*big data and cognitive computing*

MDPI

*Article*

# A Multi-Level Annotation Model for Fake News Detection: Implementing Kazakh-Russian Corpus via Label Studio

**Madina Sambetbayeva [1,2], Anargul Nekessova [2,*], Aigerim Yerimbetova [1,*], Abdygalym Bayangali [1,2], Mira Kaldarova [3], Duman Telman [4] and Nurzhigit Smailov [5,*]**

[1] Institute of Information and Computational Technologies of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Almaty 050010, Kazakhstan; sambetbayeva_ma_1@enu.kz (M.S.); bayangali.abd@gmail.com (A.B.)
[2] Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan
[3] School of Information Technology and Engineering, Astana International University, Astana 010000, Kazakhstan; kmiraj8206@gmail.com
[4] Department of Software Engineering, Satbayev University, Almaty 050013, Kazakhstan; telmand28@gmail.com
[5] Department of Electronics, Telecommunications and Space Technologies, Satbayev University, Almaty 050013, Kazakhstan
* Correspondence: aimurat_anara@mail.ru (A.N.); aigerian8888@gmail.com (A.Y.); n.smailov@satbayev.university (N.S.)

**Abstract**

This paper presents a multi-level annotation model for detecting fake news in Kazakh and Russian languages, aiming to enhance understanding of disinformation strategies in multilingual digital media environments. Unlike traditional binary models, our approach captures the complexity of disinformation by accounting for both linguistic and cultural factors. To support this, a corpus of over 5000 news texts was manually annotated using the Label Studio platform. The annotation scheme consists of seven interrelated categories: CLAIM, SOURCE, EVIDENCE, DISINFORMATION_TECHNIQUE, AUTHOR_INTENT, TARGET_AUDIENCE, and TIMESTAMP. Inter-annotator agreement, evaluated using Cohen's Kappa, ranged from 0.72 to 0.81, indicating substantial consistency. The annotated data reveals recurring patterns of disinformation, such as emotional manipulation, targeting of vulnerable individuals, and the strategic concealment of intent. Semantic relations between entities, such as CLAIM → EVIDENCE and CLAIM → AUTHOR_INTENT were formalized to represent disinformation narratives as knowledge graphs. This study contributes the first linguistically and culturally adapted annotation model for Kazakh and Russian languages, providing a robust and empirical resource for building interpretable and context-aware fake news detection systems. The resulting annotated corpus and its semantic structure offer valuable empirical material for further research in natural language processing, computational linguistics, and media studies in low-resource language environments.

**Keywords:** fake news; annotation; disinformation; Kazakh language; Russian language; semantic relations; Label Studio; corpus linguistics; automated fake news detection

check for updates

Academic Editors: Domenico Ursino and Ximing Li

Received: 5 June 2025
Revised: 8 August 2025
Accepted: 15 August 2025
Published: 20 August 2025

**Citation:** Sambetbayeva, M.; Nekessova, A.; Yerimbetova, A.; Bayangali, A.; Kaldarova, M.; Telman, D.; Smailov, N. A Multi-Level Annotation Model for Fake News Detection: Implementing Kazakh-Russian Corpus via Label Studio. *Big Data Cogn. Comput.* **2025**, *9*, 215. https://doi.org/10.3390/bdcc9080215

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In recent years, the spread of disinformation in the digital media space has reached unprecedented levels, significantly influencing public opinion, political processes, and information security. This issue is particularly pressing in multilingual societies such as

Kazakhstan and the broader post-Soviet region, where lexical, syntactic, and stylistic variability complicates the identification of fake content. Existing fake news detection models were primarily developed for English-language content and exhibit limited applicability in Kazakh–Russian contexts, underscoring the urgent need to adapt annotation approaches to local linguistic and cultural realities, unlike the English-language media landscape, which benefits from large-scale corpora and mature fake news detection models (e.g., FEVER, LIAR, FakeNewsNet), the Kazakh–Russian information space is characterized by a set of specific features not reflected in current annotation schemes. First, there are linguistic differences related to the agglutinative nature of the Kazakh language, free word order, and unique morphosyntactic structures, which hinder the direct transfer of English-based models without adaptation. Second, the Kazakhstani media environment is dominated by localized forms of disinformation, including ethnopolitical narratives and manipulations on interethnic and socioeconomic grounds, phenomena that differ substantially from Western patterns of fake news. Third, the mode of information delivery also diverges: local texts frequently employ emotional and culturally specific arguments, distort context, or target narrow social groups (e.g., rural populations or ethnic minorities), aspects that are not captured by existing English-language annotation schemes, which are predominantly tailored to the political discourse of the US or Europe. Due to these factors, the unadapted application of existing models and annotation schemes results in reduced accuracy and interpretability of outcomes. The development of a new multi-level annotation framework that incorporates the cultural, pragmatic, and discursive characteristics of Kazakh–Russian fake news is therefore essential for constructing relevant and robust models of automatic disinformation detection in this context.

As Shu et al. [1] noted, disinformation refers to false or misleading information disseminated with the intent to deceive. During critical events such as elections, pandemics, and economic crises, the volume of fake news increases, fueled by fear, uncertainty, and ideological polarization. The internet and social media serve as catalysts, accelerating the spread of unverified content and replacing professional journalism with user-generated content on social media platforms (e.g., Facebook, Instagram, VK, and Twitter) and online forums (e.g., Reddit and regional discussion boards). Given the scale of the issue, researchers in Natural Language Processing (NLP) have developed various approaches, ranging from manual fact-checking to rhetorical and source credibility analysis [2].

Singh et al. [3] emphasized that political agendas, social polarization, and economic incentives are key drivers behind the spread of fake news. Moreover, the low barrier to publication and the viral nature of digital content facilitate the rapid diffusion of disinformation in communities with limited media literacy and critical thinking.

One of the most persistent challenges remains the lack of high-quality annotated corpora, especially for low-resource languages such as Kazakh. Existing datasets typically rely on binary classification of content as fake or real, which fails to account for manipulation techniques, authorial intent, or target audience [4].

This study aims to develop and implement a multi-level annotation scheme for detecting fake news in Kazakh and Russian. The research objectives include analyzing existing annotation approaches for disinformation; identifying methodological limitations in multilingual contexts; designing a custom annotation framework implemented in Label Studio [5]; building a pilot corpus; and evaluating annotation consistency and model performance.

The object of the study is fake news discourse in Kazakh and Russian, while the subject is the annotation features of disinformation, its structure and mechanisms of influence.

The discourse of fake news represents a distinct type of media discourse characterized by the creation, dissemination, and consumption of messages containing deliberately false, distorted, or manipulative information aimed at shaping or altering public opinion,

emotional states, or audience behavior. This discourse is marked not only by factual inaccuracy (such as false claims or fabricated events) but also by the use of specific linguistic, stylistic, and pragmatic strategies, such as emotional manipulation, clickbait tactics, and contextual distortion, designed to enhance persuasive impact and reduce the recipient's critical perception. In multilingual environments, such as the Kazakh–Russian context, the discourse of fake news is further complicated by sociocultural factors and differences in linguistic registers. These complexities necessitate a comprehensive linguistic and semantic analysis to effectively capture the nuances of such content.

The study is based on the hypothesis that an adapted, multi-level annotation model provides higher interpretability and detection accuracy than monolingual, binary schemes developed for English. The hypothesis that an adapted multi-level annotation scheme enhances the interpretability and accuracy of fake news detection in the Kazakh–Russian media space is grounded in several empirical and theoretical premises. First, findings from existing studies indicate that binary annotation schemes (i.e., fake/real) fail to capture the complex pragmatic and manipulative dimensions of fake messages, particularly in multilingual and culturally diverse contexts. Second, a preliminary content analysis of Kazakh and Russian-language news texts revealed the presence of stable local disinformation patterns that are not accounted for in existing English-language models. Third, drawing on the principles of cognitive pragmatics, discourse theory, and media manipulation studies, it is posited that a multi-level annotation framework, one that incorporates authorial intent, target audiences, and manipulation techniques, can provide a more structured representation of fake news content and improve the performance of downstream automated detection systems.

The relevance of this research lies in the urgent need for linguistically and culturally sensitive annotation resources that can support effective disinformation detection in multilingual digital environments. The proposed scheme contributes to both applied and theoretical dimensions of media annotation and fake news mitigation.

*Literature Review and Conceptual Background*

In recent years, a growing body of research has emerged focusing on the detection of fake information in multilingual settings, including approaches based on transformer models. For instance, Ragab et al. [6] demonstrated the effectiveness of multilingual models such as XLM-RoBERTa and mT5 in detecting propaganda and disinformation across Arabic, English, and Hebrew. Their findings show that multilingual models outperform monolingual counterparts due to their ability to capture cross-lingual semantic patterns and facilitate knowledge transfer between languages. Pretrained multilingual transformer models, including bert-base-multilingual-cased and xlm-roberta-base, have also shown high performance across a range of natural language processing tasks and are particularly promising for the automatic detection of disinformation in low-resource language settings. These models are trained on large-scale multilingual corpora and are capable of identifying universal markers of disinformation, such as manipulative constructions, emotional tone, and rhetorical strategies that persist across languages. In the context of low-resource languages like Kazakh, multilingual models can effectively leverage knowledge acquired from high-resource languages (e.g., English) and apply it to recognize similar phenomena in different linguistic environments. This approach offers flexibility, resilience to data scarcity, and scalability, allowing a single model to be used across multiple languages without the need to develop separate solutions for each language group. Wang et al. [7] emphasized the challenges of annotation and automated disinformation detection in low-resource language contexts, highlighting the importance of manual annotation for such tasks. Similarly, the EXClaim model proposed by Dutta et al. [8] integrates entity knowledge and cross-lingual transformers for claim detection—a methodology closely aligned with our annotation labels *CLAIM* and *SOURCE*.

Multilingual corpora such as MuMiN, Multiverse, and NewsPolygraph highlight the importance of cross-cultural analysis and multi-layered annotation. However, these datasets lack a manual typological structure like the one proposed in this study, which includes not only factuality but also disinformation techniques, authorial intent, and target audience. This is primarily due to the fact that creating multilingual datasets requires substantial resources to recruit annotators proficient in multiple languages and equipped with the linguistic and cultural competence necessary to identify typological features of disinformation specific to each language. Moreover, most existing corpora have prioritized the rapid generation of large-scale training data for transformer models over detailed manual typologization, largely due to the time and financial constraints inherent in many research projects. This makes our schema particularly relevant for the multilingual media landscape of Kazakhstan.

Manual annotation is essential because automated methods for dataset construction often depend on heuristics, distant supervision, or weak indicators such as user reactions or clickbait features. These approaches tend to produce superficial labels and fail to capture the more intricate dimensions of disinformation, including implicit authorial intent or pragmatic manipulation strategies. In multilingual settings, where linguistic and cultural nuances vary widely, automated systems frequently misinterpret local idioms, pragmatic cues, and contextual meanings, leading to reduced annotation accuracy. In contrast, expert manual annotation enables the identification of complex categories and subtypes, taking into account semantic, stylistic, and cultural factors. This is crucial for developing interpretable and reliable corpora that serve as a solid foundation for training effective disinformation detection models.

The most cited resources belong to LIAR by Wang [9] and FakeNewsNet by Shu et al. [10], which enriched their annotations with social metadata. In this context, social metadata refers to elements such as the number of shares and likes on a post, user comments, information about the accounts that disseminated the news, as well as temporal and geographic data related to its spread. These parameters enable the analysis of how a piece of news circulates on social media and how audiences engage with the content, providing additional context for assessing the credibility of the information. The LIAR dataset includes the text of each claim, a six-point truthfulness scale, source, date, topic, party affiliation, speaker's position, and explanatory notes. This structure allows for nuanced analysis of the influence of source credibility, temporal factors, and political context on the spread of disinformation, and supports the development of multi-level annotation frameworks that go beyond binary to multiclass classification. FakeNewsNet complements news content with social media metadata, including the number of reposts, likes, user comments, account profiles of those sharing the news, as well as temporal and geographic information. These features enable the study of audience interaction patterns and the real-time dissemination mechanisms of fake content. Together, LIAR and FakeNewsNet serve as important reference points for the design and validation of multi-level annotation schemes and automatic fake news detection models, providing robust frameworks for both content analysis and user behavior modeling in the digital information landscape.

However, even these corpora offer annotations only at the global claim level, without addressing the semantic and structural components of the text.

Faustini et al. [11] advocate for deeper linguistic analysis and show the effectiveness of fake news detection systems when built upon well-annotated corpora. Nonetheless, these resources are limited to English and do not account for regional or cultural differences in how disinformation is perceived.

More recently, corpora in Spanish by Silva et al. [12], Portuguese by Khalil et al. [13], Arabic by Assaf & Saheb [14], and other languages have been introduced. Gruppi et al. [15]

emphasize the importance of interpretability in fake news classification, especially when considering stylistic and cultural dimensions. However, none of these resources offer a multi-level annotation structure like the one proposed here, which captures disinformation type, manipulation technique, authorial motivation, and audience targeting. A particularly important contribution of our work is the inclusion of regional features: annotation results show that a substantial portion of disinformation in Kazakh and Russian targets vulnerable groups—particularly elderly individuals and rural populations.

Since our annotation scheme draws on journalistic models such as the Inverted Pyramid, we also consider studies that apply similar methods. Norambuena et al. [16] developed an evaluation method based on event descriptors (5W1H) for analyzing journalistic structure, while Chakma & Das [17] annotated a corpus using semantic roles and structural cues for event extraction. These works support the relevance of journalistic frameworks for annotation design.

Another important dimension involves the linguistic characteristics of text as indicators of credibility. Studies by Zhang et al. [18], Santia et al. [19], Alam et al. [20] and Mottola [21] demonstrated that stylistic, syntactic, and psychological features can aid fake news classification. These findings inform the inclusion of linguistic cues in our annotation framework.

Our multi-level annotation scheme uniquely integrates detailed semantic relations, disinformation techniques, author intent, and target audience, specifically adapted for Kazakh and Russian. In contrast, existing schemes are mostly monolingual, limited in granularity, and lack structured typologies and semantic interconnections. This enables more nuanced analysis of disinformation mechanisms in multilingual and culturally specific contexts. This is further illustrated in Table 1.

**Table 1.** Comparative Analysis of Major Annotation Schemes for Fake News Detection.

| Dataset/Annotation Scheme | Multilingual | #Levels (Granularity) | Semantic Relations | Claim/ Evidence/ Source | Disinformation Techniques | Author Intent | Target Audience | Temporal Markers | Adapted for Kazakh/ Russian | Manual Typology | Application Scenario |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LIAR [9] | No | 1 (Binary/ Multiclass) | No | Yes/Partial/Yes | No | No | No | Partial (Date) | No | No | US Politics |
| FakeNewsNet [10] | No | 1 | No | Yes/No/Yes | No | No | No | Yes | No | No | News, Social Media |
| EX-Claim [8] | Yes | 2 | Yes | Yes/No/Yes | Partial | No | No | Yes | Partial | Partial | Multilingual Claims |
| NewsPolyML [22] | Yes | 2 | Partial | Yes/Yes/Yes | Partial | No | No | No | No | No | Multilingual News |
| Russian FakeNews Dataset [23] | No | 1 | No | Partial | No | No | No | No | Russian | No | Russian News |
| Multiverse [24] | Yes | 2 | Partial | Yes/Yes/Yes | No | No | No | No | No | No | News, Social Media |
| POLygraph [25] | No | 2 | No | Yes/Yes/Yes | Partial | No | No | No | No | Partial | Polish Politics |
| MuMiN [22] | Yes | 2 | Partial | Yes/Yes/Yes | Partial | No | No | Yes | No | No | Social Media |
| Our Scheme (This Study) | Yes | 3 (Multi-level + Subtypes) | Yes | Yes/Yes/Yes | Yes (6+ types) | Yes | Yes | Yes | Yes | Yes | Kazakh–Russian News |

The literature review highlights a clear research gap by identifying the following key limitations in current fake news detection approaches:

1. A predominant focus on English-language content and political topics;
2. Lack of annotation for manipulative techniques, authorial intent, and target audience;
3. Insufficient adaptation to multilingual information environments, particularly for the Kazakh language;
4. Simplistic annotation schemes that do not allow for structural or impact-based analysis of fake news.

The present study's research novelty lies in the development of a multi-level annotation framework that integrates concepts of factual accuracy, information structure, manipulation strategies, and linguistic features. The framework has been developed for

texts in Kazakh and Russian, and addresses the multilingual and sociocultural characteristics of the Central Asian media space.

Our system is distinguished, first and foremost, by its multi-level structure and deep semantic design. Unlike traditional annotation schemes limited to binary labels (fake/real), we propose an expanded typology of disinformation patterns that includes manipulative techniques, authorial intent, and target audiences. The second key advantage lies in the introduction of formalized semantic relationships between entities, enabling the modeling of the internal structure of fake news messages and their pragmatic orientation, i.e., an aspect largely absent in most existing models. The third significant distinction is local adaptation: our scheme accounts for the linguistic characteristics of both Kazakh and Russian, as well as the sociocultural contexts of Kazakhstan. This makes it highly relevant and applicable for detecting disinformation in regional media environments. Taken together, these features not only enhance interpretability but also provide a robust foundation for developing more accurate and resilient models of automatic fake news detection in multilingual settings.

The diagram shows how the annotator iterates over the input text, assigns labels and subtypes to tokens or text spans, establishes semantic relations, and interacts with both the annotation tool and moderator to ensure quality and consistency. This structured, stepwise workflow ensures reproducibility and high-quality data for downstream NLP tasks (see Figure 1).

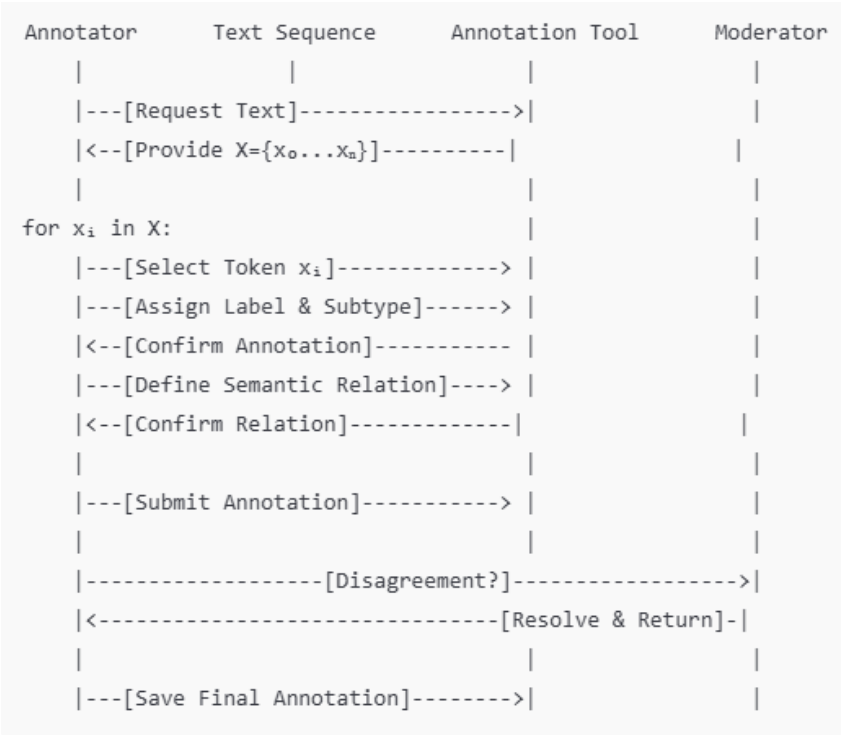



**Figure 1.** UML sequence diagram illustrating the multi-level annotation process using the proposed schema.

The proposed scheme addresses significant gaps in existing datasets and provides a robust foundation for developing more accurate and interpretable manual detection systems for fake news in multilingual environments. It is important to note that this approach encompasses not only the content itself but also the sociocultural context, thereby ensuring a greater degree of responsiveness to the unique dynamics of disinformation in under-resourced language communities.

For the present study, Label Studio was selected as the primary annotation tool, i.e., a modern platform for manual text data labeling, known for its flexible interface and support

for multi-level annotation schemes. The choice of this platform was motivated by its widespread use in academic research, its customizable annotation logic, and its ease of integration into machine learning pipelines.

## 2. Materials and Methods

The use of Label Studio enabled the implementation of the proposed annotation framework with nested label levels and subtypes, capturing structural, semantic, and pragmatic features of disinformation. Over 5000 texts were manually annotated during the study, providing a practical validation of the framework's applicability to real-world data.

To enrich the corpus, major social media platforms such as Instagram, Facebook, and Twitter were used as primary data sources, given their central role in the dissemination of news content and disinformation within the Kazakhstani and Russian-speaking media landscapes. These platforms provide access to a wide range of genres, including news articles, user-generated posts, and comments, enabling the identification of how fake news circulates among different demographic groups. The inclusion of social media data allowed the corpus to capture stylistic features characteristic of these platforms (e.g., emotionally charged headlines, hashtags, and visually oriented content) as well as patterns of audience interaction with disinformation. This, in turn, enhanced the practical value of the corpus for training models aimed at analyzing content in social media environments.

The annotation pipeline is visualized in Figure 2, illustrating the core stages of processing news data—from initial text collection to integration with machine learning models. This modular and reproducible approach ensures transparency at each stage, including manual quality control and export of annotations for further model fine-tuning. The diagram demonstrates how manual annotation can be seamlessly embedded into a full NLP workflow, ranging from data collection to optional ML training.

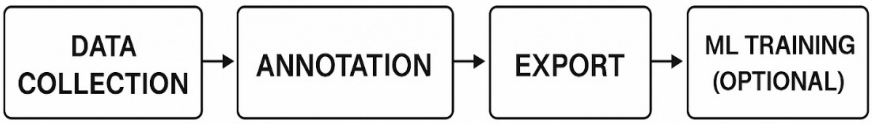



**Figure 2.** Annotation pipeline and corpus preparation for downstream machine learning.

As previously discussed, the process of manually identifying fake news in a multilingual media environment is more complex than a simple binary classification. In the context of linguistic and cultural diversity, it is imperative to annotate texts based on the structure of disinformation, its techniques, authorial intent, and target audience. This transformation of the classical detection task into a problem of structured extraction of semantic and pragmatic features associated with disinformation impact is a significant development in the field.

In this study, a multi-level annotation scheme is introduced, comprising nine primary categories and twenty-three subtypes. Each label corresponds to a semantic or pragmatic component of a disinformation text, ranging from simple elements (e.g., the presence of a source or a claim) to more complex attributes (e.g., disinformation technique or author's intent).

The UML sequence diagram illustrates the step-by-step token annotation process, where each token is selected, processed, and assigned a corresponding label within the multi-level annotation scheme (see Figure 3).

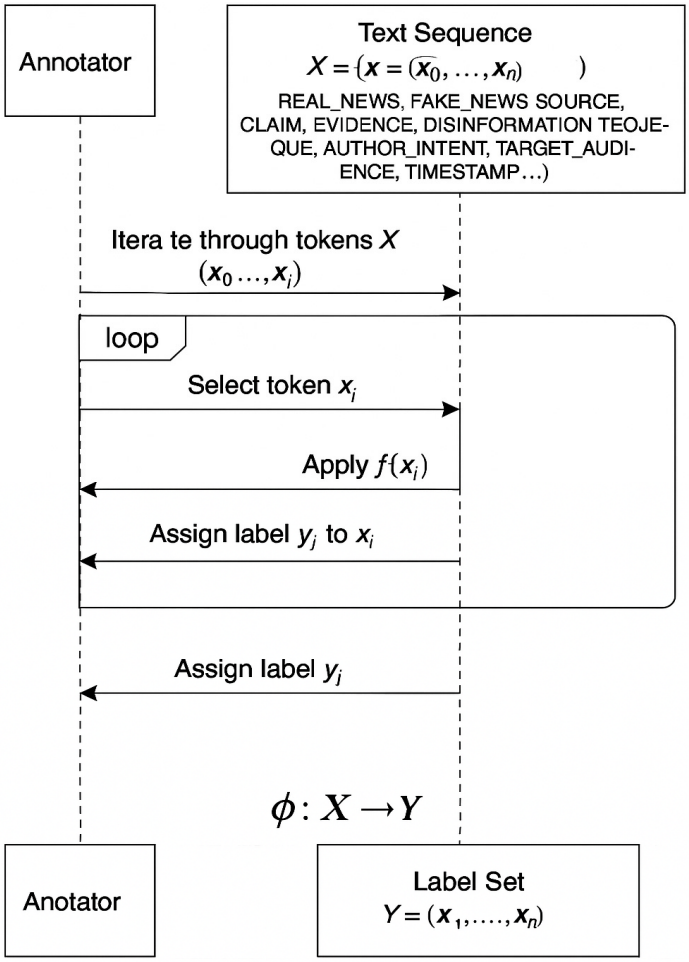



**Figure 3.** UML sequence diagram illustrating the token annotation process using a multi-level annotation schema.

The process begins with the annotator (either a human expert or an automated model) iterating through the input text sequence in Equation (1):

$$X = \{x_0, x_1, \ldots, x_n\}, \quad (1)$$

where each token $x_i$ represents a word, punctuation mark, or other textual element.

The practical significance of introducing Equation (1) lies in the representation of text as a sequence of tokens, which enables annotation at the word and phrase level, i.e., an essential requirement for implementing a multi-level scheme with precise identification of entities and subtypes in Label Studio. This approach ensures compatibility with data formats used in modern NLP frameworks (e.g., SpaCy v3.7.2 and HuggingFace v4.39.3) and facilitates downstream model training for automatic classification, as each annotation is linked to a specific token or token sequence within the text.

The set of labels Y is defined as the union of all categories and subtypes in the annotation scheme, as illustrated in Equation (2):

$$Y = \{\text{REAL_NEWS, FAKE_NEWS, SOURCE, CLAIM, EVIDENCE, DISINFORMATION_TECHNIQUE, AUTHOR_INTENT, TARGET_AUDIENCE, TIMESTAMP} \ldots\}, \quad (2)$$

where each category may be further specified by a subtype (e.g., DISINFORMATION_TECHNIQUE: clickbait or AUTHOR_INTENT: spread_fear).

Formally, the task is to construct a function:

$$\varphi: X \rightarrow Y, \tag{3}$$

as demonstrated in Equation (3), the process of mapping involves the allocation of each token (or text fragment) to its corresponding label from the annotation scheme.

In practical terms, the task reduces to constructing a classifier that, based on a training corpus of annotated texts, predicts both categorical labels and subtypes for previously unseen texts. Additionally, annotation consistency is evaluated typically using Cohen's Kappa coefficient to ensure the reproducibility of the annotation scheme and the validity of the assigned labels.

To address the task of manual fake news identification using the proposed multi-level annotation framework, we propose a method grounded in modern approaches to Natural Language Processing (NLP) and Machine Learning (ML). Let the training dataset be defined as follows, in Equation (4):

$$D = \{(X^1, Y^1), (X^2, Y^2), \ldots, (Xm, Ym)\}, \tag{4}$$

where X is a sequence of tokens and Y is the corresponding set of labels from the annotation scheme. The objective is to construct a model f: $X \rightarrow Y$ that approximates the function $\varphi$ and is capable of accurately predicting labels for new, unseen texts.

The news corpus was intentionally compiled to create a representative sample of Kazakh- and Russian-language publications, covering a wide range of sources, genres, and styles. It includes materials from official online media outlets, regional publications, Telegram channels, social media platforms, and personal blogs. Both reliable and fake news items were selected, as well as hybrid texts containing a mixture of truthful and misleading content.

The balanced distribution of the corpus across languages, comprising 50% Russian-language and 50% Kazakh-language texts, was a deliberate choice aimed at ensuring linguistic balance within the model and preventing potential bias toward one language. This consideration is particularly important in the context of Kazakhstan, a constitutionally bilingual country. Such an approach enabled comparable processing quality across both languages, ensured the validity of cross-linguistic analysis, and provided equal opportunities for subsequent model training and evaluation. Moreover, during the corpus construction phase, we encountered a limited amount of annotated data for each language group, which necessitated a symmetrical distribution to achieve stable and reproducible experimental results (see https://github.com/baiangali/fake_news (accessed on 19 August 2025)).

Annotation was carried out using the Label Studio platform. Annotators were provided with detailed annotation guidelines that outlined clear criteria for identifying entities, claim types, sources, disinformation techniques, and other components. In cases of disagreement, final decisions were made by a moderator to ensure consistency.

The annotation quality was further evaluated using Cohen's Kappa, a widely accepted metric for measuring inter-annotator agreement in sequential and categorical labeling tasks.

The proposed annotation scheme is grounded in the principles of discourse analysis and content analysis, incorporating both core categories and a hierarchical subtype structure, enabling a deeper semantic interpretation of text. The primary annotation labels and their subtypes are presented below (see Tables 2–6).

**Table 2.** Core Annotation Categories for Fake News Detection.

| Category | Description |
|---|---|
| FAKE_NEWS | Texts containing deliberately false information |
| REAL_NEWS | Verified factual news |
| SOURCE | Mention of a source (e.g., author, media outlet, and hyperlink) |
| CLAIM | Main claim or key piece of information |
| EVIDENCE | Evidence supporting truth or falsehood |
| DISINFORMATION_TECHNIQUE | Disinformation dissemination technique (e.g., emotional pressure and clickbait) |
| AUTHOR_INTENT | Author's intent behind the disinformation |
| TARGET_AUDIENCE | Target audience |
| TIMESTAMP | Timestamp of publication |

**Table 3.** Subtypes of FAKE_NEWS Category.

| Subtype | Description |
|---|---|
| conspiracy_theory | Hypothetical claims about secret forces influencing events that are not supported by evidence. |
| fabricated_news | Deliberately created false information, devoid of factual basis. |
| satire_or_parody | Ironic or satirical material, often mistakenly perceived as real news. |
| misleading_content | Deliberate distortion or omission of facts, causing erroneous conclusions in the audience. |
| manipulated_content | Genuine content that has been modified to distort its original meaning. |
| false_context | Real content presented in a distorted or irrelevant context. |

**Table 4.** Subtypes of DISINFORMATION_TECHNIQUE Category.

| Subtype | Description |
|---|---|
| clickbait | Sensational headlines or images that attract attention but distort the essence of the material. |
| emotional_pressure | The use of emotional techniques to influence the perception and behavior of the audience. |
| exaggeration | Deliberate overstatement of the scale or significance of events or facts. |
| downplaying | Deliberate understatement of the seriousness of events or facts. |
| scapegoating | An unfounded accusation in order to shift attention from the true causes or the culprits. |
| misattribution | Incorrect indication of information sources in order to create the illusion of credibility and reliability. |

**Table 5.** Subtypes of AUTHOR_INTENT Category.

| Subtype | Description |
|---|---|
| influence_opinion | Create and distribute content in order to influence the views and attitudes of the audience. |
| spread_fear | Intentionally creating content that causes anxiety, fear, or panic among readers. |
| political_gain | Creation and dissemination of information in order to gain an advantage in a political struggle or to form public opinion in the political interests of the author. |
| attract_attention | Content designed to attract maximum audience attention, often through the use of sensationalism and high-profile headlines. |
| financial_gain | Creating and distributing content with the primary goal of making a profit or economic advantage. |

**Table 6.** Subtypes of TARGET_AUDIENCE Category.

| Subtype | Description |
|---|---|
| elderly | Content aimed at the older generation, taking into account their social, economic, or psychological characteristics. |
| youth | Content aimed at young people and teenagers, often exploiting their vulnerability to certain types of information. |
| mothers | Content aimed at women with children, using themes of motherhood and family for emotional impact. |
| rural_population | Information specially prepared for an audience living in rural areas, taking into account their specific interests or needs. |
| patriots | Content specifically targeted at an audience with a pronounced patriotic attitude, using relevant emotional and value aspects. |
| ethnic_groups | Information specifically targeted at certain ethnic communities, exploiting their cultural, social, or economic characteristics to influence perception and behavior. |

To assess the reliability of the annotation scheme, inter-annotator agreement was calculated. This step confirmed the reproducibility of results and the validity of the label structure. Despite the successful implementation of the scheme, several limitations were identified, including the complexity of interpreting subjective labels (e.g., AUTHOR_INTENT) when context is limited; the relatively small volume of Kazakh-language content, due to the limited accessibility of sources; and the high cognitive load on annotators when working with a multi-level structure without automation support.

The absence of model testing on the annotated corpus at this stage. The limited volume of Kazakh-language content affects annotation outcomes by reducing the representation of certain thematic and genre-specific patterns unique to the Kazakh-language segment of the media landscape. This limitation may lead to imbalances in the distribution of categories and subtypes, as well as to lower recall in detecting specific disinformation techniques characteristic of local sources. From the perspective of generalizability, this constraint implies that the results obtained from the current dataset should be interpreted with caution when extrapolating to the broader spectrum of Kazakh-language content. Consequently, there is a clear need for further corpus expansion, including the integration of sources from regional and local media outlets.

Nevertheless, the results demonstrate the high applicability of the scheme and its potential for scaling within broader NLP tasks. The annotated data can serve as a foundation for training automated disinformation detection models, including in multilingual environments.

In developing the custom annotation scheme for fake news in Kazakh and Russian, we analyzed a range of authoritative academic sources and existing datasets containing established categories and labels. The resulting multi-level structure is grounded in both theoretical and empirical foundations, as proposed in prior research.

The binary labels FAKE_NEWS and REAL_NEWS were adapted from the LIAR dataset by Wang [9], in which political statements are annotated on a credibility scale (true, half-true, pants-on-fire, etc.).

The CLAIM label (main assertion) is widely used in corpora, such as NELA-GT, and is suggested by Gravanis et al. [26] and is formally included in the ClaimReview specification. It identifies the central statement subject to verification.

The SOURCE label was adopted from FakeNewsNet by Li et al. [27] and NELA-GT, where sources are annotated with attributes like credibility, reputation, and political orientation [28].

The EVIDENCE label, indicating support for or refutation of the claim, is used in the MuMiN corpus by Nielsen et al. [22] and in COVID-related datasets such as CoAID and ReCOVery [23].

DISINFORMATION_TECHNIQUE captures methods of dissemination (e.g., clickbait, context distortion, emotional pressure) and was developed based on contemporary research on disinformation tactics, particularly in the study by Broniatowski et al. [29] on information disorder.

AUTHOR_INTENT, which reflects the author's presumed motivation (e.g., destabilization, satire, and propaganda), draws inspiration from recent studies on strategic information operations and coordinated disinformation campaigns, such as the work by Starbird et al. [30], who analyzed intent through linguistic pattern analysis.

TARGET_AUDIENCE was derived from Bovet & Makse [31], whose work examined disinformation's impact on different social and political groups. This label was included to capture audience targeting, which is especially relevant in the Kazakhstani media landscape.

The TIMESTAMP label was included based on research in temporal annotation for event corpora and fact-checking. The idea of linking claims to specific timeframes comes from ClaimReview and Jiang et al. [32], which focuses on methods for claim detection and verification based on supporting evidence. Additionally, temporal normalization practices were adopted from the TimeML annotation framework by Jha et al. [33], widely used in corpora like TimeBank, EventTimeBank, and MEANTIME [24] for temporal event extraction in NLP.

All labels used in our annotation scheme are based on validated methodologies and have been adapted into a multi-level framework for analyzing fake news, taking into account linguistic and cultural specificities.

Figure 4 illustrates a manual annotation of a Russian-language text in Label Studio (see Figure 4). The text highlights key components such as TIME (temporal context), CLAIM (main assertion), SOURCE (information source), and EVIDENCE (supporting or refuting evidence). The annotation is further enriched with the type of disinformation (fabricated_news), the dissemination technique (clickbait), the author's intent (influence_opinion), and the target audience (general_public).




**Figure 4.** Example of manual annotation in Russian with all entities from the proposed scheme (the highlighted Russian text means "news article excerpt").

Figure 5 presents a similar annotation of a text in Kazakh. Key components such as TIME, CLAIM, and SOURCE are also annotated, and the message is classified as misleading_content with a political motivation (political_gain) targeting youth as the primary audience (see Figure 5).




**Figure 5.** Example of manual annotation in Kazakh with all entities from the proposed scheme (the highlighted Russian text means "news article excerpt").

This study employs a comprehensive methodological framework that combines both theoretical analysis and practical implementation. The primary approach used is the comparative-analytical method, which enabled a systematic examination of existing annotation schemes such as LIAR, FakeNewsNet, MuMiN, NewsPolygraph, and others. This approach allowed us to identify their strengths and weaknesses, particularly in the context of multilingual data processing.

An important tool was content analysis, which was used to study the structure of fake news in Kazakh and Russian and to identify common linguistic and thematic features of disinformation.

The data classification method was applied in constructing the multi-level annotation scheme, allowing for information to be organized by layers: from basic categorization (e.g., REAL_NEWS) to detailed subtypes describing manipulative techniques, authorial intent, and target audience.

In addition, the study employed a practical annotation method using the Label Studio platform, which enabled the implementation of the scheme as a working annotation tool and supported the creation of an annotated text corpus.

The comparative analysis method was also used to benchmark the proposed scheme against existing solutions, demonstrating its advantages in terms of granularity, flexibility, and localization for Kazakh and Russian languages.

The main outcomes of this study include the development and implementation of a multi-level annotation scheme that captures not only the presence of fake content in a text, but also reveals its underlying nature, dissemination strategies, and targeting mechanisms.

Analysis of international and local datasets revealed that most existing annotation schemes are primarily English-centric and offer a limited set of labels, which makes them less suitable for comprehensive disinformation analysis in Kazakh–Russian bilingual contexts.

To assess the consistency and reproducibility of the annotation framework, we compared the percentage distribution of annotated entities between two independent experts and calculated Cohen's Kappa coefficients for each major category. Table 7 presents the results. High Kappa values for SOURCE (0.81) and CLAIM (0.78) confirm the formal definability and stability of these labels in news texts, making them reliable anchors for automating annotation processes. Labels with lower agreement values, such as AUTHOR_INTENT and FAKE_NEWS, reflect a higher interpretative burden, often requiring additional contextual and pragmatic analysis.

**Table 7.** Percentage Distribution of Annotated Entities and Inter-Annotator Agreement (Cohen's Kappa) by Category.

| Category | Annotator 1 (%) | Annotator 2 (%) | Cohen's Kappa |
|---|---|---|---|
| FAKE_NEWS | 6.5% | 6.3% | 0.72 |
| REAL_NEWS | 7.0% | 6.9% | 0.74 |
| SOURCE | 25.3% | 25.2% | 0.81 |
| CLAIM | 28.5% | 28.4% | 0.78 |
| EVIDENCE | 8.0% | 8.1% | 0.76 |
| DISINFORMATION_TECHNIQUE | 6.5% | 6.6% | 0.74 |
| AUTHOR_INTENT | 6.2% | 6.1% | 0.73 |
| TARGET_AUDIENCE | 6.0% | 6.1% | 0.72 |
| TIMESTAMP | 5.0% | 5.2% | 0.75 |

Figure 6 illustrates the inter-annotator agreement across core annotation categories using Cohen's Kappa scores (see Figure 6). The highest levels of agreement are observed for SOURCE and CLAIM, indicating their clarity and consistent interpretability. In contrast, categories such as FAKE_NEWS and AUTHOR_INTENT show moderate agreement, reflecting the increased subjectivity and contextual dependence involved in labeling them.

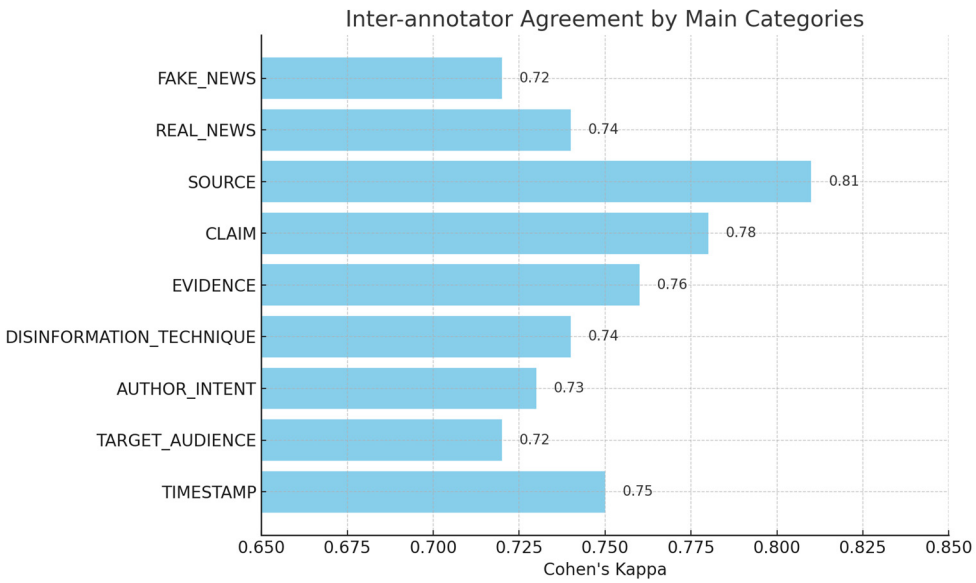



**Figure 6.** Inter-Annotator Agreement by Core Annotation Categories (Cohen's Kappa) Note. Values above 0.80 indicate almost perfect agreement, while values between 0.61 and 0.80 indicate substantial agreement, according to the classification by Landis & Koch.

A progressive improvement in accuracy, recall, and F1-score is expected at each stage (see Figure 7). These projections are based on preliminary baseline experiments using logistic regression on the top-level labels (REAL_NEWS, FAKE_NEWS) within the constructed corpus, where accuracy and F1 scores ranged between 72% and 75%. They are

further supported by findings in the literature, where fine-tuning BERT and XLM-R models on similar tasks has led to performance improvements of 8–15% compared to simpler models [6,8,9]. Additionally, analysis of transformer-based approaches in multilingual fake news detection shows that incorporating semantic relationship layers and transitioning from binary to multi-level classification enhances recall and F1 scores by enabling better differentiation of disinformation patterns. Therefore, the projected improvement presented in Figure 5 is grounded in a synthesis of preliminary experimental outcomes, empirical findings from the literature, and the structural advantages of the proposed multi-level annotation scheme.

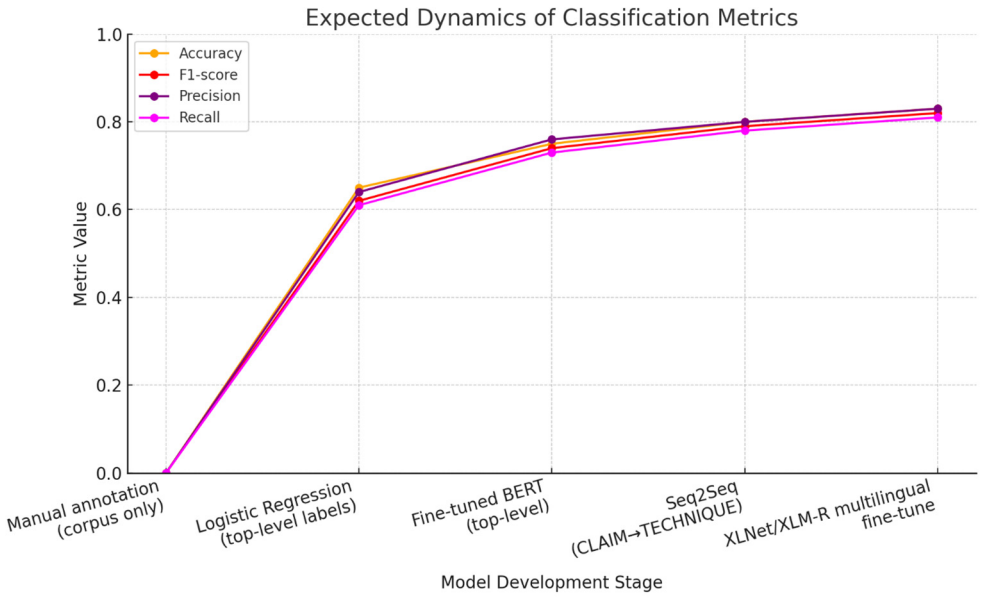



**Figure 7.** Expected Dynamics of Classification Metrics.

In addition to category and subtype levels, the proposed annotation scheme also incorporates formalized semantic relations between entities. The introduction of semantic relationships between annotated entities (e.g., CLAIM → EVIDENCE, CLAIM → DISINFORMATION_TECHNIQUE) is aimed not only at structuring the text but also at improving the quality of downstream fake news detection by enhancing contextual coherence and formalizing latent patterns. Traditional approaches often rely on isolated entities or features, without accounting for their interdependencies. This results in the loss of critical information, particularly in cases involving complex or covert forms of disinformation. Semantic links enable models to capture contextual dependencies between key components of a message. For instance, a CLAIM supported by weak or missing EVIDENCE and accompanied by a manipulative DISINFORMATION_TECHNIQUE is highly likely to be classified as fake. Such connections add an additional layer of formalized context, increasing the interpretability of the data and providing more precise rules for model training. As a result, semantic relationships contribute to greater detection accuracy by incorporating not only the content of the message but also the structural and relational dynamics between its elements (see Figure 8).

These relations enable linking of claims to sources, evidence, timestamps, dissemination techniques, and authorial intent. Table 8 below presents the complete list of such semantic relationships, which are essential for constructing logical-semantic graphs of disinformation (see Table 8). These relations include both standard links, such as a claim supported by evidence and more specific ones that capture authorial motivation and manipulative strategies.

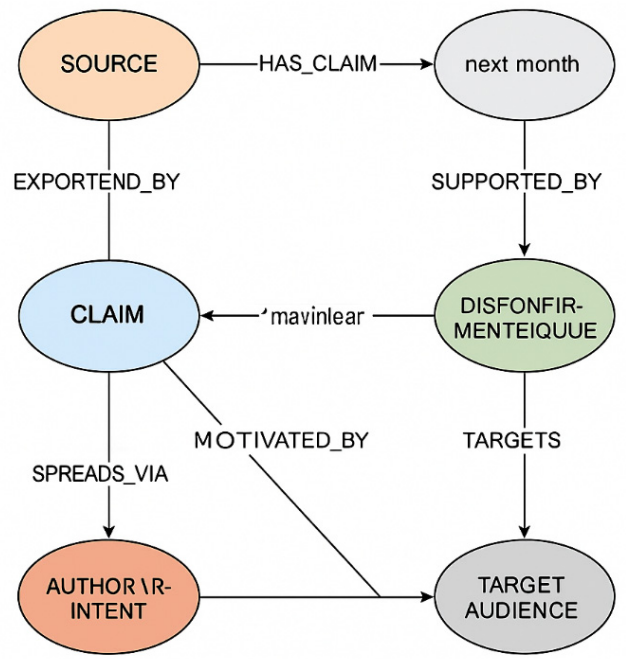



**Figure 8.** Semantic-Pragmatic Model of Entity Interaction in a Disinformation Message.

**Table 8.** Semantic Relations Between Annotation Entities.

| Relation | Description | Entity 1 | Entity 2 |
|---|---|---|---|
| HAS_CLAIM | The source makes a claim | SOURCE | CLAIM |
| SUPPORTED_BY | The claim is supported by evidence | CLAIM | EVIDENCE |
| SPREADS_VIA | A technique is used to disseminate the claim or disinformation | CLAIM | DISINFORMATION_TECHNIQUE |
| MOTIVATED_BY | The claim or disinformation is driven by the author's intent | CLAIM/FAKE_NEWS | AUTHOR_INTENT |
| TARGETS | Disinformation targets a specific audience | FAKE_NEWS/CLAIM | TARGET_AUDIENCE |
| REFERS_TO_SOURCE | The claim references an external source | CLAIM | SOURCE |
| HAS_TIMESPAN | The claim or news item is tied to a temporal context | CLAIM/FAKE_NEWS | TIMESTAMP |
| CONTRADICTED_BY | The claim is contradicted by the provided evidence | CLAIM | EVIDENCE |
| CATEGORIZED_AS | The fake news item is categorized by subtype (e.g., conspiracy, satire) | FAKE_NEWS | fake_subtype (Choices) |
| INFLUENCES | The author's intent influences the disinformation technique | AUTHOR_INTENT | DISINFORMATION_TECHNIQUE |

The semantic relations reflect the structural connections between entities within the annotation framework. They form the foundation for constructing logical-semantic graphs and enable the automated analysis of disinformation narratives.

An example of an annotated sentence in Kazakh, with full coverage of all entities and their interrelations, is presented in Figure 9.

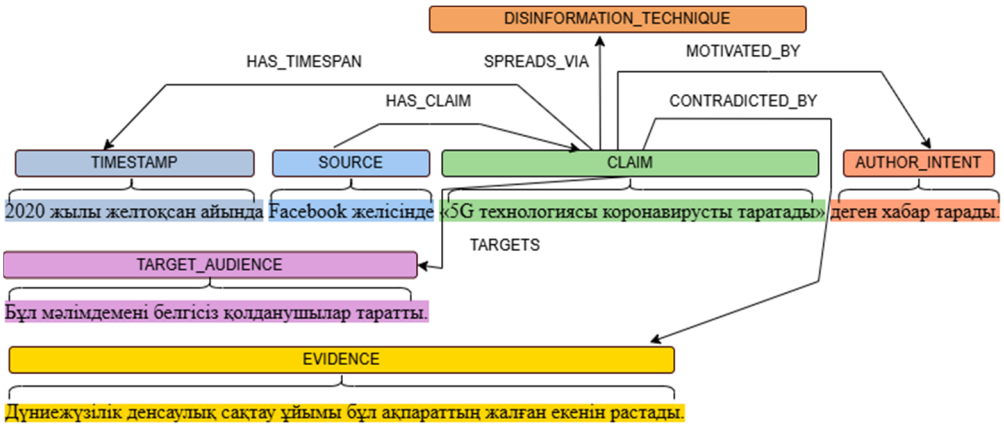



**Figure 9.** Visualization of Semantic Relations Between Entities in Kazakh Based on the Multi-Level Annotation Scheme (the highlighted Russian text means "news article excerpt").

## 3. Discussion

The proposed multi-level annotation scheme for disinformation detection represents an effort to move beyond a simple binary classification approach (fake/not fake) toward a structured, component-based representation of news texts (see Figure 10).

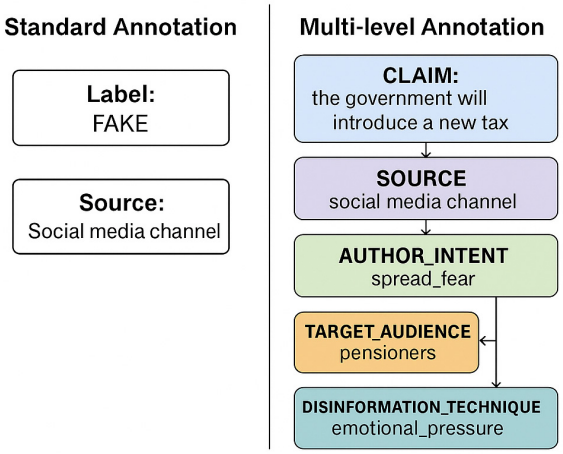



**Figure 10.** Comparison of standard annotation (binary labeling) with the proposed multi-level annotation approach for disinformation detection.

Unlike existing models such as LIAR by Wang [9] or FakeNewsNet by Shu et al. [23], which primarily focus on statement-level classification, our framework enables the identification and formalization of a broader range of linguistic and pragmatic features associated with disinformation.

A comparative analysis with other annotation frameworks—including ClaimReview, NELA-GT, and FEVER—shows that the proposed scheme integrates best practices from these projects while adapting them to the Kazakh–Russian media environment. Special emphasis is placed on elements that have been underexplored in other languages, such as the explicit annotation of the target audience (TARGET_AUDIENCE) and the author's motivation (AUTHOR_INTENT), which are crucial in contexts involving politically or socially charged content.

The high inter-annotator agreement scores (Cohen's Kappa ranging from 0.72 to 0.81 for key labels) indicate the conceptual clarity of the scheme and its suitability for large-scale annotation efforts. Moreover, the visualization of annotated examples and semantic relation diagrams highlights the framework's potential for automation, including

graph-based representations and applications in machine learning and NLP, particularly for fine-tuning transformer models.

Notably, the inclusion of Kazakh, a low-resource language in corpus linguistics and fact-checking, underscores the importance of flexible, linguistically and culturally sensitive annotation strategies. This reinforces the relevance of the research not only from an academic standpoint but also in applied domains such as media literacy and digital hygiene.

The annotation outcomes and preliminary analysis emphasize the need for a multi-layered description of fake news, capable of supporting more accurate modeling of the creation and dissemination mechanisms of disinformation in multilingual contexts.

## 4. Conclusions

The scientific contribution of this study lies in the development of the first multi-level annotation scheme for disinformation detection in the Kazakh–Russian linguistic context, the creation of a corresponding annotated corpus, and the formalization of semantic relationships between key entities. From an applied perspective, the results of this work can be used to support the automation of fake news detection tasks and the development of intelligent systems for monitoring digital media. Future work will include the training and evaluation of automated models, the development of semi-automated annotation tools, and the integration of the proposed scheme into real-world analytical platforms. The proposed scheme enables the structured representation of key components of fake news, including claims (CLAIM), sources (SOURCE), evidence (EVIDENCE), disinformation techniques (DISINFORMATION_TECHNIQUE), authorial intent (AUTHOR_INTENT), target audience (TARGET_AUDIENCE), and temporal attributes (TIMESTAMP).

One of the key findings is the effectiveness of the multi-level approach compared to traditional binary classification. High inter-annotator agreement scores (Cohen's Kappa up to 0.81) demonstrate the clarity of the guidelines and the reproducibility of the scheme. An additional advantage lies in the ability to formalize logical-semantic relationships between entities, paving the way for constructing graph-based representations of disinformation narratives.

The annotation results reveal consistent disinformation patterns, primarily targeting vulnerable audiences. Techniques used range from clickbait and exaggeration to fear-based manipulation and politically motivated content. Nevertheless, some limitations were identified, such as the interpretive complexity of subjective labels, limited availability of Kazakh-language content, and the absence of automated baseline models at the time of the study.

The observed differences in inter-annotator agreement across categories highlight the need for a differentiated approach to automation. Categories with high formal definability (e.g., CLAIM, SOURCE) can be effectively automated using standard sequential classification models. In contrast, more subjective categories (e.g., AUTHOR_INTENT, FAKE_NEWS) require the incorporation of contextual and discourse-level features, as well as additional annotation guidelines to enhance reproducibility and annotation consistency.

Future research should focus on the following areas:

- Training and testing multilingual models using the annotated corpus;
- Expanding the dataset with additional topics and sources;
- Developing semi-automated annotation tools;
- Investigating the impact of disinformation on different demographic groups.

Taken together, these findings confirm that the multi-level annotation scheme holds strong potential for both academic and applied domains. It can serve as a foundation for building more accurate and interpretable models for automated fake news detection, particularly in linguistically and culturally diverse media ecosystems.

Further research will focus on training and evaluating automated fake news detection models based on the proposed annotation scheme, as well as integrating these models into applied systems for monitoring and analyzing digital media.

In the longer term, prospective studies will explore the potential of using Large Language Models (LLMs) to automate or semi-automate the annotation process, including the preliminary generation of labels and the extraction of semantic relationships between entities. The application of such models could significantly enhance the speed and scalability of annotation. However, it also necessitates rigorous quality control and systematic validation, given the potential for automated errors, especially in socially sensitive tasks such as disinformation detection.

**Author Contributions:** Conceptualization, A.N. and M.S.; methodology, M.S. and A.N.; software, A.B. and D.T.; validation, M.K., M.S. and A.N.; formal analysis, N.S. and D.T.; investigation, M.S.; resources, D.T. and M.K.; data curation, A.B. and N.S.; writing—original draft preparation, M.S.; writing—review and editing, A.N.; visualization, A.N. and N.S.; supervision, A.N. and M.S.; project administration, A.Y.; funding acquisition, A.Y. and N.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant AP23484329.

**Institutional Review Board Statement:** This study did not require ethical approval.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Links and information about the data source can be found in Section 2 of this article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Shu, K.; Bhattacharjee, A.; Alatawi, F.; Nazer, T.; Ding, K.; Karami, M.; Liu, H. Combating disinformation in a social media age. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1385. [CrossRef]
2. Saquete, E.; Tomás, D.; Moreda, P.; Martínez-Barco, P.; Palomar, M. Fighting post-truth using natural language processing: A review and open challenges. *Expert Syst. Appl.* **2020**, *141*, 112943. [CrossRef]
3. Singh, M.K.; Ahmed, J.; Alam, M.A.; Tripathi, R. A comprehensive review on automatic detection of fake news on social media. *Multimed. Tools Appl.* **2024**, *83*, 47319–47352. [CrossRef]
4. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 1–40. [CrossRef]
5. Kowsari, J.; Meimandi, K.J.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
6. Ragab, M.I.; Mohamed, E.H.; Medhat, W. Multilingual Propaganda Detection: Exploring Transformer-Based Models mBERT, XLM-RoBERTa, and mT5. *Nat. Lang. Eng.* **2025**, *31*, 245–262.
7. Wang, X.; Zhang, W.; Rajtmajer, S. Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey. *Inf. Process. Manag.* **2025**, *62*, 102789. [CrossRef]
8. Dutta, A.; Nooralahzadeh, B.; Ribeiro, R.P.; Tan, A.H. Entity-Aware Cross-Lingual Claim Detection for Automated Fact-Checking. *Comput. Linguist.* **2025**, *51*, 123–145. [CrossRef]
9. Wang, W.Y. Liar, liar pants on fire: A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648. [CrossRef]
10. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [CrossRef][PubMed]
11. Faustini, P.H.A.; Covões, T.F. Fake news detection in multiple platforms and languages. *Expert Syst. Appl.* **2020**, *158*, 113503. [CrossRef]
12. Silva, R.M.; Santos, R.L.; Almeida, T.A.; Pardo, T.A. Towards automatically filtering fake news in Portuguese. *Expert Syst. Appl.* **2020**, *146*, 113199. [CrossRef]
13. Khalil, A.; Jarrah, M.; Aldwairi, M.; Jaradat, M. AFND: Arabic fake news dataset for the detection and classification of articles credibility. *Data Brief* **2022**, *42*, 108141. [CrossRef]

14. Assaf, R.; Saheb, M. Dataset for Arabic fake news. In Proceedings of the 2021 IEEE 15th International Conference on Appli-cation of Information and Communication Technologies (AICT), Virtual, 13–15 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4. [CrossRef]
15. Gruppi, M.; Horne, B.D.; Adali, S. An exploration of unreliable news classification in Brazil and the US. *arXiv* **2018**, arXiv:1806.02875. [CrossRef]
16. Norambuena, B.; Horning, M.; Mitra, T. Evaluating the inverted pyramid structure through automatic 5W1H extraction and summarization. In Proceedings of the Computational Journalism Symposium, Boston, MA, USA, 20 March 2020.
17. Chakma, K.; Das, A. A 5W1H based annotation scheme for semantic role labeling of English tweets. *Comput. Y Sist.* **2018**, *22*, 747–755. [CrossRef]
18. Zhang, A.X.; Ranganathan, A.; Metz, S.E.; Appling, S.; Sehat, C.M.; Gilmore, N. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In Proceedings of the Companion of the Web Conference, Virtual, 23 April 2018; pp. 603–612.
19. Santia, M.; Williams, E.; Adalı, S.; Horne, B.D. Characterizing partisan political news and misinformation on Twitter: A textual analysis. *Soc. Netw. Anal. Min.* **2021**, *11*, 53.
20. Alam, F.; Rooyen, G.; Augenstein, I. Fighting the COVID-19 infodemic in social media: A study of multilingual misinformation networks. *J. Comput. Soc. Sci.* **2021**, *4*, 1057–1083. [CrossRef]
21. Mottola, S. Las fake news como fenómeno social. Análisis lingüístico y poder persuasivo de bulos en italiano y español. *Discurso Soc.* **2020**, *3*, 683–706. [CrossRef]
22. Nielsen, D.S.; McConville, R. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. *arXiv* **2022**, arXiv:2202.11684.
23. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. FakeNewsNet: A data repository with news content, social context, and dynamic information for studying fake news on social media. *arXiv* **2018**, arXiv:1809.01286. [CrossRef]
24. Alam, A.; Rooyen, G.; Chakravarthi, S.; Augenstein, I. Multiverse: A multilingual perspective on disinformation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland, 22–27 May 2022; pp. 12345–12358.
25. Dzienisiewicz, D.; Graliński, F.; Jabłoński, P.; Kubis, M.; Skórzewski, P.; Wierzchoń, P. POLygraph: Polish Fake News Dataset. In Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA 2024), Bangkok, Thailand, 15 August 2024; pp. 250–263.
26. Gravanis, G.; Vakali, A.; Diamantaras, K.; Karadais, P. Behind the media curtain: A comprehensive overview of fake news detection on social networks. *Online Soc. Netw. Media* **2019**, *18*, 100080. [CrossRef]
27. Li, Q.; Zhang, Q.; Si, L.; Liu, Y. Rumor detection on social media: Datasets, methods and opportunities. In Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, 17 November 2019; pp. 1–9. [CrossRef]
28. Shen, X.; Huang, M.; Hu, Z.; Cai, S.; Zhou, T. Multimodal fake news detection with contrastive learning and optimal transport. *Front. Comput. Sci.* **2024**, *6*, 1473457. [CrossRef]
29. Broniatowski, D.A.; Jamison, S.; Qi, M.; AlKulaib, L.; Chen, A.; Benton, P.; Marathe, M.; Dredze, F.; Quinn, S.C. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *Am. J. Public Health* **2018**, *108*, 1378–1384. [CrossRef] [PubMed]
30. Starbird, K.; Arif, A.; Wilson, T. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. In Proceedings of the ACM on Human-Computer Interaction, Donostia Gipuzkoa, Spain, 7 November 2019; Volume 4, pp. 1–26. [CrossRef]
31. Bovet, A.; Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef]
32. Jiang, Y.; Shaar, A.; Zhang, X.; Hasanain, M.; Da San Martino, J.; Nakov, P. Claim detection and verification with evidence: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–43.
33. Jha, R.; Qazvinian, L.; Singh, M.; Roth, D. Temporal information in misinformation and fake news: A survey. *Inf. Process. Manag.* **2021**, *58*, 102712.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.