Improved Near-Duplicate Detection for Aggregated and Paywalled News-Feeds

Siddharth Tumre, Sangameshwar Patil, Alok Kumar
TCS Research
{siddharth.tumre, sangameshwar.patil, k.alok9}@tcs.com

Abstract

News aggregators play a key role in the rapidly evolving digital landscape by providing comprehensive and timely news stories aggregated from diverse sources into one feed. As these articles are sourced from different outlets, they often end up covering the same underlying event but differ in phrasing, formatting or supplemented with additional details. It is crucial for the news aggregators to identify these near-duplicates, improving the content quality and user engagement by steering away from redundant information. The problem of near-duplicate news detection has become harder with increasing use of paywalls by the news websites resulting in restricted access to the content. It is now common to get only the headline and a short snippet from the article. Previous works have concentrated on full length versions of documents such as webpages. There is very little work that focuses on this variation of the near-duplicate detection problem in which only headline and a small text blurb is available for each news article. We propose **N**ear-**D**uplicate **D**etection Using **M**etadata **A**ugmented **C**ommunities (NDD-MAC) approach that combines embeddings from pre-trained language model and latent metadata of a news article followed by community detection to identify clusters of near-duplicates. We show the efficacy of proposed approach using 2 different real-world datasets. By integrating metadata with community detection, NDD-MAC is able to detect nuanced similarities and differences in news snippets and offers an industrial scale solution for the near-duplicate detection in scenarios with restricted content availability.

1 Introduction

The digital era has brought both opportunities and challenges to the news industry. The news ecosystem has undergone significant changes, reshaping the way news is produced, distributed and consumed. News aggregator apps and portals have played a significant role in the evolution of the news industry. News aggregators[1] provide users with a one-stop platform to access news from various sources, saving time and effort[2] in browsing multiple websites or picking up physical newspapers (Lee and Chyi, 2015).

One of the key challenges faced by the news aggregators and their subscribers is redundancy due to repetitive content. Redundancy problem in news aggregators refers to the issue of users encountering duplicate or highly similar content across multiple articles within the aggregator app, web portal or the news fetched using their APIs. It can occur when the aggregators include multiple sources that all cover the same news event or topic. Many aggregator apps display content from syndicated[3] news services or wire services. These services provide the same articles to multiple news outlets. The news outlets may do a few editorial changes to the input articles. This creates some variations in the content and gives rise to near-duplicates at the news aggregator app or web-portal level.

While diversity of sources is valuable, too many similar news items from different sources can undermine the overall quality of the user experience. It affects the news consumers' engagement, retention, and perception of a news aggregator vendor's offerings. This in-turn has a potentially adverse effect on the *monetization and the financial viability* of the news aggregator app or portal itself. Further, news consumers in enterprises typically subscribe to the APIs of news aggregator vendors. These enterprises spend valuable compute and storage resources in fetching, archiving and analyzing the news they have paid for. Near-duplicate news items not only provide a cluttered user experience

[1] https://en.wikipedia.org/wiki/News_aggregator
[2] https://www.wprssaggregator.com/a-list-of-best-news-aggregators/
[3] e.g., https://en.wikipedia.org/wiki/Project_Syndicate

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 979–987

for them, but also introduces multiple inefficiencies in the enterprise infrastructure for procuring and disseminating news within their organizations. Thus, redundancy due to near-duplicate content affects the overall quality and operational efficiency of news ecosystem.

Enterprise solutions as well as the research literature for the near-duplicate detection problem have predominantly focused on input consisting of entire documents such as webpages as well as full-length news articles. However, with increasing use of paywalls by the newspaper websites and proliferation of news aggregator apps and APIs for large, enterprise-scale news procurement, it is now common to get only the headline and a small snippet of few lines from the article. As shown in Table 1, a sample news record in such news-feeds contains the headline and a snippet from the news body. For reading the full article, a reader has to follow a URL linked to the original news provider, such as a newspaper website. This *makes the problem* of near-duplicate detection *harder* compared to the previous scenario when the full body of the news article was available relatively easily. There is very little work that focuses on this variation of the near-duplicate detection problem in which only headline and a small text blurb is available for each news article.

In this paper, we propose an unsupervised approach, **N**ear-**D**uplicate **D**etection Using **M**etadata **A**ugmented **C**ommunities (NDD-MAC) to improve the efficiency for news aggregators, enterprise users, as well as the user experience for the end consumers. Using this method, we have been able to create an enterprise-wide positive impact by enabling retention and analysis of older news. Earlier this data was purged due to infrastructural and process inefficiencies. The improved system now obviates the need for data purging, provides historical continuity and empowers business analysts to observe evolution of events across longer timelines and refine their insights with contextually richer evidence.

Rest of the paper is organized as follows. In Section §2 we describe NDD-MAC approach and show how it can be used for the problem of near-duplicate news detection. Sections §3 covers the experimental setup and results. In Section §4, we briefly describe the related work. Finally, we conclude in Section §5.

Table 1: Real-life news snippets illustrating benefit of metadata for near-duplicate detection. (To avoid clutter, only key portions are highlighted.)

ID	Headline	Text
1	Time Warner, Comcast enter cable pact.	Time Warner Inc. and Comcast Corp. agreed to a deal on Monday giving Comcast an option to cut its stake in Time Warner's cable unit, opening the door for Comcast to unwind the entire partnership.
2	Comcast, Time Warner announce financial deal.	Comcast Corp. and Time Warner Inc. on Monday announced an agreement on what could be the first step of giving Comcast a way to redeem its stake in Time Warner Cable Inc.
3	Comcast and Time Warner Mulling Bid for Adelphia.	The Comcast Corporation confirmed today that it was in talks with Time Warner Inc. to make a joint bid for Adelphia Communications.
4	2 Cable Giants Set To Bid for Adelphia.	Comcast Corp. and Time Warner Inc. are planning a joint bid for Adelphia Communications Corp. as part of a deal that could lead to a broad realignment of interests in the cable industry
5	Joint bid for Adelphia?	Time Warner Inc., the world's largest media company, and Comcast Corp. said they are considering making a joint bid for bankrupt cable-television operator Adelphia Communications Corp.
6	Cable Titans Team for Adelphia.	Comcast and Time Warner yesterday announced they will make a joint bid for Adelphia Communications, jumping to the front of the pack in the widely watched auction

2 NDD-MAC: Proposed Approach

Our approach, Near-Duplicate Detection with Metadata Augmented Communities (NDD-MAC) is motivated by the observation if a pair of news articles are indeed near-duplicates of each other, then the metadata related to the news content also needs to be matching. We also use sentence-transformers[4] based semantically meaningful neural-embeddings as one of the key signals to capture the similarity between a pair of news articles. Further, we use a community detection-based graph partitioning technique to identify subsets of articles which are more cohesive within a cluster. Figure 1 gives a high-level overview of our pro-

[4]https://sbert.net/

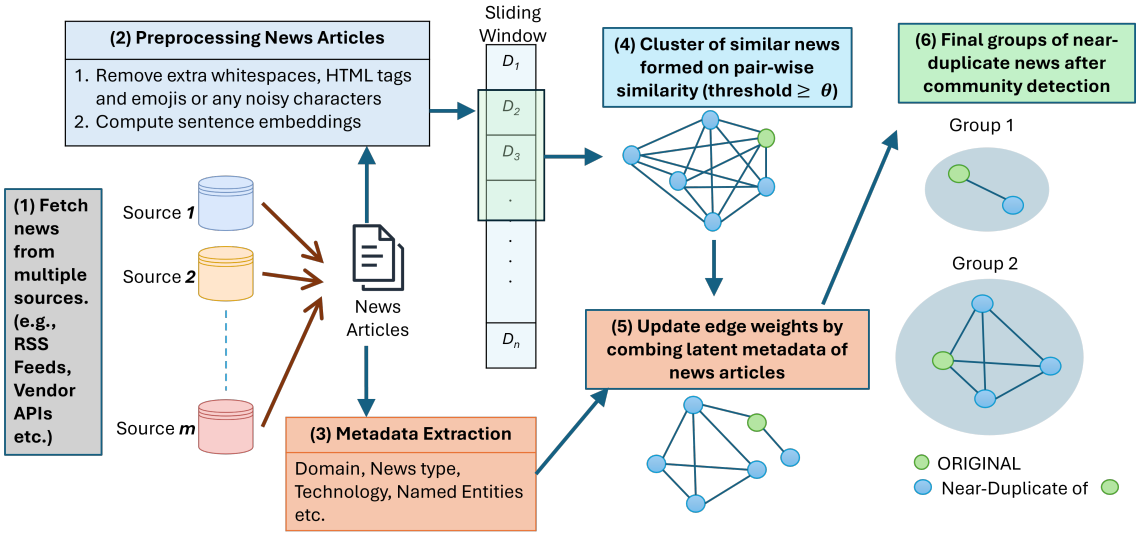


Figure 1: Block Diagram of NDD-MAC, the proposed approach for Near-Duplicate Detection with Metadata Augmented Communities

posed approach.

We infer and extract the metadata about each news article such as *news-type* (i.e., type of the key event) described (e.g., product launches, merger-acquisition, awards, financial reporting etc.), (ii) *industry domain* (e.g., finance, telecom, agriculture, healthcare etc.), (iii) *technology* (e.g., AI, cloud computing, blockchain, cybersecurity, 5G networking etc.), (iv) *types of products, services and organizations* based on the content of an article. Appendix Table A contains a sample of the metadata information extracted from the news articles using specific classifiers for each dimension. We highlight that this metadata can contain information that is not readily mentioned in the surface form of the news content. For instance, the *news-type* of news-items 1 and 2 in Table 1 is classified as *Customers & Partners* using the classification scheme in Table A where as for news items 3, 4, 5, 6 it is detected as *Mergers & Acquisitions*. Apart from the event types, the participants and their roles in the events are also different. These factors are used to update the edge weights in the initial clusters formed. Figure 2 provides the illustration of changes in the edge weights due to metadata. These updated weights benefit in the community detection stage of NDD-MAC to identify subtle differences which are missed by the sentence transformers based clustering stage.

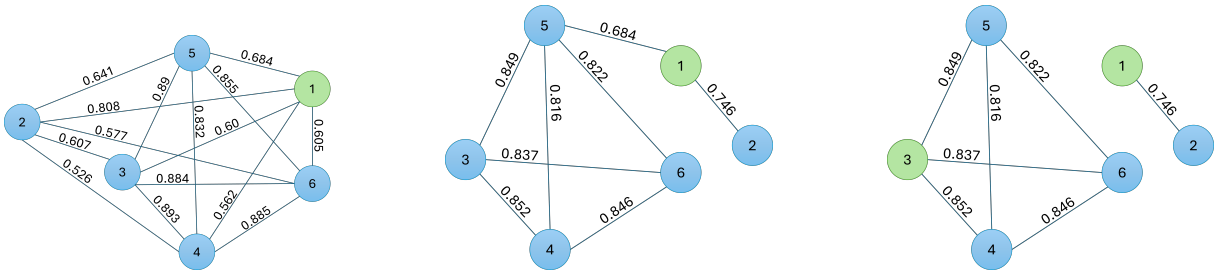
In contrast to the Locality sensitivity hashing (LSH) based approaches, the proposed approach does not restrict its focus only on the surface form of the content. To the best of our knowledge, there are no existing methods which are unsupervised and make use this metadata for the near duplicate detection task. We now describe the key steps in the proposed approach in detail.

Input pre-processing: Firstly, the news articles are pre-processed to remove any noisy characters to ensure consistent character encoding and date formatting issues are resolved. The entire news corpus is partitioned into multiple sliding windows from the start date and end date of the input. This helps to ensure that the approach can be adapted even when the resources such as compute power and memory are constrained. Then for the cleaned content of each news article snippet within each sliding window is passed through two components and discussed in detail in the following sections.

Neural Embedding Computation and Preliminary Cluster Formation: We map an input news snippet (D_i) to a high-dimensional vector embedding $C_i \in \mathbb{R}^d$ using Sentence-BERT (Reimers and Gurevych, 2019). This enables us to get the news snippets with similar meaning closer in the embedding space. This spatial relationship enables the detection of news articles on the same topic and similar content. Sentence-BERT uses of siamese and triplet neural network to modify the standard pretrained BERT network and capture better contextual embeddings compared to prior approaches. We make use of *all-mpnet-base-v2* (denoted as *MPnet*)[5] from the sentence-transformers[6]

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

[6]https://github.com/UKPLab/sentence-transformers



(a) Initial cluster of similar news using surface level embedding cosine similarity

(b) Cluster with updated edge weights using latent metadata of news articles and (threshold > 0.6)

(c) Final groups of near-duplicate news after using Louvain community detection

Figure 2: Overview of the edge weight updates in NDD-MAC approach for the example in the Table 1

library. The *all-mpnet-base-v2* model transforms input sentences into a 768-dimensional dense vector, providing semantically rich representations. It has achieved the best overall performance across semantic search and sentence embeddings benchmarks.

For every pair (D_i, D_j) of articles, we check for cosine similarity of their embedding (C_i, C_j) and form clusters. Each cluster is then represented as a graph in which the news articles are represented as nodes and the edges connecting two articles are initialized with weights as the cosine similarity among the embeddings.

Multi-Dimensional Metadata Augmentation: We notice that the news articles in different clusters may have same surface level similarity, but they may have subtle, nuanced differences and may get clubbed together. So, for every news article, we extract a set of features $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ that can capture these subtle differences. These features are augmented with the embedding based similarity between a pair of articles to further improve the near-duplicate detection task. We highlight that this metadata can contain information that is not readily mentioned in the surface form of the news content. To the best of our knowledge, the prior art does not use this metadata for the near duplicate detection task.

For this purpose, we make use of an ensemble of rule-based and machine learning classifiers $\mathcal{M} = \{m_1, m_2, ..., m_n\}$ that extract the features $(\mathcal{S})$ along multiple dimensions of the input document. We extract and reason about the metadata such as type of the events described in a news article, the participant entities and arguments of these events, as well as their realis or irrealis grammatical moods. Additional dimensions of metadata such as domain, technology, products or services, different quantities mentioned etc. are also extracted from the content of an article. Please refer to appendix A for full list of news-type, domain and technology categories used for extracting metadata from news articles. Furthermore, we identify the participants of events and facts described in a news article.

If a pair of news articles are indeed near-duplicates, then we note that their metadata also needs to match. To reinforce the similarity between news articles with matching metadata, we update their edge-weights in the cluster. The initial cosine similarity based edge weights in the preliminary clusters are updated using the jaccard index of above mentioned multi-dimensional metadata (i.e., Jaccard($\mathcal{S}_i$, $\mathcal{S}_j$)).

Cluster De-merging with Cohesive Communities: The updated edge weights bring together news articles whose metadata information is similar and hence their similarity gets reinforced. Article pairs which may be broadly related to similar entities but differ along some of the dimensions of metadata have their edge weights reduced. After this edge-weight update, we begin the process of de-merging or partitioning the clusters. We use Louvain community detection algorithm (Blondel et al., 2008; Patil, 2020) for partitioning the clusters. We also implement the Leiden community detection algorithm (Traag et al., 2019) for comparative analysis of different methods for detecting the communities. Subsets of articles which are more cohesive within a cluster compared to the rest of the cluster get partitioned in this step. After post-processing of these partitioned clusters, we get the final groups of near duplicate articles. Illustration of these steps on the real-life example in Table 1 has been shown in the Figure 2.

3 Experimental Evaluation

Datasets: We evaluate our proposed approach using two different datasets: (i) *NewsAggregator-Vendor dataset*: A large, private dataset of 34801 real-life news collected from a leading news aggregator using its subscription API, (ii) *NDD-NS* (Kumar et al., 2025): a sample of 1205 news articles extracted from the publicly available AGNews dataset[7]. As shown in Table 1, a sample news record contains the headline and a snippet from the news body. There is also the publication date of news article and a URL (which is excluded from the Table for ease of exposition) that points to the full news article.

	NewsAggregator-Vendor Dataset	**NDD-NS**
#Sources	1254	109
#Articles	34801	1205
#Sentences	136434	2560
# Words	2915389	51738
Avg. sent./article	3.92	2.22
Avg. words/article	83.77	42.94

Table 2: Dataset Statistics

Baselines and Expt. settings: We use MinHash (Rodier and Carter, 2020), SimHash (Charikar, 2002) as well as Novo and Gedikli BERT based supervised learning approach (Novo and Gedikli, 2023) as our baselines.

Rodier and Carter (2020) first convert the documents into a set of n-grams (i.e., shingles of length n). Then, they randomly sample a set of k (k=1600) shingles from the set. They generate a list of p (p=20) random numbers called permutations. For each permutation they compute minimum hash value using the fingerprints of the shingles and that permutation and assign the lowest hash value to an array of length p. This array of length p is the sketch of the document. Using these document sketches they identify near duplicate news articles. They have reported their best performance using the parameters of k=1600 shingles and p=20 permutations. We have re-implemented their approach with these parameters. We set up Simhash baseline employing an open source simhash-py [8] library in python. Best parameters (f = 64, m =3) settings are utilized as discussed in Manku et al. (2007) for

[7]https://paperswithcode.com/dataset/ag-news

[8]https://github.com/seomoz/simhash-py

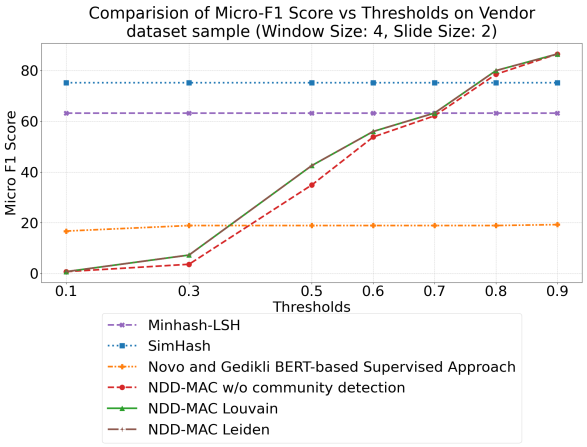


Figure 3: Comparison across different similarity thresholds on a sample from News Aggregator Vendor Dataset

an online settings. For Novo and Gedikli BERT based approach (Novo and Gedikli, 2023), we use the model trained based on the description in their paper.

We have evaluated NDD-MAC approach for various *similarity thresholds for cluster* formation {0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9}, *sliding window duration* (in number of days), viz., {1, 2, 3, 4, 5, 6, 7, 14, 21, 30, "full"} and *slide size within the sliding window duration* {1, 2, 3}.

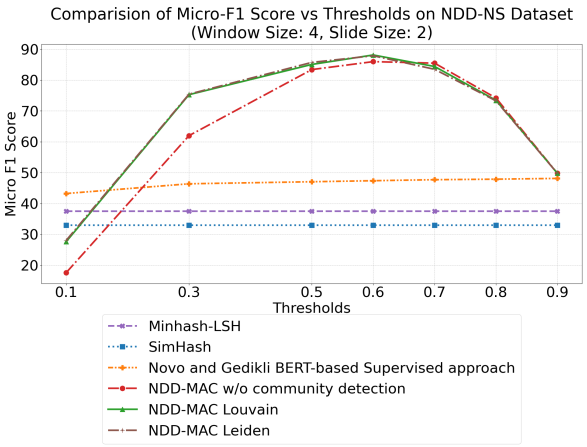


Figure 4: Comparison across different similarity thresholds on NDD-NS Dataset

Results The proposed approach achieves better performance than baselines at similarity thresholds above 0.8 for both the datasets. In the real-life dataset from News Aggregator vendor, there are multiple news which are essentially copies of each other and have high surface similarity. This results in better performance of MinHash and SimHash on the Vendor dataset compared to the more challenging NDD-NS dataset. On the NDD-NS dataset,

NDD-MAC is consistently better than MinHash and SimHash even after the small threshold of 0.3. It achieves its best performance at the similarity threshold of 0.6, window size 4 and slide size 2. From Figure 3, we can see a comparison of the NDD-MAC approach with various baselines on varying similarity threshold. We observe that SimHash is faster and memory efficient compared to the Minhash-LSH approach. This is because it stores a single hash value for a text document, while Minhash-LSH stores hash values for each of the shingles generated for a text document.

The NDD-NS dataset has very few number of words per article and very less word overlap rate (i.e., they are paraphrased very well). The average intra-cluster maximum n-gram overlap is 5.14. So, the threshold 0.6 servers good to capture the surface form of a news cluster. Similarly, for the News Aggregator Vendor dataset sample, the number of words is twice when compared with the NDD-NS. The average intra-cluster maximum n-gram overlap for the vendor dataset sample is 18.63. Due to the high word overlap rate, this makes it easier to cluster near-duplicates. As seen in the Figure 2, the performance increases with higher similarity thresholds. For real-life industrial setting, the threshold around 0.85 or 0.9 seems practically useful for real-life news-feeds from news aggregator vendors.

In addition to this, we also study the effect of varying window and slide sizes on the performance. Figure 5 shows the effect of window and slide sizes on micro-f1 scores with NDD-MAC on threshold 0.6 for the NDD-NS dataset. We note that after sliding window duration 4, the sliding length (i.e., slide size parameter) does not have a significant effect. Based on this graph, we suggest that sliding window duration can be kept around 3 or 4 days during the pre-processing stage. Although slide size has not much effect, but processing the articles in windows performs better when compared to processing the entire corpus at once and is also computationally far more effective.

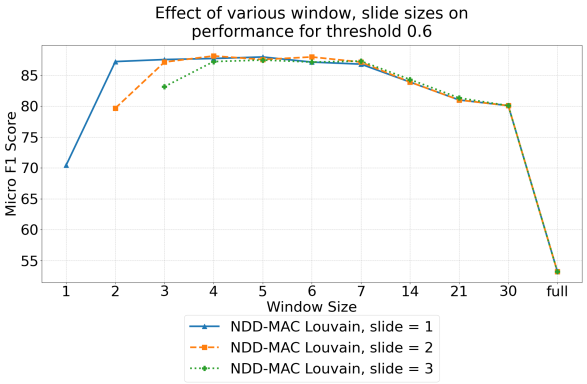


Figure 5: Effect of various window and slide sizes while processing articles on Micro-F1 scores for similarity threshold 0.6

4 Related Work

Locality Sensitive Hashing (LSH) (Leskovec et al., 2020) has been a cornerstone of the techniques used for near-duplicate detection. Multiple web search engines have applied LSH variants such as MinHash for near duplicate detection and related applications. LSH focuses on the surface form of the content. It uses only the words mentioned in the input text to form the n-grams or shingles while identifying the near-duplicate articles. Most recent adaptation of LSH based approach for the problem of near duplicate detection has been proposed by Rodier and Carter (2020). Their approach uses MinHash and is based on *shingling* proposed by Broder (2000). Shingling technique translates a document into a set of *n*-grams (i.e., shingles, a contiguous sequence of n words). Similarity of two documents can be then measured by computing set similarity. If the similarity is greater than a threshold value, documents are considered as near duplicates of one another. But this practice is costly as the number of shingles generated for a document is too large. To resolve this, they prepare document's sketch (small signatures) using MinHash technique proposed by Broder (1997). Using these document sketches they identify near duplicate news articles. In spite of being such a well-known technique, the recent adaptation of LSH based approach proposed by Rodier and Carter (2020) (MinHash-LSH) performs quite poorly on the shortened news data addressed in this paper. In Manku et al. (2007), authors implement Simhash (Charikar, 2002) fingerprint technique to identify near duplicate for web documents in an online settings or offline (batch) settings. They propose an algorithmic technique for detecting existing f-bit fingerprints that differ from a given fingerprint in at most m bit-positions, for small m. They have experimentally validated their approach on a corpus of 8B webpages.

Silcock et al. (2023) employed supervised training to develop their bi-encoder (with *MPnet* as base) and cross-encoder models, using a dataset consisting of OCR-processed text from newspa-

pers published between 1920 and 1977. Their approach generates article representations through the bi-encoder model. These representations are then used to construct a graph and identify communities to reduce the computation required for duplicate identification. The cross-encoder model then works with these clusters or communities to identify near-duplicates. We note the reliance of their method on a historical dataset (1920-1977) with large number of articles which have significantly more content per article than a news snippet available from the paywalled sources. Due to the more recent paywall constraints and evolving nature of news, using their approach may require an updated dataset that reflects the changes in news reporting style. Further, one may have to retrain the models using the updated dataset to use their approach. In contrast, the proposed approach only uses off-the-shelf *MPnet* embeddings to form preliminary clusters. Then the edge weights between pairs of news articles in these preliminary clusters are updated based on metadata augmentation. After that we perform community detection on individual clusters. In addition to this difference, we also highlight that the proposed approach does not need to train any supervised model.

Novo and Gedikli (2023) have proposed a supervised learning based approach to identify near-duplicates in which common named entities in a pair of documents are used as the key features. Firstly, they assume that if there are no common named entities in a pair of documents they are non-duplicates. Then a BERT model was fine-tuned to classify whether a given pair of articles are near-duplicates. They have evaluated their approach on a small dataset of 100 business energy news articles. Out of the resulting 4950 article pairs in their dataset, only 88 of such pairs are near-duplicates. The pairwise evaluation strategy leads to inconsistent evaluation as transitivity property among the near-duplicate documents gets violated. Further, due to the supervised learning approach, they have additional overhead of requiring labeled training data. Due to drift in the news topics and changes in the named entities mentioned in news over time, this approach tends require repeated labeling of data to update the supervised learning models. Unsupervised methods for near duplicate detection are more realistic given the practical constraints in industrial usage. Hence, we focus on unsupervised learning methods such as MinHash, SimHash etc. as relevant baselines for the near-duplicate detection task.

Near duplicate detection is an important task not only for news snippets but also it has multiple other applications (Nauman and Herschel, 2022), especially where short text snippets are common (Patil and Ravindran, 2015). The metadata augmentation idea discussed in this paper can be useful for identifying duplicate questions in technical (Silva et al., 2018; Pal et al., 2021) as well as non-technical domains (Zhang et al., 2018; Bedi et al., 2021). Detecting duplicate defect reports (Zhang et al., 2023; Patil and Ravindran, 2020; Patil, 2017) is another important application in software maintenance lifecycle.

5 Conclusion

With rise of paywalls on news websites and proliferation of news aggregators, it is now common get only the headline and a small snippet of a news article. This makes the problem of near-duplicate detection more challenging compared to when the full article was readily available. Current research has largely overlooked this problem. We introduced Near Duplicate Detection using Metadata Augmented Communities (NDD-MAC) to address this issue. Unlike the LSH-based approaches, the proposed approach does not rely solely on the surface form of the content or full article availability. It effectively detects near-duplicates using small text excerpts and incorporates Multi-Dimensional Metadata Augmentation along with community detection.

To the best of our knowledge, the prior work does not use this type of metadata for the near duplicate detection task. Evaluation on real-world datasets from a news aggregator and the AGNews dataset demonstrates that NDD-MAC significantly outperforms established baselines like MinHash-LSH, SimHash as well as a recent supervised learning based approach.

References

Harsimran Bedi, Sangameshwar Patil, and Girish Palshikar. 2021. Temporal question generation from history text. In *Proceedings of the 18th international conference on natural language processing (ICON)*, pages 408–413.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Andrei Z Broder. 2000. Identifying and filtering near-duplicate documents. In *Annual symposium on combinatorial pattern matching*, pages 1–10. Springer.

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

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388.

Alok Kumar, Siddharth Tumre, and Sangameshwar Patil. 2025. Benchmarking near-duplicate detection in the era of pay-walled news. In *Companion Proceedings of the ACM Web Conference 2025*, WWW '25. Association for Computing Machinery.

Angela M Lee and Hsiang Iris Chyi. 2015. The rise of online news aggregators: Consumption and competition. *International Journal on Media Management*, 17(1):3–24.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets*. Cambridge university press.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150.

Felix Nauman and Melanie Herschel. 2022. *An introduction to duplicate detection*. Springer Nature.

Anne Stockem Novo and Fatih Gedikli. 2023. Explaining bert model decisions for near-duplicate news article detection based on named entity recognition. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 278–281. IEEE.

Samiran Pal, Avinash Singh, Soham Datta, Sangameshwar Patil, Indrajit Bhattacharya, and Girish Palshikar. 2021. Semantic templates for generating long-form technical questions. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 235–247. Springer.

Sangameshwar Patil. 2017. Concept-based classification of software defect reports. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 182–186. IEEE.

Sangameshwar Patil. 2020. Domain-specific noisy query correction using linguistic network community detection. In *Companion Proceedings of the Web Conference 2020*, pages 126–127.

Sangameshwar Patil and Balaraman Ravindran. 2015. Active learning based weak supervision for textual survey response classification. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 309–320. Springer.

Sangameshwar Patil and Balaraman Ravindran. 2020. Predicting software defect type using concept-based classification. *Empirical Software Engineering*, 25(2):1341–1378.

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*. Association for Computational Linguistics.

Simon Rodier and Dave Carter. 2020. Online near-duplicate detection of news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1242–1249.

Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. 2023. Noise-robust de-duplication at scale. In *The Eleventh International Conference on Learning Representations*.

Rodrigo FG Silva, Klérisson Paixão, and Marcelo de Almeida Maia. 2018. Duplicate question detection in stack overflow: A reproducibility study. In *2018 IEEE 25th international conference on software analysis, evolution and reengineering (SANER)*, pages 572–581. IEEE.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Ting Zhang, DongGyun Han, Venkatesh Vinayakarao, Ivana Clairine Irsan, Bowen Xu, Ferdian Thung, David Lo, and Lingxiao Jiang. 2023. Duplicate bug report detection: How far are we? *ACM Transactions on Software Engineering and Methodology*, 32(4):1–32.

Xiaodong Zhang, Xu Sun, and Houfeng Wang. 2018. Duplicate question identification by integrating framenet with neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

A Sample of Multi-dimensional Metadata used in NDD-MAC

S. No.	News-type	Domain	Technology
1	Product Launches/Offerings	Travel and Logistics	Cloud Technology
2	Mergers & Acquisitions	Food & Beverages	AI
3	Customers & Partners	Tourism & Hospitality	Blockchain
4	Business Expansion	Manufacturing	Cybersecurity
5	Research & Innovation	Multidomain Applications of IT	ERP (SAP, ...)
6	Achievements & Recognition	Retail	IoT
7	Analyst Reports/Studies	Communications, Media, and Information Services	5G & Networking
8	Financial Reporting	Banking Finance Insurance	3D Printing
9	Legal	Healthcare	Augmented Reality
10	HR/CSR/Branding/Others	Education	Quantum Computing
11		Energy, Resources, and Utilities	Automation and Robotics
12		Public Services	Material Technology
13		Life Science	Human Computer Interface
14		Agriculture	