SPICED: News Similarity Detection Dataset with Multiple Topics and Complexity Levels

Elena Shushkevich[*], **Manuel V. Loureiro**[†],
Long Mai[‡], **Steven Derby**[†], **Tri Kurniawan Wijaya**[†]
[*]Technological University Dublin
elena.n.shushkevich@gmail.com

[†]Huawei Ireland Research Centre
manuel.loureiro@huawei.com, steven.derby@huawei-partners.com,
tri.kurniawan.wijaya@huawei.com

[‡]University College Dublin
long.mai@ucdconnect.ie

Abstract

The proliferation of news media outlets has increased the demand for intelligent systems capable of detecting redundant information in news articles in order to enhance user experience. However, the heterogeneous nature of news can lead to spurious findings in these systems: Simple heuristics such as whether a pair of news are both about politics can provide strong but deceptive downstream performance. Segmenting news similarity datasets into topics improves the training of these models by forcing them to learn how to distinguish salient characteristics under more narrow domains. However, this requires the existence of topic-specific datasets, which are currently lacking. In this article, we propose a novel dataset of similar news, SPICED, which includes seven topics: Crime & Law, Culture & Entertainment, Disasters & Accidents, Economy & Business, Politics & Conflicts, Science & Technology, and Sports. Futhermore, we present four different levels of complexity, specifically designed for news similarity detection task. We benchmarked the created datasets using MinHash, BERT, SBERT, and SimCSE models.

Keywords: dataset, news similarity, text similarity detection

1. Introduction

The rise of the internet has ushered in an era of unprecedented growth in online publishing, resulting in a deluge of news content. In this digital landscape, users often find themselves investing significant time and effort in navigating a multitude of articles, all covering the same events, especially within news aggregator platforms. This abundance of information can present a formidable challenge, as it becomes increasingly difficult to discern and access the specific and relevant content that users seek amidst the vast sea of news articles.

Publicly available training resources are scarce for developing systems for similar news article detection. Existing semantic textual similarity (STS) datasets are not suitable for news similarity detection, as they are specific to a single topic, such as MedSTS (Wang et al., 2020) and CORD19STS (Guo et al., 2020). However, news similarity detection is inherently influenced by the high degree of heterogeneity in news content and structure, which follows a well-understood taxonomy based on news categories or genres. These categories contain salient overlapping semantic features which directly affect how difficult it is to compare news articles, which can result in systems which quickly learn unproductive ruled-based heuristics rather than the complex linguistic structure relating these text documents (Wu et al., 2021). For example, sports news generally contains more distinctive features and less ambiguity than political news. Therefore, we need to compare news across different categories as well as within the same category, to better assess the performance of different models on different levels of similarity. To this end, high-quality datasets are crucial for improving news similarity detection in complex cases, where similarity within the same topic is harder to discern than between unrelated topics.

In this work, we propose SPICED (**S**cience, **S**ports, **P**olitics, **C**rime, **C**ulture, **E**conomy, **D**isasters), a multi-topic dataset addressing the mentioned problems. It includes Crime & Law, Culture & Entertainment, Disasters & Accidents, Economy & Business, Politics & Conflicts, Science & Technology, and Sports topics. The full dataset is publicly available[1].

By utilizing the original dataset, we propose four distinct approaches for creating pairs in the context of the news similarity detection task. Each approach offers a unique combination of true similar and false similar news pairs.

[1] You can download the dataset at https://zenodo.org/record/8044777

LREC-COLING 2024, pages 15181–15190
20-25 May, 2024.

Topics	CL	CE	DA	EB	PC	ST	SP
Document statistics							
# Webpages	4,419	2,129	2,757	2,320	7,675	2,064	2,423
# Source articles	7,495	3,759	4,716	3,881	14,075	3,681	3,738
Words per source							
Mean	606.9	518.4	544.7	605.1	629.6	662.1	579.3
Median	553	409	487	519	563	583	472
Minimum	34	26	37	31	34	42	43
Maximum	2,420	2,974	2,918	3,663	2,514	3,092	2,200

Table 1: The number of collected documents per topics and the mean, median, minimum, and maximum word counts per article for the seven topics: Crime & Law (CL), Culture & Entertainment (CE), Disasters & Accidents (DA), Economy & Business (EB), Politics & Conflicts (PC), Science & Technology (ST), Sports (SP) topics.

Our contributions are as follows:

- We provide an original dataset of 977 similar news pairs in English (1,954 news articles), devoted to the seven different popular news topics.

- We provide 32 datasets, all derived from an original gold-standard dataset. These datasets represent four different complexity levels for creating news pairs within the context of both single-topic and multi-topic similar news detection.

- We benchmark these created datasets using four algorithms for STS tasks which are prevalent in the literature: MinHash, BERT, SBERT, and SimCSE.

2. Related Work

Finding a suitable dataset for the news similarity detection task is challenging, as we need a dataset that uses categorized news article pairs to measure news similarity, but to the best of our knowledge there is no such dataset. However, there are still other datasets that are valuable for this task.

SemEval-2022 Task 8 is a multilingual news article similarity dataset of approximately 10,000 news article pairs encompassing 18 combinations of 10 languages (Chen et al., 2022). The similarity scores result from the averages of multiple human annotators using a four-point Likert scale over seven dimensions measuring geographic, temporal, and narrative similarities. Notice that this dataset does not contain any information regarding the classification of news articles over some taxonomy.

SentEval is a toolkit to evaluate the quality of universal sentence representations over various tasks, such as binary and multi-class classification, natural language inference, and sentence similarity (Conneau and Kiela, 2018) allowing the evaluation of sentence embeddings as features for many semantic textual similarity downstream tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017). The data for these datasets were collected from news articles, forum discussions, news conversations, headlines, image and video descriptions, and were labeled with a similarity score between 0 and 5. The target of the tasks was to evaluate the distance between two sentences using cosine distance (Rahutomo et al., 2012). SentEval tasks included some subtasks and reported both the average and the weighted average (by number of samples in each subtask) of the Pearson (Freedman et al., 2020) and Spearman (Zar, 2005) correlations.

Entailment and semantic relatedness detection tasks can also be useful in the context of similar news detection. In this area, it is important to mention the SICK dataset (Marelli et al., 2014), which consists of about 10,000 English sentence pairs, each of which is annotated for both entailment task – SICK-E (with three possible labels: entailment, contradiction, and neutral) and relatedness detection task – SICK-R (with 5-point rating scale as gold score).

In addition, the paraphrase detection task is comparable with the similar news detection task, where we highlight the Microsoft Research Paraphrase Corpus (MRPC) (Dolan et al., 2004; Dolan and Brockett, 2005), for which the goal is to identify if two sentences are variations based on synonymy and local syntactic changes. MRPC is a monolingual dataset presenting 5,801 naturally occurring paraphrase pairs extracted from over 13 million sentence pairs collected from the World Wide Web using heuristic techniques and an automatic classifier, with 67% of the paraphrase pairs deemed semantically equivalent by human annotators.

Topics	CL	CE	DA	EB	PC	ST	SP
Filters							
SimHash	76,996	8,672	24,015	30,291	123,791	8,916	14,954
Source of the same Wikinews page	511	259	316	312	822	273	334
SBERT	501	230	300	279	779	249	318
Experts' annotation	238	95	137	120	361	136	94
Duplicates removal	192	90	124	107	259	111	94

Table 2: The number of similar pairs after each sequential filtering step for Crime and Law (CL), Culture and Entertainment (CE), Disasters and Accidents (DA), Economy and Business (EB), Politics and Conflicts (PC), Science and Technology (ST), Sports (SP) topics. As we move down through each filtering step the number of articles is reduced to build our gold-standard dataset.

3. Dataset creation

This section is dedicated to discussing the process of creating the news article similarity dataset. This dataset is comprised of paired articles, each associated with a binary similarity label, serving as a fundamental resource for our research.

3.1. Collecting News Articles

WikiNews[2], a collaborative journalism initiative under the Wikimedia Foundation, adheres to specific guidelines[3] for news article creation. These guidelines stipulate that news articles must be categorized by topic and substantiated by a minimum of two independent and authoritative sources. Our selection process exclusively considered sources with valid and accessible URLs. Given that these sources collectively cover the same news events and provide substantial information, they can be reasonably deemed similar. This alignment in content served as the basis for the development of the proposed news article similarity dataset.

In the month of April 2022, our data collection efforts were directed towards gathering WikiNews articles, employing the utility of *BeautifulSoup*[4], a versatile web scraping tool. Specifically, we focused on harvesting articles belonging to the seven most populous categories, namely: Crime & Law, Culture, Disasters & Accidents, Economy & Business, Politics & Conflicts, Science & Technology, and Sports. The statistics detailing the outcome of this data collection process are documented in Table 1.

3.2. Measuring Similar News

We begin by utilizing baseline similarity models in order to query news article similarity, as a way to supplement our raw data. Because there are a combinatorically impractical number of possible document pairs, we offload some of this work to these oracles, as they enable us to efficiently identify suitable examples.

The *SimHash* algorithm[5] (Sadowski and Levin, 2007) is employed to identify pairs of news articles with high similarity. Determining similarity or dissimilarity is based on a threshold specified within the SimHash implementation. Subsequently, a validation process is conducted to ensure that both news articles in a pair originate from the same WikiNews webpage.

Next, for the subset of similar news articles (according to the SimHash filtering step) originating from the same WikiNews webpage, we utilize the transformer-based model *SBERT* (Reimers and Gurevych, 2019), specifically the paraphrase-multilingual-mpnet-base-v2[6] model, to identify the most similar news articles within the dataset. The approach of creating SimHash pairs separately for each topic is applied consistently.

3.3. Dataset Annotation

Experts review and assess rudimentary approximated news pairs to gather suitable samples for our final gold-standard annotations. Two annotators evaluate all proposed pairs and reach a consensus on their similarity. The annotators are two PhD students. Before commencing work on the dataset, these annotators, in collaboration with the authors, participated in joint reviews of several external news articles to establish a shared understanding of the task at hand. They resolve discrepancies through discussion to determine whether to retain or discard the pair. Disagreements on the annotations

[2] https://en.wikinews.org/wiki/Main_Page

[3] https://en.wikinews.org/wiki/Wikinews:Pillars_of_writing

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5] https://github.com/1e0ng/simhash

[6] We use paraphrase-multilingual-mpnet-base-v2.

were addressed on a case-by-case basis, with each annotator presenting their perspective to reach a consensus.

For experts annotation, we define the following criteria that any similar pair of news articles must satisfy:

1. Both news articles in a pair must be about the same topic and event (for example, topic – sports, event – UEFA Champions League final);
2. Both news articles should have similar lengths to avoid information asymmetry, where one article contains significantly more information than the other;
3. Opinion articles, prone to biases, should be excluded from similar news classifications. Similar news should be factual and not influenced by the authors' interpretations;
4. Any numerical values cited in the articles should be consistent. For example, if one article mentions 10 road accident victims and its pair states "more than 8 people," they should still be considered similar;
5. The time of publication must be close. News articles discussing the same event but published at significantly different times are considered dissimilar.

The last step of the filtering is to delete duplicate pairs, which can appear in cases when news articles are devoted to several topics at once. For example, a news article can be both about politics and economics. As we strive for the most balanced dataset, we remove the pair from the topic with the bigger number of samples.

In the Table3, some examples of similar news are presented. All four news examples are about events in Australia, but the first pair pertains to Sports topic, while the second pair relates to Disasters & Accidents topic.

3.4. Statistics

In Table 2 we present the number of similar news pairs at the end of each filtering step, including SimHash, the confirmation that both news in the pair is the sources of the same page, SBERT, experts' annotation, and duplicates removal. To sum up, after the news scraping we implemented sequentially four steps to confirm the news pairs' similarity, including both ML approaches (SimHash, SBERT) and manual checks (2 experts annotations). The number of similar pairs after the final step - duplicates removal - is the final number of similar pairs for each topic. We divided the datasets with a ratio of 70:30 for training and test datasets, and we used this ratio for the pairs from the general dataset for each complexity level creation. As shown in Table 1, the average number of tokens in the article ranges from 518.4 to 662.1 tokens, while the maximum can reach 3663 tokens on the EB topic. The final dataset contains 977 similar pairs. The dataset contains a much larger number of sentences than other similar datasets, which makes the task of computing their similarity more challenging. The dataset also covers a wide range of token lengths. This is beneficial for developing similarity models that can handle news articles of varying lengths.

3.5. Complexity Levels

We present an additional contribution and novelty by introducing several complexity levels within our datasets.

For a more detailed analysis of the created dataset, we divide it into four complexity levels, which allows us to explore not only different news topics, but also their degree of similarity. Table 4 provides the number of training and test instances for each dataset corresponding to each complexity levels.

Inter-Topic. This set includes similar news pairs as positive samples and dissimilar news pairs from different topics as negative samples. This approach distinguishes dissimilar pairs when they belong to different topics.

Intra-Topic. This set contains positive and negative pairs within the same topic, split into seven separate subsets corresponding to different topics. We also remove the challenging examples from the negative pairs, as they will belong to the next approach's datasets.

Hard Examples. This set consists of all positive pairs and the 3,000 most similar negative pairs, according to SimHash, within each intra-topic. The way the negative pairs are carefully chosen here makes it less obvious to distinguish similar news from the dissimilar ones. To ensure that there is no overlap between the negative pairs in the intra-topic and hard examples sets, we exclude these pairs from their corresponding intra-topic set. The results demonstrate that training a model on partitioned categories provides better improvements than hard-mining examples, though we note the number of hard examples are smaller.

Combined. While the set of news pairs in the previous three complexity levels are disjoint, this *combined* set includes all (union) of the positive and negative news pairs from the previous sets.

Thus, we have our original gold-standard dataset containing 977 pairs that belong to a variety of categories (Crime & Law, Culture & Entertainment, Disasters & Accidents, Economy & Business, Politics & Conflicts, Science & Technology, and Sports),

News 1	News 2
Sydney FC penalised for contract breach Sydney FC is in greater danger of missing the A-League play-offs after being hit with a deduction of three competition points for breaching Player Contracting Regulations. Football Federation Australia has also decided to fine the club $129,000. Sydney won't lose the points immediately and has seven days to appeal.	Sydney FC lose points Sydney FC will have three competition points deducted and be fined $129,000 after Football Federation Australia (FFA) today found the club guilty of breaching the A-League's Player Contracting Regulations. Sydney won't lose the points immediately and has seven days to appeal. If the club chooses not to appeal, the points will be deducted next Friday.
Great Barrier Reef oil disaster fear from stricken ship The Shen Neng 1 was nine miles (15 km) outside the shipping lane A Chinese ship is in danger of breaking up after running aground off north-east Australia, sparking fears of a major oil spill into the Great Barrier Reef. The Shen Neng 1, carrying 950 tonnes of oil, ran aground 70km (43 miles) off the east coast of Great Keppel Island. Some oil has already leaked and there are fears the coal-carrier may split into parts, causing a greater spillage. The Australian authorities say the ship was in a protected area, well outside the normal shipping channels. Chemical dispersants are being used to prevent the spill threatening the World Heritage-listed marine reserve...	Australia warns stranded Chinese ship could break up SYDNEY (Reuters) A stranded Chinese bulk coal carrier leaking oil into the sea around Australia's Great Barrier Reef is in danger of breaking up and damaging the reef, government officials said on Sunday. Oil is seen next to the 230-metre (754-ft) Chinese bulk coal carrier Shen Neng I, about 70 km (43 miles) east of Great Keppel Island April 4, 2010. The 230-meter (754-ft) Shen Neng I was on its way to China when it ran aground on a shoal on Saturday. It had 950 tonnes of oil on board and officials said patches of oil had been spotted in the water early on Sunday, but no major leak...

Table 3: Examples of similar news pairs.

which consists of 1954 articles. From here, we split the dataset into 679 training and 298 test pairs and create labels by generating incorrect pairs. These pairs represent a total of 1358 training articles (and 596 testing articles), resulting in 921,403 pairs for training (and 177,310 pairs for testing), encompassing all combined examples. Inter-topic pairs are those that do not belong to the same category, while intra-topic pairs are those that do. Hard examples are pairs that belong to the same category but are considered difficult because they are incorrect yet highly similar to the base article.

4. Benchmarking

This section describes the models, experiments and benchmark results of our novel similarity news dataset.

4.1. Pretrained Models

Minhash is an efficient method for estimating set similarity using the Jaccard coefficient (Broder, 1997). We used the *snapy* library[7] to obtain a simple baseline for more complex algorithms.

BERT (Bidirectional Encoder Representations from Transformers) represents a classical approach employed in this study for obtaining embeddings of news articles (Devlin et al., 2018). Subsequently, cosine similarities are calculated between these embeddings to gauge their semantic similarity. For this purpose, we leveraged the widely recognized *BERT-base-uncased* model, which is accessible through the Hugging Face Model Hub[8]. This model serves as a crucial component in the process of deriving meaningful representations of the news articles, facilitating the assessment of their similarity in a comprehensive manner.

[7] https://libraries.io/pypi/snapy

[8] https://huggingface.co/bert-base-uncased

Model	Train	Test
Inter Topic		
All	767,587	148,382
Intra Topic		
Crime & Law (CL)	33,678	5,770
Culture & Entertainment (CE)	5,526	640
Disaster & Accidents (DA)	12,606	1,950
Economy & Business (EB)	8,778	1,245
Politics & Conflict (PC)	63,241	11,190
Science & Technology (ST)	9,681	1,378
Sporting Activities (SP)	6,285	753
Hard Examples		
Crime & Law (CL)	2,234	958
Culture & Entertainment (CE)	2,162	928
Disaster & Accidents (DA)	2,186	938
Economy & Business (EB)	2,174	933
Politics & Conflict (PC)	2,281	978
Science & Technology (ST)	2,177	934
Sporting Activities (SP)	2,165	929
Combined		
All	921,403	177,310

Table 4: The number of train and test instances (news pairs) for each topic and complexity level.

SBERT (Sentence-BERT) represents a modification of the pretrained BERT architecture (Reimers and Gurevych, 2019). This variant employs siamese and triplet network structures to produce semantically rich sentence embeddings. These embeddings are designed to be directly comparable using cosine similarity, enabling effective semantic similarity assessments between sentences. In our experiments, we employed the *all-mpnet-base-v2* model[9].

SimCSE, is an innovative contrastive learning framework (Gao et al., 2021) that has consistently outperformed BERT-base on multiple Semantic Textual Similarity (STS) datasets, showcasing its robustness and effectiveness.

In our study, we employ *bert-base-uncased* model as the encoder for training SimCSE. To ensure the fidelity and reliability of our experiments, we utilize the original implementation of SimCSE provided by the authors, which is available at the following repository[10]. This decision aligns with the best practices in the field and contributes to the credibility of our findings.

In the experimental phase involving BERT, SBERT, and SimCSE, we proceed by generating embeddings for each news article and subsequently calculated the cosine similarity between these embeddings. Our aim was to identify the optimal threshold value. Conversely, in the case of MinHash, we began by generating signatures for each news article, and our threshold selection process involved the application of the Jaccard similarity coefficient to determine the most suitable threshold value.

4.2. Experiment Configurations

We systematically explore a comprehensive range of thresholds, spanning from 0 to 1, with increments of 0.01. Our objective is to identify the threshold that yields the highest F1-score on the training dataset, as this threshold selection is critical for optimal performance.

Once the threshold is determined, we utilize it to calculate the F1-score on the testing dataset, ensuring that our evaluation is consistent with the chosen threshold.

These experiments are conducted using a dedicated setup, leveraging a single V100-32GB GPU for BERT, SBERT, SimCSE, and 72 CPU cores for Minhash computations. The computational intensity of these experiments is noteworthy, with each model requiring approximately 12 hours to complete computations across all levels and topics.

[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[10] https://github.com/princeton-nlp/SimCSE

Model	MinHash	BERT	SBERT	SimCSE
Inter-Topic				
F1-score	*0.707*	*0.786*	*0.920*	*0.896*
Intra-Topic				
Crime & Law (CL)	0.816	0.851	0.957	0.957
Culture & Entertainment (CE)	0.902	0.923	0.923	0.943
Disaster & Accidents (DA)	0.742	0.853	0.935	0.853
Economy & Business (EB)	0.678	0.828	0.899	0.937
Politics & Conflict (PC)	0.650	0.776	0.911	0.875
Science & Technology (ST)	0.690	0.824	0.921	0.847
Sporting Activities (SP)	0.840	0.840	0.982	0.816
Average F1-score	*0.760*	*0.842*	*0.933*	*0.890*
Hard Examples				
Crime & Law (CL)	0.727	0.891	0.935	0.919
Cultu921403-re & Entertainment (CE)	0.833	0.906	0.902	0.943
Disaster & Accidents (DA)	0.742	0.795	0.938	0.868
Economy & Business (EB)	0.690	0.774	0.952	0.909
Politics & Conflict (PC)	0.702	0.829	0.940	0.892
Science & Technology (ST)	0.741	0.639	0.853	0.667
Sporting Activities (SP)	0.840	0.840	0.945	0.964
Average F1-score	*0.754*	*0.811*	*0.924*	*0.880*
Combined				
F1-score	*0.757*	*0.799*	*0.922*	*0.875*

Table 5: F1-scores of the experiments on various topics and complexity levels.

4.3. Results

We conducted comprehensive experiments using all the models previously described, evaluating them on datasets generated through four distinct approaches: Intra Topic, Inter Topic, Hard Examples, and Combined. The outcomes of the experiments are presented in Table 5.

Inter-Topic. SBERT achieved the highest F1-score at 0.920, followed by SimCSE at 0.896, outperforming MinHash, which achieved an F1-score of 0.707. These results suggest that determining similarity within news pairs, particularly those created using the inter-topic approach, poses a greater challenge compared to identifying similar news when utilizing datasets from the other approaches.

Intra-Topic. In intra-topic experiments, SimCSE achieved an F1-score of 0.890 in intra-topic evaluations, slightly lower than its inter-topic performance. Meanwhile, MinHash, BERT, and SBERT demonstrated varying results between the two settings. MinHash achieved an F1-score of 0.760 in intra-topic experiments, BERT scored 0.842, and SBERT outperformed with an F1-score of 0.933 in intra-topic assessments, making it the top performer within the same thematic category.

Hard Examples. In comparison to the intra-similarity approach, the hard examples approach yields lower results across all models. Among these models, SBERT demonstrates the strongest performance, achieving an impressive F1-score of 0.924. Following closely behind are SimCSE, BERT, and MinHash, with the latter displaying the lowest average performance among the group.

Combined. In the context of the combined approach, it's worth noting that SBERT stands out with the highest F1-score, 0.922. This performance aligns closely with the outcomes seen in the other approaches we've explored. On the opposite end of the spectrum, MinHash delivers the lowest F1-score, registering at 0.757, and this pattern remains consistent across the various approaches.

For all four approaches to pair creation, MinHash demonstrated lower results than BERT, SBERT, and SimCSE, highlighting the advantage of using embeddings compared to the more classical MinHash approach. Additionally, SBERT achieved the highest F1-score in all cases, which may be attributed to our use of SBERT in the news filtering process. In second place was SimCSE, which showed better results than BERT for all types of pairs, confirming the advantage of using SimCSE over BERT on the Semantic Textual Similarity (STS) datasets (Gao et al., 2021).

5. Conclusion and Future Work

In this paper, we propose a novel semantic textual similarity dataset for news data which accounts for emergent semantic categories within the text. While there are a wide variety of available textual similarity datasets, they fail to account for structural patterns that exist within text, which are generally easier for machine learning systems to classify.

We meticulously select and curate examples that pertain to distinct news topics, with the aim of creating more complex textual pairs for our study. This process led to the development of a total of 32 training and test datasets for news similarity detection. These datasets are organized based on four distinct approaches for generating news pairs: Inter-Topic Similarity, Intra-Topic Similarity, Hard Example Mining, and Combined Similarity.

The experimental results indicate that our dataset poses a significantly greater challenge for state-of-the-art models, underscoring the inherent difficulty of the task at hand. In the spirit of advancing research in this domain, we make this dataset readily accessible to the wider community. Our primary objective is to contribute valuable resources that can be instrumental in enhancing the performance and capabilities of future models.

In terms of future work, our primary objective is to expand the dataset's scope. We aim to transform it into a comprehensive resource, encompassing not only multiple topics but also multiple languages, thus fostering cross-lingual analysis and training/testing with news data in various domains.

Additionally, we intend to conduct a thorough comparative analysis, pitting our dataset against existing ones like SemEval-2022. This comparison will help us gauge the interchangeability and compatibility of these datasets.

6. Ethics Statement

The dataset used in this study comprises articles sourced exclusively from Wikinews, a publicly accessible news platform. The models employed in the dataset creation are likewise publicly available online. During the labeling process, annotators adhered to the detailed rules we specified. Consequently, it is important to note that the dataset presented here does not incorporate any confidential personal information.

The models used for benchmarking are likewise accessible online. As part of our unwavering dedication to transparency, we have made available the complete dataset we generated, along with the outcomes achieved by all the models used in the benchmarking process.

7. Bibliographical References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval@ COLING*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

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.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Luke Barrington, Antoni Chan, Douglas Turnbull, and Gert Lanckriet. 2007. Audio information retrieval using semantic similarity. In *2007 IEEE*

International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, volume 2, pages II–725–II–728.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Kaibo Cao, Chunyang Chen, Sebastian Baltes, Christoph Treude, and Xiang Chen. 2021. Automated query reformulation for efficient search based on query logs from stack overflow. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1273–1285. IEEE.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

David Freedman, Robert Pisani, and Roger Purves. 2020. Statistics: Fourth international student edition. *W. W. Norton & Company. https://www. amazon. com/Statistics-Fourth-International-Student-Freedman/dp/0393930432. Accessed*, 22.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Xiao Guo, Hengameh Mirzaalian, Ekraam Sabir, Ayush Jaiswal, and Wael Abd-Almageed. 2020. Cord19sts: Covid-19 semantic textual similarity dataset. *arXiv preprint arXiv:2007.02461*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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.

Caitlin Sadowski and Greg Levin. 2007. Simhash: Hash-based similarity detection.

Taneeya Satyapanich, Hang Gao, and Tim Finin. 2015. Ebiquity: Paraphrase and semantic similarity in twitter using skipgrams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 51–55.

Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *arXiv preprint arXiv:2109.04380*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.