Supplementary Materials

S-1 Methods to Associate Articles

Figure 6 in the main text illustrates the full article association procedure.

First, we used a rule-based algorithm using associate article bounding boxes that are under the same headline, as these are part of the same article with extremely high probability. Algorithm 1 gives pseudocode for this method. We set the parameters as $P_S = 100$, $P_T = 20$, $P_B = 50$.

For training data, where we want article pairs that are not only part of the same article, but also where they appear in the given order, we further narrow down the pairs. Specifically, we use only those pairs which are horizontally next to each other, and which have no other bounding boxes below them, as for these pairs, we can guarantee that the pair of bounding follow directly after one another (whereas for other article bounding boxes that share a headline, there may be a third bounding box in between). Algorithm 2 shows pseudocode for this procedure, and we used $P_C = 5$, and it is further illustrated in panel A of figure 6 in the main text.

For hard negatives, we used article boxes under the same headline in reverse reading order (right to left). For standard negatives, we took pairs of articles on the same page, where B was above and to the left of A, as articles do not read from right to left. One twelfth of our training data were positive pairs, another twelfth were hard negative pairs and the remainder were standard negative pairs. This outperformed a more balanced training sample.

We use this dataset to finetune a cross-encoder using a RoBERTa base model [4]. We used a Bayesian search algorithm [1] to find optimal hyperparameters on one tenth of our training data (limited compute prevented us from running this search with the full dataset), which led to a learning rate of 1.7e-5, with a batch size of 64 and 29.57% warm up. We trained for 26 epochs with an AdamW optimizer, and optimize a binary cross-entropy loss.

We evaluate these methods on a hand-labeled dataset of 214 scans, randomly selected from 1968 and 1955. These scans were labeled by a highly-trained undergraduate research assistant. Summary statistics of this dataset are given in table S-1 and evaluation results are given in the main text.

Scan count	Article bounding boxes	Headline bounding boxes	Article-article associations
214	3,803	2,805	1,851

Table S-1: Descriptive statistics of article association training data.

S-2 Methods to Detect Reproduced Content

To detect reproduced content, we use the contrastively trained bi-encoder model developed by [6], which is trained to learn similar representations for reproduced articles and dissimilar representations for non-reproduced articles. This model is based on an S-BERT MPNET model [5, 7] and is fine-tuned on a hand-labelled dataset of articles from the same underlying wire source, using S-BERT's online contrastive loss [3] implementation, with a 0.2 margin and cosine similarity as the distance metric. The learning rate is 2e-5 with 100% warm up and a batch size of 32. It uses an AdamW optimizer, and the model is trained for 16 epochs. This bi-encoder is trained and evaluated on a hand-labeled dataset, which is detailed in S-2. The results of this evaluation are given in the main text.

To create clusters from the bi-encoder embeddings, we use highly scalable single-linkage clustering, with a cosine similarity threshold of 0.94. We build a graph using articles as nodes, and add edges if the cosine similarity is above this threshold. As edge weights we use the negative exponential of the difference in dates (in days) between the two articles. We then apply Leiden community detection to the graph to control false positive edges that can otherwise merge disparate groups of articles.

Algorithm 1 Rule-based association of article bounding boxes

INPUT: $b_1, ..., b_n \in B$: set of bounding boxes that appear on the same scan, with their coordinates, denoted $left(b_i)$, $right(b_i)$, $top(b_i)$, $bottom(b_i)$, and type (headline, article, byline etc.), denoted $type(b_i)$.

PARAMETERS:

W: width of scan

H: height of scan

P_S: fraction of width, for creating side margin

P_T: fraction of height, for creating top margin

P_B: fraction of height, for creating bottom margin

OUTPUT: $ArticleArticlePairs = \{(b_i, b_j) \in B \times B | b_i, b_j$ predicted to be part of the same full article and $type(b_i) = type(b_j) = article\}$

Initialise: $M_S = W/P_S$, the side margin, $M_B = H/P_B$, the bottom margin, $M_T = H/P_T$, the top margin, $MatchedHeadlines = \{\}$, $HeadlineArticlePairs = \{\}$, $ArticleArticlePairs = \{\}$

for all b_0 in B where $type(b_0)$ is article **do**

    Create $B_0 \subset B$ where:

        a. All bounding boxes of type byline are removed

        b. b_0 is removed

        c. All bounding boxes are removed that do not share at least M_S of the horizontal axis

        d. All bounding boxes are removed whose bottom is more than M_B below the top of b_0

        e. All bounding boxes are removed whose bottom is more than M_T above the top of b_0

    if B_0 is not empty **then**

        Let b_1 be the element of B_0 that has the lowest bottom coordinate

        if $type(b_1)$ is headline **then**

            $MatchedHeadlines = MatchedHeadlines \cup \{b_1\}$

            $HeadlineArticlePairs = HeadlineArticlePairs \cup \{(b_0, b_1)\}$

        end if

    end if

end for

for all b_h in $MatchedHeadlines$ **do**

    Let $H_1 \subset HeadlineArticlePairs$ be all pairs that contain that headline, b_h

    if H_1 has at least two elements **then**

        Let A be all the bounding boxes of type article from the pairs in H_1

        Let C be all combinations of 2 elements of A

        $ArticleArticlePairs = ArticleArticlePairs \cup C$

    end if

end for

Algorithm 2 Selection of ordered article pairs

INPUT: $ArticleArticlePairs$, B, from algorithm 1.

PARAMETERS:

P_C: fraction of column width, for creating margin

OUTPUT: $OrderedPairs \subset ArticleArticlePairs$

Initialise: $OrderedPairs = \{\}$

for p in $ArticleArticlePairs$ **do**

    Let p_l be the element of p with the furthest left coordinate

    Let p_r be the other element

    if $left(p_r)$ is not further to the right of $right(p_l)$ than $width(p_l)/F_C$ **then**

        if there are no other bounding boxes below p_l **then**

            $OrderedPairs = OrderedPairs \cup \{p\}$

        end if

    end if

end for

	Positives Pairs	Negative Pairs	Reproduced Articles	Singleton Articles	Total Articles
Training Data					
Training	36,291	37,637	891	–	7,728
Validation	3,042	3,246	20	–	283
Full Day Evaluation					
Validation	28,547	12,409,031	447	2,162	4,988
Test	54,996	100,914,159	1,236	8,046	14,211
Full Dataset	122,876	113,364,073	2,594	10,208	27,210

Table S-2: Summary statistics of training and evaluation data for detecting duplicate content.

We further remove clusters that have over 50 articles and contain articles with greater than five different dates. We also remove clusters that contain over 50 articles, when the number of articles is more than double the number of unique newspapers from which these articles are sourced. This removes clusters of content that are correctly clustered in the sense of being based on the same underlying source, but are not useful for the HEADLINES dataset. For example, an advertisement (misclassified as an article due to an article-like appearance) might be repeated by the same newspaper on multiple different dates and would be removed by these rules, or weather forecasts can be very near duplicates across space and time, forming large clusters.

S-3 A Summary of Copyright Law for Works Published in the United States

Date of Publication	Conditions	Copyright Term
Public Domain		
Anytime	Works prepared by an officer/employee of the U.S. Government as part of their official duties	None
Before 1928	**None**	**None. Copyright expired.**
1928 through 1977	**Published without a copyright notice**	**None. Failure to comply with required formalities**
1978 to 1 March 1989	**Published without notice and without subsequent registration within 5 years**	**None. Failure to comply with required formalities**
1928 through 1963	**Published with notice but copyright was not renewed**	**None. Copyright expired**
Copyrighted		
1978 to 1 March 1989	Published without notice, but with subsequent registration within 5 years	70 (95) years after the death of author (corporate author)
1928 through 1963	Published with notice and the copyright was renewed	95 years after publication
1964 through 1977	Published with notice	95 years after publication
1978 to 1 March 1989	Created after 1977 and published with notice	70 (95) years after the death of author (corporate author) or 120 years after creation, if earlier
1978 to 1 March 1989	Created before 1978 and first published with notice in the specified period	The greater of the term specified in the previous entry or 31 December 2047
From 1 March 1989 through 2002	Created after 1977	70 (95) years after the death of author (corporate author) or 120 years after creation, if earlier
From 1 March 1989 through 2002	Created before 1978 and first published in this period	The greater of the term specified in the previous entry or 31 December 2047
After 2002	None	70 (95) years after the death of author (corporate author) or 120 years after creation, if earlier

Table S-3: This table summarizes U.S. copyright law, based on a similar table produced by the Cornell libraries. For concision, we focus on works initially published in the United States. A variety of other cases are also covered at `https://guides.library.cornell.edu/copyright`.

S-4 Dataset Information

S-4.1 Dataset URL

HEADLINES can be found at https://huggingface.co/datasets/dell-research-harvard/headlines-semantic-similarity.

This dataset has structured metadata following schema.org, and is readily discoverable.[1]

S-4.2 DOI

The DOI for this dataset is: 10.57967/hf/0751.

S-4.3 License

HEADLINES has a Creative Commons CC-BY license.

S-4.4 Dataset usage

The dataset is hosted on huggingface, in json format. Each year in the dataset is divided into a distinct file (eg. 1952_headlines.json).

The data is presented in the form of clusters, rather than pairs to eliminate duplication of text data and minimize the storage size of the datasets.

An example from HEADLINES looks like:

```
{
    "headline": "FRENCH AND BRITISH BATTLESHIPS IN MEXICAN WATERS",
    "group_id": 4
    "date": "May-14-1920",
    "state": "kansas",
}
```

The data fields are:

- headline: headline text.
- date: the date of publication of the newspaper article, as a string in the form mmm-DD-YYYY.
- state: state of the newspaper that published the headline.
- group_id: a number that is shared with all other headlines for the same article. This number is unique across all year files.

The whole dataset can be easily downloaded using the datasets library:

```
from datasets import load_dataset
dataset_dict = load_dataset("dell-research-harvard/headlines-semantic-similarity")
```

Specific files can be downloaded by specifying them:

```
from datasets import load_dataset
load_dataset(
    "dell-research-harvard/headlines-semantic-similarity",
    data_files=["1929_headlines.json", "1989_headlines.json"]
)
```

[1] See https://search.google.com/test/rich-results/result?id=_HKjxIv-LaF_8ElAarsM_g for full metadata.

S-4.5 Author statement

We bear all responsibility in case of violation of rights.

S-4.6 Maintenance Plan

We have chosen to host HEADLINES on huggingface as this ensures long-term access and preservation of the dataset.

S-4.7 Dataset documentation and intended uses

We follow the datasheets for datasets template [2].

S-4.7.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Transformer language models contrastively trained on large-scale semantic similarity datasets are integral to a variety of applications in natural language processing (NLP). A variety of semantic similarity datasets have been used for this purpose, with positive text pairs related to each other in some way. Many of these datasets are relatively small, and the bulk of the larger datasets are created from recent web texts; e.g. positives are drawn from the texts in an online comment thread or duplicate questions in a forum. Relative to existing datasets, HEADLINES is very large, covering a vast array of topics. This makes it useful generally speaking for semantic similarity pre-training. It also covers a long period of time, making it a rich training data source for the study of historical texts and semantic change. It captures semantic similarity directly, as the positive pairs summarize the same underlying texts.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
HEADLINES was created by Melissa Dell and Emily Silcock, at Harvard University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset was funded by the Harvard Data Science Initiative, Harvard Catalyst, and compute credits provided by Microsoft Azure to the Harvard Data Science Initiative.

Any other comments?
None.

S-4.7.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

HEADLINES comprises instances of newspaper headlines and relationships between them. Specifically, each headline includes information on the text of the headline, the date of publication, and the state it was published in. Headlines have relationships between them if they are semantic similarity pairs, that is, if they two different headlines for the same newspaper article.

How many instances are there in total (of each type, if appropriate)?
HEADLINES contains 34,867,488 different headlines and 396,001,930 positive relationships between headlines.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Many local newspapers were not preserved, and newspapers with the widest circulation tended to renew their copyrights, so cannot be included.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each data instance consists of raw data. Specifically, an example from `HEADLINES` *is:*

```
{
    "headline": "FRENCH AND BRITISH BATTLESHIPS IN MEXICAN WATERS",
    "group_id": 4
    "date": "May-14-1920",
    "state": "kansas",
}
```

The data fields are:

- `headline`*: headline text.*
- `date`*: the date of publication of the newspaper article, as a string in the form mmm-DD-YYYY.*
- `state`*: state of the newspaper that published the headline.*
- `group_id`*: a number that is shared with all other headlines for the same article. This number is unique across all year files.*

Is there a label or target associated with each instance? If so, please provide a description.

Each instance contains a `group_id` *as mentioned directly above. This is a number that is shared by all other instances that are positive semantic similarity pairs.*

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

In some cases, the state of publication is missing, due to incomplete metadata.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between instances are made explicit in the `group_id` *variable, as detailed above.*

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The data is sourced from OCR'd text of historical newspapers. Therefore some of the headline texts contain OCR errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset does not contain information that might be viewed as confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The headlines in the dataset reflect diverse attitudes and values from the period in which they were written, 1920-1989, and contain content that may be considered offensive for a variety of reasons.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Many news articles are about people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not specifically identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

If an individual appeared in the news during this period, then headline text may contain their name, age, and information about their actions.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

All information that it contains is already publicly available in the newspapers used to create the headline pairs.

Any other comments?

None.

S-4.7.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

To create HEADLINES, we digitized front pages from off-copyright newspapers spanning 1920-1989. Historically, around half of articles in U.S. local newspapers came from newswires like the Associated Press. While local papers reproduced articles from the newswire, they wrote their own headlines, which form abstractive summaries of the associated articles. We associate articles and their headlines by exploiting document layouts and language understanding. We then use deep neural methods to detect which articles are from the same underlying source, in the presence of substantial noise and abridgement. The headlines of reproduced articles form positive semantic similarity pairs.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

These methods are described in detail in the main text and supplementary materials of this paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset was not sampled from a larger set.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
We used student annotators to create the validation sets for associating bounding boxes, and the training and validation sets for clustering duplicated articles. They were paid $15 per hour, a rate set by a Harvard economics department program providing research assistantships for undergraduates.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The headlines were written between 1920 and 1989. Semantic similarity pairs were computed in 2023.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, this dataset uses entirely public information and hence does not fall under the domain of Harvard's institutional review board.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Historical newspapers contain a variety of information about people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data were obtained from off-copyright historical newspapers.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Individuals were not notified; the data came from publicly available newspapers.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The dataset was created from publicly available historical newspapers.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?
None.

S-4.7.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

See the description in the main text.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No specific software was used to clean the instances.

Any other comments?
None.

S-4.7.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?
The dataset can be used for training models for semantic similarity, studying language change over time and studying difference in language across space.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms,

legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset contains historical news headlines, which will reflect current affairs and events of the time period in which they were created, 1920-1989, as well as the biases of this period.

Are there tasks for which the dataset should not be used? If so, please provide a description.

It is intended for training semantic similarity models and studying semantic variation across space and time.

Any other comments?
None

S-4.7.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is available for public use.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is hosted on huggingface. Its DOI is 10.57967/hf/0751.

When will the dataset be distributed?
The dataset was distributed on 7th June 2023.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is distributed under a Creative Commons CC-BY license. The terms of this license can be viewed at `https://creativecommons.org/licenses/by/2.0/`

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

There are no third party IP-based or other restrictions on the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No export controls or other regulatory restrictions apply to the dataset or to individual instances.

Any other comments?
None.

S-4.7.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted on huggingface.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The recommended method of contact is using the huggingface 'community' capacity. Additionally, Melissa Dell can be contacted at melissadell@fas.harvard.edu.

Is there an erratum? If so, please provide a link or other access point.

There is no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no plans to update the dataset. If we do, we will notify users via the huggingface Dataset Card.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no applicable limits on the retention of data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We have no plans to update the dataset. If we do, older versions of the dataset will not continue to be hosted. We will notify users via the huggingface Dataset Card.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others can contribute to the dataset using the huggingface 'community' capacity. This allows for anyone to ask questions, make comments and submit pull requests. We will validate these pull requests. A record of public contributions will be maintained on huggingface, allowing communication to other users.

Any other comments?

None.

References

[1] FALKNER, S., KLEIN, A., AND HUTTER, F. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning* (10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 1437–1446.

[2] GEBRU, T., MORGENSTERN, J., VECCHIONE, B., VAUGHAN, J. W., WALLACH, H., AU2, H. D. I., AND CRAWFORD, K. Datasheets for datasets, 2021.

[3] HADSELL, R., CHOPRA, S., AND LECUN, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 1735–1742.

[4] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[5] REIMERS, N., AND GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[6] SILCOCK, E., D'AMICO-WONG, L., YANG, J., AND DELL, M. Noise-robust de-duplication at scale. *International Conference on Learning Representations* (2023).

[7] SONG, K., TAN, X., QIN, T., LU, J., AND LIU, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems 33* (2020), 16857–16867.