

Narratives at Conflict: Computational Analysis of News Framing in Multilingual Disinformation Campaigns

Antonina Sinelnik and Dirk Hovy

Bocconi University, Via Sarfatti 25, 20136 Milan, Italy
antonina.sinelnik@studbocconi.it, dirk.hovy@unibocconi.it

Abstract

Any report frames issues to favor a particular interpretation by highlighting or excluding certain aspects of a story. Despite the widespread use of framing in disinformation, framing properties and detection methods remain underexplored outside the English-speaking world. We explore how multilingual framing of the same issue differs systematically. We use eight years of Russia-backed disinformation campaigns, spanning 8k news articles in 4 languages targeting 15 countries. We find that disinformation campaigns consistently and intentionally favor specific framing, depending on the target language of the audience. We further discover how Russian-language articles consistently highlight selected frames depending on the region of the media coverage. We find that the two most prominent models for automatic frame analysis underperform and show high disagreement, highlighting the need for further research.

1 Introduction

Framing is a phenomenon grounded in political and social sciences, which specifies *how* specific topics are presented by the media. It can manifest in loaded vocabularies, like *the war on terror*, or broader phrases with implicit assumptions. Framing has long been studied as an instrument for creating a specific political image or favoring a particular point of view. While it is natural for any non-trivial argument to be framed by the presenter, its intentional (mis)use can create persistent associations and sway opinions on political issues. Many works explore framing as an instrument of propaganda and misinformation spread (Rozenas and Stukal, 2019; Munger et al., 2019; King et al., 2017). Combined with the increased velocity of disinformation in today’s media landscape, it highlights an acute need for a detection tool of persistent framing patterns.

However, while Natural Language Processing (NLP) is the most logical place for this tool, most advances in frame identification are based on English-speaking environments, in particular in the political context of the US (Tsur et al., 2015; Card et al., 2016). No single method has established itself as the state-of-the-art for multilingual data. The few existing methods vary in the best model choice and present conflicting views on the role of the target, non-English language.

However, especially in international contexts (and conflicts), (national) language (and relatedly the political position of the presenter) plays an important role in framing. Russian media present a prominent example of intentional media manipulation through framing and disinformation spread. Several studies have already examined the framing of narratives directed inside the country (Field et al., 2018; Park et al., 2022). We compare the domestic messaging to the one spread abroad and observe how the same events receive very different framing depending on the language of the target country.

Contributions: This paper contributes to the growing body of framing research in two ways. 1) We compare two prominent (English-based) frame identification approaches on a novel multilingual dataset. We establish their strengths and weaknesses, and expose the underlying assumptions. 2) by applying the best method to the newly collected data, we contribute to the body of work on framing outside of the English-speaking context. For the languages in our data, we outline the salient topics in recent disinformation campaigns.¹

¹The data and the code for reproducing the analysis will be made available at: <https://github.com/ayusinelnik/narratives-at-conflict>

2 Data

Identifying disinformation remains a matter of expert opinion and careful manual annotations, which makes it a scarce resource outside of the English-speaking world. Faced with the span and size limitations of labeled datasets on disinformation in Russian (Kuzmin et al., 2020), we decided to assemble a corpus of disinformation articles guided by the expert opinion on the subject. EUvsDisinfo emerged as one such source, part of the EU’s diplomatic service led by the EU’s High Representative, which publishes weekly reports on news articles containing pro-Kremlin disinformation. The database includes articles in 15 languages from various news outlets, and more than 15k articles have been reviewed since 2015 to date. Even though EUvsDisinfo does not assume partial or complete ownership of the media outlets by the State, it is stated that the source articles contain “*partial, distorted, or false depiction of reality and spread key pro-Kremlin messages.*” The EUvsDisinfo reporting is organized by a disinformation narrative, where a specific event or topic is at the center of the report, supported by links to source articles that reiterate the misinforming narrative. For the target corpus, we crawled all the source articles in Russian, French, Spanish, and Italian for the reporting period from 06/01/2015 to 23/05/2023. We removed any short-form pieces, articles originating from social media platforms, and any news pieces shorter than 300 characters. Table 1 shows the resulting number of articles for each language. Multilingual articles paired into the same report by EUvsDisinfo fall under *paired* category. We used subsets of paired articles for annotation tasks and hyper-parameter tuning. From the other, *unpaired* articles that were mentioned in different EUvsDisinfo reports but are closely related, we construct multilingual pairs with an approach described in the next section.

2.1 Generating Article Pairs

To construct multilingual article pairs about the same event, we produce keywords in the target language of the article, embed them in a shared space, and measure the distance. YAKE! (Campos et al., 2020) keyword algorithm was chosen for its notably high performance in a multilingual setting (Piskorski et al., 2021). As an unsupervised method, it generalized well over textual styles, domains, and, languages and provides a good fit for

Language	Paired	Unpaired	Total
Ru	200	6364	6564
Fr	105	300	405
Sp	48	566	615
It	36	440	476
Total	389	7670	8059

Table 1: Total Article Count in the Target Corpus; *Paired* are articles joined into one report by EUvsDisinfo. *Unpaired* are closely related articles from disconnected reports which we build into pairs by event

a heterogeneous collection of texts like ours. To measure the distance between keyword sets in different languages, we embedded them with MUSE (Lample et al., 2017), a state-of-the-art approach for synonym selection and contextual word similarities that aligns the embeddings in a shared space. We set the time window of ± 4 weeks from the date of the target article for which we searched a pair. The choice of a time lag was justified by two factors: the structure of the database, where the reports on disinformation appear within a week from the article publication, and the findings of Field et al. (2018), which prove agenda-setting in the Russian news within a month time from an adverse event. We searched the hyper-parameter space before applying the keywords algorithm (# of keywords, # of n-grams, deduplication threshold). The best hyper-parameter combination would be the one that results in the highest cosine similarity between keyword embeddings for the *paired* articles – those grouped under the same disinformation narrative by EUvsDisinfo reports.

3 Method and Modeling

3.1 Method Comparison Overview

The two models at the core of our comparison are both declared as well-fit for a multilingual frame identification task but vary in the architecture. The earlier model, introduced by Field et al. (2018) is a distantly supervised approach, based on constructing and contextualizing framing lexicons, fixed sets of words in a target language, that serve as indicators of framing. The later one, promoted by Park et al. (2022), is a supervised approach, based on a transformer model that performs a multi-label classification task. The two approaches will later be referenced as lexicon (-based) or LB, and transformer (-based) or TB, respectively.

In comparing the two methods, our goal is to control for as many aspects as possible. Both models, however, have inherent nuances in their setup and decision criteria, as described below.

Input Articles: Both models draw annotated articles from The Media Frames Corpus (MFC) (Card et al., 2015): To date, MFC remains the most extensive collection of annotated English-language news articles that serves as a benchmark for unsupervised, supervised, and distantly supervised framing identification methods (Khanehzar et al., 2019; Liu et al., 2019; Field et al., 2018). The current version of MFC covers 6 policy issues with 45k articles where 347k spans were annotated by multiple expert annotators with one of the fifteen frames defined by Boydston and Gross (2013). The lexicon method inputs all annotated material into training. The transformer method applies rigorous filtering to only accept annotations where 2+ annotators agree, which reduces the number of inputs by almost half;

Translation: while the lexicon method localizes and contextualizes the lexicon depending on the target language, the transformer method is English-first, based on the use of MFC in training;

Text Spans: The lexicon method identifies frames on a word level, while the transformer method extends the spans from MFC to the nearest complete sentence and produces sentence-level results.

3.2 Lexicon-based Frame Identification

3.2.1 Methodology

For each frame in the MFC, we form a base lexicon of 250 items with the highest pointwise mutual information score $I(w, F)$ (Church and Hanks, 1990), following Formula 1 below. The base lexicon is filtered to remove the words occurring in more than 98% or less than 0.5% of the articles.

$$I(F, w) = \log \left(\frac{P(F, w)}{P(F) \cdot P(w)} \right) = \log \frac{P(w|F)}{P(w)} \quad (1)$$

Equation 1 represents the Pointwise Mutual Information formula, where $P(w | F)$ denotes $\frac{(\text{wordfreq.intheframe})}{(\text{framewordcount})}$, and $P(w)$ is calculated as $\frac{(\text{word'sfreq.inthecorpus})}{(\text{corpuswordcount})}$.

At this point, we have generated one base lexicon of 250 English words per frame. This base lexicon is then translated into every target language of interest using Google Cloud Translation API. To make the lexicons in target languages more contextualized and less representative of the vocabulary specific to MFC, we train word embeddings on a large background corpus in the target language. This work proceeded with CC-100 (Wenzek et al., 2020), a dataset constructed with Common Crawl at its base, which is among the widely-used corpora for all of our target languages. While the original paper advocates the choice of any large background corpus, not the specific one used in their case, we will later see how this choice could affect the performance. In our case, the choice of CC-100 would enrich the lexicons with ample context and add regional variability to the vocabulary, given that our target corpus is composed of a variety of regional sources (*fr.sputniknews.africa* and *mundo.sputniknews.com* that covers the LATAM region are in the top-3 sources for French and Spanish, respectively). The Common Crawl-based dataset provides a common ground for method comparison: XLM-R, the model on which the transformer method is based, was also trained on Common Crawl.

For each language in the embedding training, we limit the number of lines to 1 Million randomly sampled from CC-100, where each line represents a paragraph of a text. With that, we attempt to balance training across our four languages, where the CC-100 subsets per language range from 5 GB to 40 GB. We train a 200-dimension Word2Vec model with a CBOW and a 5-word context window (Mikolov et al., 2013) for five epochs. Knowing the expanse and the mix of quality in the sources that make up the Common Crawl (Wenzek et al., 2020), we set the minimum word count to 5 to remove the infrequent words. As in the original approach, the vocabulary is restricted to 50k most frequent words. We compute a center for each translated lexicon in a target language by summing up the embeddings. We then search the background corpus and extract 500 nearest neighbors with a cosine similarity no lower than 0.5. As in the original method, we discard the base translated lexicon and only keep the neighbors in the final frame lexicon. From there, words contained in more than 98% and less than 0,5 % of documents are discarded.

Russian	French	Spanish	Italian
Yanukovich	Hollande	Maduro	Berlusconi
ONF	MRC	PSOE	PdL
DNR	Manitoba	Coahuila	napolitano

Table 2: Examples of Lexicon Generated for the Political Frame in Russian, Spanish, French, Italian

Where the resulting lexicon exceeds the expected 300 words, we only keep the 300 closest neighbors.

The cosine distance is the only parameter where we deviate from the original method. Where they use a more restrictive approach and select only neighbors with a cosine similarity no lower than 0.7 for the target language and 0.6 for English, we relax that rule to avoid instances where the lexicon equals 0 for some frames. With a background corpus as expansive as Common Crawl, we have to accept the limitation of sparse embeddings to benefit from a large variety of textual sources, which reflects the nature of the target corpus. Table 2 shows examples of how the lexicon contextualizes the political phenomena from MFC to our target languages. We can also note the representation of different regions. This point would be hard to achieve with a smaller dataset with a restricted media selection.

3.2.2 Evaluating the Lexicon

Since the resulting lexicon is in a target language for which we do not expect to have labeled data, we evaluate the lexicon’s performance on manually annotated examples from the target corpus’s paired articles, on which we also evaluate the transformer-based method. We conduct an intruder detection task commonly used in the domain. For each frame, we sample 5 random words from the lexicon, to which one word from another frame’s lexicon is added, with the condition that it is not present in the original frame lexicon. Two annotators, native or proficient in our target languages and familiar with the topic of framing, labeled 15 sets of 6 words per frame. We measure two metrics for their annotations on each language’s lexicon: *soft* accuracy, where either of two annotators identified the intruder, and *hard* accuracy, where both did, aggregated over 15 sets of annotations per language.

Two languages, Russian and French, underperform on the soft accuracy, showing several

non-overlapping frames with less than 60 % accuracy, a cutoff set in the original work. We hypothesize two factors that worsened the results: the high sensitivity of the approach to the background corpus choice and inter-annotator (dis)agreement. On average across frames, the two annotators performed with similar accuracy but diverged on which frames were confused for the others. Also seeing how varied the results of hard accuracies are across languages, we could confirm a certain level of disagreement between annotators. Having some degree of subjectivity in it, framing often exposes disagreements between annotators, even after they discuss the results (Boydston and Gross, 2013).

3.3 Transformer-based Frame Identification

3.3.1 Methodology

We train XLM-R (Conneau et al., 2020), identified by Park et al. (2022) as the best-performing model for the cross-lingual context. The model is trained on pre-filtered annotations from MFC: first, text spans are expanded to the nearest sentences, and second, only sentences with 2+ annotators are admitted to the training. Note that we do not perform hyperparameter search, as we replicate the findings of Park et al. (2022) to apply them in zero-shot scenarios to the target corpus. We trained the model until we reached results comparable to Park et al.’s (2022), or otherwise for 20 epochs. The performance grew gradually and reached Macro-F1 of 65.2, compared to 67.5 in the original paper, with the same model and settings. Contrary to the base approach, we do not train to predict the *Other* frame to be able to compare the results to those of the lexicon method and due to low annotator agreement on this frame. Additionally, some degree of variability in performance could be attributed to the changes in the MFC release versions since 2022.

3.3.2 Evaluating the Model

We perform a manual annotation task to test the model’s performance on the target corpus, just like we did for the lexicon evaluation. Here, we randomly sampled fifty sentences per language from the *paired* batch of articles in our target corpus and translated them into English for annotation. The labels were provided by an annotator familiar with news framing and sufficient knowledge of source languages to estimate that the translation to English was adequate. By checking the quality of the translation, we make sure that little

meaning is lost to the translation process, as the model takes input in English. As we do not train to predict the *Other* frame, sentences annotated as *Other* or *None* were discarded from the evaluation. Overall, testing the model on annotated examples achieved a result comparable to that of VoynaSlov (unlabeled corpus in the original paper for the transformer method) which returned macro F1 = 33,5 +- 0.72. Frames that fell significantly below the expected performance were *Capacity and Resources*, *Fairness and Equality*, *Legality*, *Crime and Punishment*, and *Public Sentiment*. While the low annotation count could explain some of the poor performance, the two frames where the count exceeds ten annotations were among the worst in evaluating the lexicon-based approach. *Capacity and Resources* was notably the worst-performing frame in the work of Park et al. (2022). Like in the previous evaluation of the annotations, we could attribute some degree of the performance to the annotators’ (dis)agreement and the subjective nature of framing. The confusion matrix, presented in the Appendix A provides more granular insight into the frames pairs with low heterogeneity between them. While the general performance is on par with the performance of the original method, the mixed performance of individual frames should be noted.

4 Evaluating and Comparing Models

4.1 Introduction

The methods of our interest produce two types of framing results: the dominant frame and all the frames present in the article, with their relative concentration. We thus decided to compare models based on both results. To bring common ground to the results, we truncated all texts in our target corpus to 225 words up to the end of the sentence, guided by the explicit text lengths in the MFC.

4.2 Analysis of Competence and Agreement on Dominant Frames

Both methods produce one dominant frame per article, identified by the most concentrated frame, with concentration counted in either the number of specific lexicon words (LB) or sentences (TB) with that frame, with a random tie-breaking. As seen in Table 3, the methods present only weak agreement in the primary frame decisions, supported by insignificant inter-method agreement scores measured by Krippendorff’s Alpha (Krippendorff, 2004), a standard method in such annotation-reliant domains as framing (Card et al., 2015; Akyürek

	Ru	Es	Fr	It
Raw Agreement	18.8	16.5	10.0	13.0
Krippendorff’s Alpha	13.7	12.6	10.2	10.3

Table 3: Dominant Frame Agreement; Raw Agreement denotes % of articles with the same dominant frame decision, out of all articles

Models’ Competence		
	Lexicon	Trans.
Binary	46.6	58.4
Positives	99.9	80.3
Positives with priors	98.8	66.8
Positives with filtered priors	93.8	63.5

Table 4: Models’ Competence measured with MACE (Hovy et al., 2013), with different data presentations

et al., 2020). In addition to high disagreement, both approaches present insignificantly low competence levels on that task. The competence here and in the following sections is measured with Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013) – an unsupervised method designed to estimate annotators’ trustworthiness with an item-response model at its core. With the methods diverging on the primary frame results, we decided to conduct competence estimation on all frames found by each method.

4.3 Analysis of Competence and Agreement on All Frames

To identify all frames present in a text, we take 1 sentence and 3 lexicon instances as one vote for the frame, as the original approaches specify. For each article, we test two settings: positive decisions (only counting the frames that were found) and binary decisions (1/0 for the presence/absence of the frame, 14 annotations per text, excluding the *Other* frame). These 14-frame representations reduce the randomness of tie-breaking and expose more granularity in how the methods perform. We additionally present priors to competence estimation. As we do not have any reliable estimation for frame distribution in the target corpus, we draw the probabilities from the MFC. Filtered priors only reflect the annotations with 2+ annotator agreement, whereas unfiltered priors account for frame probability over all annotations.

Two approaches present medium to high competence depending on the data presentation (Table 4). Introducing priors lowers the competence score for

both methods, even though their competence is still higher than with binary presentation. This hints at the possible difference in frame distribution across languages, which suggests that relying on English-language annotations, even though significantly more numerous, doesn't guarantee similar performance in other languages. It is especially notable in the performance drop for the transformer-based approach, which relies on English as both source and target language, and localizes the multilingual text by simple translation. The lower performance with the binary presentations is expected since neither of the approaches learns on negative examples with frames *Other* and *None* excluded. The methods performance by frame further suggests that the absence of certain high-presence and/or low-performance frames lowers the competence score in the binary presentation.

For the transformer-based approach, we can observe that the count of the most predicted frames is not reflective of the frame distribution in the training data: two of the top-5 frames with the highest count in training input (*Legality*, *Constitutionality*, *Jurisdiction* and *Crime and Punishment*) are coincidentally the frames with one of the lowest performances in the transformer-based approach. These two frames get consistently predicted as either *External Regulation and Reputation* or, in the case of *Crime and Punishment*, *Cultural Identity*. The latter false predictions are over-represented in the target corpus, which we assume is the reason for poor competence with binary representation.

For the lexicon-based approach, the results show less range between competence with and without priors, which is only supported by the similarity of the frame distribution in training and predictions: the target corpus results are well reflective of the training distribution. For this approach, however, some of the most numerous frames are coincidentally the ones with lower-than-chance performance even on soft accuracy: frames *Crime and Punishment* and *Public Sentiment* perform well below expected in one or even two languages, respectively. Since the lexicon-based approach, in the current comparison setup, is less restrictive (it does not require every token to be labeled, unlike in the transformer-based approach), we can attribute the poor performance to the characteristics of the background corpus,

where the sparsity or the skewness of the articles to certain topics was restrictive on the lexicon we derived.

Noted in other works in the domain (Liu et al., 2019), one point is reinforced by these results: it is crucial to note and account for the absence of frames, as much as it is essential to identify precisely their presence. To provide better accuracy, the chosen approach should be exposed to examples of no framing or *Other* frames, for which MFC had a prohibitively low count and low annotator agreement.

4.4 Results of the Method Comparison

With results over all frames, we reconfirm the low inter-method agreement, highlighted in dominant frames results: in Figure 1 we can observe the range of agreement per frame and per language. As expected *Capacity & Resources* and *Public Sentiment* frames were among the worse performing ones: both of those frames performed low across languages in either method. Even though both frames are tilting towards lower counts in training sets, we hypothesize their subjective nature, also reported by Field et al. (2018), contributes to the performance. From the preliminary results, we conclude that individual frames and language corpora should be treated on a case-by-case basis. Seeing the range of performance by each method depending on the testing corpus, we also conclude that even with extensive standardized training material such as MFC, the task of identifying frames cross-lingually remains extremely sensitive to the parameters of the chosen approach, and no method presents a one-size-fit-all solution. Despite its mixed performance, the lexicon-based approach emerges as a more confident predictor. Its drop in performance with a binary presentation could suggest that, for certain frames, the negative (not present) decision is unexpected, which could be due to limitations of the lexicon that draw from the choice of the background corpus vocabulary.

4.5 Identifying and Comparing Frames from the Majority Vote of the Models

Observing the volatility and sensitivity of the results, we proceed to analyze the frames where the majority voting (agreement between two methods) decided the frames are present. We compute the nPMI score for each language with a general PMI formula seen in Equation 1,

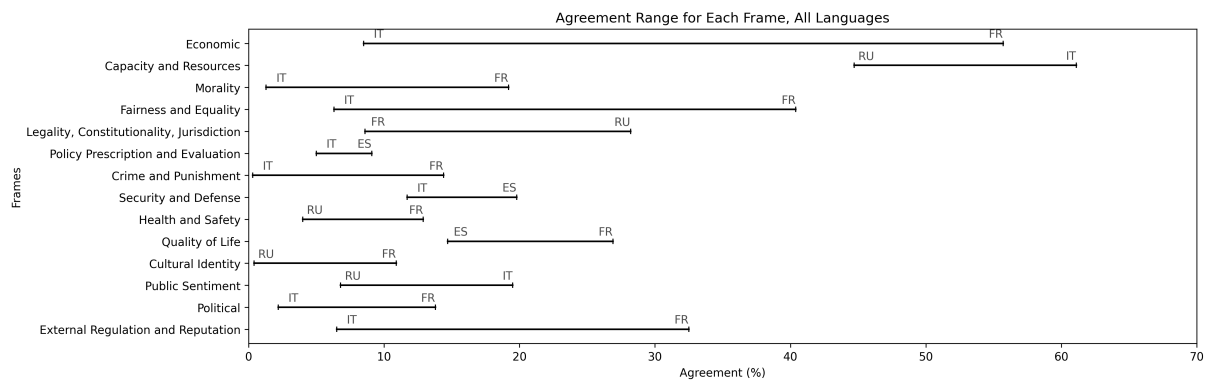


Figure 1: % of annotations where two methods reach agreement about frame's presence, by language

normalized and adapted to measure frame salience on a language level. In Figure 2, the scores are normalized to the range [-1;1], where 1 presents the complete co-occurrence of a frame with a language.

The results of the frame nPMI across four languages are varied: while articles in Italian and Spanish are the least focused on the *Political* aspects and *Quality of Life*, these two frames are at the center of attention for Russian frames. Supported by the findings of Field et al. (2018) and Rozenas and Stukal (2019), the salience of frames in Russian is not unexpected and is driven by the time frame of the target corpus, where the conflict in Ukraine and the COVID pandemic were among the key events. More interestingly, the salience of *Political* and *Quality of Life* is also strong in the French corpus, along with *Morality* and *Crime and Punishment*. The latter could be partially supported by more policy-oriented findings of Benson (2013) that note the salience of such topics as equality of immigrant treatment in French discourse.

If we follow a stricter approach and exclude the frames that performed poorly in the modeling, we see a much stronger polarization of the languages: while Russian texts stay focused on *Health and Safety*, French texts are primarily characterized by *Morality*, Italian is focused on *External Regulation and Reputation*, and Spanish puts the strongest focus on *Cultural Identity*. Below are the words most associated with each language's respective dominant frame, translated into English:

FR *Morality*: *compassion, aggressiveness, generosity, authority, injustice*;

ES *Cultural Identity*: *youth, celebrity, legend, elite, bourgeoisie*;

RU *Health and Safety*: *offspring, harmful, sick, mental, unhealthy*;

IT *External Regulation*: *containment, stabilization, integration, rebalancing, cooperation*.

To examine the Russian corpus on a more granular level, we calculate the co-occurrence of specific frames with articles in Russian released in certain regions (Figure 3). The body of articles was taken from the articles pairs assembled previously in the work and supplemented by the articles in Russian belonging to the same EUvsDisinfo reports, judged as belonging to the same disinformation topic. The countries were grouped into regions following the lists below, in descending order based on the number of articles. While we perform a simple geography-driven split to make the groups more distinct, the targeting of the disinformation campaigns might be more subtle and country-specific, depending on the set agenda.

Eastern Europe: Ukraine, Belarus, Moldova, Lithuania, Latvia, Poland;

The Caucasus: Armenia, South Ossetia, Georgia, Abkhazia, Azerbaijan;

Central Asia: Uzbekistan, Kyrgyzstan, Kazakhstan.

The resulting salient frames present a different picture from what we observed on a language level: while Russia-based media outlets have a variety of accentuations, the rest of the regions have a clear dominant focus. Most interestingly, while

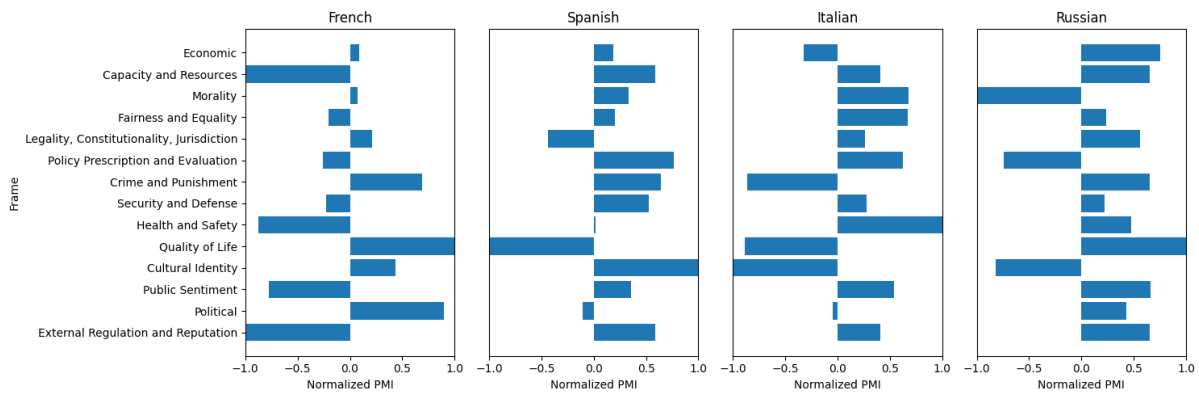


Figure 2: PMI score for four languages, normalized to [-1;1]

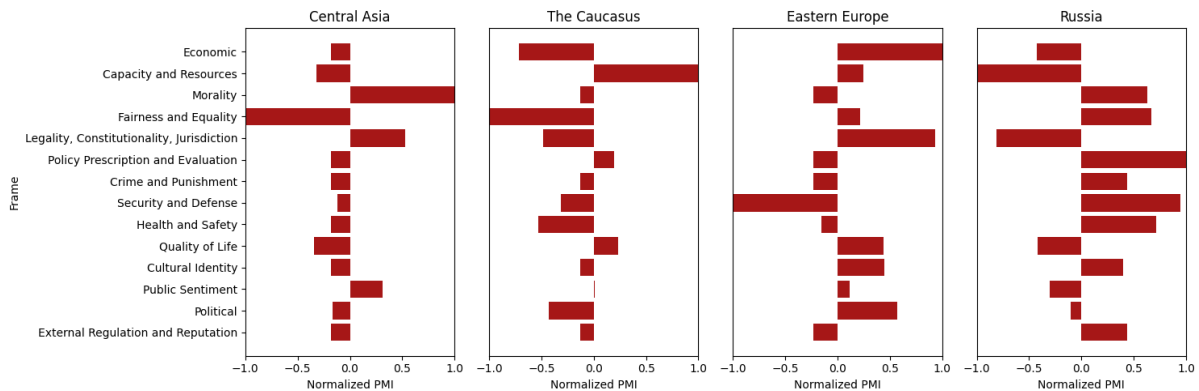


Figure 3: PMI score for four regions, normalized to [-1;1]

Central Asia presents the same dominant frame as the French corpus, in The Caucasus (*Capacity and Resources*) and Eastern Europe (*Economic* and *Legality, Constitutionality, Jurisdiction*) groups we see new dominant frames that were not prominent on a language level. Knowing that the Eastern European country group, in particular, presents a mix of countries with different political affiliations, we still observe a clear focus in the article framing. We could suspect that two almost equally prominent frames represent two country sub-groupings, which would be worth investigating in the future. The same couldn't be said about articles released in Russia: the material is more multi-focal and naturally presents a variety of topics, especially the ones covering domestic policies (*Policy Prescription and Evaluation, Crime and Punishment, Security and Defense*). This suggests a direction for further exploration and provides an example of how nuances the disinformation articles can be, depending on the language and even geography within the same language corpus of articles.

5 Related Work

The most common approaches to identifying frames treat the task as a variation of sentiment analysis or probabilistic topic modeling (Boydston et al., 2014; Tsur et al., 2015; Nguyen et al., 2013; Kwak et al., 2021). While a standardized approach, sentiment, or stance analysis presents limitations to frame identification: most articles employ multiple frames at the same time with various concentrations. Additionally, topic modeling doesn't facilitate the comparison of different corpora because of its corpus-specificity and difficulty of interpretation. The more advanced but still traditional approach is creating issue-specific manually annotated handbooks. Annotation books, though more formalized, remain a labor-intensive and issue-specific approach, which presents little opportunity for automatic text analysis and frame identification. A more common quantitative approach to frame detection, started with the work of Boydston and Gross (2013), is assembling a list of generic frames applicable across issues. Beginning with the development of Policy Frames Codebook

(Boydston et al., 2014) and the Media Frames Corpus (Card et al., 2015), a growing body of work is concerned with automating frame identification at scale. While topic modeling is a versatile approach that can be used with any language, framing analysis with Policy Frames Codebooks, and particularly MFC, relies on data written and annotated in English. This makes the state-of-the-art NLP approaches to frame identification, including the most recent findings of Mendelsohn et al. (2021), English-centric with no apparent transition to other languages. So far, no method has established itself as a standard practice in multilingual frame identifications. Two works emerge as the most prominent approaches to multilingual framing analysis. The earlier one is presented in the work of Field et al. (2018), which projects English framing onto Russian through a lexicon-based, distantly supervised approach. Their work focuses on expanding and localizing MFC annotations lexicon by creating language-specific lexicons using an extensive background corpus in the target language. The second approach, presented by Park et al. (2022), is based on translating original articles to English and evaluating them with a classifier based on large pre-trained language models. This approach emphasizes the target language less but claims to scale to low-resource languages without needing annotated material. It is advantageous when training data is insufficient, or the computations of training an entire model are prohibitively expensive. To date, these two works present the most prominent approaches to analyzing all frames in a text across languages.

6 Conclusions

We compare two approaches for frame identification on a novel dataset. The formal comparison of the two approaches brought to light a more nuanced result than expected. While the lexicon-based method produced a higher overall competence in estimating framing on multilingual pairs, the results appear mixed depending on the presentation of the data. We suspect distinct reasons for each method’s low performance. For the lexicon-based approach, the unexpected drop in performance could reflect the insufficient lexicon for specific frames. For the transformer-based approach, the poor performance on the frames overrepresented in the MFC could be either a consequence of choices in model fine-tuning setup or a direct result of heterogeneity of texts in the MFC itself. The latter

point should be investigated in the future, as the MFC data sampling decisions translate directly or indirectly into the approaches’ performance.

As both approaches present mixed performance, nuanced by language context and specific frames, we cannot conclude unequivocally the most accurate approach to be one method or the other. Further seeing low inter-method agreement scores and the range of disagreement across languages and frames, we conclude that both approaches are highly nuanced and context-sensitive, even when based on the same pre-training on MFC. Thus, neither of the prominent multilingual methods can guarantee performance in a new context, especially in low-resource languages.

Applied to our multilingual disinformation pairs, the joint decision of both methods produced various salient frames depending on the languages of the article, as we expected in the hypothesis. Our findings confirm that in disinformation campaigns, articles presenting the same event or topic focus on different aspects of the issue, depending on which audience the campaign targets. We confirm this hypothesis for four languages in the dataset and a subset of regions that are targeted with articles in the Russian language. We recognize that, while the timespan for which we collected the disinformation articles (2015-2023) provides invaluable insights into the Russia-backed disinformation campaigns, it does not allow us to generalize into an analysis of the best methods for frame absence/presence at a sentence level. A more task-focused approach, that considers aspect and the most recent studies in frame presence/absence methods is a point of future research.

7 Ethical Considerations

This study is based on publicly available models, translation services, and datasets, such as MFC and CC-100. Although we plan to release the code and the dataset collected for this work, the users should be cautious of the potential bias towards the standard version of the languages in scope, originating from the model architecture and the data collection decisions made at source (EUvsDisinfo).

8 Limitations

Since one of our goals is to compare two existing methods, their limitations also transfer to our work. First, the reliability of MFC as the training material has been contested in previous works: since articles discussing certain issues can be more or less balanced in timeframe coverage and frame concentration, it raises risks of poor performance on certain frames and skewed lexicon in lexicon-based approaches. Tied to the MFC, the question of the interpretability of issue-agnostic frames has been raised: the frames encapsulate so many associations that the issue of blurred boundaries between close frames or their lexicons can appear in certain contexts. It has been noted in the existing body of research that the current models generalize poorly to new domains, which was in part observed in our work. Second, the availability of the resources for either of the methods presents a serious limitation to their implementation: while for a lexicon-based approach, an extensive background corpus is needed to contextualize the lexicons to the target language, the transformer-based approach results in significant computational costs. The evaluation of either method remains expensive as it requires recruiting experts with domain knowledge for the annotations task. The low count of annotators, as much in this paper as in the original methods, remains a limitation. The challenge of applying current resource-heavy methods to low-resource material remains open. The assumptions under which we collected the dataset of Russia-backed disinformation present another limitation to this work. Preserving all historical material meant that some frames would be over-represented due to the nature of the topics discussed in the disinformation.

Acknowledgments

The authors thank the anonymous reviewers, especially reviewer 2, for their detailed and constructive feedback. The authors recognize the contributions of the annotators, who volunteered their time and effort to this work. The paper’s main findings are part of the Bocconi Master’s Thesis of AS, who invited other graduate students to complete the annotation task, in exchange for her equal contributions to their research works. DH is a member of the Bocconi Institute for Data Science and Analytics.

References

- Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. [Multi-label and multilingual news framing analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.
- Rodney Benson. 2013. *Shaping Immigration News*, page i–ii. Communication, Society and Politics. Cambridge University Press.
- Amber E. Boydston, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. [Tracking the Development of Media Frames within and across Policy Issues](#).
- Amber E. Boydston and Justin H. Gross. 2013. [Identifying media frames and frame dynamics within and across policy issues](#).
- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. [Yake! keyword extraction from single documents using multiple local features](#). *Information Sciences*, 509:257–289.
- Dallas Card, Amber Boydston, Justin Gross, Philip Resnik, and Noah Smith. 2015. [The media frames corpus: Annotations of frames across issues](#). 2:438–444.
- Dallas Card, Justin Gross, Amber Boydston, and Noah A. Smith. 2016. [Analyzing framing through the casts of characters in the news](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.
- Kenneth Ward Church and Patrick Hanks. 1990. [Word association norms, mutual information, and lexicography](#). *Computational Linguistics*, 16(1):22–29.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). *Preprint*, arXiv:1911.02116.
- EUvsDisinfo. EU vs Disinformation. <https://euvsdisinfo.eu>. Accessed: July 19, 2024.
- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. [Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. [Learning whom to trust](#)

- with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.
- Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. [Modeling political framing across policy issues and contexts](#). In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.
- Gary King, Jennifer Pan, and Margaret E. Roberts. 2017. [How the chinese government fabricates social media posts for strategic distraction, not engaged argument](#). *American Political Science Review*, 111(3):484–501.
- Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. SAGE Publications, Thousand Oaks, CA.
- Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov. 2020. [Fake news detection for the Russian language](#). In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDMS)*, pages 45–57, Barcelona, Spain (Online). Association for Computational Linguistics.
- Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. [FrameAxis: characterizing microframe bias and intensity with word embedding](#). *PeerJ Computer Science*, 7:e644.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.
- Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. [Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.
- Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. [Modeling framing in immigration discourse on social media](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *Preprint*, arXiv:1301.3781.
- Kevin Munger, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2019. [Elites tweet to get feet off the streets: Measuring regime social media strategies during protest](#). *Political Science Research and Methods*, 7(4):815–834.
- Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1, NIPS’13*, page 1106–1114, Red Hook, NY, USA. Curran Associates Inc.
- Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. [Challenges and opportunities in information manipulation detection: An examination of wartime Russian media](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jakub Piskorski, Nicolas Stefanovitch, Guillaume Jacquet, and Aldo Podavini. 2021. [Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up](#). In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 35–44, Online. Association for Computational Linguistics.
- Arturas Rozenas and Denis Stukal. 2019. [How autocrats manipulate economic news: Evidence from russia’s state-controlled television](#). *The Journal of Politics*, 81(3):982–996.
- Oren Tsur, Dan Calacci, and David Lazer. 2015. [A frame of mind: Using statistical models for detection of framing and agenda setting campaigns](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

A Appendix

Frame Type	Frame Description
Economic	Financial implications of an issue
Policy Capacity & Resources	The availability or lack of time, physical, human, or financial resources
Morality & Ethics	Perspectives compelled by religion or secular sense of ethics or social responsibility
Fairness & Equality	The (in)equality with which laws, punishments, rewards, resources are distributed
Legality, Constitutionality & Jurisdiction	Court cases and existing laws that regulate policies; constitutional interpretation; legal processes such as seeking asylum or obtaining citizenship; jurisdiction
Crime & Punishment	The violation of policies in practice and the consequences of those violations
Security & Defense	Any threat to a person, group, or nation and defenses taken to avoid that threat
Health & Safety	Health and safety outcomes of a policy issue, discussions of health care
Quality of Life	Effects on people’s wealth, mobility, daily routines, community life, happiness, etc.
Cultural Identity	Social norms, trends, values, and customs; integration/assimilation efforts
Public Sentiment	General social attitudes, protests, polling, interest groups, public passage of laws
Political Factors & Implications	Focus on politicians, political parties, governing bodies, political campaigns, and debates; discussions of elections and voting
Policy Prescription & Evaluation	Discussions of existing or proposed policies and their effectiveness
External Regulation & Reputation	Relations between nations or states/provinces; agreements between governments; perceptions of one nation/state by another

Table 5: List of non-issue-specific frames (Boydston and Gross, 2013) used in MFC and our annotation task

Code	Frame	Train (#)	Test (#)	Total Count (#)
1.0	Economic	9.2k	2.3k	11.5k
2.0	Capacity and Resources	2.9k	0.7k	3.6k
3.0	Morality	2.9k	0.7k	3.6k
4.0	Fairness and Equality	2.7k	0.7k	3.4k
5.0	Legality, Constitutionality, Jurisdiction	16.1k	4.0k	20.1k
6.0	Policy Prescription and Evaluation	6.4k	1.6k	8.0k
7.0	Crime and Punishment	12.5k	3.1k	15.7k
8.0	Security and Defense	4.4k	1.1k	5.6k
9.0	Health and Safety	6.8k	1.7k	8.5k
10.0	Quality of Life	2.5k	0.6k	3.2k
11.0	Cultural Identity	3.6k	0.9k	4.5k
12.0	Public Sentiment	4.6k	1.2k	5.8k
13.0	Political	19.0k	4.7k	23.7k
14.0	External Regulation and Reputation	1.5k	0.4k	1.9k
	Total	95.3k	23.8k	119.1k

Table 6: The Number of Annotations Admitted to Training XLM-R: Counts Represent Full Sentences

Code	Frame	F1	Count (#)
1.0	Economic	53.3	7
2.0	Capacity and Resources	15.4	12
3.0	Morality	74.9	5
4.0	Fairness and Equality	18.2	8
5.0	Legality, Constitutionality, Jurisdiction	22.2	6
6.0	Policy Prescription and Evaluation	16.6	9
7.0	Crime and Punishment	18.2	5
8.0	Security and Defense	31.6	17
9.0	Health and Safety	66.6	3
10.0	Quality of Life	37.5	11
11.0	Cultural Identity	55.4	24
12.0	Public Sentiment	0.0	7
13.0	Political	35.7	13
14.0	External Regulation and Reputation	41.9	26
	Macro-F1	32.9	
	Total		156

Table 7: Transformer-based Method Performance: Macro-F1

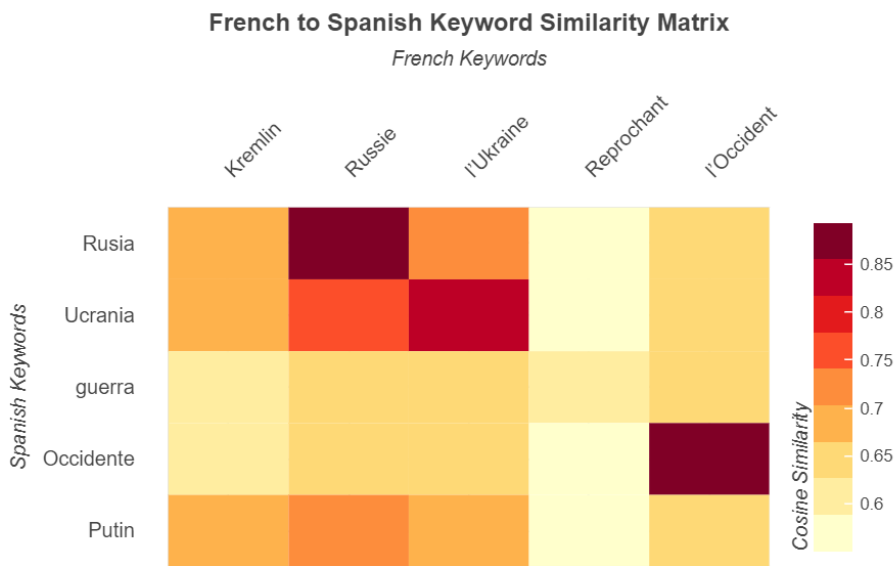


Figure 4: Keywords Cosine Similarity for a Pair of Ground Truth Articles

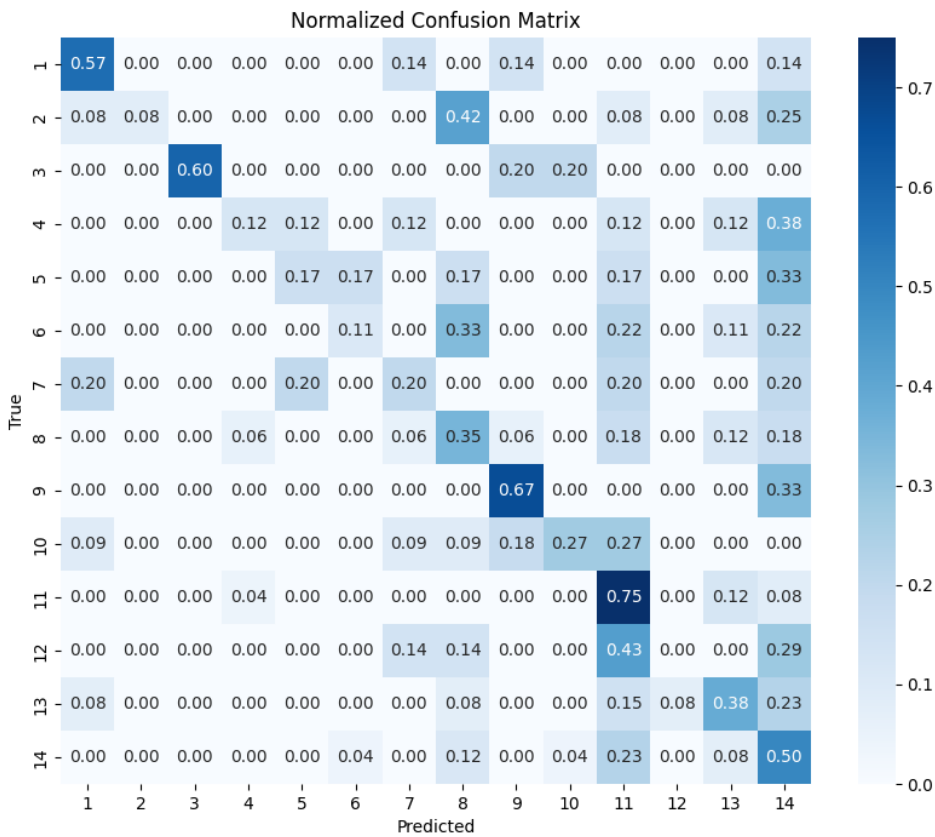


Figure 5: Normalized Confusion Matrix; the codes represent the frames, see code-frame correspondence in Table 6 or Table 7