Language Matters: Target-Language Supervision for Political Bias Detection in Turkish News

Umut Ozbagriacik[1] **and Haim Dubossarsky**[1,2,3]
[1] Queen Mary University of London
[2] Language Technology Lab, University of Cambridge
[3] The Alan Turing Institute
umutzbk11@gmail.com, h.dubossarsky@qmul.ac.uk

Abstract

We present, to our knowledge, the first systematic transformer-based outlet-ideology classification study for Turkish news. Using a topic-balanced corpus of Turkish political articles drawn from six outlets commonly perceived as left-, centre-, or right-leaning, we formulate a three-way outlet-ideology classification task. On this dataset, we evaluate a monolingual encoder (BERTurk), two multilingual encoders (mBERT, XLM-R), and a LoRA-adapted decoder model (Mistral). BERTurk achieves the best performance among individual models (70% accuracy, 71% macro-F1), reaching levels comparable to English-language studies despite operating in a lower-resource setting. Error analyses show that all encoders reliably distinguish centrist from partisan articles, but frequently confuse left- and right-leaning articles with each other. Moreover, BERTurk is relatively stronger on right-leaning content, whereas the multilingual models favour left-leaning content, suggesting an "ideological fingerprint" of their pre-training data. Crucially, models fine-tuned on an English political-bias task fail to transfer to Turkish, collapsing to near-chance performance. Taken together, these results demonstrate that effective political bias detection requires target-language supervision and cannot be achieved through naïve cross-lingual transfer. Our work establishes a first baseline for Turkish political bias detection and underscores the need for open, carefully designed Turkish (and broader Turkic) bias benchmarks to support robust and fair media analysis.

1 Introduction

The bias of the media, or the presentation of events in a political context, influences the public opinion. It influences how people perceive certain events, what they believe, and, by doing so, affects democratic outcomes. Researchers have therefore turned to automated methods that can flag partisan slant and imbalance, hoping to strengthen media literacy. However, most existing studies concentrate on English-language news, utilizing rich annotated corpora and mature NLP models, which leaves behind many low-resource languages.

Despite having more than 80 million speakers and a deeply polarised press, Turkish remains under-resourced for many tasks, including political bias detection: sizeable, task-specific annotated datasets are scarce, and off-the-shelf models have rarely been evaluated on media bias. While Turkish resources do exist for some domains (e.g. NLI datasets by Budur et al. (2020)), large-scale labelled corpora for political bias are still missing. Turkish is an agglutinative, morphologically dense language, which is quite different from most rich resource languages. Therefore, models honed on English do not transfer neatly, contributing to poor NLP support for Turkish in general.

The nature of Turkish politics is also quite unique as its concepts of left and right are not necessarily aligned with those of the West. Instead, the divide is often between secular-liberal and religious-conservative perspectives, influenced by the long-lasting centre-periphery tension (Ergil, 2010). Recent studies also report that the media environment in Türkiye has been changing considerably, which only strengthens its polarized character (O'Donohue et al., 2020). Further, the tell-tale cues are often subtle — nuances of word choice and framing rather than loud partisan slogans — while labelled datasets are practically non-existent. With that said, these labels are dynamic and may change over time as party alignments and media coalitions change (Bajec, 2023). As a result, spotting bias in Turkish news is far from straightforward.

Against this backdrop, our work offers a first systematic examination of political bias detection in Turkish news. Our primary aim is to assess how well current models can identify ideological leanings in this setting, thereby gauging the severity

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 72–81
March 29, 2026 ©2026 Association for Computational Linguistics

of the problem rather than merely assuming that existing tools generalise from English. To this end, we curate a topic-balanced corpus of Turkish political news, which enables us both to evaluate off-the-shelf models and to fine-tune them on Turkish data. We further investigate whether transfer learning from English, which is the default strategy in many low-resource scenarios, is effective for Turkish political bias detection. Although copyright constraints prevent us from releasing the underlying articles, our findings call for the urgent development and sharing of Turkish political bias datasets, as direct transfer from English proves insufficient for capturing the nuances of Turkish media.

The paper is structured as follows: Section 2 reviews related work on media bias detection and transformer-based research and implementations. Section 3 describes our dataset creation, system design, and modelling methodology. Section 4 presents experimental results and comparison of each model and ensemble with confusion matrices for visualization. Section 5 discusses the results and their implications. Section 6 provides a broader discussion, and Section 7 concludes with future work for improving and expanding bias detection in low-resource languages.

2 Related Work

2.1 Media Bias Detection

Scholars have probed media bias for decades. Early social-science work took a manual route: researchers counted instances of partisan language and imbalanced story choice to see how politics coloured reporting. A well-known example comes from Budak et al. (2016), who showed that major U.S. outlets sound more alike than expected; aside from high-profile scandals, their coverage tends to sit near the centre. When computational tools came into the picture, they started with basic tools like sentiment lexicon and surface features.

Recasens et al. (2013) mined Wikipedia edits that broke the site's neutrality rule to learn lexical cues for slanted prose. Later studies zoomed in on framing – how phrases like “taxpayer money” versus “public funds” shape an argument – and on ideological sentiment. In the context of Turkish, prior work has been scarce. An exception is Yigit-Sert et al. (2016), who clustered news stories and reader comments on polarising domestic issues; the joint view exposed latent bias but did not label articles directly. The takeaway is that bias often hides in fine detail. Across empirical studies, reliable detection depends on nuanced linguistic cues and context such as framing, lexical choice, subjectivity markers, rather than overt keywords (Recasens et al., 2013; Hamborg et al., 2019; Fan et al., 2019), which is especially difficult for Turkish given the scarcity of annotated resources. Other studies leveraged attention mechanisms to highlight bias cues in headlines; for example, Gangula et al. (2019) used headline attention to detect political bias in news articles.

2.2 Transformer-Based Approaches

Earlier bias-detection systems relied on lexicons and hand-crafted features with linear models, or on CNN/RNN encoders; these capture local patterns but struggle with long-range discourse and domain-specific phrasing. Pre-trained transformers replace manual features with contextual representations learned from large corpora and have become the state-of-the-art for framing and ideology classification. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models can detect bias signals, including the use of extreme adjectives, unilateral phrasing, or basically mentioning some political actors more than others. It was demonstrated by Horne et al. (2018), who published a dataset of news articles and tags on their political bias and demonstrated that, based on a text, it is possible to make a fairly accurate estimate of the tendencies of an outlet. Moreover, Chen et al. (2020) located biased spans and improved article-level predictions on AllSides, a widely used benchmark in which outlets are labelled Left/Centre/Right by editorial review and community input, enabling cross-outlet comparability and standardized evaluation (AllSides, n.d.). Smaller units were studied later on. Spinde et al. (2021) created MBIC, which is a collection of sentences that are tagged with bias type. Spinde et al. (2020) also investigated media bias in German news articles with a combined approach. Large language models follow that trend. In a zero-shot experiment, Menzner and Leidner (2024b) discovered that GPT-3.5, GPT-4, and LLaMA-2 detect some form of bias but improve when ruminated on task-specific data. Considering the fact that these models can be used to detect biases in real time, their BiasScanner tool demonstrates the same (Menzner and Leidner, 2024a). On the whole, large language models can be potentially successful, but they also require every domain and language to be tuned adequately.

3 Approach

We curated a corpus of articles from six Turkish newspapers and used the newspapers' political stance – left, centre or right – as labels. Outlet ideology labels are based on public reputation, and are used here as a weak supervision proxy rather than as a claim of article-level bias. While source-based labeling can introduce noise since not every article from a left-leaning paper (for example) will be overtly left-biased, we felt it provides a reasonable proxy in the absence of manual annotations.

Our study builds on the success of transformer models to detect political bias in English, and extends it to Turkish, using both multilingual and monolingual Turkish models. Given that multilingual models (e.g. XLM-R, mBERT) often underperform language-specific models on morphologically rich languages like Turkish, BERTurk's Turkish-specific pre-training gives it an advantage. We also experiment with Mistral-7B, an LLM recently tuned for Turkish instructions, to see if a generative model with broad knowledge can complement or outperform encoder-based models on this classification task. We evaluate the models on our curated political bias dataset before and after fine-tuning, comparing their performance and providing error-analysis. This tells us if detecting Turkish political bias is possible with existing transformers models, and which model performs the best. Critically, we further test if the same models, when trained on English political bias, can sufficiently transfer to Turkish, comparing their performance to the Turkish-only training. To the best of our knowledge, no prior work has tackled the problem of outlet-ideology classification in Turkish news.

4 Methodology

4.1 Dataset Collection

We collected our dataset of Turkish news articles with their political bias (left, centre, right) using a custom web scraping pipeline. To ensure a fair representation of biases, we selected six Turkish news sources: two known left-leaning, two centrist/mainstream, and two right-leaning based on public reputation.

- **BirGün** – Founded in 2004, BirGün is widely regarded as a left-leaning, secular, socialist daily newspaper. Its editorial line emphasizes labour rights, civil liberties and critical coverage of government policies.
- **Sözcü** – Established in 2007, Sözcü is commonly associated with a left-leaning and strongly secular stance. It is known for its investigative reporting and a critical approach toward ruling political actors.
- **Habertürk** – Founded in 2009, Habertürk represents a mainstream centrist position within the Turkish media landscape. Its reporting aims to maintain a relatively balanced tone across political and economic topics, following a professional news style.
- **Euronews Türkçe** – Launched in 2010 as the Turkish-language branch of Euronews, this outlet follows a centrist and internationalist editorial line. Its reporting prioritises neutrality and factual accuracy, in line with international journalistic standards.
- **Milliyet** – Founded in 1950, Milliyet is one of Türkiye's oldest mainstream newspapers and is associated with a centre-right orientation. It combines national perspectives with relatively moderate conservative framing.
- **Diriliş Postası** – Established in 2014, Diriliş Postası is a right-leaning outlet aligned with conservative and pro-government narratives. Its coverage frequently reflects religious-conservative and nationalist perspectives.

Rather than scraping arbitrary articles, we employed a keyword-balanced multi-source crawler. We defined a set of topical keywords (e.g. "Rusya" (Russia), "Suriye" (Syria), "Trump", "Erdoğan", "protesto" (protest)) covering both international and national political matters. For each source, the crawler queried its search function for each keyword and collected up to a fixed number of articles (12) per keyword. This meant that for any given news topic (e.g. the Russia-Ukraine conflict), we gathered articles from outlets across the political spectrum. This mitigates topic imbalance; that is, a left-wing outlet is unlikely to cover entirely different stories than a right-wing outlet in our data. The full keyword list used during data collection is provided in Appendix A.

For scraping, we used Python's requests library and Selenium (for dynamic sites), which parsed article content and metadata (title, date, source, keyword) from the HTML structure of each site. We ran the crawler over all the keywords (see Appendix A) and sources, resulting in an initial pool of approximately 3200 articles. Scraped articles were manually spot-checked to verify correctly extracted article content. We then applied lightweight rule-based filtering to automatically remove duplicate articles, very short texts, and pages where the main content could not be reliably parsed (e.g., malformed HTML or non-article pages). After cleaning and filtering (removing duplicates and shorter than 30 words articles), the final dataset amounted to approximately 2900 articles, with a balanced class distribution (the left-leaning and right-leaning classes each accounting for about one-third of the data, and centre slightly less). We split this dataset into training and test sets. Specifically, we held out around 20% (583 articles) as a test set, stratified so that each class is represented proportionally (left: 223, centre: 163, right: 197 in test). The remaining articles were used for model training and validation (see Table 1 for details). Text preprocessing was minimal and did not remove stop words or use stemming or lemmatization to preserve the nuanced language cues of bias, as modern transformers handle inflected forms quite well.

For English political bias we used the Kaggle News Dataset on News Bias Analysis (Articoder, 2020). The final English corpus, after transforming the data into a long format and performing some basic filtering, is 24,505 articles with a fairly even distribution of classes left (8,430), center (7,700) and right (8,375).

4.2 Model Fine-Tuning

We fine-tuned four transformer models on a three-class classification problem. For each model, we used the Hugging Face Transformers library with PyTorch. The same experimental choices were made for training the models on the English dataset. The models are:

- **BERTurk (base)** – a BERT-base Turkish uncased model pre-trained on a large Turkish corpus (35GB of text) by Schweter (2020). We used the `dbmdz/bert-base-turkish-uncased` weights as the starting point. This model has 110M parameters and an architecture identical to BERT-base. A linear classifier layer was added on top for our 3-way classification.

- **XLM-RoBERTa-base-Turkish-ner (base)** A fine-tuned multilingual RoBERTa model trained on 100+ languages, including Turkish (Conneau et al., 2020). XLM-R has strong multilingual performance. We used the `akdeniz27/xlm-roberta-base-turkish-ner` (270M parameters) on our data which is fine-tuned on a large Turkish NER dataset.

- **mBERT (base multilingual BERT)** The original multilingual BERT model (Devlin et al., 2019) with approximately 110M parameters covering 104 languages. We fine-tuned the cased version (`bert-base-multilingual-cased`). This model provides a point of comparison to BERTurk (monolingual) to see the benefit of a Turkish-specific pre-training.

- **Mistral-7B (Turkish Instruct)** The 7-billion-parameter decoder model from Mistral AI released in 2023 (Jiang et al., 2023). We worked with the community checkpoint `malhajar/Mistral-7B-Instruct-v0.2 -turkish`. Because full fine-tuning on our mid-sized corpus was unrealistic, we adopted PEFT with LoRA: only rank-32, $\alpha = 64$ adapter matrices on the projection layers were trained, while the base weights stayed frozen in bfloat16 (bf16) on a single NVIDIA A100 (Hu et al., 2021). The LoRA-augmented backbone, wrapped in `AutoModelForSequenceClassification`, outputs one of three bias labels per article. This setup treats the decoder model as a standard sequence classifier with a classification head, rather than using text generation or prompting. Training used around eight epochs in each split of a 5-fold cross-validation, ran more slowly and used more memory than the BERT counterparts, yet stayed within our resources and avoided updating all 7 billion parameters.

A stratified 5-fold cross-validation was used for fine-tuning for all models. Hyperparameters were tuned on the first fold's performance. We found that a learning rate around 1e-5 with gradient accumulation if needed, a batch size of 16 and 4–5 epochs was sufficient for the smaller BERT models. For Mistral-LoRA, a higher learning rate (3e-4) and a batch size of 8 was used and we trained for an average of 7.8 epochs. Early stopping is used with respect to validation loss. The fine-tuning objective was cross-entropy loss on the three classes.

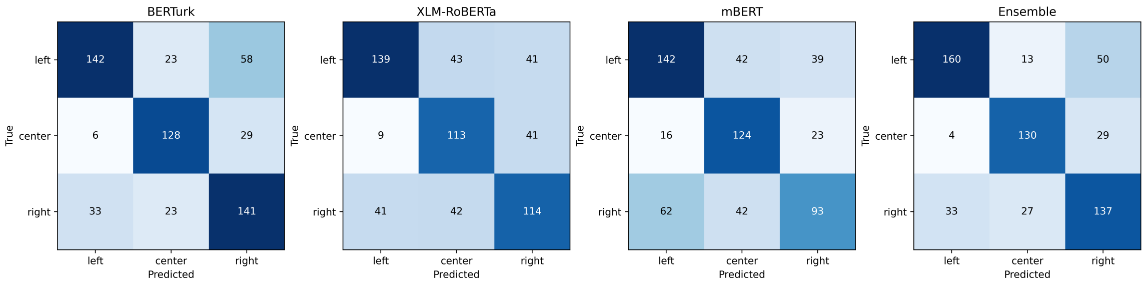

Figure 1: Confusion matrices for the three best-performing encoder-based models (BERTurk, XLM-RoBERTa, and mBERT) and the final ensemble evaluated on the Turkish test set.

4.3 Ensemble Strategy

We utilized the multiple models by using an ensemble approach at two levels. First and foremost, for each model we effectively created a group of five models by training them on different cross-validation folds. After training, we saved each fold's model and used all five to predict the bias logits of the test articles, which were then averaged. This fold-level logit averaging is applied at inference time and reduces variance, yielding a single consolidated prediction for that model. Our fold-level ensembling already averages predictions across multiple random initializations induced by different training splits, reducing variance and improving robustness. Second, we used an ensemble across the different model types. We averaged the per-class logits from BERTurk, XLM-R, mBERT, and Mistral (after the fold-wise aggregation) and selected the class with the maximum mean logit.

4.4 Evaluation

The held-out set of 583 articles was used for evaluation. We reported overall accuracy, precision, recall and macro F1-score (the average of F1 for left, centre, right) as our primary metrics, since macro-F1 is sensitive to performance on the smaller class.[1] Figure 1 shows the confusion matrices of the three most successful encoder-based models and the ensemble model where there are systemic misclassifications between articles that lean left and to the right. We provide per-class breakdowns using confusion matrices to identify biases in the models' preference choices. All experiments are fully reproducible with fixed random seeds for initialization and cross-validation fold selection. The same test set split for Turkish news articles was used to evaluate the models trained on the English dataset. [2]

5 Results

Table 1 shows BERTurk is the top individual model. At 70% accuracy and macro-F1 of 0.71 and balanced performance across the three classes, it outperforms both XLM-R (63%) and mBERT (62%), and with an even larger margin LoRA-tuned Mistral (65% accuracy and 52% F1).

Model	Acc	Prec	Rec	F1
BERTurk	0.70	0.71	0.72	0.71
XLM-R	0.63	0.63	0.63	0.63
mBERT	0.62	0.61	0.62	0.61
Mistral-LoRA	0.65	0.76	0.65	0.52
Ensemble	**0.73**	**0.74**	**0.74**	**0.74**

Table 1: Models performance trained and tested on the Turkish dataset.

Table 2 shows the top-3 models after fine-tuning on the English dataset, and their performance when they are tested on English and when transferred to Turkish. The results first show that all 3 models were able to learn political bias classification in English, with performance ranging between 50%-56%. However, none of the models was able to generalise and transfer to Turkish, as performance drops to around chance level on Turkish (33%). Interestingly, BERTurk performs slightly worse on English than XLM-R and mBERT, potentially due to its Turkish text training.

Error analyses reveal that across the three top models, centre articles are the easiest to identify, with the highest precision and recall and relatively few confusions with partisan classes from left or

[1] As macro-F1 is the arithmetic mean of the per-class F1 scores, it can differ from the value obtained by taking the harmonic mean of macro-precision and macro-recall.

[2] The code for model fine-tuning and evaluation is available at https://github.com/umutzbk/Language-Matters-Target-Language-Supervision-for-Political-Bias-Detection-in-Turkish-News

	English		Turkish	
Model	Acc	F1	Acc	F1
BERTurk	0.51	0.50	0.33	0.24
XLM-R	0.56	0.56	0.36	0.30
mBERT	0.55	0.54	0.32	0.24

Table 2: Models performance trained on English

right. BERTurk correctly classifies most articles in each class (64% left, 79% centre, 72% right), but its main difficulty lies in distinguishing left from right: it often mislabels left-leaning articles as right (26%), leading to the lowest recall on the left class, while performance on right-leaning (≈72% recall) and centrist articles (≈79% recall) is comparatively stronger.

The other two multilingual models, XLM-R and mBERT, exhibit the opposite pattern: they handle left-leaning articles relatively better but struggle markedly with the right class, where recall drops to 0.58 for XLM-R and 0.47 for mBERT, and right-leaning articles are frequently misclassified as left or centre (≈53%). This indicates the model has learned to distinguish partisan vs. non-partisan tone reliably but sometimes struggles to detect which side of the spectrum a biased article falls on.

6 Discussion

Our findings highlight both the potential and the current limitations of transformer-based political bias detection for Turkish news. At a high level, the final ensemble achieves performance comparable to the level reported for three-way ideological classification in English (around 73% accuracy, similar to the 72% reported by Baly et al. (2020)), despite operating in a substantially lower-resource setting. This demonstrates that, given a carefully curated dataset and modern transfer-learning techniques, political bias classification in Turkish is technically feasible and can reach levels that are useful for downstream analysis. At the same time, the behaviour of multilingual models, the failure of cross-lingual transfer from English, and the detailed error patterns point to the importance of treating Turkish as a first-class target language rather than relying on imported resources and models.

A first and central conclusion is the crucial role of Turkish-specific datasets for political bias. Without in-language supervision, both evaluation and training become effectively impossible. Political bias is not a generic "semantic" relation that we can expect to be captured by language models trained in unrelated contexts; it depends on culturally grounded cues, media ecosystems, and ideological cleavages that are inherently local. Our experiments show that even strong pre-trained encoders cannot be meaningfully compared or improved in the absence of a labelled corpus that reflects Turkish media realities. The dataset developed in this work therefore fills a key gap: it enables us to measure how well models do on a concrete Turkish bias detection task and to fine-tune them for that task, rather than extrapolating from English benchmarks or anecdotal examples. This conclusion is in line with recent calls for 'Democratizing AI' that highlight the importance of quality dataset curation (Dairkee and Dubossarsky, 2024; Goworek et al., 2025) as a key element in providing NLP support across many low-resource languages.

The cross-lingual transfer experiments demonstrate that "borrowing" supervision from English does not lead to a strong performance in Turkish. When we fine-tune models on an English political bias task and test them on English their performance improves substantially, confirming that task-specific supervision is effective in a high-resource setting. However, when we apply these English-fine-tuned models directly to Turkish articles, without additional Turkish supervision, their accuracy collapses to near-chance levels. This pattern holds across all three models we tested (XLM-R, mBERT, and BERTurk). Even if we do take into account differences between training domains used in the English and Turkish datasets, the observed drop is not a modest degradation; it is a near-complete loss of discriminatory power. This supports our intuition that cues of political bias are not reliably shared across languages, even when the underlying ideological families (e.g., "left" vs. "right") appear superficially similar. Instead, political bias is highly language- and culture-specific, encoded in lexical choices, idioms, framing devices, and references that do not straightforwardly map from English to Turkish. In practical terms, this means that even in a world of powerful multilingual transformers, that rely on transfer learning, effective political bias detection still requires supervision in the target language.

Turning to within-Turkish experiments, BERTurk consistently outperforms the two multilingual encoders when all are fine-tuned on the Turkish dataset. Its monolingual pre-training on Turkish text clearly helps it pick up language-specific markers of ideological leaning, yielding the strongest overall F1 and accuracy scores. However, XLM-R and mBERT, despite being top-performing multilingual models trained on vastly more data than BERTurk, significantly lag behind, particularly on the right-leaning class. Therefore, we recommend using BERTurk (or future Turkish-specific encoders) as a default choice for Turkish political bias tasks, because the gains are real and robust.

This within-language advantage of a monolingual model (BERTurk) over top-performing multilingual models in low resource transfer settings goes against the dominant view that transfer from rich to lean resource languages is inherently beneficial. However, similar findings have recently been reported also for Hindi, where MuRIL, a model trained exclusively on Indian languages outperformed XLM-R when trained on Hindi data (Goworek and Dubossarsky, 2025). Importantly, without the availability of a high-quality dataset in the target language (Turkish in our case), even the best monolingual model cannot be meaningfully evaluated or improved.

Our results on the decoder-style Mistral model reinforce this point. Even with parameter-efficient LoRA tuning, the large generative model does not match the classification performance of the encoder-based architectures. This underlines that scale and instruction tuning in Turkish do not automatically translate into superiority on specialized tasks such as bias detection. Targeted fine-tuning on a well-designed classification task remains essential. From the perspective of SIGTURK and Turkic NLP more broadly, this suggests that community effort is better spent on curating domain-specific corpora and reliable labels than on simply adopting ever larger general-purpose LLMs.

The error analyses add another layer of insight into what the models are actually learning. Across all three encoder-based systems, centre-labelled articles are consistently the easiest to identify: they show the highest precision and recall, and they are rarely confused with partisan classes. By contrast, left- and right-leaning articles are frequently misclassified as each other. This may indicate that the models have learned to distinguish between "partisan vs. non-partisan" tone, but are less reliable in determining on which side of the ideological spectrum a biased article falls. Several factors likely contribute. Our labeling scheme assigns each article the label of its outlet, regardless of whether that specific article is sharply opinionated or relatively neutral. When a left-leaning outlet runs a straight news story, the text may look linguistically centrist, so the model's prediction of "centre" is counted as an error. In addition, left- and right-wing sites in Türkiye often cover the same topics with overlapping vocabularies (e.g., "terrorist," "freedom," "economic crisis"), differing more in framing and omitted context than in surface word choice. Without deeper understanding, the model may treat both as generic "critical" language and confuse one side's attacks with the other's. The confusion matrices show exactly this pattern: direct left-right errors are common, whereas partisan-centre errors are less frequent. Another possibility is that models rely on surface cues (e.g., named entities), partially learning "which outlet is this?" rather than ideology. Fully disentangling topic, style, outlet identity, and bias remains an open challenge for future work.

Interestingly, the comparison between BERTurk and the multilingual models led us to speculate of "ideological fingerprints" of model pre-training. When fine-tuned on the Turkish dataset, BERTurk tends to perform relatively better on right-leaning articles and worse on left-leaning ones, whereas XLM-R and mBERT show the reverse pattern: they handle left-leaning Turkish articles comparatively better but struggle more with the right class, often misclassifying right-leaning content as left or centre. One plausible explanation is that multilingual models inherit subtle biases from their predominantly English training data, where media and web text may lean more toward liberal/left perspectives on average. If the underlying representations are more attuned to left-coded patterns of critique and rhetoric, this might make them more sensitive to left-leaning cues, and less calibrated for right-coded ones, when transferred to Turkish and fine-tuned with limited supervision. By contrast, BERTurk's monolingual pre-training on Turkish sources includes right-leaning and pro-government outlets which may make it more attuned to ideological markers that are specific to the Turkish right. In this sense, our analysis suggests that model choice does matter for *which* ideological voices are recognised and which are systematically under-detected. This is not primarily about a few percentage points of F1, but about representational balance and introducing political bias from the backdoor: a model that over-detects left bias and under-detects right bias (or vice versa) can potentially skew subsequent analyses of media ecosystems.

7 Conclusions

In this paper, we presented, to our knowledge, the first systematic study of transformer-based political bias detection in Turkish news. Working with a curated, topic-balanced corpus of articles from outlets spanning left, centre, and right positions, we evaluated monolingual, multilingual, and decoder-style models on a three-way ideological classification task. Despite the constraints of operating in a low-resource language and the inability to release the underlying articles for copyright reasons, our results show that it is possible to reach performance levels comparable to those reported in English, provided that models are fine-tuned on an appropriate Turkish dataset. In doing so, this work fills an important gap by establishing a first baseline for automated bias detection in Turkish media, demonstrating that even without massive, manually annotated resources, one can leverage weak labels (outlet ideology as proxy) and transfer learning to build a functional bias classifier.

Our findings have broader implications for Turkic NLP and for computational studies of media bias. They argue strongly against relying on naïve cross-lingual transfer from English and in favour of building and evaluating models in the target language. For Turkish, this means investing in open, carefully designed bias benchmarks that reflect the diversity of outlets and ideological positions, ideally with finer-grained labels than outlet identity alone. For the wider Turkic family, our methodology can be extended to other languages, enabling comparative work on how political bias is manifested across different linguistic and media environments. Beyond research, a classifier of the kind we develop here could serve as a backend for media monitoring tools, helping readers, journalists, or fact-checkers to quickly gauge the political slant of an article. Our approach also highlights good practice for low-resource settings more generally: ensuring topic diversity to prevent models from exploiting spurious shortcuts, and using cross-validation ensembles to maximise performance from limited data.

Future work should therefore prioritise (i) open Turkish political bias benchmarks that can be shared and extended, (ii) systematic extensions to other Turkic languages, (iii) more nuanced labelling schemes that explicitly separate outlet-level stance from article-level framing, and (iv) audits of how different pre-training regimes and data sources affect ideological coverage and balance. In sum, transformer-based models can already detect political bias in Turkish with reasonable accuracy, but their reliability and fairness depend critically on the availability of high-quality Turkish data and on conscious choices about which models are deployed, how they are fine-tuned, and how they are evaluated.

Limitations

For copyright reasons we cannot release the underlying news articles, which constrains reproducibility and reuse. Nonetheless, the methodology and empirical results clearly demonstrate the urgency of developing shareable Turkish political bias datasets under more permissive licensing, whether via partnerships with media organisations, the use of open-licensed sources, or carefully designed synthetic or paraphrased corpora. In addition, outlet-level ideological labels are a weak proxy of supervision, and they do not imply that all individual articles reflect explicit bias on ideology, which may introduce noise in the label into the analysis.

Acknowledgments

H.D. is supported by the EPSRC AI Hub on Mathematical Foundations of Intelligence: An “Erlangen Programme” for AI No. EP/Y028872/1.

References

AllSides. n.d. Media bias ratings. https://www.allsides.com/media-bias/media-bias-ratings. Accessed: 28 July 2025.

Articoder. 2020. News dataset for news bias analysis. https://www.kaggle.com/datasets/articoder/news-dataset-for-news-bias-analysis. Accessed: 17 December 2025.

Alessio Bajec. 2023. Turkey’s opposition opens up to the hijab. Al Jazeera. Accessed: 17 August 2025.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991. Association for Computational Linguistics.

Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Farheen Dairkee and Haim Dubossarsky. 2024. Strengthening the wic: New polysemy dataset in hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349.

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. Association for Computational Linguistics.

Doğu Ergil. 2010. Constitutional referendum: Farewell to the 'old turkey'. *Insight Turkey*, 12(4):15–22. Accessed: 28 July 2025.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Roksana Goworek and Haim Dubossarsky. 2025. Multilinguality does not make sense: Investigating factors behind zero-shot cross-lingual transfer in sense-aware tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35004–35029, Suzhou, China. Association for Computational Linguistics.

Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Benjamin Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. Lora: Low-rank adaptation of large language models. Accessed: 17 August 2025.

A. Q. Jiang, A. Sablayrolles, A. Mensch, and 1 others. 2023. Mistral 7b: a high-performing lightweight language model. Accessed: 28 July 2025.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. 2019. Roberta: a robustly optimized BERT pretraining approach. Accessed: 17 August 2025.

T. Menzner and J. L. Leidner. 2024a. Biasscanner: Automatic detection and classification of news bias to support news readers. *arXiv*.

Tim Menzner and Jochen L. Leidner. 2024b. Experiments in news bias detection with pre-trained neural transformers.

Andrew O'Donohue, Max Hoffman, and Alan Makovsky. 2020. Turkey's changing media landscape. Center for American Progress. Published: 10 June 2020.

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

Stefan Schweter. 2020. BERTurk – BERT models for turkish (version 1.0).

Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. Media bias in german news articles: A combined approach. In *ECML PKDD 2020 Workshops*, pages 581–590. Springer.

Timo Spinde, Lada Rudnitckaia, Kanishka Sinha, Felix Hamborg, Bela Gipp, and Karsten Donnay. 2021. Mbic–a media bias annotation dataset including annotator characteristics. *arXiv preprint arXiv:2105.11910*.

S. Yigit-Sert, İ. S. Altingövde, and Ö. Ulusoy. 2016. Towards detecting media bias by utilising user comments. In *Proceedings of the 8th ACM Web Science Conference (WebSci 2016)*, pages 374–375. ACM.

A Appendix A: Keyword List Used for Data Collection

A.1 Neutral Topics

rusya, suriye, abd, israil, ukrayna, iran, trump, erdoğan, akp, protesto, imamoğlu, dem parti, nato, mavi vatan, doğalgaz, asgari ücret, kalkınma planı, blockchain, fintech, e-spor.

A.2 Positive Topics

bayram, turizm rekoru, teknofest, yenilenebilir enerji, yerli otomobil, uzay programı, startup ekosistemi, güneş enerjisi, ar-ge destekleri, sağlık turizmi, savunma ihracatı, metro projesi, spor başarısı, ihracat, tarım, yapay zeka.

A.3 Negative Topics

enflasyon, işsizlik, deprem, sel, yangın, kur krizi, ekonomik kriz, kredi faizi, gıda fiyatları, mülteci krizi, siber saldırı, hava kirliliği, hak ihlali, terör saldırısı, susuzluk, firtına, kadın cinayeti.