# Incorporating Subjectivity into Gendered Ambiguous Pronoun (GAP) Resolution using Style Transfer

**Kartikey Pant** * and **Tanvi Dadu** *
Salesforce
{kartikey.pant,tanvi.dadu}@salesforce.com.

## Abstract

The *GAP* dataset is a Wikipedia-based evaluation dataset for gender bias detection in coreference resolution, containing mostly objective sentences. Since subjectivity is ubiquitous in our daily texts, it becomes necessary to evaluate models for both subjective and objective instances. In this work, we present a new evaluation dataset for gender bias in coreference resolution, *GAP-Subjective*, which increases the coverage of the original *GAP* dataset by including subjective sentences. We outline the methodology used to create this dataset. Firstly, we detect objective sentences and transfer them into their subjective variants using a sequence-to-sequence model. Secondly, we outline the thresholding techniques based on fluency and content preservation to maintain the quality of the sentences. Thirdly, we perform automated and human-based analysis of the style transfer and infer that the transferred sentences are of high quality. Finally, we benchmark both *GAP* and *GAP-Subjective* datasets using a BERT-based model and analyze its predictive performance and gender bias.

## 1 Introduction

In natural language, subjectivity refers to the aspects of communication used to express opinions, evaluations, and speculations, often influenced by one's emotional state and viewpoints. It is introduced in natural language by using inflammatory words and phrases, casting doubt over a fact, or presupposing the truth. Writers and editors of texts like newspapers, journals, and textbooks try to avoid subjectivity, yet it is pervasive in these texts. Hence, many NLP applications, including information retrieval, question answering systems, recommender systems, and coreference resolution, would benefit from being able to model subjectivity in natural language (Wiebe et al., 2004).

| | |
|---|---|
| **Objective Form** | The authors' **statements** on nutrition studies ... |
| **Subjective Form** | The authors' **exposé** on nutrition studies ... |

Table 1: Example sentence pair from the Wiki Neutrality Corpus, demonstrating the replacement of the word 'statements' into 'exposé' for inducing subjectivity in the sentence.

One of the prevalent biases induced by NLP systems includes gender bias, which affects training data, resources, pretrained models, and algorithms (Bolukbasi et al., 2016; Caliskan et al., 2017; Schiebinger et al., 2017). Many recent studies aim to detect, analyze, and mitigate gender bias in different NLP tools and applications (Bolukbasi et al., 2016; Rudinger et al., 2018; Park et al., 2018). The task of coreference resolution involves linking referring expressions to the entity that evokes the same discourse, as defined in tasks CoNLL 2011/12 (Pradhan et al., 2012). It is an integral part of NLP systems as coreference resolution decisions can alter how automatic systems process text.

A vital step in reducing gender bias in coreference resolution was the introduction of the *GAP* dataset, a human-labeled corpus containing $8,908$ ambiguous pronoun-name pairs derived from Wikipedia containing an equal number of male and female entities. This gender-balanced dataset aims to resolve naturally occurring ambiguous pronouns and reward gender-fair systems.

Text sampled from Wikipedia for the *GAP* dataset contains mostly objective sentences, as shown by the experiments performed in Subsection 3.2.1. Since subjective language is pervasive in our daily texts like newspapers, journals, textbooks, blogs, and other informal sources, it becomes essential to analyze the performance of different models for coreference resolution using subjective texts. Therefore, in this work, we introduce the subjec-

* Both authors have contributed equally to the work.



*Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 273 - 281

tivity attribute in the *GAP* dataset and analyze the performance of a BERT-based model on a newly proposed dataset.

In this work, we make the following contributions:-

1. We propose a novel approach for increasing coverage of the *GAP* dataset to include subjective text and release the GAP-Subjective dataset.

2. We outline each step in our dataset creation pipeline, which includes the detection of subjective sentences, the transfer of objective sentences into their subjective counterparts, and thresholding of the generated subjective sentences based on fluency, content preservation, and transfer of attribute.

3. We conduct automated and human evaluations to verify the quality of the transferred sentences.

4. We benchmark *GAP-Subjective* dataset using a BERT-based model and analyze its performance with the *GAP* dataset.

## 2 Related Works

### 2.1 Subjectivity Modeling and Detection

Recasens et al. (2013) conducted initial experimentation on subjectivity detection on Wikipedia-based text using feature-based models in their work. The authors introduced the "Neutral Point of View" (NPOV) corpus constructed from Wikipedia revision history, containing edits designed to remove subjectivity from the text. They used logistic regression with linguistic features, including factive verbs, hedges, and subjective intensifiers, to detect the top three subjectivity-inducing words in each sentence.

In Pryzant et al. (2019), the authors extend the work done by Recasens et al. (2013) by mitigating subjectivity after the detection of subjectivity-inducing words using a BERT-based model. They also introduced Wiki Neutrality Corpus (WNC), a parallel dataset containing pre and post-neutralization sentences by English Wikipedia editors from 2004 to 2019. They further tested their proposed architecture on the Ideological Books Corpus (IBC), biased headlines of partisan news articles, and sentences from a prominent politician's campaign speeches. They concluded that their models could provide valuable and intuitive suggestions to how subjective language used in news and other political text can be transferred to their objective forms.

The classification of statements containing biased language rather than individual words that induce the bias has been explored in Dadu et al. (2020). The authors perform a comprehensive experimental evaluation, comparing transformer-based approaches with classical approaches like fastText and BiLSTM. They conclude that biased language can be detected using transformer-based models efficiently using pretrained models like RoBERTa.

Riloff et al. (2005) explored using subjectivity analysis to improve the precision of Information Extraction (IE) systems in their work. They developed an IE system that used a subjective sentence classifier to filter its extractions, using a strategy that discards all extractions found in subjective sentences and strategies that selectively discard extractions. They showed that selective filtering strategies improved the IE systems' precision with minimal recall loss, concluding that subjectivity analysis improves the IE systems.

### 2.2 Gender Bias in Coreference Resolution

*OntoNotes* introduced by Weischedel et al. (2011) is a general-purpose annotated corpus consisting of around 2.9 million words across three languages: English, Arabic, and Chinese. However, the corpus is severely gender-biased in which female entities are significantly underrepresented, with only 25% of the 2000 gendered pronouns being feminine. This misrepresentation results in a biased evaluation of coreferencing models.

There has been considerable work on debiasing coreferencing evaluation concerning the gender attribute (Zhao et al., 2018; Webster et al., 2018). In (Zhao et al., 2018), the authors introduced a gender-balanced dataset *Winobias*, extending *Ontonotes 5.0* in an attempt to remove gender bias, containing Winograd-schema style sentences centered on people entities referred to by their occupation.

In Webster et al. (2018), the authors introduce the *GAP* dataset, a gender-balanced corpus of ambiguous pronouns, to address the gender misrepresentation problem. The dataset serves as an evaluation benchmark for coreference models containing over $8.9k$ coreference-labeled pairs containing the ambiguous pronoun and the possible antecedents. The coreference-labeled pairs are sampled from

Wikipedia and are gender-balanced, containing an equal number of instances for both male and female genders. This characteristic enables a gender-bias-based evaluation to be performed for any coreference model. They further benchmark the state-of-the-art model based on Transformers (Vaswani et al., 2017) against simpler baselines using syntactic rules for coreference resolution, observing that the models do not perform well in the evaluation.

### 2.3 Increasing Data Coverage

Asudeh et al. (2019) analyzed existing datasets to show that lack of adequate coverage in the dataset can result in undesirable outcomes such as biased decisions and algorithmic racism, creating vulnerabilities leading to adversarial attacks. For increasing coverage of the textual dataset, methods such as randomly swapping two words, dropping a word, and replacing one word with another one are heavily explored. On the other hand, generating new sentences to increase coverage via Neural Machine Translation (NMT) and style transfer remains a relatively less explored area.

Gao et al. (2019) explored soft contextual data augmentation using NMT. They proposed augmenting NMT training data by replacing a randomly chosen word in a sentence with a soft word, a probabilistic distribution over the vocabulary. Moreover, Wu et al. (2020) constructed two large-scale, multiple reference datasets, *The Machine Translation Formality Corpus* (MTFC) and *Twitter Conversation Formality Corpus* (TCFC), using formality as an attribute of text style transfer. They utilized existing low-resource stylized sequence-to-sequence (S2S) generation methods, including back-translation.

Textual style transfer has been explored extensively for the generation of fluent, content-preserved, attribute-controlled text (Hu et al., 2017). Prior works exploring textual style transfer in semi-supervised setting employ several machine-learning methodologies like back-translation (Prabhumoye et al., 2018), back-translation with attribute-specific loss (Pant et al., 2020), specialized transfer methodologies (Li et al., 2018), and their transformer-based variants (Sudhakar et al., 2019).

In a supervised setting where a parallel corpus is available, sequence-to-sequence models perform competitively. We use the OpenNMT-py toolkit (Klein et al., 2017) to train sequence-to-sequence models. Copy mechanism based sequence-to-sequence models with attention (Bahdanau et al., 2014) have been effective in tasks involving significant preservation of content information. They have been applied in tasks like sentence simplification (Aharoni et al., 2019), and abstractive summarization (See et al., 2017).

## 3 Corpus Creation

### 3.1 Preliminaries

#### 3.1.1 GAP Dataset

The *GAP* dataset, as introduced in Webster et al. (2018), is constructed using a language-independent mechanism for extracting challenging ambiguous pronouns. The dataset consists of $8,908$ manually-annotated ambiguous pronoun-name pairs. It is extracted from a large set of candidate contexts, filtered through a multi-stage process using three target extraction patterns and five dimensions of sub-sampling for annotations to improve quality and diversity. It is a gender-balanced dataset with each instance assigned one of the five labels - *Name A*, *Name B*, *Both Names*, *Neither Name A nor Name B*, and *Not Sure*.

The *GAP* is primarily an evaluation corpus, which helps us evaluate coreference models for the task of resolving naturally-occurring ambiguous pronoun-name pairs in terms of both classification accuracy and the property of being gender-neutral. The final dataset has a train-test-validation split of $4000 - 4000 - 908$ examples. Each example contains the source Wikipedia page's URL, making it possible for the model to use the external context if it may. The models are evaluated using the following two metrics: *F1 score* and *Bias* (Gender).

#### 3.1.2 Subjectivity Detection

For detecting subjectivity in sentences, we use the Wiki Neutrality Corpus (WNC) released by Pryzant et al. (2019). It consists of $180k$ aligned pre and post-subjective-bias-neutralized sentences by editors. The dataset covers $423,823$ Wikipedia revisions between $2004$ to $2019$. To maximize the precision of bias-related changes, the authors drop a selective group of instances to ensure the effective training of subjectivity detection models.

Dadu et al. (2020) shows that the RoBERTa model performed competitively in the WNC dataset achieving $0.702$ *F1 score*, with a *recall* of $0.681$ and *precision* of $0.723$. Following their work, we

train a RoBERTa-based model for detecting subjective sentences.

### 3.2 Approach

This section describes the methodology used for the creation of *GAP-Subjective*. It outlines the models used for detecting subjectivity in the original *GAP* dataset, followed by the methods used for transferring the objective sentences to their subjective counterparts. It also presents the thresholding techniques used on the generated subjective sentences based on fluency, content preservation, and transfer of attribute. Finally, it concludes by showing the results of the human evaluations conducted to verify the quality of the transferred sentences.

#### 3.2.1 Subjectivity Detection

In this section, we highlight the approach used for detecting subjectivity in the *GAP* dataset. Following the works of Dadu et al. (2020), we fine-tune the pretrained RoBERTa model using the WNC dataset for detecting subjectivity in the sentences. We randomly shuffled these sentences and split this dataset into two parts in a $90:10$ Train-Test split and performed the evaluation on the held-out test dataset. We used a learning rate of $2 * 10^{-5}$, a maximum sequence length of $50$, and a weight decay of $0.01$ for fine-tuning our model. Our trained model has $0.685$ *F1-score* and $70.01\%$ *accuracy* along with a *recall* of $0.653$ and *precision* of $0.720$. We then predict the subjectivity of the *GAP* dataset using the fine-tuned model and conclude that over $86\%$ of the sentences in the dataset are objective. Table 2 illustrates a data split wise analysis for the same.

#### 3.2.2 Style Transfer

In this section, we detail the process of performing style transfer of objective sentences present in the *GAP* dataset into their subjective variants. Our task of style transfer entails mapping a source sentence $x$ to a target sentence $\tilde{x}$, such that in $\tilde{x}$ the maximum amount of original content-based information from $x$ is preserved independent of the subjectivity attribute.

Firstly, we train a SentencePiece tokenizer on the English Wikipedia with a vocabulary size of $25000$. We consider numerical tokens as user-defined symbols to preserve them during the transfer process. Secondly, we train the style transfer model on the SentencePiece tokenized Wiki Neutrality Corpus using the OpenNMT-py toolkit. We use a $256$-sized BiLSTM layered architecture with a batch size of $16$, thresholding the gradient norm to have the maximum value of $2$ and share the word embeddings between encoder and decoder. We use the AdaGrad optimizer and use a multi-layer perceptron for global attention.

Importantly, we use the copy mechanism (Gu et al., 2016) for the sequence-to-sequence model. The mechanism has been proven beneficial in similar tasks, such as sentence simplification in a supervised setting (Aharoni et al., 2019). It is modeled using a copy switch probability over each token in the target vocabulary and each token in the context sequence at each decoding step. Hence, it allows the model to generate tokens that are not present in the target vocabulary. We hypothesized that using the copy mechanism in the models helps in preserving important entity-linked information like the associated pronoun and the names of the entities necessary for coreference resolution.

We obtain a validation perplexity of $3.10$ and a validation accuracy of $84.52\%$, implying that the model produced fluent and subjective sentences at large. To further improve the quality of the dataset, we then threshold these sentences across various metrics important for style transfer, as in recent works (Li et al., 2018; Sudhakar et al., 2019).

#### 3.2.3 Thresholding Transferred Sentences

This section details about the thresholding techniques used on the transferred sentences to maintain their quality. We perform the thresholding taking the following into consideration: fluency, content preservation, and transfer of attribute.

1. **Fluency**: We use the *OpenGPT-2* (Radford et al., 2018) as the language model to assign perplexity to the transferred sentences [1]. We compare the perplexity of the transferred sentences with the original sentences to test their fluency and discard all the sentences in which the perplexity change is more than $100$. This thresholding ensures relatively less change in the sentence structure, which is measured by the language model. Table 2 shows that $2,635$ in *development*, $603$ in *validation* and $2,641$ in *test* of *GAP-Subjective* are within the fluency threshold, comprising $44.02\%$ of the overall sentences.

[1] https://huggingface.co/transformers/perplexity.html

| Dataset Split | Total sentences | Within GLUE Threshold | Within Perplexity Threshold | Objective sentences | Final Thresholded Sentences ($A \cap B \cap C$) | Percentage of Final Thresholded Sentences |
|---|---|---|---|---|---|---|
| **Development** | 5995 | 2332 | 2635 | 5162 | 1736 | 28.9% |
| **Validation** | 1389 | 527 | 603 | 1183 | 377 | 27.1% |
| **Test** | 5971 | 2359 | 2641 | 5141 | 1800 | 30.1% |
| **Overall** | 13355 | 5218 | 5879 | 11486 | 3913 | 29.3% |

Table 2: Sentence-wise Thresholding Split

2. **Content Preservation**: We use sentence-level GLEU (Mutton et al., 2007) scores for determining the content preservation of the model. We compare the transferred sentence with their original counterparts as a reference. We consider all sentences having a GLEU less than 1.0 to ensure no sentence remains the same and more than 0.8 to provide a high level of similarity between the transferred sentence and the original sentence in terms of content information. As can be observed in Table 2, we preserve 39.07% of the overall sentences through the GLEU-based thresholding.

3. **Original Attribute**: We use the subjectivity model trained in Subsection 3.2.1 and filter out the sentences that are already subjective before transfer. Table 2 shows that 13.9% sentences in *development*, 14.83% in *validation*, and 13.90% in the *test* are subjective, corroborating that majority sentences in the dataset are objective, lacking coverage in terms of subjectivity as an attribute.

| | |
|---|---|
| **Original (Objective)** | She died the following January, aged about 22, giving birth to their only son. |
| **Transferred (Subjective)** | ***Unfortunately***, she died the following January, aged about 22, giving birth to their only son. |
| **Original (Objective)** | Her father, Philip, was a lithographic printer, and her mother, Ellen, ran a boarding house in Brixton; Kathleen was their third daughter. |
| **Transferred (Subjective)** | Her father, Philip, was a ***controversial*** lithographic printer, and her mother, Ellen, ran a boarding house in Brixton; Kathleen was their third daughter. |

Table 3: Example of transferred subjective sentences by the proposed approach

Table 3 illustrates the differences between the original objective and the transferred subjective sentences. We observe that the addition of the adverb *Unfortunately* in the original sentence makes it a subjective sentence, adding one's emotional state and viewpoints towards the event. Similarly, the addition of the adjective *controversial* changes the objective sentence to a subjective one.

| Split | Converted *GAP* Contexts |
|---|---|
| *test* | 63.85% |
| *development* | 60.60% |
| *validation* | 61.89% |

Table 4: Percentage of Converted GAP Contexts

Table 2 shows that 29.29% of the overall sentences are left after thresholding on all three metrics. We then replace the original sentences with their thresholded subjective counterparts. We observe that at least one sentence is transferred by our approach in over 60% of the GAP contexts. A data split wise analysis for the same is illustrated in Table 4.

#### 3.2.4 Human Evaluation

Although automated evaluation helps in the thresholding process for reconstructing *GAP-Subjective* and provides a significant indication of transfer quality, we perform human evaluation for a deeper analysis. We randomly sampled 68 sentences from the dataset containing 34 sentences each from the transferred sentences and original sentences in the human evaluation. The judges were asked to rank the sentence regarding its fluency and subjectivity. *Fluency* was rated from 1 (poor) to 5 (perfect). Similarly, *Subjectivity* was also rated from 1 (highly objective, factual) to 5 (highly subjective).

Table 6 illustrates the results of the human evaluation. We observe that the transferred sentences, on average, score 1.21 higher points on subjectivity than the original sentences. However, this increase in subjectivity comes with a minor 0.23 decrease in fluency.

#### 3.2.5 Offset Finding

We process each text to calculate the new offsets for the concerned pronoun and both the entities. Firstly, we determine the sentence in which the target word

| Dataset | Context |
|---|---|
| **GAP-Subjective** | *Unfortunately*, however, Stevenson suffered an injury while training and was replaced by Tyson Griffin. Gomi defeated Griffin by KO (punch) at 1:04 of the first round. Gomi would finish him with a *popular* left cross following up with a right hook causing Griffin to fall face first into the canvas where Gomi then followed up onto Griffin's back with few short punches before the fight was stopped. **He** is the first person to have stopped Griffin as all of Griffin's previous losses have gone to a decision. |
| **GAP** | However, Stevenson suffered an injury while training and was replaced by Tyson Griffin. Gomi defeated Griffin by KO (punch) at 1:04 of the first round. Gomi would finish him with a left cross following up with a right hook causing Griffin to fall face first into the canvas where Gomi then followed up onto Griffin's back with few short punches before the fight was stopped. **He** is the first person to have stopped Griffin as all of Griffin's previous losses have gone to a decision. |

Table 5: Sample Text from both datasets, GAP and GAP-Subjective

| | Fluency | Subjectivity |
|---|---|---|
| **Original Sentences** | **4.578** | 1.657 |
| **Transferred Sentences** | 4.343 | **2.872** |

Table 6: Results for Human Evaluation of the Transfer Model

was present in the original text. We then perform an exact match to find the word's position in the final transferred sentence. After finding the word's position in the sentence, we calculate the global offset for the word in the reconstructed text made of the final transferred sentences. This global offset represents the new offset for each entity.

| Dataset Split | Pronoun Found | Entity A Found | Entity B Found | All Found |
|---|---|---|---|---|
| **Development** | 99.90 | 98.65 | 99.00 | 97.55 |
| **Validation** | 99.34 | 99.78 | 99.56 | 98.68 |
| **Test** | 99.90 | 99.15 | 99.05 | 98.20 |

Table 7: Percentages of span offsets found in each data split

Table 7 represents the number of instances for which the offsets were successfully calculated as a percentage of total examples in each split. $97.55\%$ instances in *development*, $98.68\%$ instances in *validation*, and $98.20\%$ instances in *test* had correct offsets for all the three entities, thus showing that our offset finding approach was effective. To maintain the size of the dataset, we consider the original instance already present in the *GAP* dataset if the offset is not found.

Table 5 illustrates a sample context from *GAP* and *GAP-Subjective*, highlighting difference between the sentences of the context, the entity positions and the pronoun positions.

## 4 Benchmarking GAP-Subjective

### 4.1 GAP-Subjective Task

*GAP-Subjective* is an evaluation corpus that extends the *GAP* corpus by augmenting transferred subjective sentences for their objective counterparts. This dataset is segmented into *development* and *test* splits of $4,000$ examples each and *validation* split consisting of $908$ examples. The offsets for each entity and pronoun are given in the dataset. However, these offsets should not be treated as a gold mention or Winograd-style task.

We evaluate *GAP-Subjective* and compare it with *GAP* across two axes of evaluation: predictive performance, and gender bias. For assessing the predictive performance, we use an overall *F1 score*, denoted by *O*. We further calculate the *F1 score* for each of the two gendered pronouns, thus resulting in Male *F1* and Female *F1*, denoted by *M* and *F* respectively. We then calculate gender bias, indicated by *B*, which is defined as the ratio of feminine to masculine *F1 scores*, i.e., M/F.

### 4.2 Baseline Model

For benchmarking *GAP-Subjective*, we used the BERT-based architecture, introduced in Yang et al. (2019), that performs competitively in the GAP Challenge. The authors modeled the relations between query words by concatenating the contextual representations and aggregating the generated features with a shallow multi-layered perceptron. For a given query (Entity A, Entity B, Pronoun), they obtained deep contextual representations for the pronoun and each entity from *BERT*, where each entity is composed of multiple word pieces.

Following the work of Yang et al. (2019), we use the cased variant of $BERT_{Base}$ for benchmarking *GAP-Subjective*. We extract features from *BERT* using a sequence length of 128, batch size of 32, and embedding size of 768. For classification, we

| Dataset/Metric | Overall F1(O) | Precision(P) | Recall(R) | Masc-F1(M) | Fem-F1(F) | Bias(B) |
|---|---|---|---|---|---|---|
| GAP-Subjective | 0.789 | 0.772 | 0.807 | 0.786 | 0.792 | 1.007 |
| GAP | 0.796 | 0.778 | 0.815 | 0.802 | 0.790 | 0.984 |

Table 8: Results for the Benchmarking Experiments

train a multi-layered perceptron for 1000 epochs with 0.6 dropout rate, 0.001 learning rate, 0.1 L2 regularization and 32 batch size.

### 4.3 Results

Table 8 illustrates the benchmarking results for *GAP-Subjective* and *GAP* for the BERT-based architecture. We observe a significant change in the predictive performance of the BERT-based model for *GAP-Subjective* and *GAP*. We observe a decrease of $\sim 1\%$ in *F1-score*, and $\sim 2\%$ in *Masc-F1 (M)*, and a slight increase of $\sim 0.3\%$ in *Fem-F1 (F)*.

We also observe a change in the gender bias of the model between the two datasets. To understand this change, let us assume that the magnitude of deviation in bias score $m$ equals the absolute difference between the bias score and the ideal value 1 (which is obtained when there is no bias towards any of the two genders). While the model had a bias score of 0.984 in *GAP*, implying a preference towards male entities with the $m$ score of $1.6\%$. Interestingly, *GAP-Subjective* shows a minor preference towards female entities with a bias score of 1.007 and $m$ value of $0.7\%$.

## 5 Conclusion

In this work, we analyzed the addition of the subjectivity attribute in *GAP*, a widely used evaluation corpus for the detection of gender bias in coreference resolution. We utilized sentence-level supervised style transfer using sequence-to-sequence models to transfer the objective sentences in *GAP* to their subjective variants. We outlined the efficacy of our proposed style transfer approach using suitable metrics for content preservation and fluency and a human evaluation of the transferred sentences. We proposed a new evaluation corpus, *GAP-Subjective*, which consists of the reconstructed texts along with their new entity offsets. We benchmarked and analyzed the predictive performance and gender bias of BERT-based models in both *GAP* and *GAP-Subjective*. Future work may include increasing coverage of objective-heavy datasets for other downstream tasks and increasing the coverage of *GAP* using other attributes.

## Bias Statement

This paper studies two forms of biases: gender bias and subjective bias. We increase the coverage of the evaluation dataset for identifying gender bias in coreference resolution by converting objective data to its subjective counterparts. Since most of the original data were mined from Wikipedia, which has a "Neutral Point of View" policy ensuring that the data is objective, the models are evaluated for gender bias solely in a setting devoid of subjectivity. Since subjective bias is ubiquitous (Pryzant et al., 2019), adding subjectivity into the evaluation corpus becomes imperative when evaluating any form of bias. While evaluating the $BERT_{Base}$ model for the original *GAP* dataset, we found the model to prefer *Male* entities at large. In contrast, the same model trained and evaluated on the subjective counterpart *GAP-Subjective* was objective to prefer Female entities at large. Our work is based on the belief that the setting used for evaluation datasets for bias detection influences our understanding of capturing the bias in the evaluated systems.

## References


Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is

to computer programmer as woman is to homemaker? debiasing word embeddings.

Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186.

Tanvi Dadu, Kartikey Pant, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Kartikey Pant, Yash Verma, and Radhika Mamidi. 2020. Sentiinc: Incorporating sentiment information into sentiment transfer without parallel data. *Advances in Information Retrieval*, 12036:312 – 319.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Reid Pryzant, Richard Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically neutralizing subjective bias in text.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe, and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, page 1106–1111. AAAI Press.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Durme. 2018. Gender bias in coreference resolution. pages 8–14.

Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. In *EMNLP/IJCNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Linguist.*, 30(3):277–308.

Yunzhao Wu, Yunli Wang, and Shujie Liu. 2020. A dataset for low-resource stylized sequence-to-sequence generation. In *AAAI 2020*.

Kai-Chou Yang, Timothy Niven, Tzu Hsuan Chou, and Hung-Yu Kao. 2019. Fill the GAP: Exploiting BERT for pronoun resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 102–106, Florence, Italy. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.