EXPLORING EMBEDDING BIAS IN MOVIE SCENARIO DATASETS USING *WEAT*

A PREPRINT

Seongjin Kim*
Research
Aiffel Online 5

September 21, 2023

ABSTRACT

1 Introduction

The field of Natural Language Processing (NLP) in Artificial Intelligence (AI) has witnessed remarkable advancements in recent years, marked by the continual emergence of new, high-performance Large Language Models (LLMs). However, as the capabilities of these models expand, so too does the need to address a critical concern: the presence of biases within the datasets used for model training. A significant proportion of these datasets originates from internet sources, including blogs and social networking platforms, where user-generated content reflects a broad spectrum of biases encompassing gender, nationality, political affiliation, social status, and more. Consequently, AI models trained on such data inherit these biases, necessitating a meticulous examination of the types and extents of bias within these models for ethical and responsible AI deployment.

In the realm of NLP, the crux of model training involves the transformation of textual data into high-dimensional word embeddings, wherein each word assumes a position within a semantic space, replete with intricate relationships to other words. Some word pairs exhibit proximity or distance in this space, akin to "apple" and "banana" in comparison to "desk." Amid this intricate web of relationships, a subtle yet pervasive issue emerges: "embedding bias." This bias manifests itself as a distortion in the relationships between words, exemplified by associations such as men with engineering and women with art.

The crux of our research endeavors to elucidate the intricate dynamics of word relationships within two distinct categories. Specifically, we aim to shed light on the means by which biases manifest in these relationships and, subsequently, how to discern and quantify these biases. Our investigation delves into the essence of embedding bias within NLP models, offering insights into methods for systematically analyzing relationships between words across different categories. Through this exploration, we aspire to contribute to the evolving discourse surrounding bias mitigation in AI models and to enhance the transparency, fairness, and ethical utilization of AI technologies.

2 Data Collection

The dataset utilized in this research consists of movie scenarios sourced from the Korean Box Office Information System (KOBIS), a comprehensive repository of cinematic data. This dataset forms the cornerstone of our investigation, offering a wealth of textual material that spans a wide spectrum of cinematic narratives and dialogues. The dataset underwent a two-fold preprocessing procedure, each serving distinct analytical objectives. The initial processing phase involved the classification of movie scenarios into two primary categories: "Artistic" and "General." This categorization serves as a fundamental distinction within our analysis, allowing us to assess the presence of embedding bias in the context

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

of these broad attributes. The criteria for categorization were established based on the artistic and thematic qualities inherent in the movie scenarios. Subsequently, a more granular categorization was performed, wherein the movie scenarios were classified into 21 distinct movie genres. These genres encompass a diverse array of cinematic themes and styles, including Science Fiction (SF), Family, Show, Horror, Documentary, Drama, Romance, Musical, Mystery, Crime, Historical, Western, Adult, Thriller, Animation, Action, Adventure, War, Comedy, and Fantasy. These genres are referred to as "attributes" within this study, and the assignment of scenarios to specific attributes was meticulously curated to ensure accuracy and relevance. This dual categorization approach forms the foundation of this analysis, allowing here to explore embedding bias within the nuanced context of both broad artistic distinctions and specific cinematic genres.

3 Preprocessing and Representative Word Selection

In the initial stage of data processing, our focus was on measuring embedding bias between the categories of "Artistic" and "General" movie scenarios. To achieve this, we meticulously curated the words within these categories through the following steps. For precise categorization, we selected only nouns from the movie synopses. To achieve this, we employed a Korean language tokenizer, specifically Okt, known for its effectiveness in segmenting Korean text. This tokenization process yielded a structured representation of nouns within the synopses. To further refine this dataset, it is transformed the tokenized words into TF-IDF (Term Frequency-Inverse Document Frequency) vectors. This vectorization method enhances the discriminative power of words by weighing them based on their importance within the context of individual synopses and across the entire dataset. From these TF-IDF vectors, it is carefully selected representative words for both "Artistic" and "General" movie scenarios. Representative words were chosen if they were distinct to their respective category and not present in the other. Each category was assigned a set of 2,000 such representative words. This rigorous selection process ensured the uniqueness and relevance of these words within their respective categories. In the subsequent processing phase, the aim was to create representative word sets for each of the 21 movie genres present in our dataset. This process involved the following steps. Similar to the initial preprocessing step, it is tokenized the synopses associated with each genre. This tokenization enabled the extraction of meaningful linguistic elements specific to each genre. To extract representative words, it is leveraged pre-trained FastText word embedding vectors, which were made publicly available on GitHub (GitHub - Kyubyong/wordvectors: Pre-trained word vectors of 30+ languages). These embeddings allowed us to capture semantic relationships among words while avoiding the inclusion of proper nouns, such as specific character names. The selection of representative words for each genre entailed additional steps to optimize the TF-IDF vectorization process. Specifically, the study explored suitable parameters, min-df (minimum document frequency) and max-df (maximum document frequency), both ranging from 0.0 to 1.0 across 100 increments. For each parameter combination, it is computed average and standard deviation values across the genre datasets. For instance, setting min-df to 0.0 and max-df to 1.0 resulted in an average value of 0.082 and a standard deviation value of 0.22. These values indicated the presence of almost uniquely distinct words when compared to settings of min-df at 0.3 and max-df at 1.0, which yielded average and standard deviation values of 0.22 and 0.39, respectively. These meticulous parameter adjustments facilitated the extraction of representative words, ensuring their relevance and distinctiveness within each genre. The resulting representative word sets, meticulously prepared through tokenization, embedding analysis, and parameter tuning, serve as the foundation for this investigation into embedding bias across categories and genres, as elaborated upon in the subsequent sections of this study.

4 Methodology

The process consists of two primary phases: measuring bias between categories (Artistic and General) and bias within movie genres. The Word Embedding Association Test (WEAT) serves as our principal tool for quantifying bias in both scenarios. To facilitate the measurement of bias between the "Artistic" and "General" categories, it is represented to have the curated representative words from both categories using word embeddings. These embeddings capture the semantic relationships between words in a high-dimensional vector space. Word Embedding Association Test (WEAT) is performed to quantify the strength and direction of bias between the "Artistic" and "General" categories. WEAT, a widely adopted method for assessing embedding bias, measures the relative association between target and attribute word sets. Target Word Sets (The representative words for "Artistic" and "General" categories.), Attribute Word Sets (Word sets representing attributes related to bias, such as gender, occupation, or sentiment.)

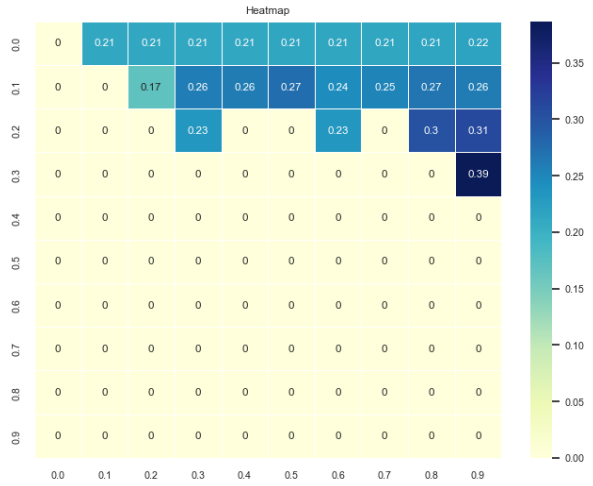

Figure 1: Sample figure caption.

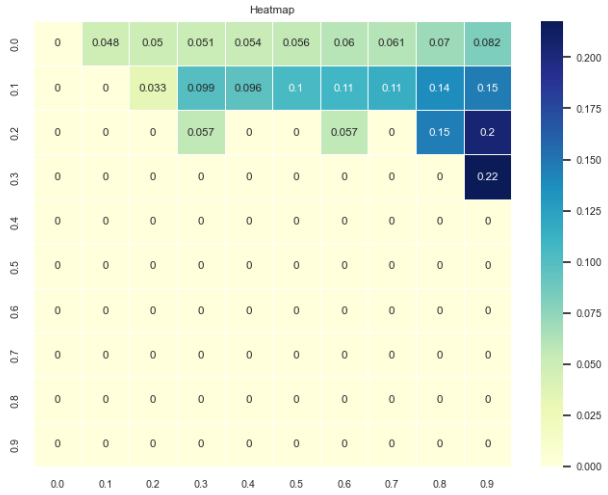

Figure 2: Sample figure caption.

5 Experimental Results

6 Discussion

7 Conclusion

References

[1] Aylin Caliskan, Joanna J Bryson, Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases *arXiv preprint arXiv:1608.07187*, 2016.

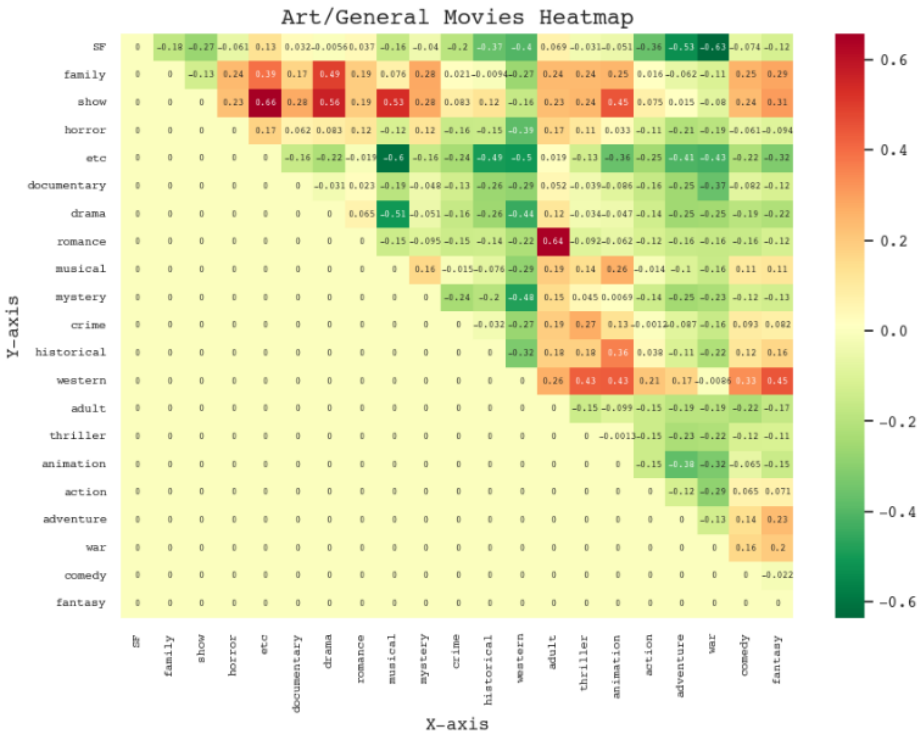

Figure 3: Sample figure caption.