MN-DS: A Multilabeled News Dataset for News Articles Hierarchical Classification

Alina Petukhova[1,*] **and Nuno Fachada**[1]

[1]Lusófona University, COPELABS, Campo Grande, 376, Lisbon, Portugal
[*]corresponding author(s): Alina Petukhova (alina.petukhova@ulusofona.pt)

ABSTRACT

This article presents a dataset of 10,917 news articles with hierarchical news categories collected between January 1st 2019, and December 31st 2019. We manually labelled the articles based on a hierarchical taxonomy with 17 first-level and 109 second-level categories. This dataset can be used to train machine learning models for automatically classifying news articles by topic. This dataset can be helpful for researchers working on news structuring, classification, and predicting future events based on released news.

Background & Summary

A news dataset is a collection of news articles classified into different categories. In the past decade, there has been a sharp increase of news datasets available for analysis[1]. These datasets can be used to understand various topics, from politics to the economy.

A few different types of news datasets are commonly used for analysis. The first is raw data, which includes all the data that a news organisation collects. This data can be used to understand how a news organisation operates, what stories are covered, and how they are covered. The second type of news dataset is processed data. This data has been through some processing, such as aggregated or cleaned up. Processed data is often easier to work with than raw data, and it can be used to answer specific questions such as providing additional information for decision making process. The third type of news dataset is derived data. This data is created by combining multiple datasets, often from different sources[2]. News datasets can be used for various purposes in a machine learning context, for example:

- Predicting future events based on past news articles.
- Understanding the news cycle.
- Determining the sentiment of news articles.
- Extracting information from news articles (e.g., named entities, location, dates).
- Classifying news articles into predefined categories.

To adequately answer research questions, news datasets should contain sufficient data points and span a significant enough period. There are many labeled news datasets available, each with specific limitations. For example, they may only cover a specific period or geographical area or be confined to a particular topic. Additionally, the categories may not be completely accurate, and the datasets may be biased in some way[3,4].

Some of the more popular news datasets include the 20 Newsgroups dataset[5], AG's news topic classification dataset[6], L33 - Yahoo News dataset[7,8], and News Category dataset[9]. Each of these datasets has been used extensively by researchers in the fields of natural language processing and machine learning, and each has its advantages and disadvantages.The 20 Newsgroups dataset was created in 1997 and contains 20 different categories of news, each with a training and test set. The data is already pre-processed and tokenised, which makes it very easy to use. However, the dataset is outdated and relatively small, with only about 1,000 documents in each category.

The AG's news topic classification dataset is a collection of news articles from the academic news search engine "ComeToMyHead" during more than one year of activity. Articles were classified into 13 categories: business, entertainment, Europe, health, Italia, music feeds, sci/tech, software & dev., sports, toons, top news, U.S., and world. The dataset contains more than 1 million news articles. However, there are several limitations to this dataset. First, it is currently outdated since data were collected in 2005. Second, the taxonomy covers specific countries like the US and Italy, but has general references like Europe or world creating overlaps in the classification (e.g., Italy and Europe), as well as potential imbalances (e.g., events in China are

likely to be underrepresented and/or underreported compared to those in the US). Finally, the dataset does not include methods for type or category description.

The L33 - Yahoo News dataset is a collection of news articles from the Yahoo News website provided as part of the Yahoo! Webscope program. The articles are labelled into 414 categories such as music, movies, crime & justice, and others. The dataset includes the random article id followed by possible associated categories. The L33 - Yahoo News dataset is available under Yahoo's data protection standards. It can be used for non-commercial purposes if researchers credit the source and license new creations under identical terms. The limitations of the L33 dataset are the license terms, restricting companies from using this dataset for commercial purposes, and the amount of data per class, with the category "supreme court decisions" having only five articles, for example. In addition, there is some overlap in the categories, which makes it challenging to train a model that can accurately predict multiple categories.

The News Category Dataset is a collection of around 210k news articles from the Huffington Post, labeled with their respective categories, which include business, entertainment, politics, science and technology, and sports. However, the dataset has several limitations. First, the dataset is not comprehensive since it only includes articles from one source. Second, news categories are not standardized, including broad categories like "Media" and "Politics" and very narrow ones like "Weddings" and "Latino voices".

We can observe that a new news dataset with up-to-date articles and additional categories would contribute to the accuracy improvement of news classification models, which is the aim of the current work.

Methods

In this paper, we present a new dataset based on the NELA-GT-2019 data source, classified with IPTC's[a] NewsCodes Media Topic taxonomy[11]. The original NELA-GT-2019 dataset contains 1.12M news articles from 260 sources collected between January 1st 2019 and December 31st 2019, providing essential content diversity and topic coverage. Sources include a wide range of mainstream and alternative news outlets.

In turn, the IPTC taxonomies are a set of controlled vocabularies used to describe news stories' content. The NewsCodes Media Topic taxonomy has been one of IPTC's main subject taxonomies for text classification since 2010. We used the 2020 version of NewsCodes Media Topic taxonomy[12]. News organisations use it to categorize and index their content, while search engines use it to improve the discoverability of news stories[13].

We observed that the first-level category of the NewsCodes Media Topic taxonomy is not accurate enough to catalogue an article. For example, the "sport" category may include different aspects, such as information about specific sports, sports event announcements, and the sports industry in general, which have more specific meanings than the first-level category label is able to convey. Therefore, we used a second-level category of NewsCodes Media Topic taxonomy to have a more specific article category. In comparison to the previously published datasets, we included in our dataset unique categories such as "arts and entertainment", "mass media", "armed conflict", "weather statistic", and "weather warning". Therefore, we created the proposed Multilabeled News Dataset (MN-DS) by hand-picking and labeling approximately 100 news articles for each second level category[b] of the NewsCodes Media Topic taxonomy.

Data Records

After manually selecting news articles relevant to each category, we obtained 10,917 articles in 17 first-level and 109 second-level categories from 215 media sources. An overview of the released MN-DS dataset by category is provided in Table 1. All data are available in CSV format at https://doi.org/10.5281/zenodo.7394851 and available for download under the Creative Commons license.

Description of columns in the data table

- id: Unique identifier of the article.
- date: Date of the article release.
- source: Publisher information of the article.
- title: Title of the news article.
- content: Text of the news article.

[a] The International Press Telecommunications Council, or IPTC, is an organization that creates and maintains standards for exchanging news and other information between news organisations.

[b] https://www.iptc.org/std/NewsCodes/treeview/mediatopic/mediatopic-en-GB.html

Table 1. The number of articles under each Level 1 category.

Categories	Count
Arts, culture, entertainment and media	300
Conflict, war and peace	800
Crime, law and justice	500
Disaster, accident and emergency incident	500
Economy, business and finance	400
Education	607
Environment	600
Health	700
Human interest	600
Labour	703
Lifestyle and leisure	300
Politics	900
Religion and belief	800
Science and technology	800
Society	1100
Sport	907
Weather	400

- author: Author of the news article.
- url: Link to the original article.
- published: Date of article publication in local time.
- published_utc: Date of article publication in utc time.
- collection_utc: Date of article scraping in utc time.
- category_level_1: First level category of Media Topic NewsCodes's taxonomy.
- category_level_2: Second level category of Media Topic NewsCodes's taxonomy.

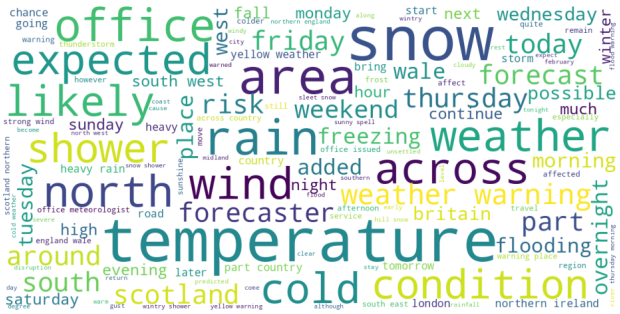
Technical Validation

To validate the dataset, we created a word cloud representation of each category, as shown in Fig. 1. The central concept of a word cloud is to visualize for each category the most popular words with a size corresponding to the degree of popularity. This representation allows us to quickly assess the quality of the text annotation since it displays the most common words of the category. In the bar chart shown in Fig. 2, we can observe that the "science and technology" first-level category contains the highest count of topic-specific words, while in more general categories, such as "weather" or "human interest", there is less variety in the texts, probably because they represent shorter and more similar articles.

We used the dataset to train the most common text classification models to extend the technical validation of the proposed dataset and establish the benchmark for multiclass classification. The following embeddings were selected:

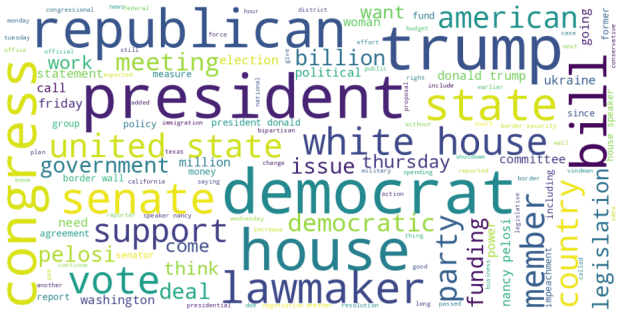
- Tf-idf embedding, where Tf-idf stands for term frequency-inverse document frequency[14]. Tf-idf transforms text into a numerical representation called a tf-idf matrix. The term frequency is the number of times a word appears in a document. The inverse document frequency measures how common a word is across all documents. Tf-idf is used to weigh words so that important words are given more weight. The dataset's news texts and categories were combined and vectorized with TfidfVectorizer[15].
- GloVe (Global Vectors for Word Representation) embeddings with an algorithm based on a co-occurrence matrix, which counts how often words appear together in a text corpus. The resulting vectors are then transformed into a lower-dimensional space using singular value decomposition[16].
- DistilBertTokenizer[17], which is a distilled version of BERT, a popular pre-trained model for natural language processing. DistilBERT is smaller and faster than BERT, making it more suitable for fast training with limited resources. The

Category level 2: crime



Category level 2: weather forecast

Category level 2: economy



Category level 2: government

Figure 1. Word cloud of MN-DS dataset for selected second-level categories.

trade-off is that DistilBERT's performance is 3% lower than BERT's. DistilBERT embeddings are trained on the same data as BERT, so they are equally good at capturing the meaning of words in context.

During dataset validation, we combined the selected embeddings with different classifiers. We tested Multinomial Naive Bayes (NB) classifier[18], Logistic Regression[19], Support Vector Classifier (SVC)[20], and DistilBERT model[21]. Since MN-DS is a multiclass dataset, we used the OneVsRestClassifier strategy for classification models[15]. OneVsRestClassifier is a classifier that trains multiple binary classifiers, one for each class. The individual binary classifiers are then combined to create a single multiclass classifier. This approach is often used when there are many categories, as it can be more efficient than training a single multiclass classifier from scratch. The tested classifiers work as follows:

- The Multinomial NB is a text classification algorithm that uses Bayesian inference to classify text. It is a simple and effective technique that can be used for various tasks, such as spam filtering and document classification. The algorithm is based on the assumption that the features in a document are independent of each other, which allows it to make predictions about the category of a document based on its individual features.
- The Logistic Regression classifier works by using a sigmoid function to map data points from an input space to an output space, where the categories are assigned based on a linear combination of the features. The weights of the features are learned through training, and the predictions are made by taking the dot product of the feature vector and the weight vector.
- The SVC classifier is a powerful machine learning model based on the Support Vector Machines algorithm. The model is based on finding the optimal decision boundary between categories to maximise the margin of separation between them. The SVC model can be used for linear and non-linear classification tasks and is particularly well-suited for problems with high dimensional data. The classifier is also robust to overfitting and can generalise well to new data.
- DistilBERTModel, a light version of the BERT classifier[17], developed and open-sourced by the team at Hugging Face. DistilBERT can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks with minimal training data.

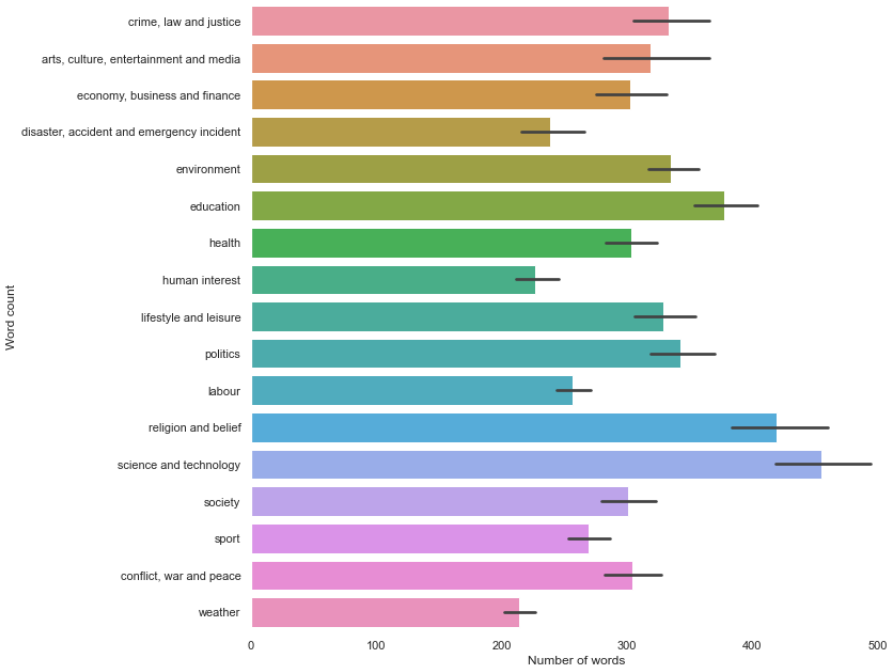


Figure 2. Mean number of non-repeated words in article body for first-level categories.

Table 2. Multilabel classification results for level 1 categories.

Embeddings Model	TFIDF			Glove			DistilBertTokenizer		
	precision	recall	f1 score	precision	recall	f1 score	precision	recall	f1 score
Multinomial NB	0.802	0.631	0.649	0.629	0.499	0.529	n/a	n/a	n/a
Logistic Regression	0.800	0.763	0.774	0.747	0.739	0.739	n/a	n/a	n/a
SVC classifier	0.808	0.796	0.799	0.768	0.762	0.760	n/a	n/a	n/a
DistilBERTModel	n/a	n/a	n/a	n/a	n/a	n/a	0.849	0.842	0.844

Classification results for level 1 and level 2 categories are presented in Table 2 and Table 3, respectively. It is possible to observe that DistilBERTModel achieves better classification results for both category levels. To improve these results in future studies, we suggest applying hierarchical classification methods as described by Silla and Freitas[22], for example.

Code availability

Code for the technical validation of the dataset is available in the GitHub repository (https://github.com/alinapetukhova/mn-ds-news-classification)

References

1. Paullada, A., Raji, I. D., Bender, E. M., Denton, E. & Hanna, A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* **2**, 100336, 10.1016/j.patter.2021.100336 (2021).

Table 3. Multilabel classification results for level 2 categories.

Embeddings Model	TFIDF			Glove			DistilBertTokenizer		
	precision	recall	f1 score	precision	recall	f1 score	precision	recall	f1 score
Multinomial NB	0.628	0.602	0.583	0.496	0.484	0.469	n/a	n/a	n/a
Logistic Regression	0.646	0.649	0.635	0.589	0.589	0.577	n/a	n/a	n/a
SVC classifier	0.645	0.646	0.628	0.581	0.595	0.571	n/a	n/a	n/a
DistilBERTModel	n/a	n/a	n/a	n/a	n/a	n/a	0.735	0.715	0.715

2. Jayakody, N., Mohammad, A. & Halgamuge, M. Fake news detection using a decentralized deep learning model and federated learning. In *Conference: IEEE 48th Annual Conference of the IEEE Industrial Electronics Society (IECON2́2)* (IEEE, 2022).

3. Stefansson, J. K. *Quantitative measure of evaluative labeling in news reports: Psychology of communication bias studied by content analysis and semantic differential*. Master's thesis, UiT, Norway's Arctic University (2014).

4. Gezici, G. Quantifying political bias in news articles, 10.48550/ARXIV.2210.03404 (2022).

5. Mitchell, T. 20 newsgroups data set (1999).

6. Ag's corpus of news articles (2005).

7. Soni, A. & Mehdad, Y. RIPML: A restricted isometry property-based approach to multilabel learning. In Rus, V. & Markov, Z. (eds.) *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, 532–537 (AAAI Press, 2017).

8. Chen, S., Soni, A., Pappu, A. & Mehdad, Y. Doctag2vec: An embedding based multi-label learning approach for document tagging. In *Rep4NLP@ACL* (2017).

9. Misra, R. News category dataset, 10.48550/ARXIV.2209.11429 (2022).

10. Gruppi, M., Horne, B. D. & Adalı, S. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles, 10.48550/ARXIV.2003.08444 (2020).

11. IPTC NewsCodes scheme (controlled vocabulary) (2010).

12. Mediatopic newscodes 2020 (2020).

13. Newscodes. https://iptc.org/standards/newscodes/#:~:text=Who%20uses%20IPTC%20NewsCodes%3F,becoming%20more%20and%20more%20popular. Accessed: 2022-11-21.

14. Sammut, C. & Webb, G. I. (eds.). *TF–IDF*, 986–987 (Springer US, Boston, MA, 2010).

15. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

16. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, 10.3115/v1/d14-1162 (Association for Computational Linguistics, Doha, Qatar, 2014).

17. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).

18. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, USA, 2008).

19. Cox, D. R. The regression analysis of binary sequences. *J. Royal Stat. Soc. Ser. B (Methodological)* **20**, 215–242, https://doi.org/10.1111/j.2517-6161.1958.tb00292.x (1958).

20. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. learning* **20**, 273–297, https://doi.org/10.1007/BF00994018 (1995).

21. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 10.48550/ARXIV.1910.01108 (2019).

22. Silla, C. & Freitas, A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **22**, 31–72, 10.1007/s10618-010-0175-9 (2011).

Author contributions statement

A.P. created the search strategy, retrieved and screened the publications, extracted the selected data for annotation, labelled dataset. A.P. and N.F. assessed the quality of the included articles and checked the data. A.P. performed the statistical analyses, created the graphics and wrote the original draft. N.F. conceived the project, provided critical comments and revised the paper. All the authors have read and agreed to the manuscript.

Competing interests

The authors declare no competing interests.