WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events

Marco Antonio Stranisci[△ ea], **Rossana Damiano**[△], **Enrico Mensa**[△]
Viviana Patti[△], **Daniele Paolo Radicioni**[△], **Tommaso Caselli**[🌷]
[△]Dipartimento di Informatica, Università degli Studi di Torino, Italy
[🌷]CLCG, University of Groningen
[ea]aequa-tech, Turin, Italy
marcoantonio.stranisci@unito.it

Abstract

Biographical event detection is a relevant task for the exploration and comparison of the ways in which people's lives are told and represented. In this sense, it may support several applications in digital humanities and in works aimed at exploring bias about minoritized groups. Despite that, there are no corpora and models specifically designed for this task. In this paper we fill this gap by presenting a new corpus annotated for biographical event detection. The corpus, which includes 20 Wikipedia biographies, was compared with five existing corpora to train a model for the biographical event detection task. The model was able to detect all mentions of the target-entity in a biography with an F-score of 0.808 and the entity-related events with an F-score of 0.859. Finally, the model was used for performing an analysis of biases about women and non-Western people in Wikipedia biographies.

1 Introduction

Detecting biographical events from unstructured data is a relevant task to explore and compare bias in representations of individuals. In recent years, the interest in this topic has been favored by studies about social biases on allegedly objective public archives such as Wikipedia. Sun and Peng (2021) developed a resource for investigating gender bias on Wikipedia biographies showing that personal life events tend to be more frequent in female career sections than in those of men. Lucy et al. (2022) developed BERT-based contextualized embeddings for exploring representations of women on Wikipedia and Reddit.

The detection of biographical events has been addressed with complementary approaches by different research communities. Projects in Digital Humanities have focused mostly on representational aspects, delivering ontologies and knowledge graphs for the collection and study of biographical events (Tuominen et al., 2018; Fokkens et al., 2017; Plum et al., 2019; Krieger, 2014). When it comes to NLP, the focus has been mainly on developing models for the detection and classification of events (Rospocher et al., 2016; Gottschalk and Demidova, 2018). Few are the works that directly target biographies and focus on identifying biographical events with varied approaches (supervised and unsupervised) across different datasets (e.g., Wikipedia *vs.* newspaper articles), making their comparison impossible (Bamman and Smith, 2014; Russo et al., 2015; Menini et al., 2017). Although not directly targeting biographies, some works focused on the identification of entity-related sequences of events (Chambers and Jurafsky, 2008) and entity-based storylines (Chambers and Jurafsky, 2009; Minard et al., 2015; Vossen et al., 2016).

Despite the above mentioned variety of approaches to biographical event detection, there are pending and urgent issues to be addressed, which limit a full development of the research area. In particular, we have identified three critical issues: *i)* the lack of a benchmark annotated corpus for evaluating biographical event detection; *ii)* the lack of models specifically designed for detecting and extracting biographical events; and finally *iii)* the lack of a systematic study of the potential representation bias of minority groups, non-Western people, and younger generations in biography repositories publicly available, such as Wikipedia (D'ignazio and Klein, 2020).

Contributions Our work addresses these issues by presenting a novel benchmark corpus, a BERT-based model for biographical event detection, and an analysis of 48, 789 Wikipedia biographies of writers born since 1808. Our results show that existing data sets annotated for event detection may be easily re-used to detect biographical events achieving good results in terms of F-measure. The analysis of the 48, 789 biographies from Wikipedia extends the findings from previous work indicating that representational biases are present in an

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12370–12384
July 9-14, 2023

allegedly objective source such as Wikipedia along intersectional axes (Crenshaw, 2017), namely ethnicity and gender.

The rest of the paper is organized as follows. In Section 2, we present WikiBio, a novel manually annotated corpus of biographical events. Section 3 presents the experiments in event detection and coreference resolution of the target entities of a biographies. Section 4 is devoted to the analysis of the biases in Wikipedia biographies. Conclusions and future work end the paper in Section 5.

Code and WikiBio corpus are available at the following url: `https://github.com/marcostranisci/WikiBio/`.

2 The WikiBio Corpus

WikiBio is a corpus annotated for biographical event detection, composed of 20 Wikipedia biographies. The corpus includes all the events which are associated with the entity target of the biography.

In this section, we present our annotation scheme, discuss the agreement scores and present some cases of disagreement. Lastly, we present the results of our annotation effort, and compare them with existing corpora annotated for event detection and coreference resolution.

2.1 Annotation Tasks

Since the biographical event detection task consists in annotating all events related to the person who is the subject of a biography, annotation guidelines focus on two separate subtasks: (i) the identification of all the mentions of the target entity and the resolution of its coreference chains; and (ii) the identification and linking of all the events that involve the target entity.

Entity annotation. The entity annotation subtask requires the identification of all mentions of a specific Named Entity (NE) (Grishman and Sundheim, 1996) of type Person, which is the target of the biography and all its coreferences (Deemter and Kibble, 2000) within the Wikipedia biography. For the modeling of this subtask, we used the GUM corpus (Zeldes, 2017), introducing different guidelines about the following aspects: *i)* only the mentions of the entity-target of the biography must be annotated; *ii)* mentions of the target entity must be selected only when they have a role in the event (Example 1, where the possessives "his" is not annotated); and *iii)* indirect mentions of the target entity must be annotated only if they are related to biographical events (Examples 2 and 3).

1. Kenule Saro-Wiwa was born in Bori [...] **His** father's hometown was the village of Bane, Ogoniland.

2. **He** married Wendy Bruce, whom **he** had known since **they** were teenagers.

3. In 1985, the Biafran Civil War novel **Sozaboy** was published.

Event Annotation. Although there is an intuitive understanding of how to identify event descriptions in natural language texts, there is quite a large variability in their realizations (Pustejovsky et al., 2003b). Araki et al. (2018) point out that some linguistic categories, e.g., nouns, fits on an event *continuum*. This makes the identification of event mentions a non trivial task. Our event annotation task mainly relies on TimeML (Pustejovsky et al., 2003a) and RED (O'Gorman et al., 2016), where 'event' is "a cover term for situations that happen or occur." (Pustejovsky et al., 2003a)

Events are annotated at a single token level with no restrictions on the parts of speech that realize the event. Following Bonial and Palmer (2016), we introduced a special tag (`LINK`) for marking a limited set of light and copular verbs, as illustrated in Example 4. The adoption of `LINK` is aimed at increasing the compatibility of the annotated corpus with OntoNotes, the resource with the highest number of annotated events.

4. Ken Saro-Wiwa `<LINK>`**was**`<LINK/>` a Nigerian `<EVENT>`**writer**`<EVENT/>` `<LINK source='be' target ='writer' />`.

Lastly, to enable automatic reasoning on biographies, we annotate the contextual modality of events (O'Gorman et al., 2016). In particular, to account for the uncertainty/hedged modality, i.e., any lexical item that expresses "some degree of uncertainty about the reality of the target event" (O'Gorman et al., 2016), we have defined three uncertainty values: `INTENTION`, for marking all the events expressing an intention (like 'try' or 'attempt'); `NOT_HAPPENED`, for marking all events that have not occured; `EPISTEMIC`, which covers all the other types of uncertainty (e.g., opinion, conditional). The uncertainty status of the

Annotation Layer	A0 & A1	A0 & A2
Event	0.72	0.86
Entity	0.65	0.86
LINK	0.76	0.64
CONT_MOD	0.71	0.64

Table 1: Inter-Annotator Agreement (Cohen's Kappa).

events is annotated by linking the contextual modality marker and the target event, as illustrated in Example 5:

5. Feeling alienated, he **decided** to **quit** college, but was **stopped** [...]

```
<CONT_MOD source
='decided' target = quit'
value='INTENTION' />
<CONT_MOD source
='stopped' target = 'quit'
value='NOT_HAPPENED' />
```

Corpus Annotation and IAA. The annotation task was performed by three expert annotators (two men and one woman - all authors of the paper), near-native speakers of British English, having a long experience in annotating data for the specific task (event and entity detection). One annotator (A0) was in charge of preparing the data by discarding all non-relevant sentences to speed-up the annotation process. This resulted in a final set of 1,691 sentences containing at least one mention of a target entity. The entity and event annotations were conducted as follows: A0 annotated the entire relevant sentences, while a subset of 400 sentences was annotated by A1 and A2, who respectively labeled 200 sentences each. We report pair-wise Inter-Annotator Agreement (IAA) using Cohen's kappa in Table 1. In general, there is a fair agreement across all the annotation layers. At the same time, we observe a peculiar behavior across the annotators: there is a higher agreement between A0 and A2 for the event and entity layers when compared to A0 and A1, but the opposite occurs with the relations layers (LINK and CONT_MOD).

For the events, the higher disagreement is due to nominal events, often misinterpreted as not bearing an eventive meaning. For instance, the noun "trip" in example 6 was not annotated by A1.

6. When Ngũgĩ **returned** to America at the end of **his** month **trip** [...]

For the entities, we observed that disagreement is due to two reasons. The first is the consequence of a disagreement in the event annotations. Whenever annotators disagree on the identification of an event, they also disagree on the annotation of the related entity mention, as in the case of the pronoun 'his' in example 6. Another reason of disagreement regards indirect mentions. Annotators often disagree on annotation spans, as in "Biafran Civil War novel Sozaboy was published" where A1 selected 'SozaBoy', while A2 'novel Sozaboy'. When it comes to LINK, problems are mainly due to the identification of light verbs. Despite the decision of considering only a close set of copular and light verbs to be marked as LINK (Cfr Bonial and Palmer (2016)), annotators used this label for other verbs, such as 'begin' or 'hold'.

7. Walker **began** to take up reading and writing.

2.2 WikiBio: Overview and Comparison with Other Resources

The WikiBio corpus is composed of 20 biographies of African, and African-American writers extracted from Wikipedia for a total amount of 2,720 sentences. Among them, only 1,691 sentences include at least one event related to the entity target of the biography. More specifically, there are 3,290 annotated events, 2,985 mentions of a target entity, 343 LINK tags, and 75 CONT_MOD links.

Corpora size and genres We compare WikiBio against five relevant existing corpora that, in principle, could be used to train models for biographical event detection: GUM (Zeldes, 2017), Litbank (Sims et al., 2019), Newsreader (Minard et al., 2016), OntoNotes (Hovy et al., 2006), and TimeBank (Pustejovsky et al., 2003b). For each corpus, we took into account the number of relevant annotations and the types of texts. As it can be observed in Table 2, corpora vary in size and genres. OntoNotes is the biggest one and includes 159,938 events, and 22,234 entity mentions. The smaller is NewsReader, with only 594 annotated events. TimeBank and LitBank are similar in scope, since they both include about $7.5K$ events, while GUM includes 9,762 entity mentions.

Text types With the exception of GUM, which includes 20 biographies out of 175 documents, all other corpora contains types of texts other than

Corpus	Size	Text types	Relevant task
TimeBank	$7,471$ events	news	Event detection
OntoNotes	$159,938$ events, $22,234$ entity mentions	frame-theory	Event & Entity detection
NewsReader	594 events	TimeML	Event detection
GUM	$9,762$ entity mentions	biographies	Entity detection
LitBank	$7,383$ events	literary works	Event Detection

Table 2: A list of five existing resources that have been employed in the biographical event detection task.

biographies such as news, literary works, and transcription of TV news. To get a high-level picture of the potential similarities and differences in terms of probability distributions, we calculated the Jensen-Shannon Divergence (Menéndez et al., 1997). Such metric may be useful for identifying which corpora are most similar to WikiBio. The results show that WikiBio converges more with GUM (0.43), OntoNotes (0.48) and LitBank (0.49) rather than with TimeBank (0.51) and Newsreader (0.54). Such differences have driven the selection of data for the training set described in Section 3.2.

Annotations of entities, events, and coreference The distribution of the target entity within biographies in the WikiBio corpus has been compared with two annotated corpora for coreference resolution and named entity recognition: OntoNotes (Hovy et al., 2006) and GUM (Zeldes, 2017). Since such corpora were developed for identifying the coreferences of all NEs in a document, we modified annotations to keep only the most frequent NEs of type 'person' in each document. The rationale was making these resources comparable with WikiBio, which includes only the coreferences to a single entity, namely the subject of each biography. After doing that, we computed the ratio between the number of tokens that mention the target entity and the total number of tokens, and the ratio between the number of sentences where the target entity is mentioned against the total number of sentences. While this operation did not impact on GUM, in which 174 out of 175 documents contain mentions of people, it had an important impact on OntoNotes, in which $1,094$ documents (40%) do not mention entities of the type Person.

Tokens mentioning the target entity are 5% on OntoNotes, 8.7% on GUM and 4% on WikiBio. Such differences can be explained by the average length of documents in these corpora, which is of 388 tokens in OntoNotes, 978 in GUM, and $3,754$ in WikiBio. As a matter of fact, if the percentage of sentences mentioning the target-entity is considered instead of the total number of tokens, WikiBio shows an higher ratio (61.7%) of sentences mentioning the target entity, than OntoNotes (20.8%) and GUM (42.6%).

The three most frequently occurring lemmas in the WikiBio corpus seem to be strongly related to the considered domain: 'write' represents 3.2% of the total, 'publish' 2.9%, and 'work' 1.8%. 'Return' (1.3%) appears to have a more general scope, since it highlights a movement of the target entity from a place to another. The comparison with other corpora annotated for event detection shows differences concerning the most frequent events. The top three in OntoNotes (Bonial et al., 2010) are three light verbs: 'be', 'have', and 'do'. This may be intrinsically linked to its annotation scheme which considers all verbs as candidates for being events, including semantically empty ones (Section 2.1). NewsReader (Minard et al., 2016) and TimeBank (Pustejovsky et al., 2003b) include two verbs expressing reporting actions among the top five, thus revealing that they are corpora of annotated news. Litbank (Sims et al., 2019), which is a corpus of 100 annotated novels, includes in its top-ranked events two visual perception verbs and two verbs of movement, which may reveal the centrality of characters in this documents. The event 'say' is top-ranked in all the five corpora.

3 Detecting Biographical Events

In this section we describe a series of experiments for the detection of biographical events. Experiments involve the use of the existing annotated corpora for two tasks: entity mentions detection

	WikiBio	**GUM**	**Litbank**	**Newsreader**	**OntoNotes**	**Timebank**
WikiBio	0.00	0.43	0.49	0.54	0.48	0.51
GUM	0.43	0.0	0.49	0.54	0.39	0.49
Litbank	0.49	0.49	0.00	0.55	0.42	0.51
Newsreader	0.54	0.55	0.54	0.00	0.48	0.45
OntoNotes	0.48	0.39	0.42	0.48	0.00	0.40
TimeBank	0.51	0.49	0.51	0.45	0.40	0.00

Table 3: The similarity between corpora for event annotation computed with the Jensen-Shannon Divergence.

(Section 3.1) and event detection (Section 3.2). In both cases we used a 66 million parameters DistilBert model (Sanh et al., 2019). In this setting the WikiBio corpus is both used as part of the training set and as a benchmark for testing how well existing annotated corpora may be used for the task. For such experiments a NVIDIA RTX 3030 ti was used. The average length of each fine-tuning session was 40 minutes.

3.1 Entity Detection

For this task we adapted the annotations in OntoNotes (Hovy et al., 2006) and GUM (Zeldes, 2017) keeping only mentions of the most frequent entities of type 'person'. As a result we obtained 870 documents from OntoNotes, 174 from GUM.

The WikiBio corpus was split into three subsets: five documents for the development, 10 for the test, and five for the training. Given the imbalance between the existing resources and WikiBio, we always trained the model with a fixed number of 100 documents, in order to reduce the overfitting of the model over the other datasets.

Experiments consist in training a DistilBert model for identifying all the tokens mentioning the target entity of a given model and were performed on six different training sets. Since the focus of our work is to develop a model for detecting biographical events, WikiBio was used as development set for better monitoring its degree of compatibility with existing corpora. Following the approach by Joshi et al. (2020), we split each document into sequences of 128 tokens, and for each document we created one batch of variable length containing all the sequences. Table 4 shows the results of these experiments. As it can be observed, including the WikiBio corpus in the training set did not result in an increase of the performance of the model. This may be due to the low number of WikiBio documents in the training.The highest performance was obtained in two experiments: one using a training set only composed of documents from OntoNotes, which obtained a F-score of 0.808, and one with a miscellaneous of 50 OntoNotes and 50 GUM documents, that obtained 0.792. To understand if the difference between the two experiments is significant, we performed a One-Way ANOVA test over the train, development, and test F-scores obtained in both experiments. The test returned a p-value of 0.44, which confirms a significant difference between the two results

3.2 Event Detection

Event Detection experiments were guided by the comparison between WikiBio and the resources for event detection described in Section 2.2. Since OntoNotes was annotated according to the PropBank guidelines (Bonial et al., 2010), which only consider verbs as candidates for such annotation, we partly modified its annotations before running the experiments. We first adapted the OntoNotes semantic annotation by replacing light and copular verbs (Bonial and Palmer, 2016) with nominal (Meyers et al., 2004) and adjectival events. Then we ran a battery of experiments by fine-tuning a DistilBert-based model using each dataset for training, and a series of miscellaneous of the most similar corpora to WikiBio according to the Jensen-Shannon Divergence metric (Table 3). Since we were concerned with both assessing the effectiveness of WikiBio for training purposes and testing how far biographic events can be extracted, we designed our training and testing data as follows. WikiBio was employed in different learning phases: in devising the training set (i.e., existing resources were employed either alone or mixed with WikiBio); additionally, the development set was always built by starting from WikiBio sentences. Finally, we always tested on WikiBio data.

As for the entity-detection experiments, the 1,691 sentences containing events annotated in the WikiBio corpus were split into three sets of equal size that were used for training (564), development (563), and testing (564). Given the disproportion

Training	Dev	Test (30 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
Gum	WikiBio	WikiBio	0.820	0.728	0.752
Gum+WikiBio	WikiBio	WikiBio	0.819	0.728	0.753
Onto	WikiBio	WikiBio	0.896	0.782	**0.808**
Onto+WikiBio	WikiBio	WikiBio	0.846	0.774	0.800
Misc	WikiBio	WikiBio	0.824	0.766	0.792
Misc+WikiBio	WikiBio	WikiBio	0.828	0.764	0.789

Table 4: Results of entity detection experiments.

between OntoNotes and other corpora, we sampled a number of sentences for training which did not exceeded $5,073$, namely three times the number of sentences annotated in our corpus. Such length was fixed also for miscellaneous training sets.

Experiments were organized in two sessions. In the first session we fine-tuned a DistilBert model for five epochs, using as training set the five corpora presented in Section 2.2 individually as well as three combinations of them: *i)* misc_01, a miscellaneous of sentences extracted on equal size from all corpora; *ii)* misc_02, in which sentences from NewsReader, the most different corpus with WikiBio (Table 3), were removed; *iii)* misc_03, a combination of sentences from OntoNotes and Litbank, namely the two most similar corpora with WikiBio. The model was fine-tuned on these training sets both with and without a subset of the WikiBio corpus for a total of 16 different training sets. In addition, we also fine-tuned and tested WikiBio alone. We then continued the fine-tuning only for the models which obtained the best F-scores.

Observing Table 5, it emerges that, differently from entity-detection experiments, including a subset of WikiBio in the training set, even if in a small percentage, always improves the results of the classifier. This especially happens for Litbank ($+0.191$ F-Score), and TimeBank ($+0.031$ F-Score).

When looking at results of finetuning for single corpora, it emerges that the model trained on the modified version of OntoNotes and TimeBank obtains the best scores. Such results are interesting for two reasons. They confirm the intuition that OntoNotes annotations may be easily modified to account for nominal and adjectival events. They also confirm the high compatibility of WikiBio and TimeBank guidelines (Sect. 2.1). Even if the latter is more divergent from WikiBio than other corpora, it seems to be compatible with it. As expected for its limited size and high divergence with WikiBio, the training set based on NewReader sentences obtains the worst results, with an F-Score below 0.5.

Results of miscellaneus training sets are interesting as well: they generally result in models with better performance, and they seem to work better on the basis of their divergence with WikiBio. Trained on misc_01, a combination of all corpora, the model scores 0.827, which is below the result obtained with the modified version of OntoNotes. If Newsreader is removed, the model obtains 0.831, and 0.832 if also TimeBank is removed. It is also worth mentioning the delta between the F-score on the training and the test sets, which is -0.054 for misc_01, -0.029 for misc_02, and -0.013 for misc_03.

After the first fine-tuning step, we performed a One-Way ANOVA for testing the significance of differences between experiments. Analyzed in such a way, the four best-ranked models never showed a p-value below 0.5, which means that there are no significant differences between them. Thereby, we kept them for the second fine-tuning step that consists on training the model for 15 epochs on these datasets. Absolute results (Table 5) show that the model trained on Timebank obtained the best F-Score. However, as for the entity detection experiments, we considered the deltas between the training and test F-scores to select the best model for our analysis. All models acquired by employing a miscellaneous training set obtained a lower delta between training and test, and scored a similar F-Score.

4 An Intersectional Analysis of Wikipedia Biographies

In this section we provide an analysis of writers' biographies on Wikipedia adopting intersectionality as a theoretical framework and the model described in Section 3 as a tool for detecting biographical events.

The concept of intersectionality (Crenshaw,

Training | Dev | Test (5 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
WikiBio | WikiBio | WikiBio	0.479	0.479	0.479
Litbank | WikiBio | WikiBio	0.847	0.640	0.622
Litbank + WikiBio | WikiBio | WikiBio	0.835	0.814	0.813
Misc_01 | WikiBio | WikiBio	0.885	0.863	0.801
Misc_01 + WikiBio | WikiBio | WikiBio	0.871	0.831	0.827
Misc_02 | WikiBio | WikiBio	0.866	0.816	0.819
Misc_02 + WikiBio | WikiBio | WikiBio	0.861	0.837	**0.832**
Misc_03 | WikiBio | WikiBio	0.850	0.811	0.817
Misc_03 + WikiBio | WikiBio | WikiBio	0.844	0.839	0.831
Onto | WikiBio | WikiBio	0.950	0.800	0.790
Onto + WikiBio | WikiBio | WikiBio	0.936	0.873	0.809
Onto_mod | WikiBio | WikiBio	0.997	0.823	0.814
Onto_mod + WikiBio | WikiBio | WikiBio	0.888	0.869	0.829
Timebank | WikiBio | WikiBio	0.89	0.801	0.790
Timebank + WikiBio |WikiBio | WikiBio	0.865	0.856	0.821
NewsReader | WikiBio | WikiBio	0.453	0.479	0.479
NewsReader + WikiBio | WikiBio | WikiBio	0.467	0.479	0.479
Training | Dev | Test (15 EPOCHS)	**F-Score_train**	**F-Score_dev**	**F-Score_test**
Misc_01 + WikiBio | WikiBio | WikiBio	0.890	0.852	0.853
Misc_02 + WikiBio | WikiBio | WikiBio	0.900	0.855	0.856
Misc_03 + WikiBio | WikiBio | WikiBio	0.896	0.859	0.855
Timebank + WikiBio | WikiBio | WikiBio	0.919	0.850	**0.859**

Table 5: Results of event detection experiments: complete table

2017) has been developed in the context of gender and black studies to account inequalities that cannot be explained without a joint analysis of socio-demographic factors. For instance, African American women workers suffer higher discrimination than their male counterpart, as Crenshaw (1989) observed in her seminal work. Therefore, the injection of different socio-demographic features for the analysis of discriminations may unfold hidden forms of inequities about certain segments of population. We adopt this framework to analyse how the representations of non-Western women writers on Wikipedia differs from those of Western Women, Transnational Men, and Western Men.

For this analysis, we gathered $48,486$ Wikipedia biographies of writers born since 1808. We define as Transnational all the writers born outside Western countries and people who belong to ethnic minorities (Boter et al., 2020; Stranisci et al., 2022). Western men's biographies are $28,036$, Western women's $12,413$, Transnational men's $5,471$, and Transnational women's $2,470$. Information about occupation, gender, year of birth, ethnic group, and country of birth was obtained from Wikidata (Vrandečić and Krötzsch, 2014), which has been used for filtering and classifying biographies.

For each biography, we first identified all the mentions of the corresponding target entity (Section 3.1). We then removed the sentences that do not contain a mention of the entity. This reduced the number of sentences to be annotated for event detection from $1,486,320$ to $1,163,475$ (-21.8%). As a final step, we annotated events (Section 3.2) in the filtered sentences.

Table 6 shows the distribution of biographical events about men, women, Western, and Transnational people. The vast majority of events are about Western men (62.2%), while at the opposite side of the spectrum there are Transnational women writers, whose representation is below 5%. Ethnicity is a cause of underrepresentation more than gender: events about Transnational men are only 11.2% of the total, while those about Western women 21.4%. The average number of events per-author shows a richness in the description of Transnational Women (50.92 events) against Western ones (43.73 events).

The analysis of event types presents a similar distribution. $27,885$ event types – intended as the number of unique tokens that occur in each distribution – are detected in Western men's biographies (44.9 *per* biography), while only $9,254$ in Transnational women's biographies (40.4 *per* biography). However, the overlap of event types between these two categories is very large (92.6%) The same comparison, conducted on the other groups, reveals a higher number of group-specific event types: 87.8% of event types about Transnational Men are shared with Western Men, and the rate is lower for Western Women (84.1%).

A comparative analysis of most distinctive events per category of people provides additional insight about the representation of women and Transnational writers in Wikipedia biographies. In order to do so, we first computed the average frequency of each event in all biographies of the four groups of writers in Table 6. We then compared these distributions with the Shifterator library (Gallagher et al., 2021), which allows computing and plotting pairwise comparisons between different distribution of texts with different metrics. Coherently with the analysis performed in previous sections, we chose the Jensen-Shannon Divergence metric, and analyzed the distribution of events about Transnational Women against Transnational Men, Western Men, and Western Women. Table 7 shows the most diverging events between Transnational and Western writers, while Table 8 shows the 20 events about Transnational women that diverge most with other distributions: Transnational men, Western men, and Western women. Events are ordered on the basis of how much they are specific to the distribution of Transnational women. In Appendix A graphs with comparisons between distributions can be consulted.

A first insight from a general overview of distinctive events about Transnational Women writers is that they seem to never die. Events like 'death' or 'died' are never distinctive for them but always for the group against which they are compared. This may be explained by the average year of birth of Transnational Women writers with a biography on Wikipedia, which is 1951, while for Western men is 1936, 1943 for Transnational men, and 1944 for Western woman.

The analysis of the most salient biographical events between Transnational women and Transnational men shows how intersectionality helps to identify gender biases. When Transnationals are considered as a single group (Table 7) against the Western counterparts, the majority of the biographical events are related to career (award, conferred) or to social commitment (activist, migrated, exile). When the comparison is made within the Transnational group (Table 8), the gender bias demonstrated by Sun and Peng (2021) and Bamman and Smith (2014) clearly emerges. In fact, 'married', 'marriage', and 'divorce' are associated to Transnational women. In addition, there is a lack of career-related events about them, while this is not the case for men (actor, chairman, politician). The comparison between Transnational women and Western men still shows a gender bias, but less prominent. Among the most salient events, only 'mother' highlights a potential bias, while events on Transnational women career ('win', 'won', 'award', 'selected'), education ('degree', 'education', 'schooling') and social commitment ('activist') are present.

Group	Events	Avg	Types
Western M	$1,57M$	56.08	$27,885$
Transnational M	$285K$	52.10	$14,057$
Western W	$542K$	43.73	$17,324$
Transnational W	$125K$	50.92	$9,254$

Table 6: The number of events and event types broken down by gender and ethnicity of writers.

Finally, the comparison between Transnational and Western women offers three additional insights. First, the only event about private life which is salient for one of the two groups is 'married'. This indicates that private life events of women - in general - are always presented in relation to their conjugal status. Second, careers and social commitments are particularly present for Transnational women. Finally, the framing of the concept "relocation" is expressed using different event triggers: the more neutral 'move' is used for Western women, while the more marked, negatively connotated term 'migrate' is associated with Transnational women.

Summarizing, Transnational Women are underrepresented on Wikipedia with respect to other groups, both in terms of number of biographies and events. The analysis of their most distinctive biographical events shows that the already-known tendency of mentioning private life events about women in Wikipedia biographies (Sun and Peng, 2021; Bamman and Smith, 2014) can be refined when coupled to ethnic origins. Indeed, the extent of the presence of gender biases is more salient when comparing the biographical entries within the same broad "ethnic" group, while is becomes obfuscated across groups, making other bias (i.e., racial) more prominent.

5 Conclusion and Future Work

In this paper we presented a novel set of computational resources for deepening the analysis of biographical events and improving their automatic detection. We found that existing annotated corpora may be successfully reused to train models that obtain good performances. The model for entity detection, trained on OntoNotes, obtained a F-score of 0.808, while the model for event detection, trained on TimeBank and Wikibio, scored 0.859. We have applied these newly developed resources to perform an analysis of biases in Transnational women writers on Wikipedia adopting intersectionality as a framework to interpret our results. In particular, we have identified that the representation of women and non-Western people on Wikipedia is problematic and biased. Using different axes of analysis - as suggested by intersectionality - it becomes easier to better identify these biases. For instance, gender bias against Transnational women are more marked when comparing their biographies against those of Transnational men rather than Western ones. On the other hand, potential racial biases emerge when comparing Transnational women to Western women. Using an intersectional framework would benefit the understanding and countering of biases of women and non-Western people on Wikipedia.

Future work will improve the model for biographical event detection, and to extend the analysis on a wider set of biographical entries from different sources.

Transnational	**Western**
poet, education, schooling, award, degree, completed, awarded, activist, obtained, professor, started, translated, conferred, migrated, exile, recipient, born, novelist, writer, lyricist	wrote, appeared, sold, illustrated, described, married, starred, met, told,illustrator, enlisted

Table 7: Comparison of biographical events between Transnational and Western writers.

Transnational Women	**Transnational Men**
defeated, daughter, actress, married, lost, appeared, marriage, deafeating, won, began, activist, loosing, divorced, raised, attended, win, featured, seeded, mother, grew	actor, son, chairman, lyricist, served, politician, critic, father, joined, death, accused, known, poet, scholar, elected, imprisoned, president, established, exile

Transnational Women	**Western Men**
activist, degree, won, actress, received, born, daughter, award, education, defeated, recipient, defeating, win, selected, mother, writer, schooling, completed, poet, lost	wrote, enlisted, service, actor, claimed, father, assigned, drafted, directed, developed, death

Transnational Women	**Western Women**
defeated, defeating, lost, activist, education, loosing, schooling, degree, poet, completed, win, seeded, injury, award, match, reach, migrated, participated, professor, loss	wrote, appeared, married, author, published, starred, death, lives, moved, died, sold, illustrator, illustrated, nominated, reviewer, write, lived, developed, spent

Table 8: Comparison of biographical events about Transnational Women *vs.* Transnational Men, Western Men, and Western Women. The tokens in the Table cells were obtained by maximizing the JSD divergence. We used the Shifterator software library (see Appendix A for details).

Limitations and Ethical Issues

This work presents some limitations that will be addressed in future work. In particular, *i)* even if the model for biographical event detection obtained good results, more sophisticated approaches may be devised to increase its effectiveness (e.g., best performing LMs, multi-task settings); *ii)* the intersectional analysis was performed on a specific sample of people, and thus limited to writers. Taking into account people with other occupations may lead to different results; finally, *iii)* only Wikipedia biographies were considered: biographies from other sources may differ in style and thus pose novel challenges to the biographical event detection task.

The research involved the collection of documents from Wikipedia, which are released under the Creative Commons Attribution-ShareAlike 3.0 license. The annotation of the experiment was not crowdsourced. All the three annotators are member of the research team who carried out the research as well as authors of the present paper. They are all affiliated with the University of Turin with whom they have a contract regulated by the Italian laws. Their annotation activity is part of their effort related to the development of the present work, which was economically recognized within their contracts with the University of Turin. A data statement for the research can be accessed at the following url: https://github.com/marcostranisci/WikiBio/blob/master/README.md

References

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable annotation of events and event relations across domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.

Claire Bonial and Martha Palmer. 2016. Comprehensive and consistent PropBank light verb annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3980–3985, Portorož, Slovenia. European Language Resources Association (ELRA).

Babs Boter, Marleen Rensen, and Giles Scott-Smith. 2020. *Unhinging the National Framework: Perspectives on Transnational Life Writing*. Sidestone Press.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.

Kimberlé W Crenshaw. 2017. *On intersectionality: Essential writings*. The New Press.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational linguistics*, 26(4):629–637.

Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2017. *BiographyNet: Extracting Relations Between People and Events*, pages 193–227. New Academic Press. Online published in: Computing Research Repository / ArXiv [v2 Wed, 26 Dec 2018].

Ryan J Gallagher, Morgan R Frank, Lewis Mitchell, Aaron J Schwartz, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. 2021. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: a multilingual event-centric temporal knowledge graph. In *European Semantic Web Conference*, pages 272–287. Springer.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Hans-Ulrich Krieger. 2014. A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in RDF and OWL. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Stefano Menini, Rachele Sprugnoli, Giovanni Moretti, Enrico Bignotti, Sara Tonelli, and Bruno Lepri. 2017. Ramble on: Tracing movements of popular historical figures. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–80.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke Van Erp, Anneleen Schoen, and Chantal Van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422.

Anne-Lyse Myriam Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *9th international workshop on semantic evaluation (SemEval 2015)*, pages 778–786.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Alistair Plum, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov. 2019. Large-scale data harvesting for biographical data. In *Proceedings of the Third Conference on Biographical Data in a Digital World 2019, Varna, Bulgaria, September 5-6, 2019*, volume 3152 of *CEUR Workshop Proceedings*, pages 66–72. CEUR-WS.org.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The TimeBank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37:132–151.

Irene Russo, Tommaso Caselli, and Monica Monachini. 2015. Extracting and visualising biographical events from wikipedia. In *BD*, pages 111–115.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Marco Antonio Stranisci, Giuseppe Spillo, Cataldo Musto, Viviana Patti, and Rossana Damiano. 2022. The URW-KG: a resource for tackling the underrepresentation of non-western writers. *arXiv preprint arXiv:2212.13104*.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360.

Jouni Antero Tuominen, Eero Antero Hyvönen, and Petri Leskinen. 2018. Bio CRM: A data model for representing biographical data for prosopographical research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*. CEUR Workshop Proceedings.

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Comparison Between Transnational Women and Men through the JS Divergence Metric

In this Section you can observe a comparative analysis of the divergence between events about Transnational women against Transnational men (Figure 1), Western men (Figure 2), and Western women (Figure 3). All divergences were computed and plotted with Shifterator (Gallagher et al., 2021).

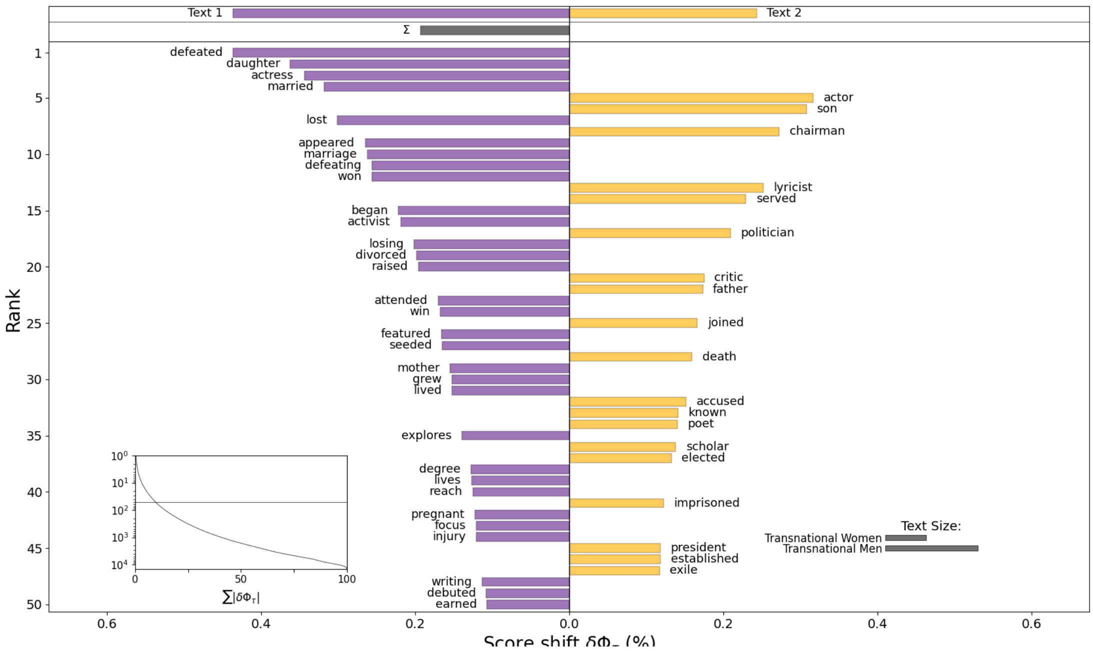

Figure 1: The comparison of events between Transnational Women biographies and Transnational Men biographies.

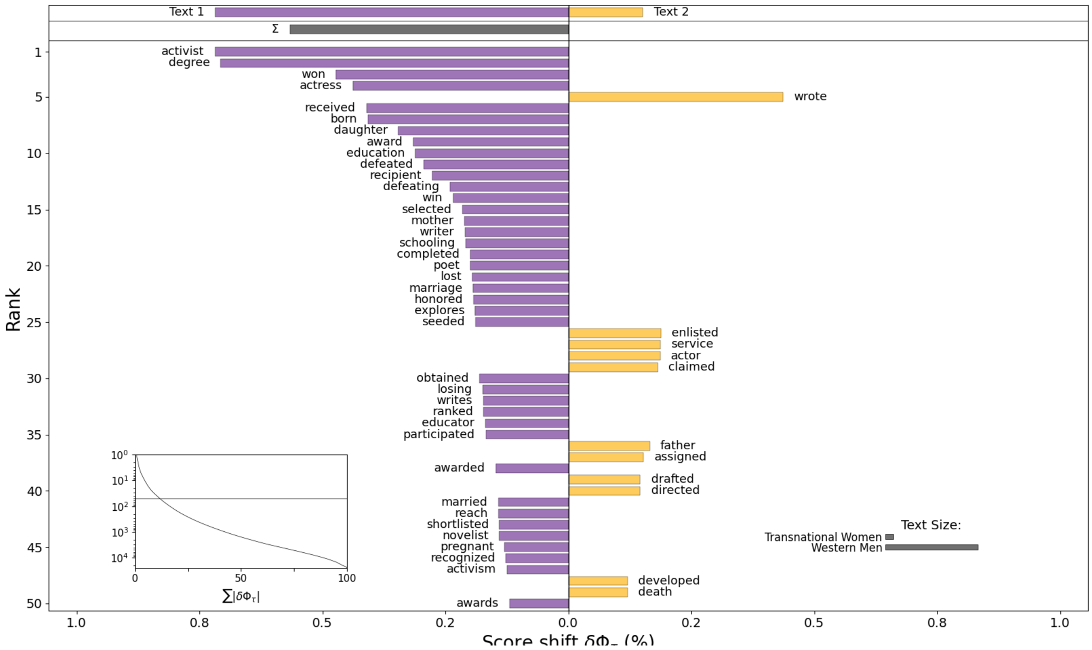

Figure 2: The comparison of events between Transnational Women biographies and Western men biographies.

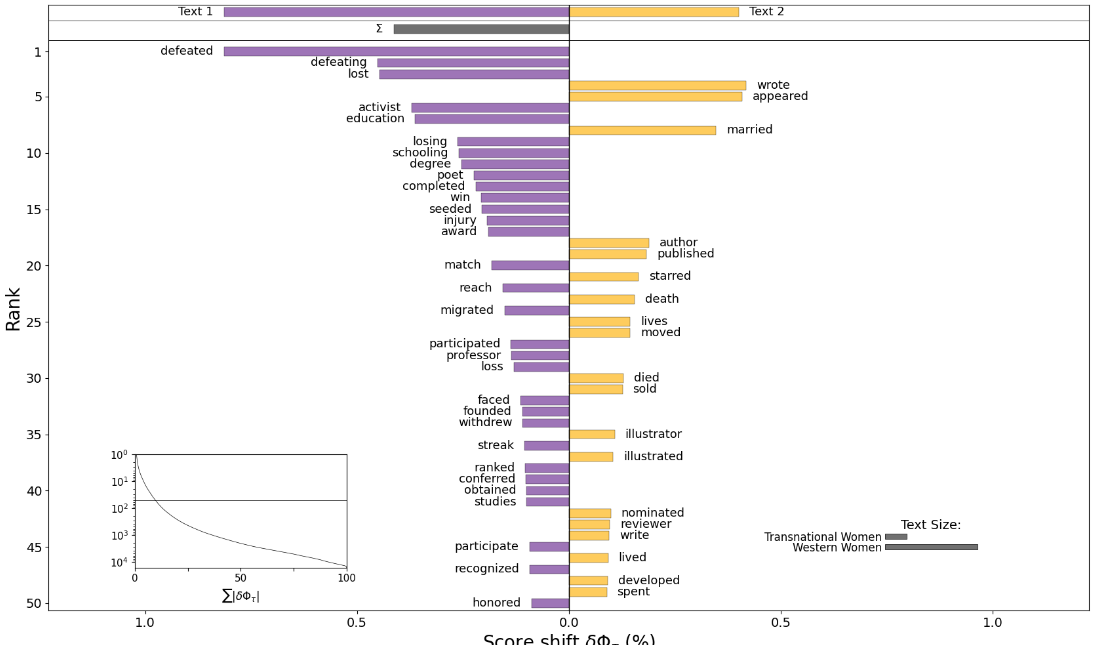

Figure 3: The comparison of events between Transnational Women biographies and Transnational Women.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

□ A2. Did you discuss any potential risks of your work?
Not applicable. For our work we handled public data from Wikipedia

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Not applicable. Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The focus of the experiments was to test the impact of different training set over the same vanilla version of a small LM like DistilBert. So we didn't provide information about that

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
we used the standard parameters of these off-the-shelf tools

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2.1 and Section "Limitations and Ethical Issues"

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
2.1