

Datasheets for Chemical Named Entity Recognition (NER) datasets

Xingmeng Zhao, Ali Niazi, and Anthony Rios
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
{xingmeng.zhao, ali.niazi, anthony.rios}@utsa.edu

April 1, 2024

1 INTRODUCTION

Chemical Named Entity Recognition (NER) is vital in many applications within the biomedical domain, serving as a fundamental step in processing and extracting relevant information from vast texts. Nonetheless, the question arises: Do these models work equally well for all demographics? This paper presents an in-depth analysis of gender bias in chemical NER models. By deploying a synthetic and real-world data framework, we measure and evaluate gender bias, revealing disparities that could lead to significant implications in downstream tasks. This work not only shines a light on the biases present but also urges the community to consider these disparities in the development and application of NER systems to achieve fairness and accuracy in biomedical text mining.

2 Template

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The CHEMNER dataset was created to assess gender-related performance disparities in chemical Named Entity Recognition (NER) systems. The

dataset’s task was to develop a framework to measure gender bias using synthetic data and a newly annotated corpus with over 92,405 words that included self-identified gender information from Reddit. This was necessary because there was a notable gap in understanding whether these NER models worked equally well across different genders, and performance disparities had the potential to cause harm

rather than good. This paper aimed to fill this gap by providing a comprehensive examination of gender-related performance disparities in Chemical NER, focusing on both synthetic and real-world data to highlight biases and encourage the development of fairer chemical NER models.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The CHEMNER dataset was developed by researchers Xingmeng Zhao, Ali Niazi, and Anthony Rios from the Department of Information Systems and Cyber Security at the University of Texas at San Antonio.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

Any other comments? No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The CHEMNER dataset is composed of synthetic template data and real-world data derived from the r/AskDocs subreddit. The synthetic part uses templates filled with male

and female names alongside chemical entities from the Unified Medical Language System (UMLS) to assess if gender-related names are misclassified as chemicals and to examine the impact of gendered names on the prediction of other entities. The real-world component includes posts from Reddit’s r/AskDocs, annotated for chemical mentions, with self-identified gender information from users.

How many instances are there in total (of each type, if appropriate)?

The CHEMNER dataset comprises a significant number of instances, split between synthetic data and real-world data from Reddit:

Synthetic Data: The dataset contains a total of 5,600,000 instances, divided equally between male and female templates, with each gender having 2,800,000 instances. These instances were generated using templates filled with names and chemical entities, designed to assess gender bias in chemical NER models.

Real-World Data (AskDocs Corpus): From the r/AskDocs subreddit, the dataset includes 9,2405 words across all posts, with annotations for chemical mentions. This data is further categorized by the self-identified gender of the post authors, including separate counts for male and female instances. Specifically, it contains 1,501 instances for “AskDoc MALE,” 1,774 instances for “AskDoc FEMALE,” and a combined total of 3,275 instances for “AskDoc ALL.”

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative

of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset represents a sample of instances from larger sets of synthetic and real-world data, crafted specifically to assess gender bias in chemical NER models. For synthetic data, the instances were generated using templates filled with the most popular male and female names associated with chemicals, derived from a broader universe of possible name and chemical combinations. This approach ensures a focused exploration of gender bias within a controlled, yet representative, synthetic environment.

The real-world data, sourced from Reddit’s r/AskDocs subreddit, constitutes a sample of posts where users have self-identified their gender. This subset of Reddit data offers a snapshot of discussions around medical issues and chemicals, providing valuable insights into gender-specific language use and concerns. However, it may not fully represent the vast diversity of medical discussions online or the entirety of gender expressions and identities, given its source and the method of gender identification.

The representativeness of the real-world sample is inherently limited by the selection criteria (posts with clear gender identification) and the platform’s demographics. While not exhaustively representative of all online medical discussions or the global population’s gender diversity, the dataset offers a pertinent window into gender biases in chemical NER within the con-

text of social media discussions, specifically used to highlight gender disparities in the recognition of chemical entities.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in the dataset consists of both “raw” data and annotation, used to analyze gender bias in chemical NER models.

For the synthetic data, the raw data includes sentences constructed from templates that incorporate gender-specific names and chemical entities. These sentences mimic real-world mentions of chemicals in textual content, artificially created to assess how well NER models identify chemicals in contexts associated with different genders. The features in this part of the dataset are the annotations that specify which parts of the text are chemical entities and which are gender-related names, allowing for a focused examination of potential biases in model performance.

The real-world data from Reddit’s r/AskDocs subreddit consists of user-generated text posts. These posts are raw, unprocessed text that includes discussions, questions, and experiences related to health, medications, and chemicals, with self-identified gender information either directly stated or implied by the users. The annotated component include labels for chemical named entities identified within the text and the classification of the instance according to the self-identified gender of the user making the post. This setup enables a detailed analysis of how gender bias might manifest in

the automatic recognition of chemical entities within natural, user-generated text.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance in the dataset comes with labels. In synthetic data, labels distinguish between chemical entities and gender-specific names to test NER models for gender bias. For real-world data from Reddit’s r/AskDocs, labels identify chemical entities and categorize posts by the user’s self-identified gender, enabling an analysis of gender bias in chemical entity recognition.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No

Are there any errors, sources of noise, or redundancies in the

dataset? If so, please provide a description.

No

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset discusses the potential leakage of Personally Identifiable Information (PII) through the Reddit data that was collected and annotated for the CHEMNER dataset. Since the dataset involves user-generated content from the r/AskDocs subreddit, where individuals may share personal health-related information or experiences, there is an inherent risk of including PII or sensitive personal health information (PHI). However, the focus of the dataset’s creation and use

is on evaluating gender bias in chemical Named Entity Recognition (NER) models, with annotations for chemical entities and gender information, rather than on identifying or discussing individual users or their personal details.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the CHEMNER dataset relates to people in several ways. First, it includes synthetic data designed to evaluate gender bias in chemical Named Entity Recognition (NER) models, utilizing names typically associated with specific genders to assess the models' performance. Second, the real-world component of the dataset consists of data extracted from Reddit's r/AskDocs subreddit, comprising user-generated posts where individuals share personal health concerns, questions, and experiences, often including self-identified gender information. These aspects make the dataset directly relevant to human subjects, focusing on their interactions, personal health information, and the potential biases in NER systems that may impact them.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The CHEMNER dataset specifically identifies subpopulations by gender,

utilizing self-identified gender information from Reddit posts in the r/AskDocs subreddit to explore gender bias in chemical Named Entity Recognition (NER) models. This gender information is part of the annotations for the real-world data component of the dataset, where posts by users include mentions of their gender, often in the context of seeking medical advice or discussing health-related experiences. The synthetic data component is also designed with gender in mind, incorporating gender-associated names in the templates to assess the NER models' performance in recognizing chemical entities across different gender contexts. For the AskDoc, there are 1,501 chemical mentions associated with males and 1,774 with females, while the synthetic data maintains an equal number of instances for both genders.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The CHEMNER dataset is designed for analyzing gender bias in chemical Named Entity Recognition (NER) us-

ing synthetic data and real-world Reddit posts from r/AskDocs, focusing on health-related discussions. While not intentionally gathering sensitive information, user-generated content on health issues could include sensitive health data. The dataset’s aim is to improve NER models, not to collect data on personal attributes like racial or ethnic origins, sexual orientations, or other sensitive categories.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The CHEMNER dataset consists of synthetic data, created using templates filled with gender-specific names and chemical entities to simulate scenarios for evaluating gender bias in chemical Named Entity Recognition (NER) models, and real-world data from Reddit’s r/AskDocs subreddit, which includes user-generated posts discussing health-related questions and experiences at <https://www.reddit.com/r/AskDocs>.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How

were these mechanisms or procedures validated?

The collection of real-world data from Reddit’s r/AskDocs subreddit.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Two graduate student was involved in the CHEMNER dataset’s data collection and annotation process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset is created in 2021, consists of data gathered from the Reddit’s r/AskDocs subreddit over the period from January 1, 2005, to June 20, 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset includes posts from Reddit with self-identified gender information, used to assess gender biases in chemical Named Entity Recognition (NER) systems.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected from the Reddit community r/AskDocs, utilizing posts with self-identified gender mentions. This method does not involve direct collection from individuals but rather uses information available from a third-party source, Reddit.

Were the individuals in question notified about the data collection?

If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

For the newly developed corpus using data from the Reddit community r/AskDocs, posts with self-identified gender mentions were collected. A simple regular expression was utilized to identify self-identified gender, and the chemical entities in the dataset were annotated in two stages to increase the reliability of the labels. Initially, two graduate students annotated chemicals, resulting in an inter-annotator agreement of .874. Subsequently, a graduate student manually reviewed all disagreeing items to adjudicate the label and generate the gold standard. The annotation followed guidelines developed for the CHEMDNER corpus.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g.,

to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

No

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The data preprocessing, cleaning, and labeling were conducted using Flair, a powerful NLP library. You can find the software and more information on its capabilities at <https://github.com/flairNLP/flair>.

Any other comments?

No

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No

What (other) tasks could the dataset be used for?

The dataset, originally used to study gender biases in chemical Named Entity Recognition (NER), has potential for broader applications. It can support research on gender differences in health discussions, adverse drug reaction detection, pharmacovigilance, sentiment analysis, healthcare policy making, and linguistic studies.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might

impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset, centered on gender disparities in chemical Named Entity Recognition (NER) from Reddit posts, can raise considerations for future use. Because Its focus on binary gender identification may not fully capture the gender spectrum, posing a risk of reinforcing stereotypes or excluding non-binary identities. Additionally, there’s a potential for quality of service issues if models trained on this data perform unequally across genders. To mitigate these risks, future users should consider including more diverse gender identities, apply fairness-aware modeling techniques, ensure transparency about the dataset’s limitations, and address privacy and consent issues related to using publicly sourced data

Are there tasks for which the dataset should not be used? If so, please provide a description.

Given the dataset includes data from Reddit posts with potentially identifiable information, it should not be used for tasks aimed at identifying or profiling individuals, respecting user privacy and confidentiality.

Any other comments? No

Distribution

Will the dataset be distributed to third parties outside of the entity

(e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be publicly available on zenodo.org website.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is distributed as a compressed (.zip) file that contains annotations in the CoNLL format and synthetic templates, designed for the analysis of gender biases in chemical Named Entity Recognition (NER) systems.

When will the dataset be distributed?

The dataset will be distributed after the paper published.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the authors of the paper unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: A Comprehensive Study of Gender Bias in Chemical Named Entity Recognition Models. Xingmeng Zhao, Ali Niazi, and Anthony Rios. Proceedings of NAACL, 2024.

Have any third parties imposed IP-based or other restrictions on

the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No

Any other comments?

No

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Xingmeng Zhao and Anthony Rios will supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The curators of the dataset, Anthony Rios, can be contacted at <https://anthonyrios.net>.

Is there an erratum? If so, please provide a link or other access point.

No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

There are no plans to update the dataset at this time.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no limits on the retention of the information in the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

No

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?

No