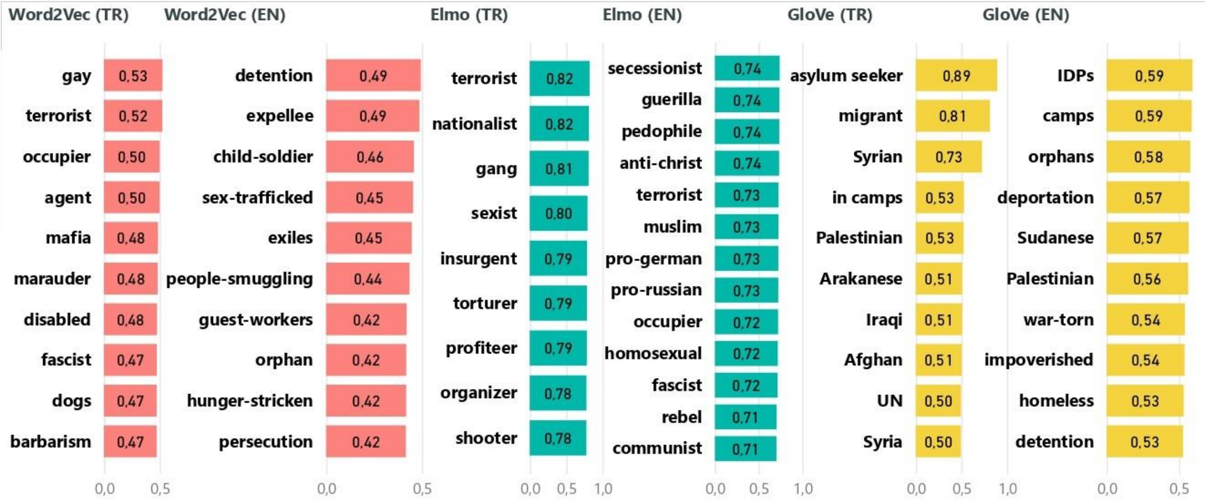

Figure 8: Word2Vec, Elmo, and GloVe results for **Refugee**

The word "homosexual" is frequently used with stereotypical expressions related to homosexuality as well as with unrelated derogatory expressions in both languages and all three tools (see Fig. 9). The fact that the word "pedophilia" is the most frequently mentioned word related to homosexuality is a disturbing association. Additionally, in ELMo's English dataset, unlike the others, derogatory expressions related to politics, religion, and race stand out. The existing tables show that society has serious problems with refugees, immigrants, and homosexuals. The extent of the problems that may arise in the use of artificial intelligence algorithms based on this table is quite worrying, considering that the data is publicly available.

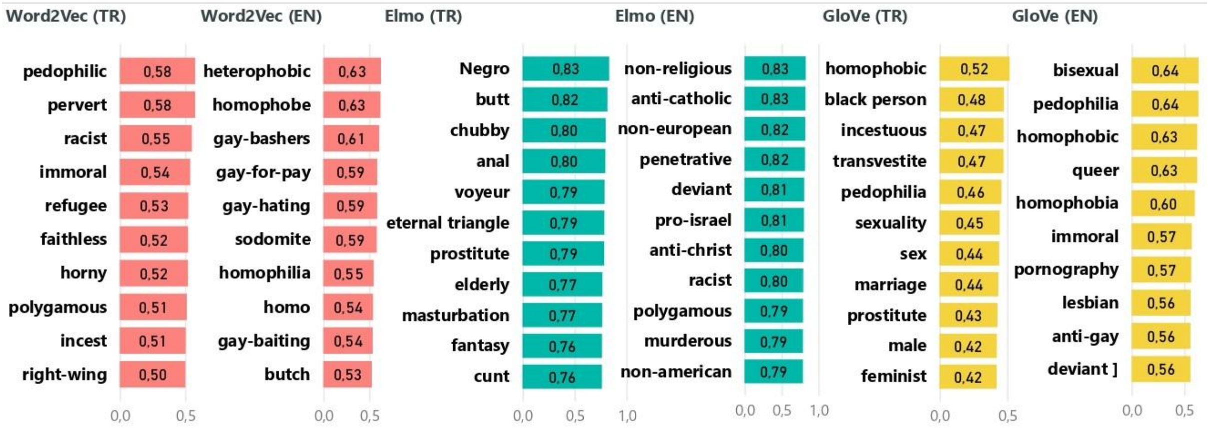

Figure 9: Word2Vec, Elmo, and GloVe results for **Homosexual**

When examining words such as "homeless," "sick," and "poor" in both languages and all three tools, clichéd emphasis, and extremely derogatory expressions stand out (see Fig. 10, 11, and 12). Especially the word "homeless" is frequently accompanied by hate speech unrelated to the subject. According to this table, it is possible to say that society's attitude towards "homeless people" is worrisome. However, it was observed that the expression "poor" is relatively more common with expressions of pity. Another noteworthy point is that the word "orphan" appears together with politically oriented expressions based on race, religion, and ethnicity in ELMo's English dataset, which is completely different from other datasets (please see Fig. 13). A similar noticeable difference applies to the word "sick" as well. In the English dataset of all three tools, unlike the Turkish dataset, the word "sick" contains severe derogatory expressions. When comparing the tools, it can be seen that ELMo's Turkish dataset also has similar derogatory expressions that are relatively weighty compared to other Turkish datasets. However, regarding the word "sick," it can be said that GloVe's Turkish dataset has the least derogatory expressions and the most realistic expressions related to the word "sick."

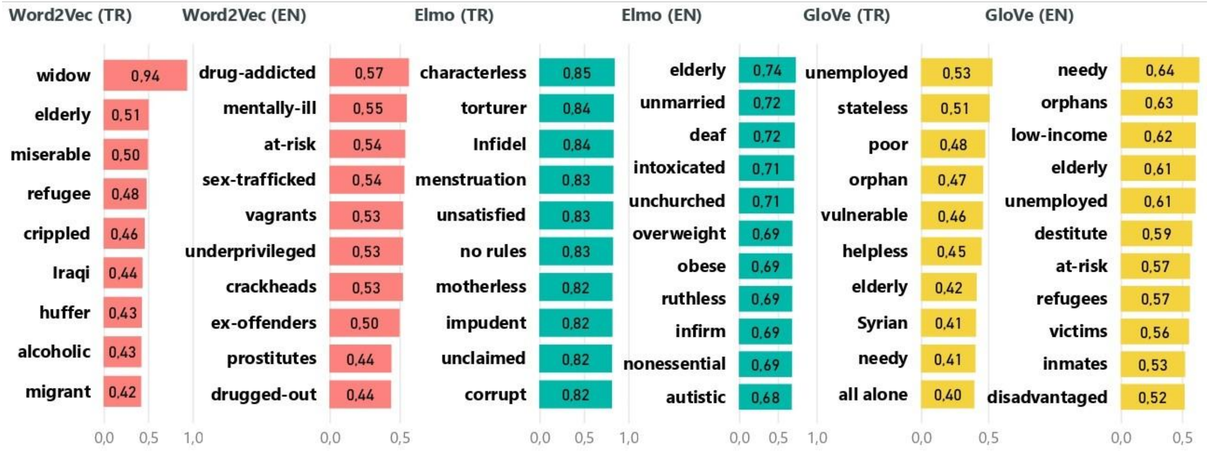

Figure 10: Word2Vec, Elmo and GloVe results for **Homeless**

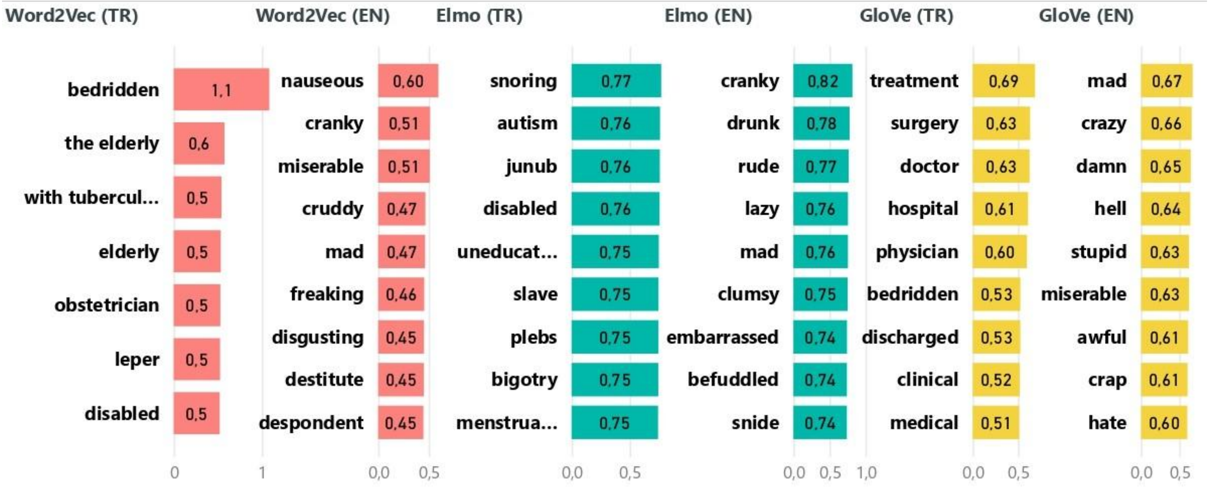

Figure 11: Word2Vec, Elmo and GloVe results for **Sick**

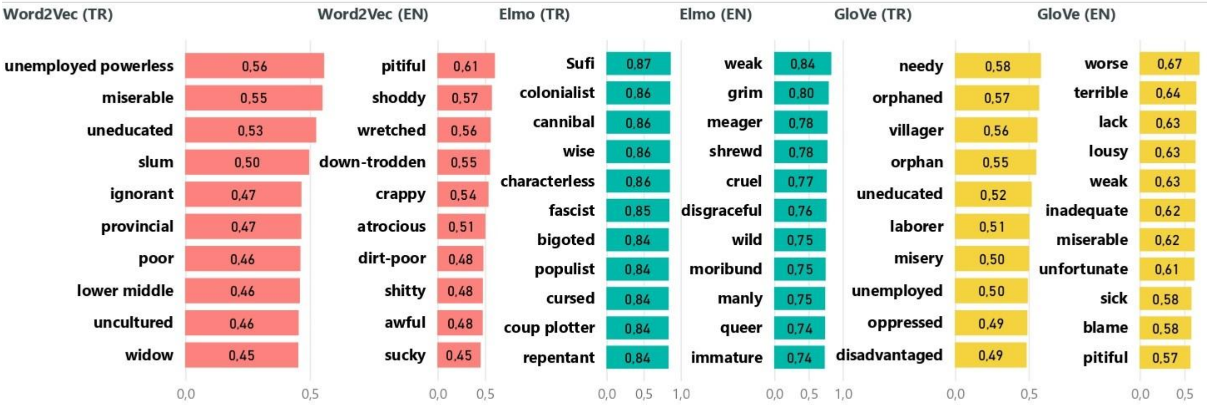

Figure 12: Word2Vec, Elmo and GloVe results for **Poor**

In summary, the Turkish dataset contains highly aggressive content, particularly with regard to terms related to migration. On the other hand, when words evoke distressing situations such as illness or poverty, the observations show more cliche expressions rather than insults. Similarly, in the English dataset, derogatory attitudes towards minority groups such as immigrants and refugees are noticeable, while highly aggressive expressions are encountered for words that describe deprivation and hardship, such as homeless, sick, and orphan. This situation may indicate less anger and more compassion towards vulnerable situations and people at the societal level. However, the situation for minority immigrants in the Turkish dataset is much more worrying, which may suggest that the effects of migration and social unrest are strongly felt at the societal level. Moreover, according to the English dataset, minority elements cause relatively less concern at the societal level, but the frequency of aggressive expressions towards situations of helplessness, such as illness, poverty, and homelessness, is striking.

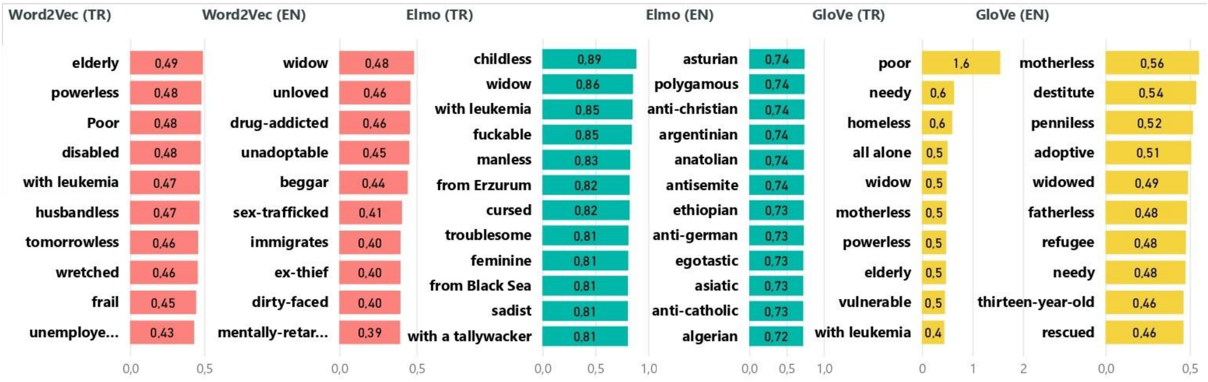

Figure 13: Word2Vec, Elmo, and GloVe results for **Orphan**

To summarize, the current situation shows, as previously noted, that the datasets used relatively commonly in artificial intelligence contain biased expressions that can generally be dangerous for vulnerable and disadvantaged groups. However, these can vary from culture to culture. It is expected that the current picture will contribute to anticipating the problems that could arise due to artificial intelligence technologies developed using these datasets facing society. The danger we have presented in this section is analogized to the masses receiving television data through intelligent systems created with current data. The ethical aspects of this are discussed in the next section.

4. CULTURAL CONTEXT

In this section, we reinterpret the cultivation theory (Gerbner et al., 1986), which is the most common mass communication theory ever, in the perspective of AI learning, to draw attention to the possible social and mass problems that AI learning may cause due to the data it uses.

Cultivation theory, conceptualized by George Gerbner in the 1960s and 1970s, argues that television has significant effects on individuals' perceptions of social reality. The theory is one of the three most referenced sources in communication sciences between 1956 and 2000 (Bryant & Miron, 2004). It is also the most cited theory in 16 academic journals published between 1993 and 2005 (Potter & Riddle, 2007).

The theory has been questioned with every technological advancement of the media. In the last sixty years, the mass media space has witnessed the spread of cable, satellite, video games, and, most recently, social media. Since 2000, over 125 studies have endorsed the theory, pointing to its ability to adapt to the ever-changing media landscape. However, to date, most studies on the effects of cultivation theory on current mass media technologies have focused on the effects of these technologies on individuals and communities (Haans 2000; Dahlstrom and Scheufele 2010; Hwang and Jeong 2020; Stein et al. 2021). This study opens a different perspective on the theory in question for discussion.

In our opinion, individuals who were positioned passively in the face of mass messages in the age of television are in the position of themselves responsible for mass messages in the age of AI. On the other hand, in a way, AI technologies have taken the place of passive audiences in the television age. In other words, the cultural ecosystem built on the masses with television messages -by "sowing" in Gerbner's words is being built today in the training of AI technologies.

Cultivation theory claims that watching television affects people's view of the world and typically leads to a heightened sense of risk and insecurity (Gerbner & Gross, 1976). The theory is based on two interrelated statements: (1) television programs represent a coherent but strikingly distorted view of the real world, and (2) over-watching of these coherent and formulated distorted representations mediates the shaping of viewers' worldviews (Morgan et al., 2015; Shanahan et al., 1999; Shrum, 2009). In other words, in his research, Gerbner tries to analyze culture through the media by using the content analysis method and claims that the cultural values transmitted from television create a total misconception. According to him, television sows and cultivates attitudes and constitutes the common source of the daily culture of the population. The desired environment or individual takes a long time to form. At the end of the process, an artificial culture is created through television.

Based on all these, we think data can be associated with culture and politics in two ways specific to our study. The first is that the things that exist in the daily life and culture of the society are produced over and over again in the internet environment and have turned into data stacks and presented to the use of the society again through AI algorithms fed with this data. The second is that the things that are not in daily life and culture turn into data stacks in fictional forms through the liberation and virtuality provided by the internet environment and are included in the social process through AI algorithms fed with these data and direct the cultural transformation. Thus, data-centered socialization has emerged.

For a similar objective reason, Gerbner has shaped his research around the television.

Techno-cultural structure of the period. In other words, unlike other mass media, individuals do not need to wait, research or go out on television because the television is waiting for them directly in the middle of the house. As a family member, television constantly tells its stories patiently and persistently. The perception of the social reality of individuals in contact with only a small part of their physical and social environments in their daily lives is greatly influenced by representative experiences based on what they hear, see, and read without direct experiential stimuli (Gerbner, 1998). As individuals' image of reality becomes subject to the symbolic environment of the media, the influence of the symbolic environment on individuals increases (Bandura, 2001). In summary, Gerbner's theory is closely related to the techno-cultural structure of his period and its concrete indicators, such as increasing violence and discrimination.

Looking at the present, similar techno-cultural indicators are striking (See: Code Bias Documentary, Amazon recruiting robot, etc.). As the world transforms into data-centric socialization, a “network society,” the media's role in shaping public discourse will become increasingly important (Castells et al., 2005). Also affecting the importance of this role is the prevalence of the media, thanks to online channels and 24-hour cable news channels. One of the most influential and useful definitions of 'network society' equates the term inextricably with the rise of digital networks, stating that 'the formation and exercise of power relations have been decisively transformed by the rise of global digital communications networks as the core force. the symbol processing system of our time' (Castells et al., 2005). The result is that the capacity of media bias to influence discourse is stronger than ever. In other words, while the internet environment forms and reinforces our attitudes (Castells, 2014); On the other hand, it provides a ground on which we leave our digital footprints on these attitudes, especially with the luxury of being "anonymous''. It is known that anonymity triggers cyber aggression (Smith, 2015). Therefore, the footprint we leave in the digital ecosystem contains more radical and dirty attitudes than it can be in real life.

Discrimination and violence are problems that shape social attitudes and discourses due to television messages. They are also present in the data and algorithms that comprise autonomous systems such as smart cities, vehicles, and factories. problems such as smart cities, vehicles, factories, etc. It is also present in the data and algorithms that make up autonomous systems such as Because there is ample evidence of the violent elements contained in the wider discourse of the network society (Helbing, 2019). Unless regulations are made regarding whether the data full of violent elements is fair or which data can be used more accurately in which context, it will be inevitable for such intelligent ecosystems to host today's cultural problems. Because it is impossible for a structure that we have grown with our own seeds not to bear traces of us. In this sense, it seems essential for a healthy social structure in the long run to raise artificial intelligence, which will always carry a piece of us, as sensitively as raising a child correctly.

In today's version of the Gerbner-inspired sowing and cultivation theory, we are in the position of planting and scattering our seeds with our digital footprints, not the minds in which the seeds are planted. This analogy requires that the seeds were sown before, that is, we have grown and matured. Although we have been shaped by the biased and unhealthy messages of the media, we think that it is important to progress by eliminating the unhealthy ones of the seeds, namely the data, to get yield from the seeds we will grow, to avoid the problems we experienced while growing, and to strengthen the data. Thus, it will be possible for the seeds sown in today's techno-cultural civilization to yield the desired product over time.

This theory of Gerbner basically states that television is a narrator from a family of people, and accordingly, broadcasts on television create a monopoly on thoughts and consciousness. In short, Gerbner explains the phrase "too much television watching makes you stupid" scientifically with this theory. The same is true for data. Too much and unorganized data from multiple perspectives will create unintelligent systems and people.

For this reason, we think that the organization and standardization of data are vital today and in the future. Moreover, we believe that social sciences and computer sciences should be tightly integrated with implementing these procedures. Finally, we think this standardization must be auditable and transparent, and we hope to draw attention to the importance of the subject and create insights for future studies.

5. DISCUSSION AND CONCLUSION

In recent years, social problems such as racism, sexism, and ageism caused by algorithm errors have become remarkable. Since such systems can be used in many safety-critical environments where life-changing, crucial decisions are made. Therefore, many new approaches have been proposed, and some approaches have been modified. In this study, we establish the idea that the evaluation of algorithms independently from data can cause problematic results in today's environment, where algorithms are used to make important decisions at important points. For solving problems caused by algorithm errors, we present two perspectives. First, the importance of a holistic-driven approach, and second, the importance of an interdisciplinary assessment.

Within this context, first, an experiment was conducted using three different popular algorithms and datasets. These algorithms are Word2Vec (Mikolov, Tomas; et al., 2013) with the fastText dataset (https://fasttext.cc/) from the Facebook (Meta) company, GloVe algorithm with its dataset from Stanford University (Pennington, J.; et al. 2014), and ELMO (Embeddings from Language Model) algorithm with its dataset from Allen Institute at the University of Washington (Peters, M. E.; et al., 2018). They are to define the function itself as data and present both perspectives. At this point, our aim was to show the existing data and the potential problems of this data rather than the data that should be. The results demonstrate that the keywords we searched in the datasets were often found in conjunction with other highly offensive and stigmatizing words. When we interpret our results with a social scientific approach other than a computer scientist's perspective, we conclude that the current problem is not a new problem, and we show that the old problem in question can be handled from a different perspective from the current technological tools. We exhibit that the problems that arise due to the interaction of humans and television, which Gerbner put forward in the theory of cultivation and cultivation, are still up-to-date for artificial intelligence and human interaction in today's artificial intelligence technologies, by changing the objects that affect and are affected. As a result, we emphasize the importance of "handling the algorithm and data pair holistically" and "performing this type of research from an interdisciplinary perspective" in order to observe and prevent possible socio-cultural problems.

Hence one of the contributions of the paper is to develop a perspective that is compatible with the current social context at the point of understanding and solving the identified problem. Finally, we predict that if the current situation is not taken seriously, large-scale social problems that are difficult to overcome may arise.

As a future study, we intend to explain and outline the requirements for such a field in order to direct the academic productivity of scientists capable of conducting interdisciplinary studies in order to produce solutions to the problem we have defined. Simultaneously, it seeks to mediate the dismantling of preconceived notions about the integration of social and computer sciences, thereby assisting those investing in this field in overcoming numerous financial and moral impediments.

References

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black box fairness testing of machine learning models. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 625–635.

Bandura, A. (2001). Social cognitive theory of mass communication. *Media Psychology*, *3*(3), 265–299.

Bryant, J., & Miron, D. (2004). Theory and research in mass communication. *The Journal of Communication*, *54*(4), 662–704.

Buolamwini, J., & Gebru, T. (23--24 Feb 2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. *81*, 77–91.

Castells, M. (2014). The impact of the internet on society: a global perspective. *Change*, *19*, 127–148.

Castells, M., Cardoso, G., & Others. (2005). *The network society: From knowledge to policy*. Center for Transatlantic Relations, Paul H. Nitze School of Advanced ….

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, *3*, 81.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv Preprint arXiv:1609. 07236*.

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: testing software for discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510.

Gerbner, G. (1998). Cultivation Analysis: An Overview. *Mass Communication and Society*, *1*(3-4), 175–194.

Gerbner, G., & Gross, L. (1976). Living with television: the violence profile. *The Journal of Communication*, *26*(2), 173–199.

Gerbner, G., Gross, L., Morgan, M., & Signorielli, N. (1986). Living with television: The dynamics of the cultivation process. *Perspectives on Media Effects*, *1986*, 17–40.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(4), 1143–1185.

Helbing, D. (2019). Societal, Economic, Ethical and Legal Challenges of the Digital Revolution: From Big Data to Deep Learning, Artificial Intelligence, and Manipulative Technologies. In D. Helbing (Ed.), *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution* (pp. 47–72). Springer International Publishing.

Hertweck, C., Heitz, C., & Loi, M. (2021). On the Moral Justification of Statistical Parity. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 747–757.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Morgan, M., Shanahan, J., & Signorielli, N. (2015). Yesterday's New Cultivation, Tomorrow. *Mass Communication and Society*, *18*(5), 674–699.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., & Clark, C. (2018). Deep contextualized word representations. arXiv 2018. *arXiv Preprint arXiv*.

Potter, W. J., & Riddle, K. (2007). A content analysis of the media effects literature. *Journalism & Mass Communication Quarterly*, *84*(1), 90–104.

Shanahan, J., Shanahan, J., James, S., & Morgan, M. (1999). *Television and Its Viewers: Cultivation Theory and Research*. Cambridge University Press.

Shrum, L. J. (2009). Media consumption and perceptions of social reality: Effects and underlying processes. In *Media effects* (pp. 66–89). Routledge.

Smith, P. K. (2015). Cyberbullying and Cyber Aggression. In *Handbook of School Violence and School Safety* (pp. 111–121). Routledge.

Takan, S., Ergün, D., & Katipoğlu, G. (2023). Gamified Text Testing for Sustainable Fairness. *Sustainability: Science Practice and Policy*, *15*(3), 2292.

Townsend B., Paterson C., Arvind T. T. , Nemirovsky G., Calinescu R., Cavalcanti A., Habli I.,Thomas A. (2022) From Pluralistic Normative Principles to Autonomous-Agent Rules. Minds Mach. 32(4): 683-715.

Udeshi, S., Arora, P., & Chattopadhyay, S. (2018). Automated directed fairness testing. *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering - ASE 2018*, 98–108.

Yang, L., Tang, S., & Li, K. (2021). The Influence of the Disadvantaged Mindset on System-Justifying Beliefs. *Frontiers in Psychology*, *12*, 787417.