A Benchmark for Text Classification in News Recommendations

Xinyi Li, Edward C. Malthouse

[1]*Northwestern University, Evanston, Illinois, USA*
[2]*Northwestern University, Evanston, Illinois, USA*

Abstract

Text classification is an important task in natural language processing. In the current era, people mainly obtain information from online news resources. It is then important to have an automatic and accurate news classifier to categorize every day's news stories such that readers can find articles of interested more easily. We use news story data from the McClatchy organization to establish benchmarks on how accurately stories can be classified by multiple existing deep learning classifiers. Among the models we evaluated, Bidirectional Encoder Representations from Transformers (BERT) provides the best accuracy, macro-averaging precision, micro-averaging precision, macro-averaging recall and micro-averaging recall. Different from many other benchmark news data set, McClatchy provides both headline and full-text for each news story. We compare the performance of every deep learning-based classifier using headlines versus full-texts—the top three predicted categories include the labeled value 95% of the time with full-texts training and 92% with headlines only. Furthermore, the defined topics in McClatchy are not mutually exclusive. Some predictions identified as inaccurate are in fact classified into reasonable topics. We further provide a visualization of stories from various defined topics. The predicted results and the visualization of news stories illustrate the untrustworthiness of labeled classes and the intrinsic difficulty of categorizing news stories.

Keywords

News recommendations, news taxonomy, news topic categorization, text classification, deep learning

1. Introduction

The newspaper industry has seen a steady and steep decline over the past decade in part because traditional, ad-supported revenue models are no longer viable. There have been widespread layoffs and closures, resulting in 'ghost newspapers' and 'news deserts', where almost 200 out of 3,143 counties in the U.S have been left with no daily newspaper and 1,540 counties with only one weekly newspaper [1]. The demise of local newspapers is not only a business problem, but also a public-good and societal problem. Communities without news organizations have seen an increase in government spending due to a lack of accountability [2]. Citizens who consume less news are unable to evaluate elected officials and less likely to vote. Reading news is one way for people to gain knowledge and to become more open-minded. It is therefore important

Joint Proceedings of 10th International Workshop on News Recommendation and Analytics (INRA'22) and the Third International Workshop on Investigating Learning During Web Search (IWILDS'22) co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22), July 15, 2022, Madrid, Spain

XinyiLi2024@u.northwestern.edu (X. Li); ecm@northwestern.edu (E. C. Malthouse)
0000000205411795 (X. Li); 0000000170770172 (E. C. Malthouse)

to help local news organizations find a viable revenue model, which relies on a larger portion of subscriber revenue [3]. Retaining a subscriber depends on developing a reading habit [4], and recommmender systems (RS) can play a vital role in helping readers find stories of interest and creating the habit.

News RS are different from other RS in eCommerce [5] because only focusing on accuracy might lead to adverse social effects [6]. If a news RS keeps recommending items based on a user's preference, readers might have narrow opinions, lack certain information because they are in a filter bubble [7], and be in an echo chamber [8]. The spread of information online is rapid and wide [9]. Once an echo chamber is generated, the reader can become a spreader of online fake news [10], which will cause serious social issues and misinformation. The collapse of local journalism also exposes the public to the risk of receiving and spreading misinformation [11]. Raza stated that improving the diversity of recommendations can solve these problems to some extent [6].

Diversity can be measured from the content perspective[12]. In general, improving the diversity of recommended items can prevent users from falling into a similarity hole [13, 14]. First, in the domain of news, improving the diversity can relieve echo chambers. Beam *et al.* [15] and Hannak *et al.* [16] found that news diversity plays an important role in shaping people's political and public opinion. Lee *et al.* [17] showed that diversity in news recommendations would increase users' satisfaction. Therefore, it is necessary to have a taxonomy to understand a text. A second reason to automatically classify news topics is to help newsrooms allocate resources. News organizations can count the number of stories of different types read by subscribers and use the counts in churn models to understand what types of stories are associated with retaining a subscriber versus driving one to churn [3]. These insights can help news managers assign reporters to cover stories that will engage and retain paying customers, as opposed to stimulating page views to generate dwindling ad revenue, often with sensational content. Such churn models require a reliable content taxonomy, and therefore it is necessary to develop an automatic news classifier [18]. Third, a well-performed taxonomy is also expected to distinguish the fake and real news stories such that some social issues caused by the spreading of fake news can be alleviated. Lee *et al.* [17] proposed the trustworthy of recommended news would also increase users' satisfaction. Therefore, it is necessary to have an automatic text classifier to filter out fake articles. Finally, newspaper websites often make some non-personalized news recommendations [19], which also requires a good news taxonomy.

This paper studies news articles from 33 categories provided by the McClatchy news organization. Different from most existing benchmark news datasets, the topics are not mutually exclusive. For example, news articles belonging to the topics 'localOpin' and 'localGovt' also belong to the topic 'local'. Our contribution is to establish a benchmark for how accurately news stories can be classified by existing deep learning (DL) techniques. Furthermore, most news organizations can provide a text headline for each story, but many have great difficulty in providing the full text of stories. It is therefore of interest to know how accurately stories can be classified only from headlines versus with full texts, which we evaluate. Lastly, we provide a visualization of news articles to illustrate the intrinsic difficulty of classifying news stories.

2. Literature Review

Many DL techniques have been applied to news classification, but most studies only evaluated binary outcomes. Jang [20] evaluated Convolutional Neural Network (CNN) based models using Word2Vec to classify news articles a either relevant or irrelevant. Detecting fake news is getting more attention and Liu *et al.* [21] proposed a hybrid model of Long Short-Term Memory (LSTM) and CNN for detecting it. Jadhav *et al.* [22] proposed a hybrid of Recurrent Neural Network (RNN) and Deep Structured Semantic models to identify important features of fake news. Qasim *et al.* [23] studied nine BERT models such as BERT-base, BERT-large, RoBERTa-base, and RoBERTa-large, etc. on detecting COVID-19 fake news. In this paper, we focus on multi-class news topic categorization.

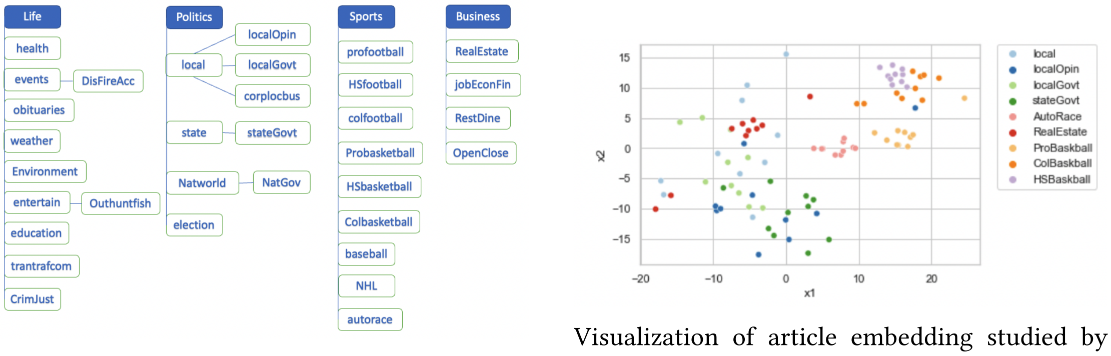
There is some extant research on multi-class news categorization. Zhang [24] studied multi-class news categorization and proposed a customized DCLSTM-MLP model. Different from the existing CLSTM-MLP model, this model uses a discrete vector to represent the probability variance of each word and considers the contribution of each word in a classification. Due to the innovated discrete vector representation of each word, Zhang's model had better prediction accuracy than CNN-MLP and CLSTM [24]. However, their news text sample is small. Huang and Jiang [25] studied the classification of Chinese news articles with machine learning (ML) and DL techniques including LSTM, Bi-LSTM, CNN Naïve Bayes, TFIDF-SVM, and Word2Vec-SVM. They showed that ML and DL methods have similar accuracy if the texts are pre-processed properly. Deng *et al.* [18] proposed a model that performs better than CNN-BiLSTM, attention CNN, attention BiLSTM, and BERT-DCNN-BiGRU to classify long Chinese texts. Pre-processing Chinese texts is different from English ones because Chinese characters can either be morphemes or words and there are no spaces to mark word boundaries. Deng's work did not evaluate short texts (e.g., headline only) or languages other than Chinese. We evaluate several the most popular and well-known DL techniques for English news topic categorization.

3. Methodology

Minaee *et al.* [26] proposed a comprehensive review of DL techniques for multiple text classification tasks. In this paper, we establish a benchmark of how DL classifiers perform on McClatchy news topic classification.

3.1. McClatchy dataset

The McClatchy news dataset consists of 74,672 news items from 25 geographically dispersed US markets. It provides headline and full-text for each story. The average headline length is around 13 words, the the average story length is around 163 words. The training, validation, and testing data set is split into an 8:1:1 ratio. The dataset is imbalanced, with the most common topic 'health' (around 24%) and the least common 'NHL'. The ratio between the number of 'health' and 'NHL' is around 65:1, which is acceptable. McClatchy news has 33 topics (shown in Figure 1); however, different from other benchmark news datasets such as *AG News* [27] and *20 Newsgroups* [28], the boundary between some defined topics is blurred. For example, it is reasonable and accurate to classify 'stateGovt' articles into 'state'.



Defined topics of *McClathy* articles.

Visualization of article embedding studied by BERT.

Figure 1: In the left figure, 33 topics in white boxes are those defined by McClathy; topics in blue boxes are those defined by us to organize given topics.

3.2. Evaluated models

We investigate the performance of a spectrum of deep models applied to the McClathy news classification task. These models include feed-forward networks (FFN), convolutional neural networks (CNN), recurrent neural networks (RNN), attention-based deep models, and their interpolated variants.

Feed-forward networks such as FastText [29] have been a workhorse for industry-level NLP applications owning to their computational efficiency in producing quality sentence representations from character-level features; yet their feed-forward nature thwarts them from explicitly capturing sequential dependencies. The RNNs leverage a recurrent bias to alleviate this issue. We primarily study LSTM [30], an RNN variant that address a typical vanishing/exploding gradient bottleneck of vanilla recurrent architectures. Empirical studies demonstrate that LSTMs and its bi-directional variant have achieved state-of-the-art performance in many natural language applications [31].

CNNs are mainly used in the area of computer vision [32] but have recently been adopted for natural language understanding [33, 34, 35]. While RNNs excel in capturing sequential structure, CNNs demonstrate better capacity in detecting local and position-invariant patterns to detect key phrases in a sentence [18]. When applying a CNN for text classification, we obtain the embedding with a fixed dimension for each word, and then a context can be represented as a matrix (i.e. each row is a wording embedding). In our experiments, we adopt textCNN [36], which represent the context as a matrix and then applies 1D convolution along the time dimension (i.e. temporal convolution) to obtain a hidden representation. We then apply a classification head to this hidden representation for news classification. Recent works have also studied causal convolution, a technique that prevents information leakage from future to past, and achieves competitive results on sequence modeling benchmarks against recurrent counterparts [37]. In experiments we do not find this variant to be beneficial, and defer further investigation of such causal variants to future work.

The attention mechanisms was first proposed as an improvement for RNNs [38] in machine translation tasks; subsequently full attention-based models [39] were introduced. Intuitively, the attention mechanism assigns weights to each word or sentence based on its importance in a context [39]. In this paper, we evaluate the BiLSTM-attention model proposed by Zhou *et al.* [40]—an attention layer is added such that the final context vector is a weighted sum of feature vectors studied by BiLSTM.

RNNs and CNNs have their own issues. RNNs easily forget the information from the beginning, while textCNN lacks interpretability, ignores dependencies among local features, and is difficult to weight the importance of each feature. Therefore, hybrid models that combine LSTM and CNN architectures have been proposed. We will evaluate Zhu *et al.*'s C-LSTM [41]. The key phrases studied by CNN are fed in Bi-LSTM to obtain the sentence representation. We further evaluate a hybrid model Bi-LSTM-CNN-attention [42] that combines Bi-LSTM, CNN and the attention mechanism.

Lastly, we evaluate Google's BERT [43]. Different from RNNs and CNNs, which require sequential text inputs and embedding for each word, BERT is a transformer model pre-trained on tasks masked language modeling and next sentence prediction. It is trained by randomly masking 15% words and predicting the masked token based on the other independent tokens, and relying on the self-attention mechanism, BERT understands each word by connecting it to every other words [43].

4. Experimental Results

Tuning parameters such as the batch size, dropout rate, embedding dimension, etc. are chosen based on the model's performance on the validation dataset. For BERT models we apply its pre-trained tokenizer and for other models we apply the 'Spacy' tokenizer and 'glove.6B.100d' word representations. The 33 topics are imbalanced and the news topic categorization belongs to the multi-class classification task. Therefore, besides the accuracy metric we also apply macro-averaging precision (Macro-P), micro-averaging precision (Micro-P), macro-averaging recall (Macro-R) and micro-averaging recall (Micro-R) for the model evaluation [44]. Macro-averaging of a measure (i.e., precision, recall) is the average of the same measure calculated for each class, and micro-averaging is computed by first summing up counts of true positives, true negatives, false positives and false negatives for all classes [44]. For each model, we evaluate these metrics with headlines only and separately using full texts.

Tables 1 and 2 show the numerical performances of selected DL techniques mentioned in Section 3. To make better comparisons, models' performances with respect to each metric are visualized in Figure 2. In general, models trained with full-text outperform those trained with headlines only. Furthermore, BERT has much better performances than all the other models models. However, compared to BERT performance on the *DBpedia* [45] data set with 14 categories reported in [26], BERT has worse accuracy on McClatchy.

As mentioned in Section 3.1, the categories defined by McClatchy are not as mutually exclusive as the other benchmark datasets. Therefore, we first take a look at some predictions made by our best model BERT. Among the 191 news articles belonging to the defined category 'localGovt', 112 are predicted correctly, and then 12 are predicted to be 'local'. From Figure 1, we believe that it is reasonable and accurate to categorize 'localGovt' articles into 'local'. These observations imply that if we apply broader topics, then the models' performances should be improved.

Table 1
Testing performances of models using headline information.

Metric	BiLSTM	TextCNN	BiLSTM-attent	C-LSTM	BiLSTM-CNN-attent	FastText	BERT
Accuracy	0.71	0.70	0.69	0.69	0.71	0.66	0.76
Macro-P	0.64	0.65	0.64	0.62	0.66	0.60	0.71
Micro-P	0.71	0.69	0.70	0.67	0.71	0.65	0.76
Macro-R	0.64	0.57	0.58	0.65	0.62	0.52	0.71
Micro-R	0.71	0.70	0.69	0.72	0.71	0.66	0.76
Accuracy (top-3)	0.87	0.87	0.85	0.87	0.88	0.84	0.92
Macro-P (top-3)	0.84	0.86	0.81	0.82	0.84	0.83	0.89
Micro-P (top-3)	0.87	0.87	0.85	0.87	0.88	0.84	0.92
Macro-R (top-3)	0.81	0.78	0.78	0.80	0.83	0.74	0.88
Micro-R (top-3)	0.87	0.87	0.85	0.87	0.88	0.84	0.92

Table 2
Testing performances of models using full-text information.

Metric	BiLSTM	TextCNN	BiLSTM-attent	C-LSTM	BiLSTM-CNN-attent	FastText	BERT
Accuracy	0.75	0.73	0.72	0.75	0.76	0.71	0.81
Macro-P	0.70	0.69	0.66	0.68	0.71	0.65	0.78
Micro-P	0.75	0.72	0.72	0.74	0.76	0.70	0.81
Macro-R	0.68	0.62	0.65	0.67	0.69	0.54	0.77
Micro-R	0.75	0.73	0.72	0.75	0.76	0.71	0.81
Accuracy (top-3)	0.89	0.91	0.88	0.91	0.92	0.90	0.95
Macro-P (top-3)	0.86	0.90	0.85	0.90	0.89	0.89	0.93
Micro-P (top-3)	0.89	0.91	0.88	0.91	0.92	0.90	0.95
Macro-R (top-3)	0.84	0.84	0.83	0.86	0.88	0.81	0.92
Micro-R (top-3)	0.89	0.91	0.88	0.91	0.92	0.90	0.95

Second, instead of classifying news into the topic with the highest probability, we decide to evaluate models' performances by making top-3 predictions. When the defined class is in the top-3 probabilities, we count it as a 'correct' prediction, and then we compute top-3 accuracy, macro-precision, micro-precision, macro-recall and micro-recall for each model. Tables 1 and 2 show that performances are very good using top-3 prediction—the top three predicted categories include the labeled value 95% of the time with full-text training and 92% with headline only. Figure 2 also shows that if top-3 prediction is applied, BERT using only headlines performs better than all the other models using full-texts. These computational results imply that models understand texts well but the boundaries among defined categories must be more clear.

We lastly project the vector representations for some headlines from some selected topics to 2D using t-SNE and visualize the spread of these news stories (shown in Figure 1). We observe that the defined topics 'local', 'localOpin', 'localGovt' and 'stateGovt' are intrinsically hard to be separated from each other, and the 'local' category has the widest spread, indicating it is the most ambiguous category. The fact that some identified inaccurate predictions are actually classified into reasonable topics, models' performances are dramatically improved if

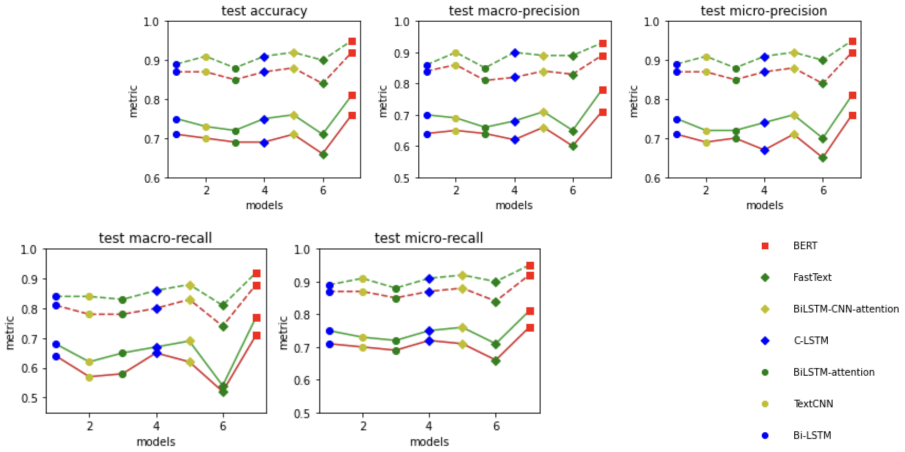


Figure 2: Red lines are testing performances using headlines; green lines are those using full-texts. Testing performances applying top-3 predictions are plotted as dotted lines, while those applying top-1 predictions are solid lines.

top-3 prediction is adopted, and the visualization of some stories illustrate the untrustworthy of labeled classes and the intrinsic difficulty of categorizing McClatchy news stories.

5. Discussion

Having a reliable news taxonomy can improve the context understanding, manage news resources for news organizations, prevent the spread of fake news, and study the components of non-personalized news RS. We use data from McClatchy to establish a benchmark of how accurately stories from ambiguously defined topics can be classified by some popular DL news classifiers. Among the models we evaluated, BERT performs the best. We further evaluate the importance of having full-texts by comparing each model's performances with headlines only and full-texts. From our experimental results, full-texts provide more contextual information for better classification performances. Our computational experiments also show that (1) some predictions identified as being inaccurate are in fact reasonable; (2) models have much better performances if we apply top-3 predictions, indicating that we could not completely rely on labeled classes; and (3) our visualization of news stories from some selected topics further confirm the intrinsic difficulty in classifying news items.

There are some limitations in our study. First, in this paper, we just establish a benchmark on how some selected DL news classifiers perform on McClatchy news stories whose defined topics are not mutually exclusive. The techniques that we evaluated are not completed and all of them have better variants. Therefore, we expect to provide a more comprehensive model evaluations on McClatchy. Second, we have identified that the defined categories in McClatchy are intrinsically hard to be classified. Some predicted classes are in fact accurate. This observation motivates us to think of the necessity of having the current detailed topics and want to identify a better way to categorize news articles by applying news RS algorithms.

Acknowledgments

We would like to express our gratitude to Oliver Zhuoran Liu from Industrial Engineering and Management Science department at Northwestern University for providing his valuable and constructive advice during the course of this research. We thank the Northwestern Local News Initiative for providing the data. We would also like to thank all the anonymous reviewers for reviewing this work and providing insightful comments.

References

[1] P. M. Abernathy, The expanding news desert, Center for Innovation and Sustainability in Local Media, School of Media and Journalism, 2018.

[2] P. Gao, C. Lee, D. Murphy, Municipal borrowing costs and state policies for distressed municipalities, Journal of Financial Economics 132 (2019) 404–426.

[3] S. J. Kim, Y. Zhou, E. C. Malthouse, Y. Kamyab Hessary, In search for an audience-supported business model for local newspapers: Findings from clickstream and subscriber data, Digital Journalism (2021) 1–21.

[4] Y. Zhou, B. J. Calder, E. C. Malthouse, Y. K. Hessary, Not all clicks are equal: detecting engagement with digital content, Journal of Media Business Studies 0 (2021) 1–18.

[5] T. Zhu, P. Harrington, J. Li, L. Tang, Bundle recommendation in ecommerce, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Association for Computing Machinery, New York, NY, USA, 2014, pp. 657–666.

[6] S. Raza, A news recommender system considering temporal dynamics and diversity, arXiv preprint arXiv:2103.12537 (2021).

[7] R. Fletcher, R. K. Nielsen, Automated serendipity: The effect of using search engines on news repertoire balance and diversity, Digital Journalism 6 (2018) 976–989.

[8] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proceedings of the National Academy of Sciences 118 (2021).

[9] O. Stitini, S. Kaloun, O. Bencharef, Towards the detection of fake news on social networks contributing to the improvement of trust and transparency in recommendation systems: Trends and challenges, Information 13 (2022) 128.

[10] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2020) 102025.

[11] R. Picard, Media and communications policy making, Springer, 2020.

[12] M. Ziegler, F. Iida, R. Pfeifer, Cheap" underwater locomotion: Morphological properties and behavioral diversity, in: IROS05 Workshop on Morphology, Control and Passive Dynamics, 2005.

[13] L. Candillier, M. Chevalier, D. Dudognon, J. Mothe, Diversity in recommender systems, in: Proceedings: The Fourth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services. CENTRIC, 2011, pp. 23–29.

[14] P. Castells, N. J. Hurley, S. Vargas, Novelty and diversity in recommender systems, in: Recommender systems handbook, Springer, 2015, pp. 881–918.

[15] M. A. Beam, G. M. Kosicki, Personalized news portals: Filtering systems and increased news exposure, Journalism & Mass Communication Quarterly 91 (2014) 59–77.

[16] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, C. Wilson, Measuring personalization of web search, in: Proceedings of the 22nd International Conference on World Wide Web, arXiv, 2013, pp. 527–538.

[17] S. Y. Lee, S. W. Lee, Normative or effective? the role of news diversity and trust in news recommendation services, International Journal of Human–Computer Interaction (2022) 1–14.

[18] J. Deng, L. Cheng, Z. Wang, Attention-based bilstm fused cnn with gating mechanism model for chinese long text classification, Computer Speech & Language 68 (2021) 101182.

[19] A. Chakraborty, S. Ghosh, N. Ganguly, K. P. Gummadi, Optimizing the recency-relevance-diversity trade-offs in non-personalized news recommendations, Information Retrieval Journal 22 (2019) 447–475.

[20] B. Jang, I. Kim, J. W. Kim, Word2vec convolutional neural networks for classification of news articles and tweets, PloS one 14 (2019) e0220976.

[21] Y. Liu, Y.-F. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2018.

[22] S. S. Jadhav, S. D. Thepade, Fake news identification and classification using dssm and improved recurrent neural network classifier, Applied Artificial Intelligence 33 (2019) 1058–1068.

[23] R. Qasim, W. H. Bangyal, M. A. Alqarni, A. Ali Almazroi, A fine-tuned bert-based transfer learning approach for text classification, Journal of Healthcare Engineering 2022 (2022).

[24] M. Zhang, Applications of deep learning in news text classification, Scientific Programming 2021 (2021).

[25] C.-M. Huang, Y.-J. Jiang, An empirical study on the classification of chinese news articles by machine learning and deep learning techniques, in: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, 2019, pp. 1–6.

[26] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: a comprehensive review, ACM Computing Surveys (CSUR) 54 (2021) 1–40.

[27] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in neural information processing systems 28 (2015).

[28] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[29] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016).

[30] F. A. Gers, E. Schmidhuber, Lstm recurrent networks learn simple context-free and context-sensitive languages, IEEE Transactions on Neural Networks 12 (2001) 1333–1340.

[31] S. Merity, N. S. Keskar, R. Socher, Regularizing and optimizing lstm language models, arXiv preprint arXiv:1708.02182 (2017).

[32] S. Khan, H. Rahmani, S. A. A. Shah, M. Bennamoun, A guide to convolutional neural networks for computer vision, Synthesis Lectures on Computer Vision 8 (2018) 1–207.

[33] Y. Chen, Convolutional neural network for sentence classification, Master's thesis, Univer-

sity of Waterloo, 2015.

[34] J. P. A. Vieira, R. S. Moura, An analysis of convolutional neural networks for sentence classification, in: 2017 XLIII Latin American Computer Conference (CLEI), IEEE, 2017, pp. 1–5.

[35] J. Du, L. Gui, Y. He, R. Xu, X. Wang, Convolution-based neural attention with applications to sentiment classification, IEEE Access 7 (2019) 27983–27992.

[36] Y. Kim, Convolutional neural networks for sentence classification, CoRR abs/1408.5882 (2014). URL: http://arxiv.org/abs/1408.5882. arXiv:1408.5882.

[37] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018).

[38] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[40] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), 2016, pp. 207–212.

[41] C. Zhou, C. Sun, Z. Liu, F. Lau, A c-lstm neural network for text classification, arXiv preprint arXiv:1511.08630 (2015).

[42] Y. Zhu, X. Gao, W. Zhang, S. Liu, Y. Zhang, A bi-directional lstm-cnn model with attention for aspect-level text classification, Future Internet 10 (2018) 116.

[43] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[44] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427–437.

[45] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic web 6 (2015) 167–195.