SHYN: Multilingual Multiaspect Media Profiling

Daniil Orel and **Rakhat Khamitov** and **Abylaikhan Kazymbetov** and **Shynar Mars**
and Ben Tyler and Adnan Yazici
Nazarbayev University
Astana, Kazakhstan
{daniil.orel, rakhat.khamitov, abylaikhan.kazymbetov,
shynar.mars, btyler, adnan.yazici}@nu.edu.kz

Abstract

With the rapid increase in the volume of information generated on the web, the issue of misinformation has become more pronounced. Consequently, the ability to assess the factuality of reporting by various web sources, on par with understanding their political bias and attempts to manipulate the reader's decisions has emerged as a critical need. Presently, the majority of existing systems predominantly cater to English language sources, leaving a significant gap in multi-language coverage. To address this, we are introducing **Shyn** (derived from the Kazakh word **Shyn** meaning "the truth") - an innovative online tool designed for media profiling across ten major languages: English, Chinese, Hindi, Kazakh, Italian, French, Korean, Spanish, German and Russian. This tool represents a significant step forward in supporting diverse linguistic contexts in the critical task of discerning accurate information on the internet.

1 Introduction

Fake news stands out as a significant challenge in our digital information era (Lévy, 1999). Its ability to manipulate collective intelligence has led to severe consequences such as riots (Chuai and Zhao, 2020), damage to public property (Ahmed et al., 2020), and abnormal behavior in financial markets (Clarke et al., 2020). What is even more troubling is that the identification of fake news often occurs only after it has circulated in the media for a considerable period of time (Iyengar and Massey, 2019). That is why factually incorrect information should be identified as soon as it get posted.

At the same time, readers must be aware of the political bias in the media they consume, as it is crucial for understanding the context of presented information. For instance, engaging with a pro-governmental source may involve a lack of criticism and potential exaggeration of positive government impacts. In addition to that, recognizing language in reporting is vital, as certain words can evoke specific emotions. Media outlets manipulate language for effect, emphasizing or downplaying incidents. Attending to sentence framing and the underlying message is essential for informed reading.

To address the aforementioned issues, various systems have been developed. While some assess the truthfulness of information by considering the speaker's personality and emotions (Fitzpatrick et al., 2021), others focus on the content of articles themselves. For instance, Prta (Da San Martino et al., 2020) concentrates on recognizing propaganda techniques, while Tanbih (Zhang et al., 2019) evaluates factuality and political bias by continuously analyzing articles from approximately 1,000 media sources in English and Arabic. Tools like NewsScan (Kevin et al., 2018) provide readers with information about news articles, including lexical properties (e.g., ease of reading) and intrinsic properties (e.g,. sentiment score, political bias).

Existing web tools, however, can be somewhat inflexible, in that many are centered around the English language and are not applicable to articles and media sources in other languages. Additionally, these tools may be limited to specific websites, and struggle to process new ones. They are also oriented on limited topics and do not give a holistic representation of the web-sites.

In response to these limitations, we introduce **Shyn**—a tool designed to dynamically profile web-sources. **Shyn** is capable of working with ten major languages: English, Chinese, Hindi, Kazakh, Italian, French, Korean, Spanish, German and Russian (the video with demonstration of web-site processing in these languages is available via the link). In the development of **Shyn**, we have achieved the following, which we will discuss in this paper:

- Curated a synthetic multilingual corpus to train models for factuality, political bias, genre, framing, and persuasion techniques detection.

- Demonstrated that training on translated sentences does not compromise the model's performance on texts originally written in the target language.
- Developed a web-platform, which provides multi-aspect media profiling in near real-time fashion.

2 Data

2.1 Factuality, bias, and media freedom estimation

We utilized the CLEF 2023 CheckThat! lab's open-source dataset (Nakov et al., 2023), consisting of articles from 1189 websites across training (947), development (120), and test sets (122). Each website includes factuality labels from www.mediabiasfactcheck.com, categorizing sources into **Low** (unreliable), **Mixed** (potentially biased or mixed-factual), and **High** (credible and low-biased).

Simultaneously, for these web-sites we extracted political bias labels from the mediabiasfactcheck website. These labels encompass major political bias classifications such as **Right, Left**, and **Least biased**, as well as sub-categories like **Far left**, **Far right**, **Left-center**, and **Right-center** bias.

Beyond gathering political bias labels, our focus extended to collecting data on the freedom of websites. The freedom of a website reflects the degree to which its articles are influenced by the government. In this context, web sources fall into distinct levels, ranging from **completely** and **mostly free** to those with **moderate** or **limited freedom**, and finally, those experiencing **total oppression**.

Given the dataset's availability exclusively in English, we aimed to expand its breadth by translating the existing articles into other languages. Employing Neural Machine Translation (NMT), a technique harnessing Deep Neural Networks to automate text translation across languages (Zhang and Zong, 2020), we obtained translations for the provided articles.

To streamline the translation process and mitigate lack of computations resources, we opted to employ a translation API suitable for large-scale NMT systems. Our choice of Google Translate was guided by its superior performance compared to other translation APIs (Kadaoui et al., 2023).

Split	# of sources	# of articles
Train	9,470	79,480
Dev	1,200	10,490
Test	1,220	10,540
Total	**11,890**	**100,510**

Table 1: Updated CLEF 2023 dataset statistics

As a result, we obtained a new corpus, nearly ten times as large as the original one, with two more additional labels (political bias, and web-site's freedom ranking). The statistics on this dataset is given in Table-1. This corpus can be accessed by the following link.

2.2 News genre, framing, and persuasion techniques

The data used for predicting news genre, framing, and persuasion techniques was obtained from SemEval 2023 Task 3 (Piskorski et al., 2023b). It is a multilingual dataset comprising training, validation, and test data in English, French, German, Italian, Polish, and Russian (Piskorski et al., 2023a). Additionally, it includes test datasets in Spanish, Greek, and Georgian.

Most of these languages align with our 10 target languages. For our target languages not present in the SemEval dataset (Korean, Chinese, Hindi, Kazakh), we translated English SemEval data, as English is the most represented language in SemEval's training set, as shown in Figure 2. For Spanish, we only translated the training and development sets from English to Spanish, as the test set was provided.

Genre identification. The dataset encompasses three types of news genres: opinion pieces, objective news reporting, and ridiculed writings of real organizations or individuals, commonly referred to as satire.

Framing. There are 14 framing domains, identified by Card et. al. (Card et al., 2015): Economic, Capacity and resources, Morality, Fairness and equality, Legality, constitutionality and jurisprudence, Policy prescription and evaluation, Crime and punishment, Security and defense, Health and safety, Quality of life, Cultural identity, Public opinion, Political, External regulation and reputation. Every document in the dataset was assigned one or many labels from these categories.

Persuasion techniques. Totally 23 types of persuasion techniques are used in the dataset: name calling, guilt by association, casting doubt, appeal to hypocrisy, questioning reputation, flag waiving, appeal to authority, appeal to popularity, appeal to values, appeal to fear, strawman, red herring, whataboutism, casual oversimplification, false dilemma, consequential oversimplification, slogans, conversation killing phrases, appeal to time, loaded language, confusion, exaggeration or minimisation, repetition (Piskorski et al., 2023b). None, one, or many such labels are assigned to paragraphs of articles in the dataset.

3 Modeling

3.1 Metrics

To evaluate our models, we used F1-score and MAE. F1-score covered all tasks, while MAE focused on ordinary categories: factuality, bias, and freedom. Factuality levels were treated as ordinal variables (0 for low, 1 for mixed, 2 for high), bias ranged from left (low values) to right (high values), and freedom labels used lower numbers for less freedom (totally oppressed sources were assigned label of 0, while completely free ones were assigned 4).

F1-score represents the harmonic mean of precision and recall, balancing the trade-off between correctly identifying the minority class and avoiding false positives in the majority class. The higher the F1-score, the better the model's performance.

MAE requires further clarification. In our application, we prioritize avoiding significant errors, such as misidentifying a credible website as a source of fake information (case 1). However, we allow tolerance for minor errors, like indicating that a website with mixed factuality has low credibility (case 2). To explain the use of MAE: we assign labels **0, 1, 2** for low, mixed, and high factuality, respectively. In the first case, if the model predicts **0** instead of **2**, the MAE is 2; in the second case, if **0** is predicted instead of **1**, the MAE is 1. Essentially, MAE facilitates unequal treatment of misclassifications among different ordinal classes. The lower the MAE, the better the model's performance.

3.2 Baseline Model

We opted to leverage pre-trained State-of-the-Art multilingual models as a foundation and subsequently fine-tuned them to suit our specific task. As such a model, we selected **mDeBERTa**. It is a multilingual version of the DeBERTa model, which incorporates disentangled attention mechanisms and a mask decoder with information on word's absolute position (He et al., 2021). For all tasks this model was trained on 100 epochs with adjusted class weights to address data imbalance.

3.3 Predicting factuality, bias and freedom

The performance of models - namely their weighted F1 score and MAE score - trained on our translated dataset for factuality, bias, and freedom data, is depicted in Table 3. From the table, it is evident that our model demonstrates notably strong performance in estimating the factuality and freedom of a website. Unfortunately, we lack references for performance comparison across all languages. Nevertheless, it is noteworthy that our model surpasses baseline models and could potentially rank in the top three in the CLEF-2023 competition (Nakov et al., 2023), based on its MAE score of **0.46** for the English language.

In evaluating the model's bias estimation, we noted its accurate identification of the main political bias types ("Bias (3)" in the table): left, right, and least biased sources. However, when discerning more nuanced bias types ("Bias (7)" in the table), such as determining if a website is far-left, left-center, or simply left, the model's performance significantly decreases.

3.4 News Genre Prediction

In this analysis, we compare our approach with a support vector machine (SVM) model based on 5-character n-grams, as provided by the organizers of SemEval-2023 Task 3.

As the organizers did not evaluate Chinese, Hindi, Kazakh, and Korean languages, baseline scores are unavailable for these languages. Table 4 demonstrates that our solution consistently outperforms the baseline. Notably, in the case of Spanish, where we lacked specific training data provided by the organizers and utilized a translation of the English training set, our model still exhibits commendable performance. As highlighted in 2.2, for Chinese, Hindi, Kazakh, and Korean, we utilized translated English data, and the results indicate comparable performance to what was achieved for English data.

3.5 News Framing Prediction

In this instance, the SemEval 2023 Task 3 organizers employed a uni- and bi-gram SVM as a baseline. Table 5 illustrates that our model surpasses the

Split	English	French	German	Italian	Polish	Russian	Georgian	Greek	Spanish
Train	446	158	132	227	145	143	-	-	-
Dev	90	53	45	76	49	48	-	-	-
Test	54	50	50	61	47	72	29	64	30

Table 2: Number of documents per language in SemEval 2023 dataset

baseline for all languages. It's noteworthy to highlight that, in comparison to the results in 3.4, the performance gap has narrowed for all languages except Spanish. Remarkably, for Spanish, our model trained on English translations exhibits outstanding performance.

We acknowledge that being aware of the topics covered by a specific web resource alone may not provide comprehensive information. Therefore, we enhance the user's understanding by incorporating each web resource's stance on crucial topics. These include gun rights, workers' rights, environmental policies, immigration, healthcare, military spending, abortion, and education.

To identify the positions of web resources on these topics, we instruct GPT-4 to identify their opinions based on mediabiasfactcheck's bias indicators[1].

3.6 Persuasion Techniques Identification

Table-6 shows that for this task performance of our models has significantly dropped, but it is still beyond uni- and bi-gram SVM, used as a baseline. We believe that this drop in performance is related to the high number of labels (23), and the fact that model predicted these labels on paragraph, not article level, having less context.

4 System Design

Our system's architecture is depicted in Figure-1. When a user accesses **Shyn**, they input the web page or article address into the search bar. The frontend of the system, developed using React.js[2], initiates a request to the FastAPI[3] backend. Initially, we verify if this web source has been previously checked within the last three months; if so, precomputed results are directly returned.

Alternatively, our backend attempts to parse up to 100 articles from the web source (see section 4.1) and utilizes the API of our models hosted on HuggingFace[4] (detailed in section 4.2). Upon evaluating the content of the web source, the backend transmits the result to the frontend. Due to potential longer processing times, we implement short polling (outlined in section 4.3) to manage this delay.

Finally, the frontend displays the aggregated predictions.

4.1 Article Parsing

To extract information from articles, the newspaper3k[5] library in Python is used. This library is a valuable tool for web scraping, because it can parse articles from various websites, even though they have different HTML structures. While newspaper3k excels in its parsing capabilities, it's worth noting that some websites may block such activities. Additionally, the library has not seen recent updates (last change being two years ago). Despite this, it continues to function effectively for our purposes in most cases.

4.2 Integration with ML Model

The backend integrates with our models from 3 hosted on Hugging Face, communicating with them over HTTP.

4.3 Handling Response Time Challenges

Since the scoring process may take longer than anticipated, a solution has been implemented to address potential user experience issues. Given that a single HTTP request might not be sufficient, the capability of short polling has been added. The frontend initiates a request to submit the task and then periodically queries the backend for the task status. This ensures that even if a user switches tabs, closes the browser, or experiences a connection disruption, the results of the fact-checking process can still be retrieved. This design choice enhances user satisfaction and usability by providing a resilient and responsive system.

[1] https://mediabiasfactcheck.com/left-vs-right-bias-how-we-rate-the-bias-of-media-sources/

[2] https://react.dev/

[3] https://fastapi.tiangolo.com/

[4] https://huggingface.co/

[5] available at https://newspaper.readthedocs.io/en/latest/

Model	Languages									
	English	Chinese	Hindi	Kazakh	Italian	French	Korean	Spanish	German	Russian
Bias(7)	0.41	0.43	0.40	0.37	0.41	0.42	0.40	0.39	0.42	0.41
	0.95	0.90	1.03	1.02	0.96	0.95	1.0	0.98	0.90	0.95
Bias(3)	0.59	0.61	0.56	0.54	0.58	0.58	0.59	0.60	0.60	0.58
	0.50	0.46	0.54	0.57	0.52	0.51	0.52	0.49	0.48	0.53
Factuality	0.58	0.62	0.61	0.59	0.61	0.60	0.61	0.59	0.62	0.60
	0.46	0.43	0.43	0.48	0.44	0.43	0.45	0.45	0.42	0.44
Freedom	0.85	0.85	0.85	0.86	0.85	0.85	0.86	0.86	0.85	0.86
	0.32	0.32	0.32	0.26	0.32	0.32	0.26	0.26	0.32	0.26

Table 3: F1 (upper row) and MAE (lower row) for factuality, political bias, and freedom estimation

Model	Languages									
	English	Chinese	Hindi	Kazakh	Italian	French	Korean	Spanish	German	Russian
Baseline	0.29	-	-	-	0.39	0.57	-	0.15	0.63	0.40
Ours	**0.49**	**0.51**	**0.48**	**0.50**	**0.62**	**0.71**	**0.48**	**0.43**	**0.77**	**0.68**

Table 4: Multilingual evaluation for News Genre prediction (macro F1 score).

Model	Languages									
	English	Chinese	Hindi	Kazakh	Italian	French	Korean	Spanish	German	Russian
Baseline	0.35	-	-	-	0.49	0.33	-	0.12	0.49	0.23
Ours	**0.52**	**0.48**	**0.47**	**0.51**	**0.57**	**0.48**	**0.47**	**0.62**	**0.55**	**0.39**

Table 5: Multilingual evaluation for News Framing prediction (micro F1 score).

5 Conclusion and Further Work

Our system excels in profiling web sources across 10 languages. Through our multilingual experiments, we have observed that models trained on translated data exhibit robust generalization capabilities in regards of texts originally written in the target language.

The distinctiveness of **Shyn** in comparison with other systems, mentioned in 1 extends beyond its multilingual capabilities. Our analysis encompasses a broader spectrum of web-site related features: we evaluate various aspects such as factuality, political bias, the freedom of web-sites, genres of news materials, their framing, and persuasion techniques altogether. Additionally, there are plans to incorporate multimodal evaluation features in the next version of **Shyn**. This advanced evaluation will go beyond analyzing textual information alone and extend to assessing cross-references and audience overlaps of web-sources.

6 Limitations

Our system currently faces a limitation that needs to be acknowledged. As discussed in 4.1, certain web-resources hinder parsing by deploying popups that restrict crawler access to site content. We are actively developing a solution to circumvent these blockers and aim to implement this enhancement in the forthcoming version of **Shyn**.

7 Ethical Considerations

It is important to note that the ratings provided by our model for media sources aren't absolute truths or definitive measures of truth or falsehood. Rather, they offer a thoughtful and systematic evaluation of potential truthfulness, derived from a distribution that, while not flawless, has been meticulously developed and rigorously scrutinized to offer a meaningful assessment. Therefore, the primary responsibility for making decisions regarding the information rests with the users themselves.

References

Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, Francesc López Seguí, et al. 2020. Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. *Journal of medical internet research*, 22(5):e19458.

Model	Languages									
	English	Chinese	Hindi	Kazakh	Italian	French	Korean	Spanish	German	Russian
Baseline	0.20	-	-	-	0.40	0.24	-	0.25	0.32	0.21
Ours	**0.34**	**0.30**	**0.31**	**0.30**	**0.43**	**0.31**	**0.30**	**0.30**	**0.48**	**0.33**

Table 6: Multilingual evaluation for Persuasion Techniques detection (micro F1 score).

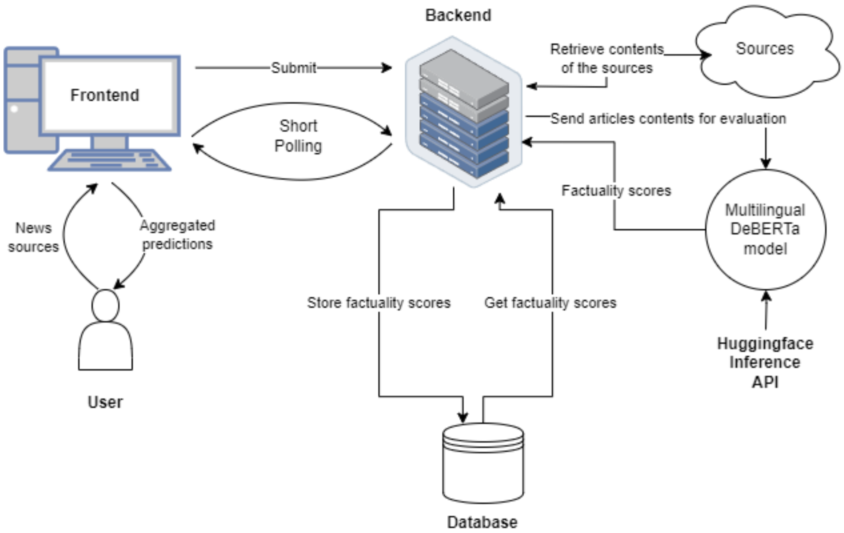

Figure 1: Diagram of **Shyn** system

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 438–444, Beijing, China.

Yuwei Chuai and Jichang Zhao. 2020. Anger makes fake news viral online.

Jonathan Clarke, Hailiang Chen, Ding Du, and Yu Jeffrey Hu. 2020. Fake news, investor attention, and market reaction. *Information Systems Research*, 32(1):35–52.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.

Benjamin Fitzpatrick, Xinyu "Sherwin" Liang, and Jeremy Straub. 2021. Fake news and phishing detection using a machine learning trained expert system.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Shanto Iyengar and Douglas S Massey. 2019. Scientific communication in a post-truth society. *Proceedings of the National Academy of Sciences*, 116(16):7656–7661.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties.

Vincentius Kevin, Birte Högden, Claudia Schwenger, Ali Şahan, Neelu Madan, Piush Aggarwal, Anusha Bangaru, Farid Muradov, and Ahmet Aker. 2018. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 28–33, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lévy. 1999. *Collective intelligence: Mankind's emerging world in cyberspace*. Perseus Publishing.

Preslav Nakov, Firoj Alam, Giovanni Da San Martino, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Panayot Panayotov. 2023. Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy).

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2020. Neural machine translation: Challenges, progress and future.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. 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): System Demonstrations*, pages 223–228, Hong Kong, China. Association for Computational Linguistics.

A Application Interface

This section showcases the user interface of our system. As seen in Figure 2, the main page features a search bar where users can input either a specific article URL or a website address for evaluation. The interface also includes a feature for users to access their history of previous checks.

In Figure 3, we present an example of our model's analysis. This includes an article from *sayasat.kz*, a well-regarded independent media outlet in Kazakhstan, and a comprehensive evaluation of *thegrayzone.com*, known for content that has been rated as having mixed factuality by mediabiasfactcheck. The results from these checks are displayed in an intuitive diagram, accompanied by a detailed textual description, offering users a clear understanding of the model's assessment.

B Evaluation on Real Data

We realise that training on translated texts we fit our models to a particular domain, which can be different from what is really written in our target languages. To check if our models are able to generalize well on texts, originally written in languages other than English, we decided to test them on a small hand-crafted dataset. To do so, we tried to find web-resources with various factuality, bias and freedom labels on mediabiasfactcheck web-site, and then searched for their versions in target languages. Fortunately, most of mediabiasfactcheck sources have such versions, so for every language we selected 10 web-resources that originally write news in our target languages. Then we applied **Shyn** to evaluate them. Then we predicted the 3 aforementioned criteria using our models, and constructed Table-7.

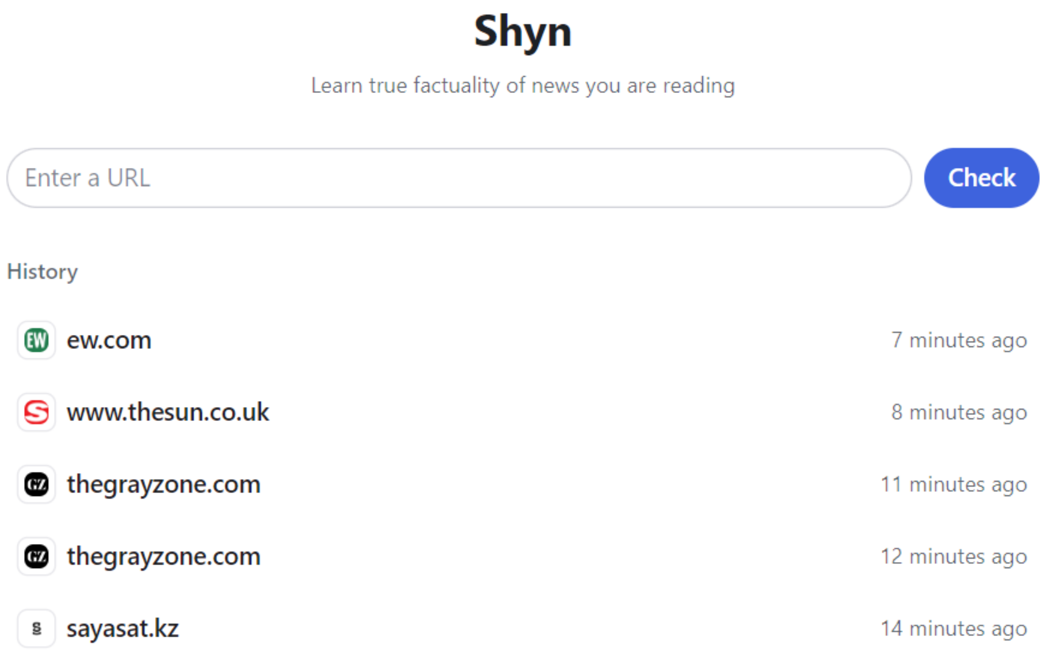

Figure 2: **Shyn** app's main page

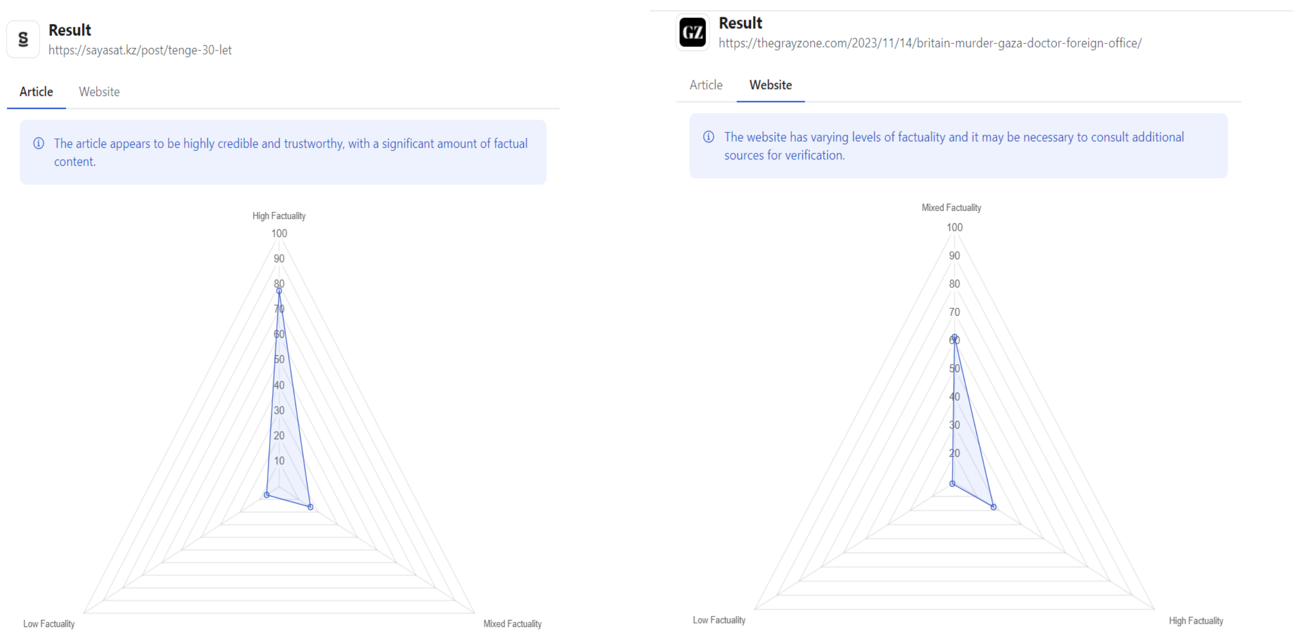

Figure 3: Example of aggregated predictions for high and mixed factuality sources

Model	Languages									
	English	Chinese	Hindi	Kazakh	Italian	French	Korean	Spanish	German	Russian
Bias(7)	0.41	0.43	0.40	0.37	0.41	0.42	0.40	0.39	0.42	0.41
	0.95	0.90	1.03	1.02	0.96	0.95	1.0	0.98	0.90	0.95
Bias(3)	0.59	0.61	0.56	0.54	0.58	0.58	0.59	0.60	0.60	0.58
	0.50	0.46	0.54	0.57	0.52	0.51	0.52	0.49	0.48	0.53
Factuality	0.58	0.62	0.61	0.59	0.61	0.60	0.61	0.59	0.62	0.60
	0.46	0.43	0.43	0.48	0.44	0.43	0.45	0.45	0.42	0.44
Freedom	0.85	0.85	0.85	0.86	0.85	0.85	0.86	0.86	0.85	0.86
	0.32	0.32	0.32	0.26	0.32	0.32	0.26	0.26	0.32	0.26

Table 7: F1 (upper row) and MAE (lower row) for factuality, political bias, and freedom estimation on real world data