Multilingual News Recommendation using Graph Neural Networks

Marat Bekmyrza [1]

Abstract

News Recommendation System is a vital tool for readers and news providers in the era of information overload. Finding relevant news articles is a time consuming process, especially for people who read news from various sources in different languages which is a common case for international students. However, the topic of multilingual news recommendation is not well researched in the literature. Text processing is critical in understanding news content, but most of the available tools mainly adopted only for English. Therefore, this work explores how we can build a recommendation system that understands multiple languages. Proposed method combines Pretrained-Language Models (PLMs) to encode article titles and Graph Neural Networks (GNNs) to learn user interactions. Experimental results on open source datasets, such as MIND, Adressa and CI&T, show 10+ improvement in AUC and F1 scores compared with previous works.

1. Introduction

It is important to research how we can analyze news provided in different languages for two reasons. The first is more personal and relates to people who like to read news from multiple sources in different languages. This situation is common among international students in English speaking countries who want to get up-to-date news about their current residing country and also about their home country. Therefore, we were interested in how to build news recommendation system that can take news articles from different sources and suggest the ones that user would like to read. The second reason is that most of the existing models cannot be directly applied to build a recommendation system for various languages. Text processing is critical in news recommendation and most of the available tools adopted for English only. In this work, we analyze whether it is sufficient to just translate the data to English to build the recommendation models.

The traditional recommendation systems can be broadly classified into two types (Javed et al., 2021): *content-based filtering* is based on product features and user's past activity, it suggests a new item by comparing its features with the features of user's previously liked items; *collaborative filtering* approach recommends a new product that has been liked by other users with similar preferences. In contrast to conventional recommendation tasks, news recommendation has the following challenges. First, online news articles update very quickly. New articles are posted continuously and existing ones expire in short time. Thus, news recommendation faces a severe *cold-start problem* (Das et al., 2007). It occurs when there is insufficient information about interactions of a new user or a new product that has recently entered the system. Second, news articles contain rich textual information and it is important to correctly understand the content meaning from a text (Kompan & Bieliková, 2010). Text representations highly vary depending on the choice of language processing technique. Third, there is no explicit rating of news articles posted by users on news platforms. Hence, it is necessary to find indirect ways of understanding user interest. It is usually inferred by their implicit interaction, such as their clicking behavior on the news article links (Ilievski & Roy, 2013).

The research question that we want to answer in this project is how to build a recommendation system that understands multiple languages. We use different publicly available datasets: MIND in English (Wu et al., 2020a), Adressa in Norwegian (Gulla et al., 2017), and CI&T in English and Portuguese (Moreira, 2017). These datasets contain click history of users and textual information about articles. We apply Pretrained- Language Models (PLMs) for representation of textual information and build a model based on Graph Neural Networks (GNNs) for user behaviour prediction. We encode texts in their original language using language specific PLMs, as well as using multilingual PLMs. We also translated all non-English data to English and applied English-based PLMs to explore whether translation is sufficient to build a recommender and how it compares with language specific PLMs.

[1]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. Correspondence to: Marat Bekmyrza <marat.bekmyrza@uwaterloo.ca>.

CS886 Advanced Topics in Artificial Intelligence: Graph Neural Networks, University of Waterloo, Waterloo, Canada. Winter 2023.

Main Results

- Translated models could produce comparable results as the models designed using the data in its original language. Thus, in case where the specific language under consideration has no original PLM and also is not supported by the multilingual PLMs, one can translate their data to English and apply classic English PLMs, such as BERT (Devlin et al., 2018), without experiencing significant performance drops.
- Our model produced significantly higher AUC and F1 scores compared to previous models from (Hu et al., 2020) and (Wu et al., 2021) which is the result of combining two current powerful tools, PLMs and GNNs. Code is publicly available [1].
- From our experiments, mixing data in different languages for the same users showed higher scores than the models trained with separate data for each language.

2. Related Work

Over the past years, there were many survey papers that review the techniques of news recommendation and ways to improve them. (Wu et al., 2022) reviewed the methods and challenges of personalized news recommendation systems. They found that traditional recommendation systems used collaborative filtering for news modeling but it has certain challenges such as cold-start problem because of continuous dynamic changes to news. There have been many attempts to tackle the cold-start problem. For example, (Zihayat et al., 2019) developed a framework that combines utility and probabilistic models to address the lack of initial information.

Current approach for news recommendation consists of two parts: understanding news text and modelling user clicking behaviour. Recently, many focused on news modeling with Natural Language Processing (NLP) techniques (Wang et al., 2020), which is an important step in news recommendation, to understand the content of candidate news and a prerequisite for inferring user interests from clicked news. For example, (Wang et al., 2018) proposed to use a knowledge-aware CNN network to learn news representations from embeddings of words and entities in news title.(Okura et al., 2017) proposed to use autoencoders to learn news representations from news content. PLMs have achieved great success in NLP due to their strong ability in text modeling (Devlin et al., 2018).

Another important aspect to build a good news recommendation system is to have a high-quality benchmark dataset. For our work, we have used open source dataset from different languages. (Wu et al., 2020b) presented a large-scale dataset named MIND and showed that the performance of news recommendation highly relies on the quality of news content understanding and user interest modeling.

Similar to text representation methods, user behaviour modeling can be also classified into two categories: feature-based and deep learning-based. (Garcin et al., 2012) proposed to use Latent Dirichlet Allocation (LDA) to extract topics from the concatenation of news title, summary and body. There are several methods that use neural networks to learn user click behaviors. For example, (Okura et al., 2017) proposed to use a Gated Recurrent Unit (GRU) network to learn user behaviour.

All of these works focused mainly on monolingual recommendation systems. One of the few papers that work with multilingual data is (Wu et al., 2021). They propose to use PLMs for textual data representation and collected multilingual dataset from MSN News platform on user impressions from 7 countries with different language codes. However, this dataset is not publicly available.

3. Methodology

3.1. Problem Formulation

For the context of our problem, let's say we have an app that aggregates daily news articles from multiple platforms in different languages. It also can be a single platform that provides news in different languages, such as Google News. The key challenge here is having news in multiple languages. In order to design recommendation system, we need to understand different reading behaviours of users. There are users with mixed preferences, i.e., people who like to read political news on the same topic in different languages. On the other hand, users might have disjoint preferences on different platforms. This can be people who like to read about sports in one language (e.g., in Spanish), and about business in another language (e.g., in English). Complicated reading behaviours suggest to train separate recommendation model for each language since we are assuming that user histories on different platforms might have disjoint reading behavioural patterns. For example, Aisha is from Kazakhstan and likes to read articles about political and economic situation in her home country from local news website in Kazakh language, while she prefers to read about Artificial Intelligence (AI) and Machine Learning (ML) from an English source. She does not want to get suggestions about news in Kazakhstan from an English source since they might have insufficient coverage and lack local expertise. In the same way, she believes that most of the AI and ML advancements are published in English and are updated with higher frequency. Looking at Aisha's reading

[1] https://github.com/mbekmyrz/newsrec

behaviour, we see that above two patterns can be considered disjoint and thus treat her as two separate users in Kazakh and English languages. So, we can train one model to understand Kazakh language and recommend news to Aisha about Kazakhstan, and train another model to work with English data which will capture Aisha's interest in AI and ML news.

The news recommendation problem in our paper can be illustrated as follows. We have the click histories for K users $U = \{u_1, u_2, u_3, ..., u_k\}$ over M news items $I = \{d_1, d_2, ..., d_M\}$. The $Y \in R^{KXM}$ is defined according to users implicit feedback, where $y_{u,d} = 1$ indicate the user u clicked news d, otherwise $y_{u,d} = 0$. In this work, we process only the news titles to generate article features. Each news title T contains a sequence of words $T = \{w_1, w_2, ..., w_m\}$. Each word is tokenized and supplied to PLM, which returns a feature vector $f = \{x_1, x_2, x_3, ...x_{768}]\}$ of standard size of 768 for each news title. So, the PLM returns a feature matrix F^{MX768}.

Given the user-item interaction matrix Y and feature matrix F, we aim to predict whether a user u has potential interest in a news item d as a link prediction task on a bipartite graph.

3.2. Proposed Solution

For the news modelling, we decided to use only titles of news articles for two reasons. The first reason is that news websites and mobile apps usually present news on their pages as a list of article titles with some of them containing images and little abstract. Therefore, we argue that users decide to click on the article link mainly by looking at article titles. Even though incorporating all the news features gives better results compared to using only title information, the difference is not significant. This was shown in (Wu et al., 2020b), where AUC score by using only title information was 66.22 compared to 67.6 produced by including all the features, such as title + abstract + body + category + entities. The second reason is the computation limitations in terms of hardware resources and time. Processing of all the features would have increased the complexity of the models and require more time for training.

Titles of the Adressa Norwegian and the CI&T Portuguese datasets were translated into English using OpenNMT[2] (Klein et al., 2017) and Google Translate tools respectively.

For the processing of the textual data, we applied the base versions of the following PLMs available on Hugging Face. For monolingual data:

- English: *bert-base-uncased* denoted as BERT in (Devlin et al., 2018).
- Norwegian: *NbAiLab/nb-bert-base* denoted as nbBERT from (Kummervold et al., 2021).
- Portuguese: *neuralmind/bert-base-Portuguese-cased* denoted as ptBERT from (Souza et al., 2019).

For multilingual data:

- BERT multlingual: *bert-base-multilingual-cased* denoted as mBERT from (Devlin et al., 2018).
- InfoXLM: *microsoft/infoxlm-base* denoted as InfoXLM from (Chi et al., 2020).

All PLMs were available in the python *transformers* library from Hugging Face. Each word in a title was first tokenized and then the list of title tokens supplied to the PLM model to generated text features. All the PLMs provide features vector of size 768 for each token. Standard output from the PLMs is two tensors: last hidden state values and pooler output. Last hidden state contains features for each token in a sample. Whereas, pooler output is the features of the first token transformed with a linear layer. We selected the pooler output since it encapsulates all the context information within a sample. Other alternative approaches include averaging of all the tokens in the last hidden state or applying another learn able attention layer. Attention network showed better results compared to the first token representation and averaging in (Wu et al., 2021). However, we decided to work with already available pooler output due to time limitations.

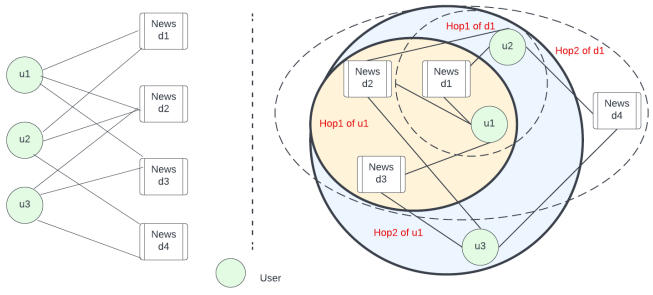

Figure 1. Heterogeneous bipartite user-article graph (left) and two GNN convolution layers (right). The first and the second hop vertices are shown for the user u_1 and article d_1.

For the user interest modelling, heterogeneous bipartite graph was constructed from the user-article interactions provided in the datasets. The graph contains two types of vertices: users and articles. An edge connects user vertex with a clicked article vertex. Set of neighbour vertices of a user is their complete reading history. Reverse edges from articles to users were added in order to let a GNN be able

[2] https://github.com/OpenNMT/OpenNMT-py

to pass messages in both directions. Graph contains all the users and articles available in a dataset since most of GNNs are designed for static graphs and do not support addition and deletion of a node. The problem now can be considered as a link prediction in a graph which is a common strategy for recommender systems with user-item like structure. Therefore, we divide data into train, validation and test sets by masking the existing edges. We also add negative edges that represent non-existing edges, i.e., connect users with articles they did not read. Edge label is added to differentiate positive and negative edges. Task now is a binary classification to predict edge labels: 1 - user read the article, 0 - user did not read the article. Existing edges are randomly divided into train, validation and test by 80%, 10%, 10% ratio respectively. During training, we use 70% of edges for message passing, i.e., for construction of the training graph. Other 30% is used for supervision. Negative edges are generated on-the-fly with the ratio of 1:1 or 2:1 to the number of positive edges chosen as a hyper parameter. Article vertex in the graph has a feature vector initialized with values derived by PLMs. User vertex feature is learned during training by aggregating information from neighbours.

Model structure contains the following parts: input layer, two GNN convolution layers with ReLU non-linearity in-between, classifier. For the input layer we chose embedding and linear layers which are similar, but embedding layer does not take initial features and learns node id based representations. In the first model denoted as Embedding in Table 2, both embedding and linear layers applied to the article node and only embedding layer to the user node. In the second model denoted as Linear in Table 2, user vertices are initialized randomly and only linear layer was applied to both user and article nodes. GraphSAGE (Hamilton et al., 2017) and GAT (Veličković et al., 2017) were chosen for the GNN convolution layers since they can be adopted to work with bipartite graphs. GraphSAGE samples neighbour vertices and applies mean aggregation. Whereas, GAT applies attention based aggregation in which learnable attention coefficients are multiplied with the node feature and to the neighbour vertices before summation. Final classification layer applies dot-product between embeddings of the candidate news article and the user vertices.

Figure 1 shows heterogeneous bipartite graph constructed from the user-article relations. For example, u_1 has read articles d_1, d_2, and d_3. In the first GNN convolution layer, information from the direct neighbours, hop 1, is gathered and the second layer collects information from the neighbours of neighbours, hop 2. In this way, information is passed along the graph containing complex user-article interactions.

4. Results

4.1. Datasets

In the news recommendation domain there are very few datasets that are openly accessible for research. The following ones are the most commonly used datasets: MIND[3] in English (Wu et al., 2020a) , and Adressa in Norwegian[4] (Gulla et al., 2017) , and CI&T[5] in English and Portuguese (Moreira, 2017).

MIND was collected from the user behavior logs of Microsoft News of 1 million users who had at least 5 news click records during 6 weeks from October 12 to November 22, 2019. Adressa dataset, was constructed from the logs of the Adresseavisen website in ten weeks. However, we used the version for one week due to large data sizes. Each click event contains several features, such as session time, news title, news category and user ID. Each news article is associated with some detailed information such as authors, entities and body. CI&T is a rich and rare dataset contains a real sample of 12 months logs (Mar. 2016 - Feb. 2017) from CI&T's Internal Communication platform (DeskDrop). This dataset features some distinctive characteristics Item attributes: Articles' original URL, title, and content plain text are available in two languages (English and Portuguese). The uniqueness of this dataset is that it contains users who read news in different languages. There are samples where user has read both English and Portuguese article. Dataset also contains contextual information, like users visit date/time, client (mobile native app / browser) and geolocation. Table 1 summarizes the dataset contents.

Dataset	Language	Users	News	Clicks	News data
MIND	en	1M	160K	15M	tit + bod + abs + cat + ent
Adressa	no	535K	15K	1.8M	tit + bod + abs + cat + ent
CI&T	pt	1.6K	821	14K	tit + bod + cont + cat
CI&T	en	1.6K	2.1K	26K	tit + bod + cont + cat
CI&T	pt + en	1.8K	2.9K	40K	tit + bod + cont + cat

Table 1. Datasets information about the number of unique users and articles: en - English, no - Norwegian, pt - Portuguese, cat, ent and con stand for category, entities and context.

4.2. Evaluation

We implemented our model using PyTorch Geometric (PyG) library and run Google Colab GPU with batch size of 128. For the training, binary cross entropy is used to compute the loss. AUC and F1 scores were used as classification metrics to evaluate the different model performances.

Table 2 shows the scores on the test set of the best models on the validation set. mBERT and InfoXLM are applied

[3] https://msnews.github.io

[4] http://reclab.idi.ntnu.no/dataset/

[5] https://www.kaggle.com/gspmoreira/datasets

to extract features from the versions of the datasets in their respective original language. BERT is applied to English datasets and to the translated data.

Dataset	Language	PLM	GNN	AUC	F1	AUC	F1
-	-	-	-	Embed		Linear	
Adressa	no	nbBERT	GAT	98.6	95.6	97.1	92.1
Adressa	no	mBERT	SAGE	98.9	95.2	98.6	94.5
Adressa	no	InfoXLM	SAGE	**99.1**	**96.6**	98.4	94.4
Adressa-Tr	en	BERT	SAGE	98.8	96.5	97.6	92.3
CI&T	pt	ptBERT	SAGE	82.4	70.7	79.8	73.5
CI&T	pt	mBERT	GAT	81.0	70.2	80.8	73.5
CI&T	pt	InfoXLM	SAGE	80.7	71.6	78.4	72.5
CI&T-Tr	en	BERT	SAGE	**83.0**	**73.6**	76.2	71.7
CI&T-En	en	BERT	SAGE	**87.2**	**78.2**	83.3	77.0
MIND	en	BERT	SAGE	94.6	82.6	86.1	79.4
MIND	en	mBERT	SAGE	**95.3**	**84.4**	84.3	74.5
MIND	en	InfoXLM	SAGE	95.2	83.2	86.0	75.3

Table 2. Results of our model on different datasets with the feature extracted by the specified PLM. Adressa-Tr and CI&T-Tr denote the translated versions of the datasets. CI&T-En is the portion of the dataset which was originally in English.

Table 3 shows the best results of the models from (Hu et al., 2020) for the Adressa and from (Wu et al., 2021) for the MIND. MSN is the proprietary multilingual dataset collected in (Wu et al., 2021). (Hu et al., 2020) uses Convolutional Neural Networks (CNN) to extract text features and applies two-layer Graph Convolution Network (GCN) for the user modelling. (Wu et al., 2021) uses different PLMs, such as RoBERTa (Liu et al., 2019), UniLM (Bao et al., 2020) and InfoXLM (Devlin et al., 2018), to represent textual data and got the best results for each PLM by using EBNR (Okura et al., 2017) and NRMS (Wu et al., 2019) techniques for the user modelling. They also applied multilingual InfoXLM for the data which was originally in English.

Dataset	Language	News Model	User Model	AUC	F1
Adressa	no	CNN	GCN	84.0	83.9
MIND	en	–	EBNR	68.2	–
MIND	en	BERT	EBNR	69.6	–
MIND	en	RoBERTa	EBNR	69.7	–
MIND	en	UniLM	NRMS	70.6	–
MSN	en	InfoXLM	NRMS	64.3	–

Table 3. Results of the models taken from the literature. MSN is a proprietary multilingual dataset.

Models in the Table 4 are trained and tested on the different combinations of the CI&T dataset. Mixed data means that we concatenated both Portuguese and English parts of the dataset and supplied them to the multilingual PLMs, such as mBERT and InfoXLM, in their original languages. Since multilingual PLMs can understand multiple languages we do not need translation here. Whereas, translated test data contained all the users in both parts of the dataset, however only the features of the English articles and their respective edges were used for training and Portuguese data was translated and concatenated with the English data for testing. Translated train-test uses concatenation of the translated data and originally English portion of the dataset for both train and test.

We also explored the importance of including article features extracted by PLMs by trying different initializations. We used feature tensor of containing only zeros and random data as the article nodes features fed to the GNN model. Here, and in some of the later results, we used only the Adressa and the CI&T due to the large size of the MIND.

Data	Train	Test	PLM	AUC	F1
Mixed	pt + en	pt + en	mBERT	84.2	77.9
Mixed	pt + en	pt + en	InfoXLM	84.3	76.6
Translated test	en	tr + en	BERT	93.7	73.3
Translated train-test	tr + en	tr + en	BERT	**95.8**	**95.1**

Table 4. Results of the SAGE model on different combinations of the CI&T dataset. Here *tr* denotes the translated data from Portuguese to English.

Dataset	Features	AUC	F1
Adressa	zero	91.0	88.0
CI&T	zero	81.2	61.9
CI&T	random	81.3	59.7

Table 5. SAGE model results without PLM processing and different initialization of the article feature matrix.

4.3. Hyperparameter Tuning

We varied different hyperparameter values and recorded their influence on the model performances. Negative sampling ratio denotes the ratio of artificial non-existing edges added with respect to the number of positive edges. Sampled neighbours has to values and the first shows how many neighbours were sampled in the first hop, and similarly the second one is for the second hop. Hidden channels of the network is dimension of the weight matrices used by GNN Convolution layers. All hyperparameter tuning results were recorded using SAGE convolution.

Hyperparameter	Values
Negative sampling ratio	1, 2
Batch size	128
Sampled neighbors	[10,5], [10,10], [20,10], [20,20]
Hidden channels	16, 32, 64
GNN convolution layer	SAGE, GAT

Table 6. Hyperparameter values used for training.

5. Discussions

The first observation is that translation models could produce comparable results as the models with the original data. For example, from the Table 2 we see that the best AUC and F1 scores were achieved by the multilingual InfoXLM

PLM	ΔAUC	ΔF1
ptBERT	0.54	8.35
mBERT	0.82	8.76
InfoXLM	0.72	11.3

Table 7. Influence of negative sampling ratio r shown as the difference of AUC and F1 scores on the CI&T dataset. Results for $r = 2$ subtracted from the baseline $r = 1$.

PLM	ΔAUC	ΔF1
ptBERT	2.02	-0.48
InfoXLM	-0.57	1.42

Table 8. Influence of sampled neighbors n shown as the difference of AUC and F1 scores. Results for the baseline $n = [10, 5]$ are subtracted from $n = [20, 20]$ so that positive values show the impact of increasing number of sampled neighbors.

Data	PLM	Size	ΔAUC	ΔF1
Adressa	nbBERT	32	-0.15	0.74
Adressa	nbBERT	64	0.04	0.80
Adressa	mBERT	32	0.18	0.58
Adressa	mBERT	64	0.03	0.55
CI&T	ptBERT	32	0.21	-5.89
CI&T	ptBERT	64	-1.09	-21.1
CI&T	mBERT	32	1.38	-8.98
CI&T	mBERT	64	0.96	-19.2

Table 9. Influence of hidden channels size s shown as the difference of AUC and F1 scores. Results for the baseline $r = 16$ are subtracted from $s = 32$ and $s = 64$ so that positive values show the impact of increasing hidden size.

and SAGE for the Adressa. However, it can be noted that the translated version applied to BERT and SAGE has also comparable scores and the difference is insignificant and might be affected by different sampling of train and test edges. Whereas, BERT applied to the CI&T-Tr, which is the translated to English portion of the CI&T Portuguese, outperformed other models based on the original portion of the dataset. This suggests that in cases where the specific language under consideration has no original PLM and also is not supported by the multilingual PLMs, one can translate their data to English and apply BERT without experiencing significant performance drops. This is already shown for different NLP tasks. However, to our best knowledge, this is the first time, it was applied to news recommendation problem.

Another key observation is that our model produced significantly higher AUC and F1 scores compared to the CNN+GCN model (Hu et al., 2020) on the Adressa dataset as shown in Table 3. There might be two reasons for this. The first is the introduction of the PLM model for the feature extraction compared to the CNN model. However, this assumption was undermined by the results in Table 5. We found out that even without PLM, with zero initialization of the features for Adressa, the AUC score was reaching 90+ and F1 higher than 85 which are higher than the best result of the model in (Hu et al., 2020) which also uses GNN. Therefore, we think there is another reason behind this performance which might be based on the differences of the SAGE and GAT convolution from the GCN and on the difference of data preprocessing. Our data preprocessing contains random shuffle of the whole click records, whereas (Hu et al., 2020) uses day-by-day separation, i.e., use the first 6 days for training and the last day of for validation and testing. It was shown by (Cheng, 2020) that day-by-day approach produces lower AUC scores and it is better to random shuffle data containing all the days.

The reason why model was showing high scores despite removing PLM processing shows that GNN can capture inherent implicit user-article interactions passing information within all the graph. It generates user embedding not only based on direct article reading history, but also collects information from the similar users who read the same article in the second hop of the convolution. In this way, model can cluster users with similar interests with close embeddings. This is supported by the fact that our model was outperforming models with PLMs but with non-GNN user modelling in (Wu et al., 2021). This can be tested in the future work by analyzing suggested articles to the users with close distances between their embeddings.

Regarding models with translated models on mixed data, Table 4 shows the results of the CI&T datasets with different mix of its parts. Here we see significant improvement from the baseline models in the Table 2. Baseline models were trained using only one portion of the data, i.e., only in one language. Whereas Table 4 shows that having all the data about user history in different languages shows significant performance improvements. This suggests that the most of the users have mixed preference reading behaviour. In other words, users tend to read news with similar characteristics even if the news articles are in different languages.

Hyperparameter tuning shows that having 1:1 negative sampling ratio shows better performance in Table 7. This is probably because models get unbalanced negative classes for the 2:1 ratio and tend to have biased classification. Results by varying the number of sampled neighbours is not conclusive, since the we have negative and positive values in both columns in Table 8. However, having $[10, 5]$ would decrease the model complexity and require less time for training. Similarly, in Table 9 we see both negative and positive results when channel size is increased from the baseline of 16. However, the distinction here lies in the size of the datasets. So, for the smalle CI&T dataset smaller hidden channel size might prevent the model from over fitting and have better generalization than higher sizes. Whereas, for the larger Adressa having higher hidden channel size tend to benefit the performance.

Due to time limitations, we could not conduct experiments

which might be interesting for future work. For example, including all the news features, such as title, astract, body and entities, performance comparison with the fixed word embedding models, such as GloVe, increasing of the number of GNN convolution layers and try more complex convolution types. We also think that adding attention layers to the output of the GNN layers could produce better performance results. Another approach might be to introduce separate aggregation functions for user and article nodes, and also try to have different edge types to indicate that user read one article multiple times.

6. Conclusion

In this paper we present our work on multilingual news recommendation system. We built and conducted extensive experiments on original language and translated datasets. The results showed that incorparting GNN shows great performance improvements and that translation to English is sufficient to build a model and get comparable results with the original language.

References

Bao, H., Dong, L., Wei, F., Wang, W., Yang, N., Liu, X., Wang, Y., Gao, J., Piao, S., Zhou, M., et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pp. 642–652. PMLR, 2020.

Cheng, H. Mind report, 2020. https://msnews.github.io/assets/doc/1.pdf.

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., and Zhou, M. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*, 2020.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pp. 271–280, 2007.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Garcin, F., Zhou, K., Faltings, B., and Schickel, V. Personalized news recommendation based on collaborative filtering. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pp. 437–441. IEEE, 2012.

Gulla, J. A., Zhang, L., Liu, P., Özgöbek, Ö., and Su, X. The adressa dataset for news recommendation. In *Proceedings of the international conference on web intelligence*, pp. 1042–1048, 2017.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Hu, L., Xu, S., Li, C., Yang, C., Shi, C., Duan, N., Xie, X., and Zhou, M. Graph neural news recommendation with unsupervised preference disentanglement. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4255–4264, 2020.

Ilievski, I. and Roy, S. Personalized news recommendation based on implicit feedback. In *Proceedings of the 2013 international news recommender systems workshop and challenge*, pp. 10–15, 2013.

Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., and Luo, S. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, 2021.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

Kompan, M. and Bieliková, M. Content-based news recommendation. In *International conference on electronic commerce and web technologies*, pp. 61–72. Springer, 2010.

Kummervold, P. E., De la Rosa, J., Wetjen, F., and Brygfjeld, S. A. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 20–29, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.3.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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.

Moreira, G. Articles sharing and reading from CI&T DeskDrop. https://www.kaggle.com/gspmoreira/datasets/, 2017.

Okura, S., Tagami, Y., Ono, S., and Tajima, A. Embedding-based news recommendation for millions of users. In

Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1933–1942, 2017.

Souza, F., Nogueira, R., and Lotufo, R. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019. URL http://arxiv.org/abs/1909.10649.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Wang, H., Zhang, F., Xie, X., and Guo, M. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, pp. 1835–1844, 2018.

Wang, H., Wu, F., Liu, Z., and Xie, X. Fine-grained interest matching for neural news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 836–845, 2020.

Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., and Xie, X. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 6389–6394, 2019.

Wu, C., Wu, F., Qi, T., and Huang, Y. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1652–1656, 2021.

Wu, C., Wu, F., Huang, Y., and Xie, X. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems (TOIS)*, 2022.

Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., and Zhou, M. Mind: A large-scale dataset for news recommendation. In *ACL*, pp. 3597–3606, 2020a.

Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3597–3606, 2020b.

Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., and An, A. A utility-based news recommendation system. *Decision Support Systems*, 117:14–27, 2019.