Measuring Presence of Women and Men as Information Sources in News

Muitze Zulaika and **Xabier Saralegi** and **Iñaki San Vicente**
Orai NLP technologies
Zelai Haundi 3, 20170 Usurbil, Spain
{m.zulaika, x.saralegi,i.sanvicente}@orai.eus}

Abstract

In the news, statements from information sources are often quoted, made by individuals who interact in the news. Detecting those quotes and the gender of their sources is a key task when it comes to media analysis from a gender perspective. It is a challenging task: the structure of the quotes is variable, gender marks are not present in many languages, and quote authors are often omitted due to frequent use of coreferences. This paper proposes a strategy to measure the presence of women and men as information sources in news. We approach the problem of detecting sentences including quotes and the gender of the speaker as a joint task, by means of a supervised multiclass classifier of sentences. We have created the first datasets for Spanish and Basque by manually annotating quotes and the gender of the associated sources in news items. The results obtained show that BERT based approaches are significantly better than bag-of-words based classical ones, achieving accuracies close to 90%. We also analyse a bilingual learning strategy and generating additional training examples synthetically; both provide improvements up to 3.4% and 5.6%, respectively.

1 Introduction

Text mining in general, and Natural Language Understanding (NLU) in particular, are being successfully used as support tools on various areas of the humanities. In many studies, evidence is encoded in natural language, either in text or audio collections, and NLU techniques help significantly in the task of finding such evidence.

Within the humanities, one of the fields that has gained momentum in recent years is gender studies, which has come to cover a wide range of topics. This paper focuses on gender studies aimed at analysing the presence of women in the narratives constructed by the press. The objective of this kind of research is to quantify the presence of women compared to men in the news according to various indicators. The Global Media Monitoring Project (GMMP) is a long running international project carrying out such studies. It has a consolidated methodology that evaluates a wide number of indicators (Macharia, 2020), such as:

- Sex of presenters, reporters and news subjects & sources in newspaper, television and radio news.
- Subject and source selection by sex and by sex of reporters in print, television and radio stories.
- Function of subjects & sources in newspaper, television and radio news.
- Subjects & sources quoted directly in newspapers.

All of those indicators are analysed manually, which implies a great effort that limits the frequency and the size of the sample on which the studies are carried out. It is therefore appropriate to research whether such indicators can be measured using artificial intelligence. This paper focuses on those that require language comprehension. Specifically, we tackle the indicator dealing with the presence of women and men as information sources in the news, by means of NLU techniques. We approach the task using a binary gender identification schema, due to technical reasons. This decision should not be interpreted as a denial of a more complex reality.

This paper presents an attempt to measure the presence of women as sources of information for news stories, in the line of work started by (Asr et al., 2021). We approach the problem by using a simple strategy whose core is a multiclass supervised classifier. The main task consists on identifying sentences including quotes and detecting the gender of the sources of those quotes. This is no trivial task. The structure of the quotes is variable, in some languages there is no gender marking, and,

Proceedings of the 6th LaTeCH-CLfL Workshop, pages 126–134
October 16, 2022. ©2022 International Conference on Computational Linguistics

moreover coreferences are often used and thus the source of the quote is not explicit.

Example 1
[EU] Sagarduik[1] ohartarazi du «kostatuko» dela kutsatze-tasa 60 puntura jaistea.
[ES] Sagardui advierte de que «va a costar» reducir la tasa de contagio a 60 puntos.
[EN] *Sagardui warns that 'it'll be hard' to reduce the contagion rate to 60 points.*

A manual analysis of a sample of news items (see Section 3.1) showed that it is not necessary to solve all cases of coreference in order to measure the presence of women and men as sources of information. This fact allows us to tackle the task with a relatively simple pipeline. The pipeline consists of two steps; first lexical substitutions involving surnames (See Example 1) are resolved, and afterwards a multiclass classifier detects the sentences in the news that correspond to utterances as well as their corresponding gender. To address the multiclass classification task, different strategies based on fine-tuning neural language models have been compared to Bag-of-Word paradigm based approaches.

The contributions of this work are the following:

- This is the first work addressing the task measuring the presence of women as sources of information for news stories using a supervised approach.
- We provide the first datasets for Basque and Spanish, including a bilingual benchmark for the task[2].
- Study of approaches based on pretrained language models to address the task.
- Study of cross-lingual learning and data-augmentation strategies to cope with the scarcity of training data for this task.

From here on, the paper is organized as follows: the next section reviews the state of the art. In section 3 we present an analysis about the measurement of the presence of women and men as information sources, how datasets were prepared and what criteria we followed in the annotation. Section 4 presents the approaches analysed to tackle the task as well as the results obtained. Finally, some conclusions are drawn.

2 State of the Art

When developing policies to address social challenges, evidence-based diagnostics are necessary, and the digital press is a source of such evidence, or more specifically, the narratives of reality that they offer. This type of analysis based on NLP techniques has already been carried out for several social challenges. (Lee, 2019) applies several text mining techniques such as collocations, word co-occurrences and topic-modeling (LDA) for extracting information about immigrant workers in Korea. (Chouliaraki and Zaborowski, 2017) study the narratives in the press about the refugees during the 2015 refugee crisis across eight European countries. (Lansdall-Welfare et al., 2017) focus on gender equality, using an approach based on counting n-gram frequencies and extracting Named Entities on a large British news corpus. (de Oliveira et al., 2021) present a comprehensive survey of the challenges fake news detection in social media pose, including NLP approaches. In the same field, (Khaldarova and Pantti, 2016) analyse fake news narratives generated about the Ukrainian conflict and Twitter users' reactions to those stories.

Regarding gender related research, we can find numerous works that make use of NLP. (Hu et al., 2021) apply sentiment analysis techniques to identify sexist attitudes in Chinese social media. (Kozlowski et al., 2020) study gender stereotypes in male-oriented magazines and female-oriented magazines. They use Topic Modelling techniques to analyse the topics associated with each gender and their evolution over time. (Nagaraj and Kejriwal, 2022) present a large scale analysis of the gender disparity in literature published in the pre-modern period by means of NERC and NED techniques. (Underwood et al., 2018) use similar techniques, namely the BookNLP pipeline (Bamman et al., 2014), to research the evolution of the meaning of gender in works of fiction published between the 18th and 21st century. (Sims and Bamman, 2020) deals with the task of modelling the propagation of information in literature by paying attention to gender dynamics and their relationship to the representation of men and women in novels. To do so, they use, among others, coreference resolution and speaker attribution techniques. Other authors (Garnerin et al., 2019; Lebourdais et al., 2022; Doukhan

[1] *Sagardui* is the surname of the source, the full name is *Gotzone Sagardui*.

[2] Datasets are publicly available under CC-BY license at `https://github.com/orai-nlp/news-src-gender`.

et al., 2018) study the presence of gender on audio content in order to analyse the presence of men and women in audiovisual media.

(Asr et al., 2021) present a system for the analysis of gender bias in the news. This is the work in the literature closest to ours. The system allows measuring indicators of the presence of men and women who are mentioned in the news and as sources of quotes included in the news. The following pipeline is used to measure the presence of men and women as authors of quotes: 1) extraction of quotes, 2) identification of their sources, and 3) prediction of their gender. For quote extraction, a strategy combining a symbolic approach based on syntactic dependencies and regular expressions is used. For source identification, they apply a NERC model to detect person names and a coreference model to link names with quotes. The gender of the person's name is determined by searching a name database.

Quote extraction and speaker attribution are tasks that have been addressed in the literature. Three types of quotes are distinguished in the literature: direct quotes, indirect quotes, and mixed quotes. Early approaches to citation extraction focused on direct quotes and made use of symbolic strategies (Pouliquen et al., 2007; Glass and Bangay, 2007). (Elson and McKeown, 2010) also focus on direct quotes and propose a more complex strategy combining a NERC process, linguistic rules based on syntactic information and a supervised classifier. (Pareti et al., 2013) presents the first large-scale experiments on direct quotes, indirect quotes, and mixed quotes extraction. They deal with the task with a supervised approach based on Conditional Random Field (CRF) models and maximum entropy classifiers.

The work closest to ours is (Asr et al., 2021). As we have already mentioned, they deal with the measurement of women as source of information as well, but unlike ours, it is based on a symbolic approach that also requires a more complex pipeline than the one we propose. On the other hand, there have also been published works (Pareti et al., 2013; Elson and McKeown, 2010; Pouliquen et al., 2007) on quote extraction, which are somehow related to the measurement of this indicator. These approaches, supervised in some cases, also present pipelines with a complexity that exceeds what is necessary to address the task we propose.

3 Presence of women as source information in news

The measurement of the indicator of presence of women and men as sources of information in the news can be approached as an aggregation of the measurement of that indicator on the sentences that make up a news article. Therefore, the NLU tasks to be addressed would be to identify the sentences that correspond to quotes and then classify the gender of the authors of these quotes. We can define the whole challenge as follows:

Given a document $D = \{s_1, ..s_i, ..s_n\}$ detect the sentences $\{s_i\}$ that correspond to quotes and assign the corresponding gender to the source of the quote $gen(s_i) = \{m, f\}$.

Unfortunately, the presence of coreferences, as well as the lack of explicit gender information in some languages, makes this task very difficult to solve. However, since the ultimate goal is the measurement of a macro indicator, it may not be necessary to resolve all quotes. In order to clarify these two points, we present a study carried out on a manually annotated sample in section 3.1. On the one hand, we analyse the complexity of the task in Spanish and Basque, Basque being a language without gender marks, and on the other hand, whether resolving all quotes is necessary to measure the presence of women and men as sources of information.

Once the analysis was done, we further annotated a large number of examples in order to construct the datasets for training and evaluating the supervised approaches proposed in this work. Details about annotation guidelines and datasets preparation are given in section. 3.2.

3.1 Analysis of the task

In order to analyse the difficulty of the task of measuring the presence of women and men as sources in the news, an annotated sample was created from a collection of news items. In total, the sample contained 400 news in Basque and 400 news in Spanish, randomly selected from a collection of news articles crawled from various digital press websites in the Basque Country[3]. All sentences of each news item were manually analysed, marking those corresponding to quotes as well as the

[3] News were collected between February 2021 and March 2022. Spanish news were crawled from El Diario Vasco, El Correo and Noticias de Alava, and Basque news from Argia and Berria.

	ES		
	F	**M**	**%**
All quotes	401	401	100%
No coreference	97	91	22.69%
Coreference	304	310	77.56%
Ellipsis	154	174	43.39%
Surname lex. sub.	77	63	15.71%
Other lex. sub.	73	73	18.20%
Gender mark	170	164	40.90%
No gender mark	231	237	59.10%

Table 1: Statistics of the Spanish sample.

	EU		
	F	**M**	**%**
All quotes	302	302	100%
No coreference	38	50	12.58%
Coreference	264	252	87.42%
Ellipsis	145	151	48.01%
Surname lex. sub.	91	70	30.13%
Other lex. sub.	28	31	9.27%
Gender mark	42	51	13.91%
No gender mark	260	251	86.09%

Table 2: Statistics of the Basque sample.

source's name and gender in the quote. Coreferences were also solved manually.

To annotate the examples we follow the criteria used in the GMMP methodology guide (GMMP, 2020). We have annotated the gender of each person in the story who is quoted, either directly[4] or indirectly[5]. Only quotes by individual people are annotated, quotes from sources such as groups, organizations or collectives are not considered.

The final samples are composed of an equal number of quotes by women and men, specifically, 401 quotes per gender for Spanish and 302 per gender for Basque. In order to reach this equality, we had to collect news that explicitly contained quotes by women, since the initial random sample yielded an imbalanced number of quotes toward male sources (74.63% and 67.59% for Basque and Spanish, respectively). The manual effort required for this annotation is high: on average, a quote is found in the 9% of the Spanish sentences analysed (up to 14% in Basque). A total of 5274 sentences were annotated in Spanish, and 2667 in Basque. The number of annotated examples is lower in Basque due to time constraints and limited resources.

Tables 1 and 2 present the statistics of the samples. Regarding sentences containing quotes, in most cases the sentence does not contain information regarding the gender of the source (**No gender mark** row), especially in Basque (86.09%). Coreferences are also abundant in both languages. Most of those coreferences are ellipsis (**Ellipsis** row), 43.4% in Spanish and 48% in Basque, a type of coreference that is very difficult to resolve (Soraluze et al., 2017). The rest of the coreferences detected in the quotes correspond to lexical substitutions, including a significant number of substitutions of the full name for the surname (**Surname lex. sub.** row) which is a type of coreference very easy to resolve. The rest of lexical substitutions (**Other lex. sub.** row) correspond to pronoun and job position related substitutions.

Being the objective of this work the measurement of the indicator of presence of women as sources in large-scale news, we have analysed whether to obtain an estimate of this indicator is necessary the resolution of all types of coreference.

As a formula for calculating the presence indicator in a collection of n news items, we have established the following ratios for each gender:

$$FQ = \frac{\sum_{i=1}^{n} \frac{\#F\_quotes_i}{\#F\_quotes_i + \#M\_quotes_i}}{n} \tag{1}$$

$\#F\_quotes_i$ is the number of quotes in news item i whose authors are women, and $\#M\_quotes_i$ is the number of quotes in news item i whose authors are men.

We analysed whether the ratios (at news article level) calculated taking into account only the quotes that include gender information and/or easily treatable surname type coreferences (**Quotes with gender marks** in Table 3) correlate with the ratios that take into account all quotes (**All_Quotes** in Table 3). The Pearson correlation values obtained are 0.940 and 0.938 for Basque and Spanish news of the sample respectively, meaning that to measure the indicator it is sufficient to consider only citations that include gender or surname coreference information. It remains for future work to

[4] A person is quoted directly if their own words are printed in the story - e.g. 'I am disappointed and angry about the continued use of drugs in sport' said the President of the Olympic Committee.

[5] A person is quoted indirectly if their words are paraphrased or summarised in the story - e.g. The President of the Olympic Committee today expressed anger at the incidence of drug use.

check whether this correlation also holds for subsets of news items with different attributes such as time period, subject area or news source. These specifications significantly simplify the pipeline required to estimate the indicator and the classification task to be solved by the multiclass classifier.

	FQes	FQeu
All_Quotes	0.33	0.26
Quotes with gender marks	0.35	0.30

Table 3: Presence of women as source in news sample estimated by taking into account all quotes (**All_Quotes**) and taking into account only quotes including gender information and/or surname type coreferences (**Quotes with gender marks**).

3.2 Dataset

The sample annotated in the previous section is limited and created with the objective of analysing the task. A dataset to train and test supervised classifiers for our use case must fulfill certain requirements: it has to be large enough, maintain a balance between female and male categories, and have no gender bias. In addition, we set to make development and test sets as similar as possible between languages. Thus, datasets presented in section 3.1 were increased. Examples were further annotated in Basque and Spanish, meeting the aforementioned conditions. In order to make the Basque and Spanish datasets equal, examples were translated from one language to the other and added to the respective dataset. Thus, Basque and Spanish datasets will have the same content in all sentences. However, it should be kept in mind that the Basque language is a language without brand gender, so some Spanish female and male sentences, in Basque will be tagged as ’Other’ (see section 4 for details about annotation scheme). Therefore, although the datasets of the two languages are made up of the same sentences, the evaluations are not comparable between languages, since a number of sentences have different labels in each language.

To correct the gender bias, an equivalent example was generated for each example quote but with the opposite gender of the source. Let’s take the example "*Partidaren atarian, **Axier Arteagak** onartu zuen norgehiagoka «zaila» izango zuela.*[6]". The equivalent example is generated by selecting the name of a real person from the same domain (Basque pelota in the example) but with the opposite gender: "*Partidaren atarian, **Ane Mendiburuk** onartu zuen norgehiagoka «zaila» izango zuela.*[7]".

In addition, we added more F, M and Other (see section 4) sentences to increase the training dataset. To do this, we process news sentences with a quote classifier. This classifier detects whether sentences contain quotes or not. Gender detection of sources (F and M labels) was performed manually. The classifier detected quotes in 2,000 randomly selected news for each language. In total, we added 792 female expressions, 792 male ones and 2,922 corresponding to the ’Other’ category in Spanish. The respective numbers for Basque were 666, 666 and 3,770.

The statistics of the final datasets constructed are shown in Tables 4 (Spanish) and 5 (Basque), including all the aforementioned improvements. In this task, positive examples are quotes with gender marks (F and M), and the rest of the sentences (Other) are negative. That is, the ’Other’ category includes quotes without gender marks and sentences without quotes. For the sake of the experiments, we assume that substitution type coreferences can be solved automatically, hence, we’ve added their manual resolutions and classified them as positive.

	F	**M**	**Other**	All
Train	884	884	5,323	7,091
Dev	184	184	1,022	1,390
Test	125	125	1,049	1,299
All	1,193	1,193	7,394	9,780

Table 4: Statistics of Spanish monolingual datasets, number of sentences per class.

	F	**M**	**Other**	All
Train	695	695	3,690	5,080
Dev	184	184	1,022	1,390
Test	89	89	1,121	1,299
All	968	968	5,833	7,769

Table 5: Statistics of Basque monolingual datasets, number of sentences per class.

[6] Before the match, **Axier Arteaga** accepted that it would be a “difficult” competition.

[7] Before the match, **Ane Mendiburu** accepted that it would be a “difficult” competition.

4 Identification of quote and source's gender

We propose to measure the indicator of presence of women as a source in news by dealing with the task at sentence level. Once solved at the sentence level, we can aggregate sentence results to compute the indicator at news article level, and subsequently at the collection level. We have shown in section 3.1 that the task at the sentence level can be simplified and not all types of coreference need to be taken into account. It is enough to consider only those citations that include gender marks and/or surname-type coreferences. The proposed approach to address the task has two steps: (i) lexical substitutions (surnames) are resolved at the news item level, and (ii) sentences from the news item are processed by a multiclass classifier that determines whether the sentence contains a quote and the gender of the source of the quote.

We approach the problem of identifying quotes and the gender of their sources as a single sentence classification task. Each sentence of the news article is classified based on three categories:

- **F**: Quote made by a woman including gender marks.
- **M**: Quote made by a man including gender marks.
- **Other**: Non-quote or quotes without gender marks.

To implement the multiclass classifier, two approaches have been compared: a) bag-of-words representation and SVM (Support Vector Machine) and LR (Logistic Regression) classifiers, and b) dense fine-tuned representation approach using a pretrained BERT neural models.

To implement the first approach we used the vocabulary with minimum absolute document frequency of 4 and maximum relative document frequency of 0.6. We used the TFIDF statistic as the weight in the vector representation.

To implement the neural approach, we adopted the fine-tuning strategy proposed by (Devlin et al., 2019), using various BERT models and fine-tuning them over the datasets presented in sections 3.1 and 3.2. We have analysed the following BERT models:

- **BERTeus** (Agerri et al., 2020) is a BERT-base-cased language model for Basque pretrained on a corpus containing 224.6M words, including news articles from online newspapers and the Basque Wikipedia.
- **IXAmBERT** (Otegi et al., 2020) is a multilingual language model pretrained with English, Spanish and Basque texts. The model was trained on a corpus composed of Wikipedia dumps of the three languages, and Basque news articles from online newspapers.
- **BETO** (Cañete et al., 2020) is a Spanish language model pretrained on a 3B token corpus from various sources. It is similar to BERT-base-cased, although its vocabulary contains 31k BPE subword tokens and the model was trained for 2M steps.

All the fine-tuning experiments were carried out using an Nvidia Titan RTX3090 GPU card. Initial learning rate was set to 3e-5 and the best model was chosen over the results obtained in the development set, after fine-tuning up to ten 10 epochs. We report the best result out of 5 random initializations. The Transformers library (Wolf et al., 2020) was used.

	Precision		**Recall**		**F-score**		
	F	M	F	M	F	M	AVG
LR	0.33	0.27	0.44	0.35	0.38	0.31	0.35
SVM	0.35	0.30	0.35	0.34	0.35	0.32	0.34
BETO	0.86	0.83	0.85	0.90	0.85	0.87	0.86

Table 6: Monolingual results for Spanish quote and gender detection.

	Precision		**Recall**		**F-score**		
	F	M	F	M	F	M	AVG
LR	0.21	0.21	0.46	0.38	0.29	0.27	0.28
SVM	0.28	0.23	0.43	0.35	0.34	0.27	0.30
BERTeus	0.87	0.84	0.92	0.89	0.90	0.86	0.88

Table 7: Monolingual results for Basque quote and gender detection.

Tables 6 and 7 present the results of the monolingual experiments. For each system, we report precision, recall and F-score results over F and M categories. We leave the category "other" out, since F and M are the relevant ones for measuring the indicator of the gender presence as information source. As a general metric of the systems' performance, we report the average of the F and M categories' F-score values (see the last column of the tables 6 and 7). Analysing the results, we arrive at two conclusions that hold for both languages: (i) neural language models perform significantly better than bag-of-words based classical algorithms, up to 51 points F-score for Spanish and 58 points for Basque; and (ii) using neural language models, the classifier detects with a high F-score the quotes of the news and the gender of its sources.

Regarding the classical algorithms, results obtained with the two algorithms are very similar, both for LR and SVM the recall is higher than the precision, achieving a F-score of 0.35 and 0.30 for Spanish and Basque, respectively. On the other hand, neural language models classify the gender of sources with a high F-score average value, 0.86 and 0.88 for Spanish and Basque, respectively.

As for gender, it is observed that female and male sources are not detected with the same F-score, however, the difference is small and the classifiers perform similarly for both genders.

4.1 Multilingual training

One of the strategies proposed in the literature to cope with the shortage of training examples is to combine the examples available for different languages and use a multilingual model as a base pretrained model. The logic behind this approach is that multilingual models are able to generalize across languages, and thus they will benefit from training examples in different languages. We performed experiments combining the Basque and Spanish training datasets and optimizing the number of epochs with the development dataset of the evaluation language. We constructed a combined dataset maintaining language balance, which includes 5,080 sentences per language[8]. The full bilingual training dataset consists of 1,390 female quotes, 1,390 male quotes and 7,380 other quotes.

	Multilingual (IXAmBERT)						
	Precision		**Recall**		**F-score**		
	F	M	F	M	F	M	AVG
ES	0.90	0.85	0.90	0.89	0.90	0.87	0.89
EU	0.88	0.89	0.99	0.88	0.93	0.88	0.91

Table 8: Multilingual training results for quote and gender detection.

The results of this experiment are shown in Table 8. With the multilingual model we have managed to improve monolingual results, we have achieved a 3 point improvement in both languages.

4.2 Synthetic examples

Error analysis of both monolingual and multilingual experiments surfaced a few cases where the classifier predicts the opposite gender, although the source name is written directly in the sentence. Our hypothesis is that this error may be related to the number of names of sources that the model has seen in training, because the training examples contain only a limited number of names. This implies that the model may not know the gender of the nouns present in the test, because they are missing in the train dataset. In order to tackle the problem of Out-Of-Vocabulary (OOV) names, we include synthetically generated examples in the training. Specifically, we generate new examples from examples that exist in training, replacing name occurrences with other names included in a list.

The aim of this experiment is to test whether adding examples including OOV names to the training set directly influences the detection of the gender of information sources. Hence, we performed the experiment under ideal settings.

The name list includes names that appear in the test but not in the training data. Our error analysis shows that sentences with source's names that appear once or not at all in the training dataset, are correctly classified with a 17.91% accuracy, while names that appear two or more times achieve an0 89%. Therefore, taking into account these statistics, we've created two synthetic examples for each OOV name. For example, using the OOV name Denisa and random training quotes we have generated two examples:

Example 2
[EU$_1$] «Ahal den bezain azkarren eutsiko diogu berriro horri», adierazi du Denisa Urtiagak.
[Translation$_1$] *"We will get back to it as soon as possible," says Denisa Urtiaga.*

[EU$_2$] Joera aldaketa horren atzean kontzientziazio lan handia dagoela uste du Denisa Molinak.
[Translation$_2$] *Denisa Molina believes that there is a great awareness-raising work behind this change of trend.*

Tables 9 and 10 present the results of the synthetic examples experiment. If we compare this results with the previous experiment, we observe that both for monolingual and multilingual models,

[8] Basque language training dataset is used as reference, since it is the smaller one. Spanish examples are selected randomly.

	Synthetic Monolingual						
	Precision		**Recall**		**F-score**		
	F	M	F	M	F	M	AVG
BETO	0.95	0.92	0.83	0.88	0.89	0.90	0.89
BERTeus	0.90	0.90	0.90	0.90	0.90	0.90	0.90

Table 9: Synthetic training results, monolingual model.

	Synthetic Multilingual (IXAmBERT)						
	Precision		**Recall**		**F-score**		
	F	M	F	M	F	M	AVG
ES	0.97	0.94	0.91	0.94	0.94	0.94	0.94
EU	0.87	1.00	1.00	0.85	0.93	0.92	0.92

Table 10: Synthetic training results, multilingual model.

the use of synthetic examples has a beneficial effect on gender detection.

Both monolingual and multilingual models benefit from using synthetic examples. For Spanish, the monolingual model performs three points higher (first row in Table 9) and the multilingual model performs five points higher (first row in Table 10). Regarding Basque, the same behavior is observed, the use of synthetic examples brings up the performance of the monolingual model two points (2nd row in Table 9) and one point with the performance of the multilingual model (2nd row in Table 10).

5 Conclusion

This work addresses the task of automatically measuring the presence of women and men as sources of information in the news. This is a standard indicator in media monitoring processes for gender balance.

We have shown that the large-scale measurement of this indicator can be automated using NLU techniques. To the best of our knowledge, this is the first work proposing a supervised approach to tackle this problem. The experimentation has been validated on two languages with different characteristics, Spanish and Basque.

According to the analysis of our datasets, in order to estimate the presence indicator at the collection level it is not necessary to solve all the cases of coreference associated with the quotes, which simplifies the pipeline required for the measurement of the indicator.

Experiments show that the tasks of citation detection and author gender classification can be tackled jointly by means of a supervised multiclass classifier based on neural language models. Fine-tuning a pretrained neural model provides significantly better results than supervised approaches based on bag-of-words paradigm.

The supervised approach based on neural language models can achieve better results if they are trained with examples from both languages and a multilingual pretrained model is used. Further improvement can also be achieved by adding synthetic examples to the training set, generated for person names not included in the training.

6 Acknowledgements

This work has been supported by Emakunde, the Basque Institute for Women. Muitze Zulaika is also funded by the Investigo programme included in the framework of the European Union's NextGenerationEU plan.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*, 16(1):e0245533.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Lilie Chouliaraki and Rafal Zaborowski. 2017. Voice and community in the 2015 refugee crisis: A content analysis of news coverage in eight european countries. *International Communication Gazette*, 79(6-7):613–635.

Nicollas R de Oliveira, Pedro S Pisa, Martin Andreoni Lopez, Dianne Scherly V de Medeiros, and Diogo MF Mattos. 2021. Identifying fake news on social networks based on natural language processing: trends and challenges. *Information*, 12(1):38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. 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 4171–4186.

David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5214–5218. IEEE.

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-fourth AAAI conference on artificial intelligence*.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender representation in french broadcast corpora and its impact on asr performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, pages 3–9.

Kevin Glass and Shaun Bangay. 2007. A naive salience-based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA'07)*, pages 1–6.

GMMP. 2020. *Global Media Monitoring Project (GMMP) Methodology Guide*, pages 8–32.

Bing Hu, Fang-Ling Luo, Zeng-Wen Peng, and Shi-Qi Lin. 2021. Sexism and male self-cognitive crisis: Sentiment and discourse analysis of an internet event. *Journal of Broadcasting & Electronic Media*, 65(5):679–698.

Irina Khaldarova and Mervi Pantti. 2016. Fake news: The narrative battle over the ukrainian conflict. *Journalism practice*, 10(7):891–901.

Diego Kozlowski, Gabriela Lozano, Carla M Felcher, Fernando Gonzalez, and Edgar Altszyler. 2020. Gender bias in magazines oriented to men and women: a computational approach. *arXiv preprint arXiv:2011.12096*.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465.

Martin Lebourdais, Marie Tahon, Antoine Laurent, Sylvain Meignier, and Anthony Larcher. 2022. Overlaps and gender analysis in the context of broadcast media. In *LREC 2022*.

Changsoo Lee. 2019. How are 'immigrant workers' represented in korean news reporting?—a text mining approach to critical discourse analysis. *Digital Scholarship in the Humanities*, 34(1):82–99.

Sarah Macharia. 2020. *Global Media Monitoring Project (GMMP) 2020-2021 final report*, pages 1–6.

Akarsh Nagaraj and Mayank Kejriwal. 2022. Robust quantification of gender disparity in pre-modern english literature using natural language processing. *arXiv e-prints*, pages arXiv–2204.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz de Ilarraza. 2017. Enriching basque coreference resolution system using semantic knowledge sources. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 8–16.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 3(2):11035.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.