Reconstruct to Retrieve: Identifying Interesting News in a Cross-Lingual Setting

Boshko Koloski
Jožef Stefan Institute
Jožef Stefan IPS
boshko.koloski@ijs.si

Blaž Škrlj
Jožef Stefan Institute
Jožef Stefan IPS
blaz.skrlj@ijs.si

Nada Lavrač
Jožef Stefan Institute
nada.lavrac@ijs.si

Senja Pollak
Jožef Stefan Institute
senja.pollak@ijs.si

Abstract

An important and resource-intensive task in journalism is retrieving relevant foreign news and its adaptation for local readers. Given the vast amount of foreign articles published and the limited number of journalists available to evaluate their interestingness, this task can be particularly challenging, especially when dealing with smaller languages and countries. In this work, we propose a novel method for large-scale retrieval of potentially translation-worthy articles based on an auto-encoder neural network trained on a limited corpus of relevant foreign news. We hypothesize that the representations of interesting news can be reconstructed very well by an auto-encoder, while irrelevant news would have less adequate reconstructions since they are not used for training the network. Specifically, we focus on extracting articles from the Latvian media for Estonian news media houses. It is worth noting that the available corpora for this task are particularly limited, which adds an extra layer of difficulty to our approach. To evaluate the proposed method, we rely on manual evaluation by an Estonian journalist at Ekspress Meedia and automatic evaluation on a gold standard test set.

1 Introduction

Media houses often report relevant foreign news and adapt them to the local readership. With the ever-rising number of published articles and the limited number of people retrieving and curating the stories, the task becomes harder for media houses. The media houses often need to allocate scarce resources available, such as translators of specific languages and journalists, to curate and adapt the stories. In this work, we propose an approach that, given a handful of articles in a given language (Estonian), automatically suggests a set of potentially interesting news in a chosen foreign language (Latvian), employing a deep auto-encoder network to reconstruct and retrieve the relevant foreign articles. The task of identifying foreign interesting news is defined by the Estonian media house, interested in retreival of Latvian articles. For example, articles covering international politics (e.g. American elections) are not interesting, as the Estonian house would have other sources for these news. Also, many local articles are not interesting, as they are irrelevant for Estonians. However, very specific articles are of their interest, including the ones, covering Estonians in Latvia, topics relevant to Estonian readership (e.g. discussion on electronic scooters, doping affairs. These topics are however not predefined, but there is a small dataset of retrieved interesting news. The ratio of interesting to non-interesting news is very small, suggesting the task to be considered as imbalanced classification. In many imbalanced classification tasks (such as phishing detection (Douzi et al., 2017), software defect prediction(Tong et al., 2018), wind-turbine (Roelofs et al., 2021) anomaly detection), auto-encoders models have been utilized due their ability to reconstruct subgroups of examples well. Zhang and Zhu (2020) used Wasserstein auto-encoders for document retrieval, while Liou et al. (2014) used word-based auto-encoders for document retrieval.

This work extends the previous work on interesting cross-border news retrieval by Koloski et al. (2021), where the authors define a custom metric – SNIR *(Seed news of interest ratio)*. First, the method embeds both the set of *interesting* articles and the set of candidate articles into a multilingual space (Conneau et al., 2020). Next, the SNIR score of each *candidate* is calculated as the fraction of

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 81–89
September 11–12, 2023. ©2023 Association for Computational Linguistics

the *interesting* articles in a neighborhood of m articles. This metric follows the nearest-neighbor-based approach, where they check the ratio of interesting versus non-interesting news in the neighborhood for a given article. If the ratio is bigger than a given threshold, then the article is considered as interesting and thus relevant for translation and adaptation. They define an article as interesting if it is highly relevant to the Estonian readership at the time of publishing. To our knowledge, this is the only related work for the addressed task. We extend this work by proposing a novel method, as well as by proposing an automated evaluation setting for our task.

In the rest of this paper, Section 2 analyzes related work, Section 3 describes the data used, followed by the explanation of the proposed method in Section 4, and its evaluation in Section 5. Conclusions and further work are presented in Section 6.

2 Related work

In the field of journalism, one of the crucial responsibilities is to search for and gather captivating news stories from neighboring countries. Recent research by Asim et al. (2019) examines the use of ontologies, a type of language technology, in the domain of news retrieval. According to their findings, ontologies are primarily used for semantic search in news retrieval systems. Additionally, the collaboration between translation and journalism is essential in the process of news retrieval (Conway and Davier, 2019; Valdeón, 2020). Machine translation plays a significant role in automatically converting news stories in different languages to a language that is familiar to the news media curator (Utiyama and Isahara, 2003; Kumano et al., 2002; Eck et al., 2004; Bielsa and Bassnett, 2008; Almahasees, 2018).

Large Language Models (LLMs) are currently at the forefront of the field of machine translation. There are mainly two types of LLMs: autoregressive and autoencoding. Autoregressive models generate text by predicting the next word in a sequence given all the previous words. Examples of autoregressive models include GPT-3 (Brown et al., 2020) - model based on the Casual Language Modeling (CLM) task and BERT (Devlin et al., 2018) - model trained with the Masked Language Modeling (MLM) objective. Autoencoding models, on the other hand, are trained to reconstruct the original input given a corrupted version of it. These models learn to represent the input in a compact form that captures the most important information. Examples of autoencoding models include T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). Both types of LLMs are highly effective for a wide range of tasks in interest to journalism: genre identification (Kuzman et al., 2023), text classification (Sun et al., 2019; Koloski et al., 2022a), sentiment analysis (Shirsat et al., 2017; Godbole et al., 2007; Bautin et al., 2008; Balahur et al., 2013; Keivandarian and Carvalho, 2023), machine translation (Zhu et al., 2020; Clinchant et al., 2019; Weng et al., 2020; Hendy et al., 2023), keyword extraction (Martinc et al., 2022; Koloski et al., 2022b) and more. Moreover, multilingual variants of these models (such as XLMR(Conneau et al., 2020)) have been developed to support multiple languages, making them even more useful for cross-lingual NLP tasks.

Autoencoder networks have found widespread use for input retrieval via reconstruction in various domains. For instance, Lu et al. (2021) developed a Siamese autoencoder for dense text retrieval, Xu et al. (2021) benchmarked a network consisting of an autoencoder and a generative adversarial network for zero-shot cross-modal retrieval, reporting promising results. Additionally, Ma et al. (2022) investigated the effect of contrastive pre-training for dense retrieval via autoencoder networks and achieved highly favorable outcomes. In this paper, we apply autoencoders to discover *interesting* news by reconstructing documents. We define *Interesting news* (based on prior work (Koloski et al., 2021)) as news that readers relate to and originates from foreign countries.

3 Data

The data used in this work consists of Estonian and Latvian articles (published in the period between *01.01.2018* until *01.12.2019*) by media houses belonging to the Ekspress Meedia Group. We used the following corpora from the EMBEDDIA news archives data set (Pollak et al., 2021).

- The collection of **Estonian** news articles from the archives of Ekspress Meedia, resulting in 17,148 articles
- The collection of **Latvian** news articles published by the DELFI portal - a Latvian subsidiary of the Ekspress Meedia Group. We

used the data before 1.12.2019 for training (29,178 articles) and the data after for testing (1,339 articles). We split the data in this manner to assess the method's ability to generalize over unseen news and events for the future.

- The set of **21 Latvian** news, consisting of articles published (between 01.01.2019 and 31.12.2019) in the Latvian journal and identified by an Estonian journalist as being interesting for the Estonian public. We also dispose of their aligned **Estonian** counterparts, which are the news that was published in the Estonian newspaper after translation and adaptation.

4 Method

4.1 Data Acquisition

Automated Acquisition of Estonian Ground Truth Our method follows the work by (Koloski et al., 2021) consists of two steps. In the first step, we use exact string matching to extract Estonian articles that mention Latvian Delfi[1] *(Läti Delfi, Lati Delfi, Delfi.lv)* in the article body text as a source of news. The hypothesis is that these articles were identified as significant for translation/adaptation from their original Latvian counterparts. In this manner, we acquired 100 Estonian articles, and we denote them as **Estonian$_{\textbf{ground}}$**.

4.1.1 Cross-Lingual Mapping

We hypothesize that the potentially interesting Latvian news are the ones that are in a joint multilingual space of Estonian and Latvian articles, gravitating closer to the surrounding of each article of the **Estonian$_{\textbf{ground}}$**. To do so, we follow the (Zosa et al., 2020) methodology for extracting articles in a multilingual setting:

1. We use sentence-transformers (Reimers and Gurevych, 2019) *XLM-r-distilRoBERTa-base-paraphrase-v1* embeddings to embed the articles from Estonian$_{\text{ground}}$ and the Latvian$_{\text{train}}$ articles in a common, multilingual space.

2. For each article $E_i \in$ Estonian$_{\text{ground}}$ collection, we select $k \in \{1, 100\}$ closest Latvian articles (based on the Euclidean distance, efficiently computed via a KD-tree (Bentley, 1975) structure), obtaining a collection of Latvian articles $LE_{i,k}$ for each article of the **Estonian$_{\textbf{ground}}$** articles.

3. Finally, we join all of the sets $LE_{i,k}$ from the previous step, obtaining the final **Latvian$_{\textbf{extracted@k}}$** - Latvian extracted set of articles.

At the end of this step, for a given k, we obtain a collection of training articles. The number of articles in the **Latvian$_{\textbf{extracted@k}}$**, for a chosen k is shown in Figure 2.

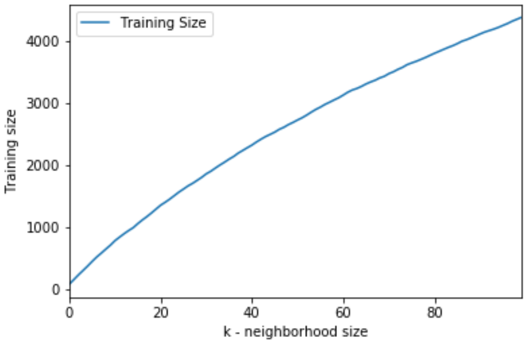

Figure 2: Distribution of articles for given k-neighborhood.

To evaluate the mapping, the Mean Reciprocal Rank (MRR) between the mappings of Estonian to Latvian articles, and vice-versa, were computed for the 21 pairs, where we obtained an average MRR of 66.67%. Even if the linking is incorrect, we assume that even when we do not retrieve the exact match, the articles in the identified neighborhood k still represent a neighborhood of potentially interesting source articles.

4.1.2 Validation Set of Manually Labeled Positive and Negative Examples

For positive examples, we used the 21 manually identified interesting Latvian news $21P$ (see Section 3). However, no negative examples were provided. Therefore, we extracted five random articles for every Latvian article in the $21P$ collection, obtaining a list of 105 articles. A journalist from Ekspress Media manually checked the list and identified 38 articles as unimportant for retrieval. We denote these articles as **NL**. We combined the 21 Latvian examples from the $21P$ collection with the 38 negative articles from the **NL** set, forming a validation set **V**.

[1] Delfi is one of the biggest news portals in Estonia and Latvia, many other media outlets (some of which contributed to the original dataset) often cite this source.

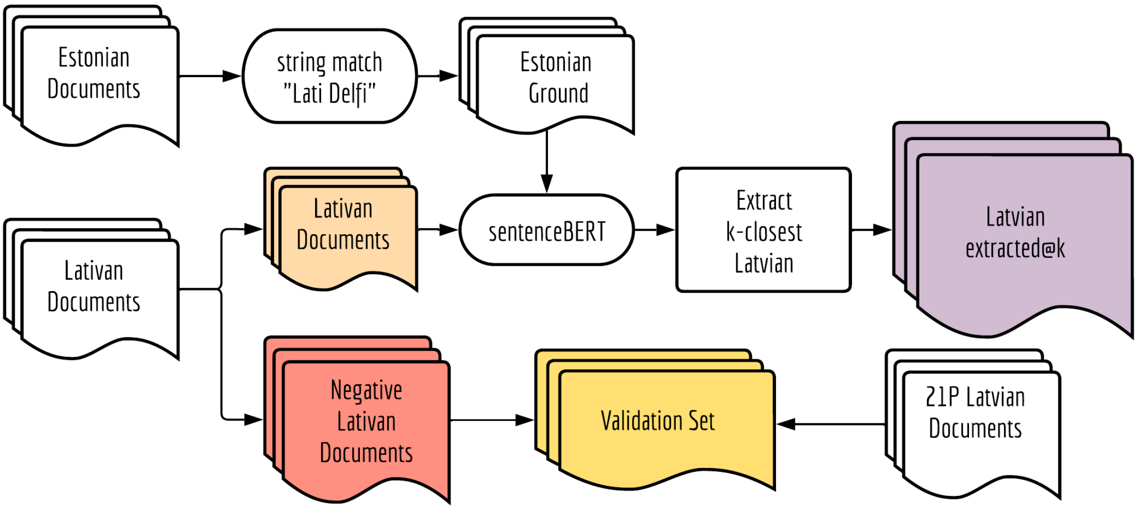

Figure 1: Summarization of our data-acquisition approach.

4.1.3 Experimental Data

We used the following experimental data sets, constructed as explained above:

- The training set $\textbf{Latvian}_{\textbf{extracted@k}}$ consisting of the mapped Latvian **k**-neighborhood articles obtained for every $\text{Estonian}_{\text{ground}}$ article. Figure 2 represents the distribution of articles per various k.

- The validation set **V** consists of 21 positive and 37 negative Latvian examples. The validation set was used to set the classification threshold and evaluate the auto-encoder network, as presented in Section 4.2.

4.2 Learning

We postulate that articles of interest share similar representation patterns. To investigate this hypothesis, we use a set of k Latvian articles from the $\text{Latvian}_{\text{extracted@k}}$ set to learn representations using deep auto-encoder network architectures. We experiment with several deep auto-encoder network architectures to identify the most effective approach. The core concept of the network is to take the original representation of an article, denoted by L_i, and encode it into a lower dimension, obtaining a compressed intermediate representation denoted by C_{L_i}. The encoder part of the network performs the encoding, while the decoder learns to reconstruct the code back to the original representation, yielding a reconstructed representation denoted by L_i^*. By learning these representations, we can better understand the common patterns shared by articles of interest and use this knowledge to improve our retrieval method.

4.2.1 Hyperparameters

We consider using two types of networks for our auto-encoder-based neural network: regularized and non-regularized. To embed the articles, we use the *XLM-r-distilRoBERTa-base-paraphrase-v1* model from sentence-transformers (Reimers and Gurevych, 2019), which converts them to 768-dimensional vectors that serve as input. Our encoder architecture has five layers with 512, 256, 128, 64, and 32 dimensions, while the decoder reverses the same architecture. We use the ReLU (Nair and Hinton, 2010) activation function between layers for all architectures. Figure 3 illustrates the architecture setup.

We optimize our network by using the Mean Squared Error between the reconstructed (L^*) and original (L) representations as the loss function, with the Adam optimizer (Kingma and Ba, 2014) and a learning rate of 0.001. We train for up to 1000 epochs and stop early if we don't improve the validation score in 10 consecutive epochs.

4.2.2 Classification Settings

The auto-encoder outputs the reconstructions of the original input and cannot be used directly for classification. However, in many imbalanced classifications (Zhang et al., 2016) and outlier detection (Chaurasia et al., 2020) problems, the auto-encoder

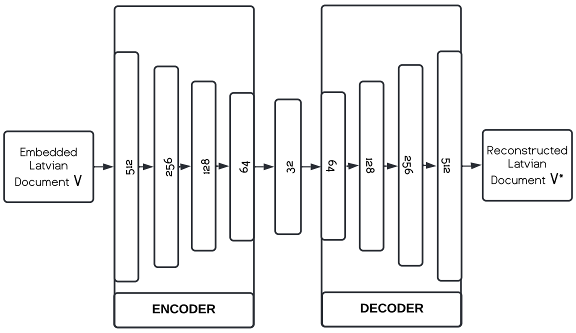
Embedded Latvian Document V
512
256
128
64
32
64
128
256
512
Reconstructed Latvian Document V*
ENCODER
DECODER

Figure 3: Architecture configuration. The encoder and decoder consist of the same architecture.

is used to prioritize outputs based on its reconstruction error (via thresholding). We use the following scoring function:

$$g(L^*, L, t) = \begin{cases} 1 & \text{cosineSimilarity}(L^*, L) \geq t; \\ 0 & \text{otherwise;} \end{cases}$$

where L^* is the reconstructed and, L is the original representation. The classification threshold is denoted by t. To classify an example after a network is trained, we first reconstruct it through the network and apply the classifying function g.

4.2.3 Threshold Learning

In each learning epoch, we reconstruct the validation examples from the set V, which includes 21 positive and 37 negative gold standard examples. This produces a list of reconstructed articles, denoted by $V*$. Then, we measure the reconstruction errors and create a list of errors $R_{k,e}$, where k denotes the population size and e denotes the epoch.

To determine the classification threshold, we search the grid with a step size of 0.01, denoted by stepRange $= [\min(R_{k,e}), \max(R_{k,e})]$. We test each step value as a potential threshold value t. We apply the classifying function g with t and compute the weighted F1-score of the classified reconstructions. We select the t value that yields the optimal F1 score. Formally, we choose t such that:

$$\underset{t \in \text{stepRange}}{\operatorname{argmax}} \left[\text{F1-score}\Big((g(V*, V, t), \text{gold-standard}) \Big) \right]$$

This process enables us to determine the classification threshold that maximizes the F1-score for the reconstructed articles in the validation set, thus providing an effective means for classifying the reconstructed articles.

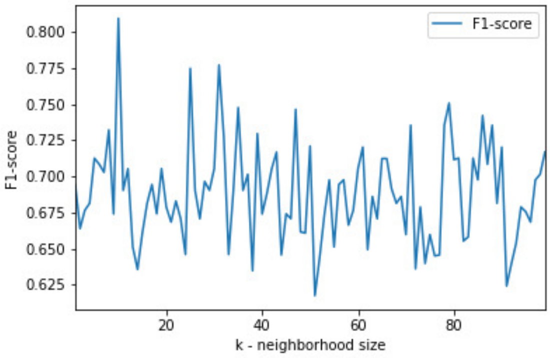

Figure 4: Distribution of F1-scores for the optimal threshold parameter at given k-neighborhood.

The non-regularized *Model32* outperformed the regularized model. Table 1 lists the parameters and evaluations. The model achieved[2] weighted F1-score of 0.81, recall-score of 0.8103, precision score of 0.8087 and accuracy of 0.8103. Figure 4 represents the effect of the training size to the validation score. The confusion matrix for the best-performing validation is listed in Figure 5.

Although the approach necessitates the utilization of negative examples for acquiring the optimal threshold, its purpose is to "regularize" the auto-encoder within the latent space. This ensures that the method doesn't retain specific events in memory but instead contributes to a more effective regularization process.

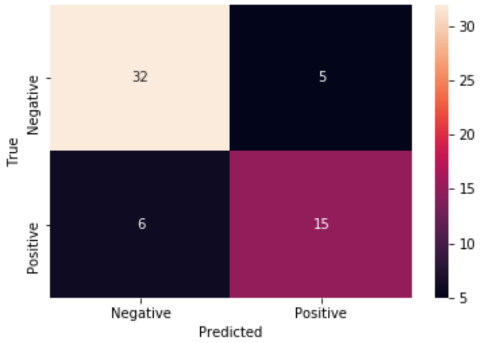

Figure 5: Confusion matrix of the best-performing validation.

5 Evaluation

We evaluate the method in two scenarios, manual and automated. In both systems, we use the test-

[2] True-negatives = 32, False-negatives = 6, False-positives = 5, True-positives = 15

Name	Type	train-size	k-neigh	threshold	epoch	F1-score	Yes	Maybe	No
Model32	Non-regularized	712	10	0.6035	11	**0.8093**	2	2	6
Model32D	Regularized	1951	32	0.5961	5	0.7608	0	2	8
Baseline	Randomized	x	x	x	x	0.4967	0	0	10

Table 1: Summary of the settings and evaluations for the best-performing networks. The optimal threshold is shown in the *threshold* column, followed by the number of epochs trained in the *epoch* column. Finally, the F1-score represents the validation score, followed by the manual evaluations (*YES*/*MAYBE*/*NO*). The human evaluator carried out the evaluations.

ing data for retrieving the top-ranked articles as interesting and relevant.

5.1 Manual Evaluation

We retrieved the top 10 articles (20 in total) in two different network settings and compared them to a baseline (10 randomly chosen articles). To assess the task, we use two different network configurations and a baseline:

- **Model32**, non-regularized network
- **Model32D**, regularized network
- **Baseline**, a random selection of articles sent to the evaluator. We consider the majority *NO* in the F1-score.

The testing data set consists of 1339 articles, which are input to our network and the reconstruction error is measured. The top-10 reconstructed articles with the smallest reconstruction errors are considered potentially interesting and sent to a journalist expert.

A journalist at Ekspress Media manually evaluates the retrieved articles in the categories introduced in (Koloski et al., 2021), i.e., *YES* - the article is definitely relevant, *MAYBE* - the article is relevant to some extent and *NO* - the article is of no relevance. The results are described in Table 1. The journalist found two articles of definitive relevance and 2 of possible relevance for retrieval in the best settings. Given that the problem is difficult, i.e., retrieving very special articles from a large set of all articles, the results still indicate that for Model32, 40% of the articles are potentially interesting. This is slightly lower than the results of (Koloski et al., 2021), wherein the best setting, one more article, was labeled as MAYBE. Of the 30 articles we sent for evaluation to the human evaluator, two were chosen as interesting, four as MAYBE, and the remaining as not interesting.

5.2 Automated Evaluation

This subsection demonstrates that our method performs better than random article retrieval. We first create a test set comprising of $21P$ labeled Latvian articles and the Latvian$_{\text{test}}$ set for automatic evaluation. Next, we run an auto-encoder and measure the reconstruction errors without applying threshold classification. Then, we sort the articles by their reconstruction scores and search for the $21P$ relevant articles while retrieving the top-k articles. We use $Model32$ to calculate the recall@k to assess the performance, treating the $21P$ articles as the gold standard. We also establish a baseline using random scoring of articles, where we randomly shuffle the articles in the test set and conduct 10^6 random evaluations. As shown in Figure 6, the results suggest that our method outperforms the random retrieval method for identifying interesting articles for Estonian readers. Therefore, our method shows promise for further investigation and improvement in the future.

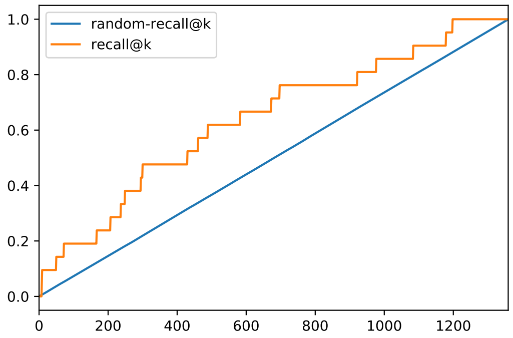

Figure 6: Recall comparison of the distributions. X-axis showcases the number of documents k, while y-axis shows the cumulative recall ($recall@k$).

6 Conclusion and further work

In this work, we have developed an auto-encoder-based approach for detecting and retrieving cross-border news. The method is trained unsupervised, given news datasets in two languages, relevant and non-relevant articles, and potential media houses hot words. The approach is shown to retrieve articles with 40% relevance, as evaluated manually by a media expert, and outperforms random-based approaches through recall@k evaluation.

For further work, we suggest exploring the significance of certain topics and keywords in a given time window, hypothesizing that story/topic relevance is time-dependent. We also propose exploring a term-matching approach that considers named entities and keyword matching to rank the relevance of an article. Lastly, we suggest investigating how combining the SNIR and auto-encoder as a weighted rank score could improve retrieval quality. To improve the relevance of retrieved articles, future work could explore the use of user feedback and relevance feedback mechanisms (such as RLHF). By incorporating user preferences and feedback, the system may be able to better tailor its results to the needs and interests of individual users.

Availability

The code required to replicate the experiments can be found at the following link: https://github.com/bkolosk1/reconstruct_to_retrieve.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding (No. P2-0103) and the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581). A Young Researcher Grant PR-12394 supported the work of the first author.

Limitations

While the method is promising, it has limitations, such as a small evaluation set and the fact that tokens are not masked during retrieval, which may require retraining on a temporal basis. These limitations may affect the generalizability of the method to larger datasets and other languages. In addition, further investigation into the significance of topics and keywords in a given time window and using a term-matching approach could also enhance the method's effectiveness. Additionally, our investigation is limited to comparing the method solely against unsupervised methodologies, which restricts the scope of our work and opens up possibilities for further improvement.

Ethics Statement

The authors have used only existing datasets and do not identify any elements for ethical considerations.

References

Zakaryia Mustafa Almahasees. 2018. Assessment of google and microsoft bing translation of journalistic texts. *International Journal of Languages, Literature and Linguistics*, 4(3):231–235.

Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Nasir Mahmood, and Waqar Mahmood. 2019. The use of ontology in retrieval: A study on textual, multilingual, and multimedia retrieval. *IEEE Access*, 7:21662–21686.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 2, pages 19–26.

Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517.

Esperança Bielsa and Susan Bassnett. 2008. *Translation in global news*. Routledge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Siddharth Chaurasia, Sagar Goyal, and Manish Rajput. 2020. Outlier detection using autoencoder ensembles: A robust unsupervised approach. In *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pages 76–80.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of bert for neural machine translation. *arXiv preprint arXiv:1909.12744*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kyle Conway and Lucile Davier. 2019. Journalism and translation in the era of convergence. *Journalism and Translation in the Era of Convergence*, pages 1–217.

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*.

Samira Douzi, Meryem Amar, and Bouabid El Ouahidi. 2017. Advanced phishing filter using autoencoder and denoising autoencoder. In *Proceedings of the International Conference on Big Data and Internet of Thing*, BDIOT2017, page 125–129, New York, NY, USA. Association for Computing Machinery.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *LREC*.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *Icwsm*, 7(21):219–222.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Neda Keivandarian and Marco Carvalho. 2023. A survey on sentiment classification methods and challenges. In *The International FLAIRS Conference Proceedings*, volume 36.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022a. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, 496:208–226.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2022b. Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 400–409, Marseille, France. European Language Resources Association.

Boshko Koloski, Elaine Zosa, Timen Stepišnik-Perdih, Blaž Škrlj, Tarmo Paju, and Senja Pollak. 2021. Interesting cross-border news discovery using cross-lingual article linking and document similarity. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 116–120.

Tadashi Kumano, Isao Goto, Hideki Tanaka, Noriyoshi Uratani, and Terumasa Ehara. 2002. A translation aid system by retrieving bilingual news database. *Systems and Computers in Japan*, 33(8):19–29.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. 2014. Autoencoder for words. *Neurocomputing*, 139:84–96.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791.

Xinyu Ma, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. A contrastive pre-training approach to discriminative autoencoder for dense retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4314–4318.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2022. Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, 28(4):409–448.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, et al. 2021. Embeddia tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Cyriana M.A. Roelofs, Marc-Alexander Lutz, Stefan Faulstich, and Stephan Vogt. 2021. Autoencoder-based anomaly root cause analysis for wind turbines. *Energy and AI*, 4:100065.

Vishal S Shirsat, Rajkumar S Jagdale, and SN Deshmukh. 2017. Document level sentiment analysis from news articles. In *2017 international conference on computing, Communication, Control and Automation (ICCUBEA)*, pages 1–4. IEEE.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Haonan Tong, Bin Liu, and Shihai Wang. 2018. Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning. *Information and Software Technology*, 96:94–111.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 72–79.

Roberto A. Valdeón. 2020. On the interface between journalism and translation studies: A historical overview and suggestions for collaborative research. *Journalism Studies*, 21(12):1644–1661.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9266–9273.

Xing Xu, Jialin Tian, Kaiyi Lin, Huimin Lu, Jie Shao, and Heng Tao Shen. 2021. Zero-shot cross-modal retrieval by assembling autoencoder and generative adversarial network. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–17.

Chenggang Zhang, Wei Gao, Jiazhi Song, and Jinqing Jiang. 2016. An imbalanced data classification algorithm of improved autoencoder neural network. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pages 95–99.

Yifei Zhang and Hao Zhu. 2020. Discrete wasserstein autoencoders for document retrieval. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8159–8163.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Elaine Zosa, Mark Granroth-Wilding, and Lidia Pivovarova. 2020. A comparison of unsupervised methods for ad hoc cross-lingual document retrieval. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 32–37, Marseille, France. European Language Resources Association.