SemEval-2022 Task 8: Multilingual news article similarity

Xi Chen
U. Mass. Amherst
xchen4@umass.edu

Ali Zeynali
U. Mass. Amherst
azeynali@umass.edu

Chico Q. Camargo
University of Exeter
f.camargo@exeter.ac.uk

Fabian Flöck
GESIS
f.floeck@gmail.com

Devin Gaffney
Meedan
devin@meedan.com

Przemyslaw A. Grabowicz
U. Mass. Amherst
grabowicz@cs.umass.edu

Scott A. Hale
University of Oxford and Meedan
scott.hale@oii.ox.ac.uk

David Jurgens
University of Michigan
jurgens@umich.edu

Mattia Samory
GESIS
mattia.samory@gesis.org

Abstract

Thousands of new news articles appear daily in outlets in different languages. Understanding which articles refer to the same story can not only improve applications like news aggregation but enable cross-linguistic analysis of media consumption and attention. However, assessing the similarity of stories in news articles is challenging due to the different dimensions in which a story might vary, e.g., two articles may have substantial textual overlap but describe similar events that happened years apart. To address this challenge, we introduce a new dataset of nearly 10,000 news article pairs spanning 18 language combinations annotated for seven dimensions of similarity as SemEval 2022 Task 8. Here, we present an overview of the task, the best performing submissions, and the frontiers and challenges for measuring multilingual news article similarity. While the participants of this SemEval task contributed very strong models, achieving up to 0.818 correlation with gold standard labels across languages, human annotators are capable of reaching higher correlations, suggesting space for further progress.

1 Introduction

Consider the following question: Given a pair of "hard" news articles,[1] are they covering the same news story? Answering this question likely requires knowing specific aspects of the events covered: what happened, where and when it happened, who was involved, and why and how it happened (Pan and Kosicki, 1993; Klein and Martínez, 2009; Dijk, 1988).

Effectively modeling the similarity of news stories holds substantial practical benefits in structuring the content of the hundreds of thousands of news articles generated every day.[2] Given the volume of articles, an effective measure of news story similarity enables clustering and identification of event coverage in news media (Rupnik et al., 2016; Bisandu et al., 2018). Commercial news aggregation services, as provided by, e.g., Google News or Apple News, perform a similar clustering approach, yet are primarily monolingual and have not been made openly available or extensively researched beyond proprietary solutions. In addition, quantifying news article similarity allows the comparison of news outlets in terms of their coverage, understanding which stories consume much of the media agenda, as well as tracking the diffusion of news stories through a media ecosystem and over time. Being able to measure these aspects is important for a host of research questions in media and communication studies including, for example, agenda setting (McCombs, 2005). Another highly desirable property of such methods is to be applicable in multilingual settings, to detect news stories covered across languages in an increasingly globalized news ecosystem (Rupnik et al., 2016).

Assessing the similarity of two news articles introduces new challenges not found in traditional semantic textual similarity. Most importantly, methods for semantic textual similarity typically measure the extent to which two arbitrary documents are "the same," without concretely specifying the meaning of this similarity, or only do so in broad strokes (cf. Agirre et al., 2012; Lee et al., 2005; Nguyen et al., 2014). One byproduct of this vague application domain and under-specification is that

[1] "Hard news" is characterized as having a high level of newsworthiness demanding immediate publication (Tuchman, 1972). In our use, we aim to exclude opinion, features, and other forms of journalistic pieces not mainly concerned with covering current events as in Flaxman et al. (2016).

[2] For example, the source we use for metadata, Media Cloud, collects 629K articles per day.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1094 - 1106
July 14-15, 2022 ©2022 Association for Computational Linguistics

agreeing on gold labels is notoriously difficult, at least at the full-length document level (Nguyen et al., 2014; Westerman et al., 2010), as even human labelers are dependent on specific instructions and/or knowledge of the absolute space of documents to label, to understand the relative concept of "similar" (Bär et al., 2011). News article similarity is thus more related to attempts to compare narratives (Chambers and Jurafsky, 2009; Miller et al., 2015; Chaturvedi et al., 2018), which require understanding the structure and content to assess similarity.

Here, we introduce SemEval 2022 Task 8 for the task of quantifying news similarity, a hard discourse-level task. Stories often include a variety of descriptions, people, and entities that may appear in another, dissimilar story. Further, the temporal nature of news means that as real-life events evolve, stories describing the same event may include new details or entities—possibly becoming a new news story altogether. For this task, we create a high-quality dataset by annotating pairs of news article for similarity in 10 different languages on several dimensions, e.g, geographic, temporal, and narrative similarity. Participants in this task were given a large collection of news articles, with 4,918 pairs receiving ground-truth similarity labels, and were asked to estimate the overall similarity of 4,902 news article pairs given to participants without labels for any dimension. The task is also challenging due to its large language diversity: the training data consists of 8 language combinations, while the evaluation dataset has 18 language combinations including three languages not appearing at all in the training data.

2 Data

2.1 Data Collection

The metadata and full text of news articles was collected from Media Cloud, an open-source platform aggregating millions of stories published online (Roberts et al., 2021). We collected the metadata and full text of all news articles from January 1, 2020 to June 30, 2020 in 10 languages, thanks to white-listing by Media Cloud. Overall, this collection includes news articles in the following languages: English (31M articles), Spanish (8.2M), Russian (7.2M), German (3.2M), French (3.2M), Arabic (2.9M), Italian (2.4M), Turkish (1.1M), Polish (595K), and Mandarin Chinese (342K). We hired and trained annotators with fluency in these languages. The total dataset consists of about 60M articles. Article metadata includes dates, headlines, and URLs of articles. To filter, match, and sample pairs of news articles for annotation, we apply a series of processing steps to the dataset (Figure 1), described below. The annotated data is available on Zenodo while the full text of most webpages annotated is available in a special collection at the Internet Archive. We have also created a Python package to crawl and process the webpages.

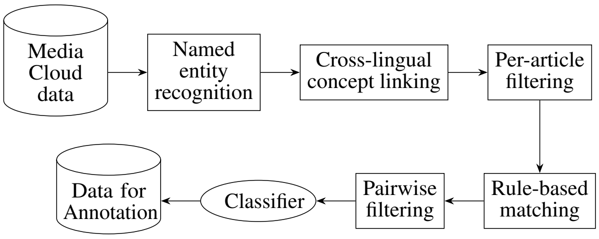

Figure 1: Filtering and pair matching pipeline.

Filtering. We applied a series of filtering steps to clean the data. First, we filtered out articles that miss one of basic metadata attributes: story ID, URL, title, or text. Second, we dropped data points that do *not* correspond to news articles of social or political importance[3] and very short articles whose word count is less than 100. Third, we filtered out articles that have titles or URLs that exactly match a newer news article. After applying these filtering steps, the numbers of articles per language are: English (10M articles), Spanish (4.6M), Russian (1.8M), German (1.3M), French (1.2M), Arabic (1.8M), Italian (1.5M), Turkish (655K), Polish (369K), and Mandarin Chinese (205K).

Matching and Sampling of News Pairs. Randomly sampled pairs of news articles are unlikely to be related. Therefore, a major design point in our pilot work was to identify meaningful candidate pairs. We experimented with document embeddings (Cr5: Josifoski et al., 2019), sentence embeddings (Sentence BERT: Reimers and Gurevych, 2019) applied to headlines and lead paragraphs, and named entities (spaCy, polyglot, and Babelfy; Moro et al., 2014) to identify similar articles. With extensive pilot study, we devised an efficient sampling pipeline (Figure 1). First, the named entities of each article are extracted using spaCy and

[3] Irrelevant websites include: "reddit.com," "facebook.com," "twitter.com," "fb.com," "wikipedia.org," "epochtimes.com," "youtube.com," "slideshare.net". We also dropped any url with 'sport' in it.

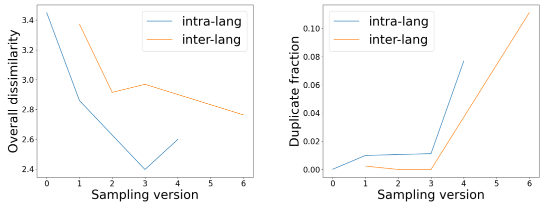

Figure 2: The average dissimilarity of news article pairs (left) and the fraction of duplicates (right) per sampling version.

polyglot.[4] For monolingual pairs, we select pairs of articles having high Jaccard similarity of these named entities. For cross-lingual pairs, we attempt to match the named entities to Wikipedia article titles and store the Wikidata concept ids of matching Wikipedia articles, which are language agnostic. We then select cross-lingual pairs of articles having high Jaccard similarity of these Wikidata concept ids.

To remove duplicate articles (i.e., articles that have the same or nearly the same text, but are published with different titles and URLs), we drop all pairs of articles that share one or more long sentences (of 40 or more characters) or where the Jaccard similarity of article text is higher than a certain threshold. Once the training set of news article pairs was annotated, we trained a logistic regression classifier that was used for further sampling. The features included in the classifier are: the word counts of both articles, the number of common words, the number of common named entities, cosine similarity of the named entities with BM25 embeddings (Robertson and Zaragoza, 2009), text Jaccard similarity, and an exponentially decaying function of publication date difference.

The pipeline was updated over time to increase the fraction of OVERALL similar pairs among samples (Figure 2, left). Version 1 of our sampling pipeline selects pairs based solely on the Jaccard similarity of named entities without any classifier, since initially no labeled data was unavailable. Version 2 introduces a temporal window where only articles published within a few days from each other are considered. Version 3 introduces a minimal threshold for Jaccard similarity of named entities. Versions 4 and 5 count the reappearance of words for Jaccard similarity and implement a more efficient similarity computation in Cython, respectively. Version 6 removes the word reappearance counts after an evaluation of its effectiveness. We note that while improvements to matching and sampling increased the fraction of similar news articles, we also experienced a small increase in the fraction of duplicate news articles (Figure 2, right).

2.2 Annotation

Annotation guidelines were developed through an iterative process, grounded in media studies literature on news. After several pilot annotation rounds, we formed a detailed codebook for seven dimensions of similarity. The questions were:

GEO How similar is the geographic focus (places, cities, countries, etc.) of the two articles?

ENT How similar are the named entities (e.g., people, companies, organizations, products, named living beings), excluding previously considered locations appearing in the two articles?

TIME Are the two articles relevant to similar time periods or describing similar time periods?

NAR How similar are the narrative schemas presented in the two articles?

OVERALL Overall, are the two articles covering the same substantive news story? (excluding style, framing, and tone)

STYLE Do the articles have similar writing styles?

TONE Do the articles have similar tones?

Annotators answered each question using a four-point Likert scale with the options, “Very Dissimilar,” “Somewhat Dissimilar,” “Somewhat Similar,” and “Very Similar.” In this paper, we represent these ordinal labels as numbers from 4 (Very Dissimilar) to 1 (Very Similar). In addition, each question can be answered with the option “Other”, which is used mainly for marking pairs of duplicate news articles and unavailable articles, e.g., due to a paywall or take-down (annotators were asked to report such cases via a free-text comment). The annotation codebook defines each dimension and gives examples with explanations of labeled news article pairs (e.g., Table 1).

To achieve the desired linguistic diversity and magnitude of news annotation we trained 25 annotators hired across 3 institutions (GESIS, UMass, UMich), out of which 10 labeled over 1,000 news article pairs during the course of roughly six months (Table 2). Annotators were compensated €12 per hour at GESIS and $15 per hour at UMass and UMich.

We implemented a custom annotation interface in Ruby and MongoDB that assigns news articles pairs at random within the language abilities of

[4] To scale the pipeline to tens of millions of articles, we use the efficient, simple language models rather than the transformer models in spaCy version 3.

Article 1	Article 2	GEO	ENT	TIME	NAR	STY	TONE	OVERALL
NYC testing two more people for coronavirus	New York City Reports 2 Additional Suspected Cases of Coronavirus	VS	VS	VS	VS	VS	VS	VS
Video of a man beating his girlfriend mercilessly goes viral	Curry house in Worcester improves from one-star to five food hygiene rating	VD	SD	SD	VD	SS	SD	VD
All with flu symptoms to be tested for Covid-19 in Chandigarh	ICMR study points towards possible community transmission of coronavirus COVID-19 in India	SD	SD	SS	SD	SS	SS	SD

Table 1: Example annotated pairs. The pairs were annotated based on the full-text of the articles. Each of the seven dimension is annotated with a Likert scale with four options: Very Similar (VS), Somewhat Similar (SS), Somewhat Dissimilar (SD), and Very Dissimilar (VD). The articles in the first pair released very similar information about two people tested Coronavirus positive at New York City. The second pair is very dissimilar since one article described the violence against a women while the other one reported the rating improvement of a restaurant. They shared nothing in common. The final pair overlapped somewhat in terms of GEO (India), ENT (Indian Council of Medical Research, ILI, severe acute resparatory illness), and TIME. The two articles, however, still refer to different events.

id	items	shared	seconds/item	correlation
1	1,657	809	296	0.88
2	2,311	495	344	0.86
3	1,197	611	237	0.85
4	1,178	794	213	0.85
5	134	98	153	0.84
6	1,036	626	128	0.84
7	1,302	345	220	0.84
8	466	398	224	0.84
9	787	208	233	0.84
10	887	368	506	0.83
11	1,062	466	387	0.82
12	361	321	311	0.81
13	262	213	165	0.81
14	139	135	235	0.79
15	1,076	716	71	0.77

Table 2: Annotators and their statistics: the number of labeled items (news article pairs), the number of shared items (also annotated by another annotator), median number of seconds to label an item, and Pearson correlation of their OVERALL labels with the mean labels of other annotators. Only annotators with at least 100 labels are shown.

each annotator. To engage and motivate annotators, the interface also provides feedback to annotators in the form of basic statistics such as the number of annotations, and the inter-rater agreement of the top annotators. The interface also shows past annotations and highlights disagreements, which were discussed at biweekly video conference meetings.

Codebook, Training, & Annotation. The annotation process has multiple stages (Figure 3). All annotators read the codebook and attended a training session at which the codebook and annotations of example pairs were discussed. The annotators then completed 30 practice annotation pairs independently. After completing each practice pair, annotators were able to view the gold standard labels that had been agreed by all of the SemEval task authors. Each gold standard label was accompanied by a written explanation of why the label was assigned. Question and answer sessions were held at which the annotations were discussed as well.

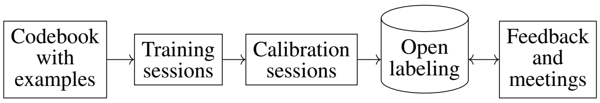

Figure 3: Annotator workplan.

Annotators then labeled another 30 gold standard pairs having detailed explanations, which we used to calibrate annotators' understanding of the codebook. Any disagreements were discussed until agreement was reached. The practice and calibration pairs were all English-language articles, which was a shared language ability between all our annotators. All news article pairs with gold standard labels and explanations, as well as the codebook, are released on Zenodo.

After these practice and calibration activities, pairs were annotated by a variable number of annotators, usually 1, 2, or 3, in the "open labeling" phase, where news article pairs were served to annotators continuously. Annotators were given feedback in the annotation interface on their agreement

languages	annotations	mean(OVERALL)
en	5,189	2.92
de	2,166	2.56
es	955	2.40
zh	866	2.24
de-en	863	3.20
tr	817	2.79
pl	584	2.36
ar	572	2.41
es-en	504	2.79
it	411	2.65
es-it	320	2.29
ru	289	2.78
zh-en	253	3.04
fr	184	2.39
de-fr	116	1.88
pl-en	77	2.38
de-pl	35	1.69
fr-pl	11	1.91

Table 3: The number of annotations and the mean OVERALL label (the higher, the more dissimilar) across the 10 languages and their combinations.

	GEO	ENT	TIME	NAR	OVERALL	STYLE	TONE
Krippen.	0.73	0.69	0.57	0.69	0.77	0.46	0.38
Gwet	0.81	0.67	0.75	0.69	0.76	0.69	0.67

Table 4: Inter-rater agreement measures, Krippendorf's α and Gwet's AC$_1$, for each labeled dimension.

with other annotators, met regularly to discuss disagreements, and had an open channel for communication on a shared Slack instance. After annotating English-language pairs, non-English pairs were introduced and discussed with annotators. Finally, cross-language pairs were also introduced. The total number of annotations and average OVERALL label per each language pair is shown in Table 3.

Inter-annotator Agreement. The inter-rater agreement on the OVERALL similarity dimension is very high, with a Krippendorff's α of 0.77. We note that the distribution over labels is generally not uniform, e.g., the labeled news article pairs are skewed towards "Very Similar" in TIME, STYLE, and TONE (Figure 4). Gwet's AC$_1$ is known to be less sensitive to non-uniform marginal label distributions (Gwet, 2008), and it suggests a good agreement in all dimensions (Table 4).

Annotators vary in terms of the quantity and quality of the provided annotations. To compare the performance of annotators to the performance of models, we measure the inter-rater agreement of each annotator in a way that corresponds to the score of models, i.e., as a Pearson correlation between the labels of that annotator and a series of average labels from other annotators. Note, however, that this correlation is measured over labels contributed by the given annotator to both training and evaluation datasets, it is biased by the language-abilities of the annotator, and the mean label does not include the ego annotator. Our top 5 annotators consistently reach very high agreement scores of 0.85–0.88, whereas bottom 5 annotators reach agreement scores of 0.73–0.80 (Table 2).

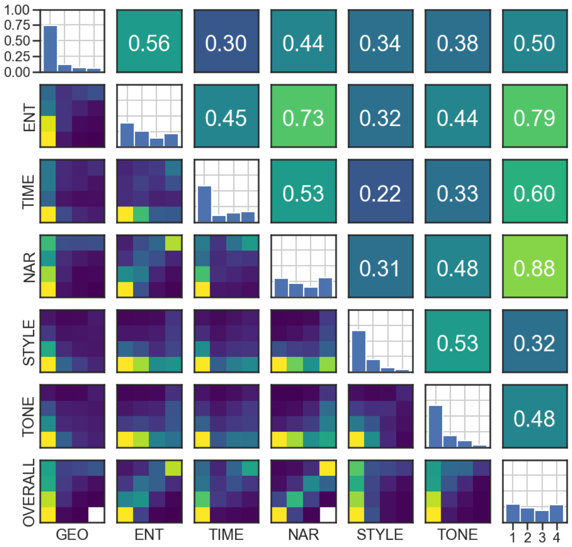

Figure 4: Histograms and Pearson correlations of every pair of scores in the labelled data.

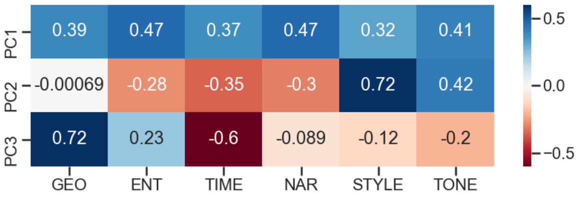

Figure 5: Heatmap showing the coefficients of the first three principal components of variation in the scores.

2.3 Statistics of the Labeled Dataset

Figure 4 illustrates the relationship between the multiple similarity dimensions in the annotated dataset. The bar charts on the diagonal represent the distribution of annotations, from 1 (Very Similar) to 4 (Very Dissimilar). Panels below the diagonal represent two-dimensional histograms, and panels above the diagonal report the Pearson correlation between different dimensions, namely GEO, ENT, TIME, NAR, STYLE, TONE, and OVERALL.

language(s)	train	eval	mean(OVERALL)
ar	274	298	2.41
de	857	608	2.57
de–en	531	185	3.18
de–fr		116	1.88
de–pl		35	1.69
en	1,800	236	2.86
es	570	243	2.34
es–en		496	2.79
es–it		320	2.29
fr	72	111	2.39
fr–pl		11	2.00
it		411	2.65
pl	349	224	2.35
pl–en		64	2.35
ru		287	2.78
tr	465	275	2.74
zh		769	2.22
zh–en		213	3.07
Totals	4,918	4,902	2.62

Table 5: The number of news article pairs in the training and evaluation datasets by language and the mean OVERALL label (the higher, the more dissimilar).

NAR and ENT show the highest correlation with OVERALL (0.88 and 0.79 respectively). These two dimensions also provide the largest contributions to the variation in annotations, as indicated by the first component of the PCA shown in Figure 5.

There is no significant difference between the training and evaluation datasets with respect to the labels of any similarity dimension, and these results are also found when the dataset is disaggregated by language pair.

3 Task

3.1 Task Description & Rules

The Task was created on CodaLab.org[5] and advertised with alongside the other SemEval 2022 tasks. Participants were told, "The task is: Given a pair of news articles, are they covering the same news story? This SemEval task aims to develop systems that identify multilingual news articles that provide similar information. This is a document-level similarity task in the applied domain of news articles, rating them pairwise on a 4-point scale from most to least similar."

Participants were given 60 English-language pairs for trial data in August 2021. The training data was released to participants in two batches: the first batch was released on September 15, 2021, and the second was released on November 4, 2021. The training data consisted of article pairs in 8 different language combinations (Table 5).

Due to copyright restrictions, we were unable to release the raw text of the news articles included in the training data. In lieu of this, we developed and shared a Python package to download the text of news articles. For the training data, the downloader tries to fetch the articles from the Internet Archive or the live web and parse them with `newspaper3k`.[6] This mirrored the actions of annotators who were given links to the articles on the Internet Archive and live web.

The evaluation data was released on January 10, 2022, and consists of 4,902 pairs of news articles across 18 languages. For the evaluation data, we only included pairs of articles where both news articles were available on the Internet Archive.[7]

For both the training and evaluation datasets, we removed any article pairs where one or more annotators labeled the OVERALL similarity as "Other". This usually indicated that the pair was unavailable or not a news article.

The evaluation period ran from January 10 to February 3, 2022. This date reflected the extra time needed to download the articles as well as a short extension due to technical issues with the Codalab system. Participants were allowed to submit up to 5 submissions per day and 1,000 submissions overall.

3.2 Baselines

Our baseline models use SVC with linear kernel, logistic regression, random forest, and XGBoost (Chen and Guestrin, 2016). For feature selection, we found positive correlation between the fraction of "Very Similar" news article pairs and their Jaccard similarity in terms of named entities, as well as full text (Figure 6). Thus we evaluate three sets of features in the baseline models: set-A (Jaccard similarity of named entities), set-B (set-A and text Jaccard similarity), and set-C (set-B and word count difference).

3.3 Evaluation and Ranking

The teams were evaluated using Pearson's r correlation with the mean OVERALL labels on the

[5] `https://competitions.codalab.org/competitions/33835`

[6] Some participants found better success at parsing the articles using `trafilatura` (Barbaresi, 2021)

[7] A special collection on the Internet Archive now includes most webpages.

evaluation data. In ranking teams, we were inspired by recent work (e.g., Dodge et al., 2019) on estimating model performance while recognizing that not all systems solving a task are on equal footing. Specifically, some teams may have submitted more or fewer submissions due to time, computational budget, or model performance. Having varied numbers of submissions for each team/system creates an opportunity for rethinking how to estimate how well the system actually does.

Our approach ranks teams by bootstrapping their expected rank under certain constraints. We assume that for most teams, submissions are an exploration of the hyperparameter/model configuration space of their system. Each submission's score is then informative of the distribution of its expected performance. To create the official Task rankings, we bootstrap the expected rank from all teams' submissions. Specifically, we bootstrap rankings by sampling an equal number of submissions (n=5) from the most-recent 50 submissions of each team and then use the maximum score from each team's sampled submissions to compute one ranking of all teams. To get our final ranking, we repeat this process to sample n=10,000 rankings and take the average rank for each team across these samples. In essence, this process measures the expected ranking if each team was given the same number of hyperparameter/configuration searches.

In practice, our new ranking approach largely does not change the ranking from simply ordering teams by their highest-performing submission. However, a handful of teams did shift positions. The relative stability suggests that models were not affected by different hyperparameter/configuration selections.

4 Results

The task received over 500 public submissions from over 30 participants. Next, we provide an overview of the baselines and the approaches that have been adopted by the 19 teams who participated in the competition's leaderboard and submitted a description of their systems.

4.1 Summary of the Approaches

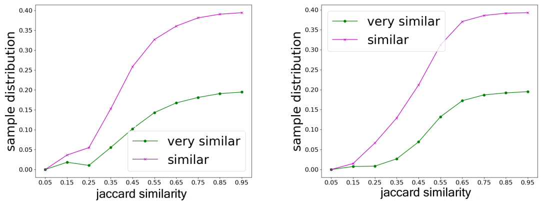

Figure 6: Sample distribution within different Jaccard similarity of named entities (left) and text (right). The "Similar" class includes both the "Very Similar" and "Somewhat Similar" labels

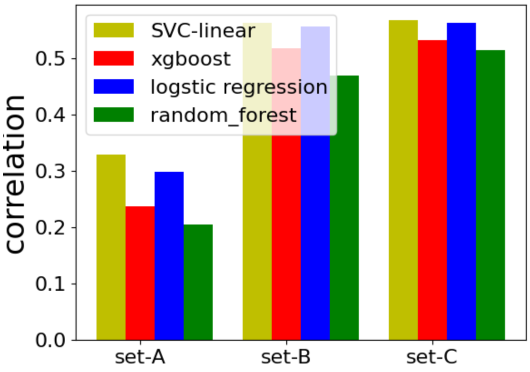

Figure 7: Baseline performances for feature sets.

The teams explored a staggering range of approaches, including multimodal systems that encode the articles' images and knowledge-based features (Zosa et al., 2022). Systems were evaluated on their ability to assess news similarity of pairs of cross-lingual news articles, and on a secondary task involving only pairs of articles in English. While some teams developed dedicated systems for the two evaluations, cross-lingual and English-only, the great majority of such systems were variations of a single design. For the sake of conciseness, in the remainder of the section we will restrict discussion to the cross-lingual news similarity task and to the best-performing systems. Table 6 reports salient characteristics of these systems, in order of their ranking using Pearson's correlation coefficient. The participant describe each system in finer detail and offer valuable insights on adapting them to the English-only subtask and on negative results (Nai et al., 2022; Wangsadirdja et al., 2022; Pisarevskaya and Zubiaga, 2022; Zosa et al., 2022; Singh et al., 2022; Giovanni et al., 2022; Hajjar et al., 2022; Chen et al., 2022; Joshi et al., 2022; Heil et al., 2022; Sandeep et al., 2022; Xu et al., 2022; Kuimov et al., 2022; Ishihara and Shirai, 2022; Luo et al., 2022; Jobanputra and Rodriguez, 2022; Dufour et al., 2022; Bhavsar et al., 2022; Stefanovitch, 2022).

TEAM	TRANSFORMERS							CROSS-LANG.		DATA HANDLING						TECHNIQUE	
	XLM-RoBERTa	**MPNet**	**SBERT**	**mBERT**	**other**	**bi- or cross-emb.**	**finetune/pre-train**	**multilang. emb.**	**translation**	**fields used**	**splices text**	**intermediate sim**	**NER**	**DA/external data**	**others**	**ensemble/stacking**	**additional tasks**
HFL	x					cross	yes	x	x	TB	x			x		x	**
GateNLP-UShef					LaBSE	bi	yes	x	x	TB			x	x		x	
DataScience-Polimi		x				bi	yes	x	x	TBD				x		x	
ITNLP2022	x				infoxlm	cross	yes	x	x	TBK				x		x	****
EMBEDDIA			x			bi	yes	x		TB				x			
HuaAMS	x					cross	yes	x	x	B				x		x	
WueDevils			x		USE	bi	no	x		TBP		x				x	
DartmouthCS				x		neither	yes	x	x	B						x	**
Nikkei				x	bert	bi	yes	x	x	TB		x		x		x	
YNU-HPCC				x		cross	yes	x		B							
SkoltechNLP					xlm-mlm	bi	yes	x		B							
Team Innovators					DeBERTa	cross	no		x	TBD	x			x			***
TCU				x		cross	yes	x	x	B							
OversampledML		x				neither	no	x	x	TB	x	x	x				**
BL.Research			x		NER-tf, BART	neither	no		x	TB	x	x	x		*		
LSX_team5			x			neither	no		x	B		x					
TMA					LASER	neither	no	x		TBDP	x	x					
dina	x					cross	no		x	B	x		x				
IIIT-MLNS					distilbert	bi	yes	x		TBDK			x	x		x	

Table 6: A summary of submitted models ordered by their performance. For each TEAM, the table reports common choices in terms of TRANSFORMER architecture, approaches for tackling CROSS-LINGUAL input, DATA HANDLING such as feature engineering and augmentation, and learning TECHNIQUE. Legend: T = title, B = body, D = description, K = keywords, P = publication date, * = sentiment, topics, geocoding, ** = 6 subdimensions, *** = semantic similarity, hyperpartisan news, **** = 3 subdimensions

4.2 Rankings and Variation Across Languages

The final rankings for the multilingual task as well as the English-language only subset are shown in Supplemental Table A1. Overall performance on the multilingual task (as measured with Pearson's r) ranged from 0.35 to 0.82 with a mean of 0.66 and a median of 0.72.

The highest single-language performance was achieved on French (median 0.84, max 0.87) and French–Polish (median 0.82, max 0.95) pairs. The worst performance was on German–French pairs (median 0.60, max 0.72). Supplemental Figure A1 shows the distribution of the best scores achieved by each team in each language.

Among the baseline models, we find that the SVC performs best, while the Jaccard similarities of named entities and text matter more than word count difference (Figure 7). However, the majority of submitted models perform significantly better than the baseline models.

4.3 Nuanced Inputs: Multiple Fields, Fine-tuning, & Feature Engineering

In addition to the main body of the articles, most systems leveraged information from multiple fields such as their titles and descriptions. All systems involved deep neural embeddings of those fields, with all but one team using Transformer-based architectures. The top-ranking system used several techniques to optimize an XLM-RoBERTa-based model without further feature engineering. *Accurately embedding multiple fields of the articles appears a crucial source of performance. Systems that engaged in fine-tuning or continued pre-training the embeddings scored higher on average.* Yet, there was no clear pattern on which architecture would produce performant representations for the task. In particular, the teams offered mixed evidence on the superiority of bi-embedding over cross-embedding approaches for the task. For example, teams Nikkei and SkoltechNLP found bi-encoders to outperform cross-encoders, whereas team HFL found the opposite (Ishihara and Shirai, 2022; Kuimov et al., 2022; Xu et al., 2022).

To improve upon the baseline of the sole embeddings (albeit often marginally), 10 teams experimented with additional feature engineering. Several teams explored forms of keyword and named-entity extraction. These approaches were arguably promising in that they mirrored the process of sampling,[8] though the results offer no conclusive ev-

[8] The sampling process was not shared with teams.

idence. Similarly, teams also tested strategies to focus the input around the most informative parts of the articles. This was due to multiple factors: first, the limitations of Transformer-based architectures which can only handle limited-length input; second, a small but consistent number of errors in automatically parsing the articles adding noise to the text; and last, the nature of the task: according to the “inverted pyramid” writing style, the start of a news article often summarizes the most important information. Thus, participants experimented with splicing the article body, which led to performance improvements.

4.4 Tackling Generalization: Multilingual, Augmentation, & Learning Strategies

A challenging characteristic of the task is the presence of cross-lingual pairs of articles—with several new language combinations introduced first in the evaluation data. The teams approached the challenge by resorting to multilingual embeddings or machine translating the articles to a high-resource language. *The best-performing systems employed a combination of both approaches, multilingual embeddings and translation, as part of a broader strategy for data augmentation.* Furthermore, the best-performing systems *resorted to forms of ensemble learning* such as stacking, which offered a further way to improve the generalization of the models (with the exceptions of teams TCU and dina (Luo et al., 2022; Pisarevskaya and Zubiaga, 2022)). With few exceptions (see Jobanputra and Rodriguez, 2022), optimizing for multiple tasks also seems to improve performance—e.g., the top-ranking system jointly learns all seven dimensions of similarity provided in the training data.

4.5 Simplicity–Performance Trade-Offs

While the best-performing systems explore sophisticated designs and techniques, the teams also suggested simpler methods that prove surprisingly effective. In fact, several teams found that simple systems outperformed more complex approaches in their experiments. A system that relies on pretrained embeddings without fine-tuning achieved a performance of 0.759 Pearson’s correlation coefficient (Wangsadirdja et al., 2022) vis-à-vis the top score of 0.818 (Xu et al., 2022). Similarly, a baseline regressing over two features—shared named entities and cosine similarity between the article embeddings—scored as high as 0.677 (Sandeep et al., 2022). Finally, several teams reported performance improvement by using lexical features without particular adaptation to the cross-lingual settings of this task (e.g., teams Nikkei & TMA: Ishihara and Shirai, 2022; Stefanovitch, 2022). In a nutshell, carefully reflecting the characteristics of the task into the system design can lead to good performance even with simpler models.

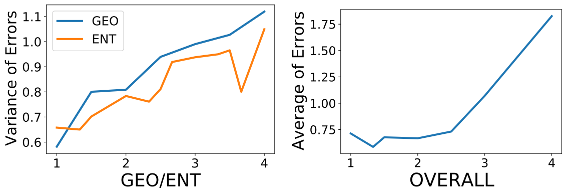

Figure 8: The variance of models’ error against pairs with the same GEO/ENT as a function of GEO/ENT (left) and an average of models’ error against pairs that were at least somewhat similar in GEO, ENT, TIME, and NAR (right).

4.6 Error Analysis

In this section, we analyze the errors of submitted models. Twenty-one teams achieved an accuracy of at least 0.70, and we focus our error analysis on these teams.

First, we compute the correlation between each model’s error (absolute difference between the predicted OVERALL and the OVERALL reported by the annotators) and the sub-dimensions (GEO, ENT, TIME, and NAR) for each pair. We find a strong Pearson correlation between the variance of errors and the GEO and ENT sub-dimensions: the correlation for GEO is 0.97, while for ENT it is 0.88 (Figure 8, left).

We hypothesized that if there is a pair with high similarity in terms of GEO, ENT, TIME, and NAR but dissimilar in terms of the OVERALL label, then models will have difficulties against this pair. To test this hypothesis, we select only pairs that are Somewhat/Very Similar in terms of GEO, ENT, TIME, *and* NAR dimensions. Then, we report how the average error of models varies for different OVERALL ratings. We expect that the average error will be higher for pairs with higher OVERALL values (i.e., pairs that are more dissimilar overall). Figure 8 (right) shows the result of this analysis. We can see that there is a strong correlation between the average of error and the OVERALL label. The Pearson correlation between the error and OVERALL for the selected pairs is 0.88.

5 Discussion

Multilingual news article similarity is a challenging problem despite sharing some characteristics with Semantic Text Similarity. Participants in this SemEval task tried a number of innovative approaches to the problem. Systems that used multiple parts of the article (headline, body, publication date) and systems that fine-tuned or otherwise trained embeddings generally performed better than those that did not. The best-performing systems generally combined multilingual embeddings and translation. Nonetheless, there was no clear consensus as to the best architectures, embedding models, or preprocessing to perform on the data.

There were clear variations across languages, and more work is needed to create multilingual systems that work across diverse language combinations. Errors were particularly common when the news articles shared some similarity in terms of their geographic focus, temporal focus, named entities, and narratives but were nonetheless dissimilar overall. While the best-submitted model achieved a very high correlation of 0.82 with gold standard labels, the best human annotator reached 0.88 correlation, which suggests ample space for further progress.[9]

Our dataset is drawn from the first half of 2020 and covers several geopolitical events (e.g., BlackLivesMatter) as well as the first wave of the COVID-19 pandemic. The nearly 10,000 annotated pairs of news articles across 18 combinations of 10 different languages will enable exciting developments in natural language processing methods as well as social science studies of how the global media reported on this unique period.

Acknowledgments

This research has received funding through the Volkswagen Foundation. We thank Media Cloud for data access. We thank the Internet Archive which made it possible for participants to all have access to the same data. We are deeply grateful to the annotators and task participants: thank you.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Nidhir Bhavsar, Rishikesh Devanathan, Aakash Bhatnagar, Muskaan Singh, and Tirthankar Ghosal. 2022. Team Innovators @ SemEval-2022 for Task 8 : Multi-Task Training with Hyperpartisian and Semantic Relation for Multi-Lingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation.*

Desmond Bala Bisandu, Rajesh Prasad, and Musa Muhammad Liman. 2018. Clustering news articles using efficient similarity measure and n-grams. *International Journal of Knowledge Engineering and Data Mining*, 5(4):333–348.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have i heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Zhongan Chen, Weiwei Chen, Yunlong Sun, Hongqing Xu, Shuzhe Zhou, Bohan Chen, Chengjie Sun, and Yuanchao Liu. 2022. ITNLP2022 at SemEval-2022 Task 8 : Pre-trained Model with Data Augmentation and Voting for Multilingual News Similarity. In *The 16th International Workshop on Semantic Evaluation.*

[9] Note that these two correlations are computed on different sets of news article pairs, since the performance of a model is estimated on the evaluation set, while the correlation score of humans is computed on all their annotations.

Teun A van Dijk. 1988. *News as discourse*. University of Groningen.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194.

Sébastien Dufour, Mohamed Mehdi Kandi, Karim Boutamine, Camille Gosset, Mokhtar Boumedyen Billami, Christophe Bortolaso, and Youssef Miloudi. 2022. BL . Research at SemEval-2022 Task 8 : Using various Semantic Information to evaluate document-level Semantic Textual Similarity. In *The 16th International Workshop on Semantic Evaluation*, pages 1–8.

Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320.

Marco Di Giovanni, Thomas Tasca, and Marco Brambilla. 2022. DataScience-Polimi at SemEval-2022 Task 8 : Stacking Language Models to Predict News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Joseph Hajjar, Weicheng Ma, and Soroush Vosoughi. 2022. DartmouthCS at SemEval-2022 Task 8: Predicting Multilingual News Article Similarity with Meta-Information and Translation. In *The 16th International Workshop on Semantic Evaluation*.

Stefan Heil, Karina Kopp, Konstantin Kobs, Albin Zehe, and Andreas Hotho. 2022. LSX _ team5 at SemEval-2022 Task 8 : Multilingual News Article Similarity Estimation based on Pre-Trained Transformers , ConceptNet Embeddings and Word Mover ' s Distance. In *The 16th International Workshop on Semantic Evaluation*.

Shotaro Ishihara and Hono Shirai. 2022. Nikkei at SemEval-2022 Task 8 : Exploring BERT-based Bi-Encoder Approach for Pairwise Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Mayank Jobanputra and Lorena Martin Rodriguez. 2022. OversampledML at SemEval-2022 Task 8 : When multilingual news similarity met Zero-shot approaches. In *The 16th International Workshop on Semantic Evaluation*.

Sagar Joshi, Dhaval Taunk, and Vasudeva Varma. 2022. IIIT-MLNS at SemEval-2022 Task 8 : Siamese Architecture for Modeling Multilingual News Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Martin Josifoski, Ivan S. Paskov, Hristo S. Paskov, Martin Jaggi, and Robert West. 2019. Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM.

Christian Klein and Matías Martínez. 2009. Wirklichkeitserzählungen. felder, formen und funktionen nicht-literarischen erzählens. In *Wirklichkeitserzählungen*, pages 1–13. Springer.

Mikhail Kuimov, Daryna Dementieva, and Alexander Panchenko. 2022. SkoltechNLP at SemEval 2022 Task 8 : Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Michael D Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the annual meeting of the cognitive science society*, volume 27.

Xiang Luo, Yanqing Niu, and Boer Zhu. 2022. TCU at SemEval-2022 Task 8 : A Stacking Ensemble Transformer Model for Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Maxwell McCombs. 2005. A look at agenda-setting: Past, present and future. *Journalism studies*, 6(4):543–557.

Ben Miller, Jennifer Olive, Shakthidhar Gopavaram, and Ayush Shrestha. 2015. Cross-document non-fiction narrative alignment. In *Proceedings of the First Workshop on Computing News Storylines*, pages 56–61.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Zihan Nai, Jin Wang, and Xuejie Zhang. 2022. YNU-HPCC at SemEval-2022 Task 8 : Transformer-based Ensemble Model for Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 321–330.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Dina Pisarevskaya and Arkaitz Zubiaga. 2022. Team dina at SemEval-2022 Task 8 : Enhancing Pre-trained Language Models for Semantic Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M. Malik, Cindy Bishop, Emily Ndulue, Aashka Dave, Justin Clark, Bruce Etling, Rob Faris, Anushka Shah, Jasmin Rubinovitz, Alexis Hope, Catherine D'Ignazio, Fernando Bermejo, Yochai Benkler, and Ethan Zuckerman. 2021. Media Cloud: Massive Open Source Collection of Global News on the Open Web. In *Proceedingsofthe Fifteenth InternationalAAAIConferenceonWeb andSocial Media (ICWSM2021)*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jan Rupnik, Andrej Muhic, Gregor Leban, Primoz Skraba, Blaz Fortuna, and Marko Grobelnik. 2016. News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, 55:283–316.

Sai Sandeep, Sharma Chittilla, and Talaat Khalil. 2022. HuaAMS at SemEval-2022 Task 8 : Combining Translation and Domain Pre-training for Cross-lingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Iknoor Singh, Yue Li, Melissa Thong, and Carolina Scarton. 2022. GateNLP-UShef at SemEval-2022 Task 8 : Entity-Enriched Siamese Transformer for Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Nicolas Stefanovitch. 2022. Team TMA at SemEval-2022 Task 8 : Lightweight and Language-Agnostic News Clustering. In *The 16th International Workshop on Semantic Evaluation*.

Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology*, 77(4):660–679.

Dirk Wangsadirdja, Felix Heinickel, Simon Trapp, Albin Zehe, Konstantin Kobs, and Andreas Hotho. 2022. WueDevils at SemEval-2022 Task 8 : Multilingual News Article Similarity via Pair-Wise Sentence Similarity Matrices. In *The 16th International Workshop on Semantic Evaluation*.

Steve J Westerman, Timothy Cribbin, and Julie Collins. 2010. Human assessments of document similarity. *Journal of the American Society for Information Science and Technology*, 61(8):1535–1542.

Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. HFL at SemEval-2022 Task 8 : A Linguistics-inspired Regression Model with Data Augmentation for Multilingual News Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Elaine Zosa, Emanuela Boros, Boshko Koloski, and Lidia Pivovarova. 2022. EMBEDDIA at SemEval-2022 Task 8 : Investigating Sentence , Image , and Knowledge Graph Representations for Multilingual News Article Similarity. In *The 16th International Workshop on Semantic Evaluation*.

Appendix

Team	Rank	Max Score	Mean Score
HFL	1	0.818	0.788
GateNLP-UShef	2	0.801	0.781
cyk1337	3	0.792	0.682
ITNLP2022	4	0.784	0.633
EMBEDDIA	5	0.784	0.685
L3i	6	0.783	0.688
DataScience-Polimi	7	0.790	0.656
HuaAMS	8	0.771	0.759
WueDevils	9	0.759	0.711
DartmouthCS	10	0.748	0.509
aim	11	0.748	0.686
Nikkei	12	0.743	0.718
SkoltechNLP	13	0.734	0.596
Andi	14	0.726	0.723
Team Innovators	15	0.733	0.690
BUT	16	0.726	0.588
sebduf	17	0.706	0.701
BL.Research	18	0.703	0.688
OversampledML	19	0.701	0.679
TCU	20	0.715	0.511
Ormus	21	0.701	0.567
LSX_team5	22	0.572	0.572
dina	23	0.507	0.228
Elena_Shu	24	0.492	0.332
naizihan	25	0.475	0.411
TMA	26	0.507	0.352
IIIT-MLNS	27	0.441	0.301
rahul19266	28	0.350	0.268
EAS	29	0.391	0.163

(a) Multilingual Setting

Team	Rank	Max Score	Mean Score
HFL	1	0.872	0.839
EMBEDDIA	2	0.855	0.704
L3i	3	0.855	0.786
WueDevils	4	0.857	0.822
DataScience-Polimi	5	0.873	0.770
DartmouthCS	6	0.845	0.647
cyk1337	7	0.837	0.725
ITNLP2022	8	0.833	0.777
aim	9	0.839	0.773
GateNLP-UShef	10	0.833	0.813
SkoltechNLP	11	0.871	0.716
BL.Research	12	0.828	0.820
sebduf	13	0.824	0.821
OversampledML	14	0.814	0.794
Team Innovators	15	0.829	0.764
HuaAMS	16	0.804	0.792
naizihan	17	0.783	0.676
BUT	18	0.779	0.685
Andi	19	0.771	0.762
Nikkei	20	0.765	0.742
Ormus	21	0.767	0.673
TCU	22	0.755	0.743
LSX_team5	23	0.683	0.683
TMA	24	0.740	0.557
Elena_Shu	25	0.623	0.421
dina	26	0.624	0.306
EAS	27	0.659	0.346
IIIT-MLNS	28	0.542	0.350
rahul19266	29	0.366	0.299
us241077	30	0.226	0.226

(b) English-only Setting

Table A1: Rankings for each team in the official mulingual setting and in the optional English-only setting, in which one additional team participated. The final team rankings shown here were computed through the bootstrapping process described in §3.3. We additionally report the maximum and mean scores (Pearson r) for each team, which largely correspond to the same ranking as our bootstrapping process.

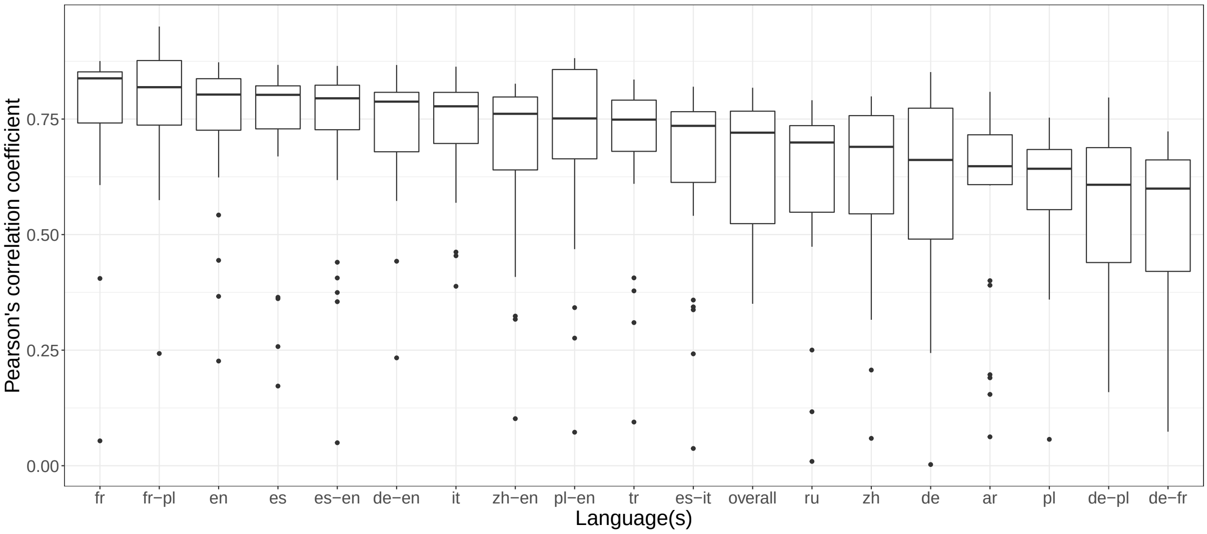

Figure A1: Distribution of the highest Pearson's correlation coefficients achieved by each team per language.