big data and cognitive computing

MDPI

Article

Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text

Andrey Bogdanchikov [1], Dauren Ayazbayev [1,*] and Iraklis Varlamis [2]

[1] Department of Computer Science, Suleyman Demirel University, Kaskelen 040900, Kazakhstan
[2] Department of Informatics and Telematics, Harokopio University of Athens, 17779 Athenes, Greece
* Correspondence: dauren.ayazbayev@sdu.edu.kz

Abstract: The rapid development of natural language processing and deep learning techniques has boosted the performance of related algorithms in several linguistic and text mining tasks. Consequently, applications such as opinion mining, fake news detection or document classification that assign documents to predefined categories have significantly benefited from pre-trained language models, word or sentence embeddings, linguistic corpora, knowledge graphs and other resources that are in abundance for the more popular languages (e.g., English, Chinese, etc.). Less represented languages, such as the Kazakh language, balkan languages, etc., still lack the necessary linguistic resources and thus the performance of the respective methods is still low. In this work, we develop a model that classifies scientific papers written in the Kazakh language using both text and image information and demonstrate that this fusion of information can be beneficial for cases of languages that have limited resources for machine learning models' training. With this fusion, we improve the classification accuracy by 4.4499% compared to the models that use only text or only image information. The successful use of the proposed method in scientific documents' classification paves the way for more complex classification models and more application in other domains such as news classification, sentiment analysis, etc., in the Kazakh language.

Keywords: convolutional neural network; document classification; word embedding

Citation: Bogdanchikov, A.; Ayazbayev, D.; Varlamis, I. Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. *Big Data Cogn. Comput.* **2022**, *6*, 123. https://doi.org/10.3390/bdcc6040123

Academic Editors: Carson K. Leung and Min Chen

Received: 9 August 2022
Accepted: 19 October 2022
Published: 24 October 2022

1. Introduction

Document classification plays an important role in natural language processing tasks. When documents are clustered in semantically coherent groups or organized in thematic categories they can help reduce time needed to search and retrieve information from large collections [1]. A predefined grouping of documents using thematic or other criteria allows to focus on the groups that are considered more relevant to the user request and thus to avoid comparing the query with all the documents of the collection [2]. Additionally, the classification of documents in thematic categories, their labeling with one or more topics, and the extraction of aspects from each document can facilitate the browsing of the collection and can be the basis for multi-faceted search and retrieval. For example, news sites organize the news articles into topics in order to allow their readers to quickly browse the news of interest to them or to register only for specific topics. Similarly, when scientific papers are classified in various disciplines, categories and topics following classifications systems such as ACM's Computing Classification System or NLM's Medical Subject Headings it is easier for researchers to locate newly published research in their field of expertise.

Since a document can belong to one or more disciplines at a time, the classification algorithms must also consider this multi-label classification task [3]. Among the most popular algorithms for document classification are k-nearest neighbors, support vector machines, Naive Bayes, Artificial Neural Networks, etc. [4]. The performance of the

algorithms is not always the same and their prediction accuracy may vary with the task or the target collection. The accuracy of the document classifier also depends on the representation model it employs which in some cases depends on the language of the text [5]. The long list of representation models comprises discrete text representations and the Vector Space Model (e.g., one-hot encoding with 0 or 1 used as weights, Bag-of-words representation with occurrence count or frequency metrics such as TF-IDF used as weights and n-gram models), probabilistic models that provide a semantic representation (e.g., Latent Semantic Indexing and Probabilistic Latent Semantic Indexing), graph-based model [6], neural network based models (e.g., continuous bag of words, Word2Vec, GloVe, Doc2Vec, bidirectional encoder representations from transformers, etc.) [7]. Most of the aforementioned representation models perform well in document classification tasks only when a large training corpus is available, or when the models have been pre-trained on huge generic text corpora in order to capture the syntactic and semantic patterns of a language [8]. The effect of the language can be huge, since the syntactic rules significantly vary among languages and word polysemy and synonymy can further increase the complexity of the sense disambiguation task. For instance, in many languages one word can have several meanings, and by adding suffixes the word can dramatically change its meaning.

In the current work, we focus on scientific papers written in Kazakh language and attempt to classify them into predefined thematic topics. It is important to clarify here that in more represented languages such as English, French, or Chinese the textual information is sufficient in order to achieve a good classification performance. This is mainly due to the abundance of linguistic resources for these languages, which includes vocabularies, ontologies, and pre-trained deep learning models which allow to understand the meaning of any text and classify it accordingly. However, this is not the case with the Kazakh language. In order to tackle this problem and at the same time to take advantage of the state-of-the-art linguistic resources that are available for the language, we use both the textual and image content of documents, combining the two contents using a neural network architecture.

The motivation for our work is the need to properly organize and search scientific documents written in languages other than English (or other resource-rich languages) and provide the infrastructure for the development of a scientific document search engine for the Kazakh language. The effective and efficient classification of documents using their textual and visual content is of utmost importance, and thus we examine how existing techniques and models can be combined for this purpose.

The documents in our collection are research papers published in the bulletin of Kazakh National University and they belong to one of three different series. In order to classify the documents using their text only, we aggregate the Word2Vec representation of words in the text and apply the Naive Bayes algorithm. For the classification of documents using their image content only, we employ a convolutional neural network architecture. Finally, we combine the two pieces of information by feeding them to a composite neural network that learns to classify using both the textual and image content. The results achieved with the combined method improve the previous results in the dataset, which is quite promising for the fusion of information in the classification task.

The contributions of this work can be summarized as follows:

- A classification model for scientific documents that fuses information from the text and images and outperforms in performance the models that use either text or images only.
- An application of the model in scientific documents written in Kazakh.

According to the literature survey we performed, this is the first work that combines text with image content in the classification of scientific documents written in a language with limited linguistic resources, such as the Kazakh language. The successful results can be a guide for other researchers to combine multimodal content in their tasks in order to improve performance.

In Section 2 that follows, we perform a review of the related literature on scientific document classification and also on documents of any kind that are written in Kazakh. Section 3 details on the proposed architecture and explains how image and text context are fused to the neural network. Section 4 illustrates the experimental setup and the results obtained so far. Finally, Section 5 concludes the article with our main findings and conclusions and the next steps in this work.

2. Literature Review

2.1. NLP Solutions for the Kazakh Language

Many works have been recently devoted to the developing of word or document embeddings for various languages. One of them is the EMBEDDIA project [9,10] which attempts to develop high quality ELMo embeddings for less-resourced languages such as Estonian, Latvian, Lithuanian, Slovenian, Croatian, Finnish and Swedish. Authors also developed language tools that can be used to filter user comments when they detect hate speech, to extract keywords from articles and use them for tagging, and to generate text on a specific topic. These tools can be also combined or adapted to solve many more similar problems in these languages.

Another interesting work in the Tatar language has been presented by [11]. The Tatar language is a member of the Turkish language family, and as such is highly related to the Kazakh language. The authors have developed three benchmark datasets for evaluating word embedding techniques in Tatar. More specifically, the datasets cover the word analogy, word relatedness, and word similarity tasks. What is even more interesting is that the authors considered the morphological richness of Tatar language as well as geographical and cultural aspects. To evaluate the performance in the benchmark datasets, they recommended the use of metrics such as accuracy, Spearman's rho, and rank correlation coefficient. These benchmark datasets of word analogy, relatedness, and similarity can be used to fine-tune word embedding techniques, which in turn can be employed to make dictionaries of synonyms or related terms to measure text similarity, to expand search engine queries, and so on.

As far as it concerns the Kazakh language, the list of works is limited. Among them we have to mention the work of [12] who developed the KazNLP library. The library was written in Python programming language and can be used to define initial normalization of texts, tokenization of word-sentences, morphological analysis, etc. Apart from the KazNLP library, the authors compiled a dataset from news websites which comprised of Kazakh (63%) and Russian (34.4%) texts. They also employed the Kazakh Language Corpus (KLC) [13] that contains more than 135 million words in more than 400 thousand documents from different genres written in literary, official, scientific, publicistic and informal language, and the Kazakh Dependency Treebank [14], which relabeled part of the KLC with lexical, morphological, and syntactic annotation.

In a slightly different context, Yelibayeva et al. [15] proposed a model that can be used in semantic searches and Q&A systems in the Kazakh language. The proposed ontological model was defined by a set of syntactic descriptions which are used to extract nominative word combinations that are consequently useful for machine translation and multilingual search tasks.

A search for pre-trained language models in Kazakh reveals that FastText embeddings and Word2Vec embeddings are already available online. A recent work from [16] has demonstrated that pre-trained word embeddings can be beneficial for the Named-Entity Recognition task in Kazakh documents, and this is a positive finding for using Kazakh word embeddings in more NLP tasks. Consequently, Word2Vec embeddings are our first choice for handling the representation of the textual content of our documents.

2.2. Classification of Scientific Documents

In order to find the best choice for combining text with images, we performed a review of recent works that focus on the classification of documents. The concept of combining textual and audiovisual information in classification tasks has been employed in the past by researchers; however, not in the context of scientific documents. In the following, we highlight the main works, explain the task they examine, the dataset, method and architecture they employ and the performance improvement they achieved.

Early works on scientific document classification employed Naive Bayes, k-nearest neighbors [17] and feed forward neural networks [18]. Researchers employed the text contents of the documents as features to predict the initial class labels of the documents and citation links are employed to refine the labels. In [19], it was attempted to classify scientific documents using Support Vector Machines and BoW representation (using TF/IDF-based weights) in order to better handle the high dimensionality of the vectors.

In [20], the authors also relied on TF-IDF for their research paper clustering system but they also extracted representative keywords from the abstract and applied topic modeling techniques (i.e., Latent Dirichlet allocation) to extract topics. The final clustering was based on the output of a k-Means clustering algorithm that takes into account the similarity of document representation vectors. In their experiment, they employed papers in English published in the Future Generation of Computer Systems (FGCS) journal from which they removed all stopwords and extracted only nouns.

A promising approach that builds on language embeddings has been presented by [21], who trained a large-scale academic paper embedding called Paper2Vec (or P2V). The authors used a dataset that contains 46.64 million papers and 528.68 million citation links and the resulting embeddings boost the performance in paper classification, paper similarity, and paper influence prediction task. However, all the papers in the dataset are written in English and the evaluation was on English paper collections. In the same direction, reference [22] proposed the use of fastText word embeddings, which they trained on a large data set of more than 5 million patents. They combined the embeddings with a bidirectional gated recurrent unit (GRU) to automatically classify patents and reported a maximum micro-average precision of 72% that improved the performance of vector-based representations by 17% and the performance over the Wikipedia-based generic embeddings by 9%, giving evidence that domain- and language-specific training can be beneficial for embedding models. Once again, the patents dataset used for training and evaluation were in English.

In a very recent research work [23] the authors have used graph embeddings in scientific paper classification, taking advantage of the knowledge graphs that have been made behind the collaboration between co-authors boosting the classification accuracy to 98%. However, they assume that the scientific papers are already listed in the major literature databases and author information is available and without ambiguity.

2.3. Fusion of Text and Images

The attempts that rely on the image content of the documents either examine the whole structure of the document or focus on the images only.

In [24], the authors focused on the classification of document attributes in a large collection of documents (scientific articles of single or double column, programming code, novels, legal texts, etc.) using only their scanned images and a deep neural network architecture. They focused on attributes such as font emphasis, font type, font size and scanning resolution and employed the ResNet50 model which was trained on the ImageNet dataset in order to classify attributes using single- or multi-task learning. In the single task learning task, the attributes were classified separately, whereas in the multi-task setup the neural network provided a simultaneous prediction for all attributes. Multi-tasking learning showed higher accuracy than single task learning, with the accuracy being above 0.94 for all attribute types.

Among the approaches that fuse image and text information in order to improve classification accuracy, the work of [25] evaluated several algorithms (kNN, Naïve Bayes, reverse DBSCAN) in the classification of emails from the Enron corpus. They passed images through the Tesseract OCR library to extract text and then classified the text content as a whole. Although this may work on business emails and their image attachments, this is not applicable to scientific articles and the images they contain.

In [26], the authors have also attempted to classify scanned documents combining the textual features extracted using Tesseract with the visual features extracted from a CNN (MobileNet v2 more specifically). They evaluated their method on a dataset of legal documents (the Tobacco3482 dataset) and a second dataset that contains emails, letters, invoices, scientific reports, etc. The resulting classification accuracy was at 84% and improved the performance of the baseline methods that used only text (a multi-layer perceptron neural network) or only images (CNN 1D) by more than 10%.

Table 1 summarizes the main approaches that we found in this study and lists their main characteristics. From the lists we can see that our approach is the first that combines text and image content in scientific documents written in the Kazakh language and employs pre-trained language models (word embeddings) and an image classification neural network.

Table 1. The main approaches in document classification and their main features.

Task(s)	Dataset	Model	Language of Text	Reference
Keyword extraction, comment moderation, text generation	CoSimLex, cross-lingual analogy	ELMo embeddings, CroSloEngual BERT, LitLat BERT, FinEst BERT, SloRoberta and Est-Roberta	Estonian, Latvian, Lithuanian, Slovenian, Croatian, Finnish, Swedish	[9,10]
Word analogy relatedness and similarity	Custom	Skip-gram with negative sampling SG, FastText, GloVe	Tatar, English	[11]
Text normalization, word-sentence tokenization, language detection, morphological analysis	The Kazakh language corpus, Kazakh dependency treebank	Graph-based parser	Kazakh	[12]
Nominative word combination extraction	-	Syntactic descriptions	Kazakh	[15]
Named entity recognition	Tourism gazetteers	WSGGA model	Kazakh	[16]
Patent classification	Patents provided by the National Institute of Informatics from 1993 to 2002	Shared Nearest Neighbor	Japanese, English	[17]

Scientific document classification	Cora scientific paper corpus	Feed forward Neural Networks	English	[18]
Classification	Papers from International Conference on Computing and Applied Informatics (ICCAI) and Springerlink collection	Support Vector Machines and BoW representation (using TF/IDF-based weights)	English	[19]
Classification	Papers from Future Generation of Computer Systems (FGCS) journal	K-means clustering based on TF-IDF	English	[20]
Paper classification, paper similarity, and paper influence prediction	Academic papers	Paper2Vector	English	[21]
Patents classification	Patents	fastText	English	[22]
Semantic similarity between papers, citation relationship between papers and the journals	The Microsoft Academic Graph, the Proceedings of the National Academy of Sciences, the American Physical Society	Decision tree, multilayer perceptron	English	[23]
Classification of document attributes	L3iTextCopies	Multi-Task learning, single task learning	Images	[24]
Classification	Enron corpus	k-Nearest Neighbors, Naïve Bayes and Reverse density-based spatial clustering of applications with noise	English	[25]
Classification	Tobacco 3482, RVL-CDIP	Document embedding, sequence of word embeddings, convolutional neural network, MobileNetV2	English	[26]

Classification	GXD2000, DSP	Random Forest, CNNBiLSTM, HRNN, CNN	English	[27,28]
Clustering, Classification	Custom	DBScan, Random Forest, SVM, Logistic Regression	Hinglish, Marglish, Devanagari	[29,30]

3. Methods and Materials

The proposed method capitalizes on the fusion of textual and image content of scientific documents in order to improve the classification performance. The papers used in the experiments are taken from the bulletin of Kazakh National University and are thus written in the Kazakh language. In order to compare the proposed method with existing state of the art, we classify the papers first using only the image content, then using only the text content, and finally using both image and text in a common model. Our main hypothesis is that the visual content, when combined with the textual content, can improve the classification accuracy of scientific documents written in Kazakh. Consequently, the research questions that emerge are:

RQ1: How does the accuracy of classification methods that use only the visual content of scientific documents compare with that of methods that use only the text content?

RQ2: What effect does the use of the visual content of scientific documents have on classification accuracy?

The application of the classifier based on the images of each paper requires pre-processing to bring all images to the same dimensions since the images in the documents may have various sizes. All images were resized to 128 × 128 px size. The next step in pre-processing is to handle documents with a different number of images. Since our image classifier is a simple CNN, we decided to classify each image separately. However, we also have the option to classify all the images of a document in a batch. In this case, we assume a maximum number of images per document, and we pad with black images (all pixels are 0) all the documents that contain less images than this maximum number. After pre-processing, the image classification model (CNN) takes action. The CNN model we trained and employed had the following parameters: batch_size = 64, activation function = softmax trained for 20 epochs. The output of the CNN layer was fed to a dense layer with three neurons (corresponding to the 3 classes). Figure 1 shows an architecture of the proposed CNN network.

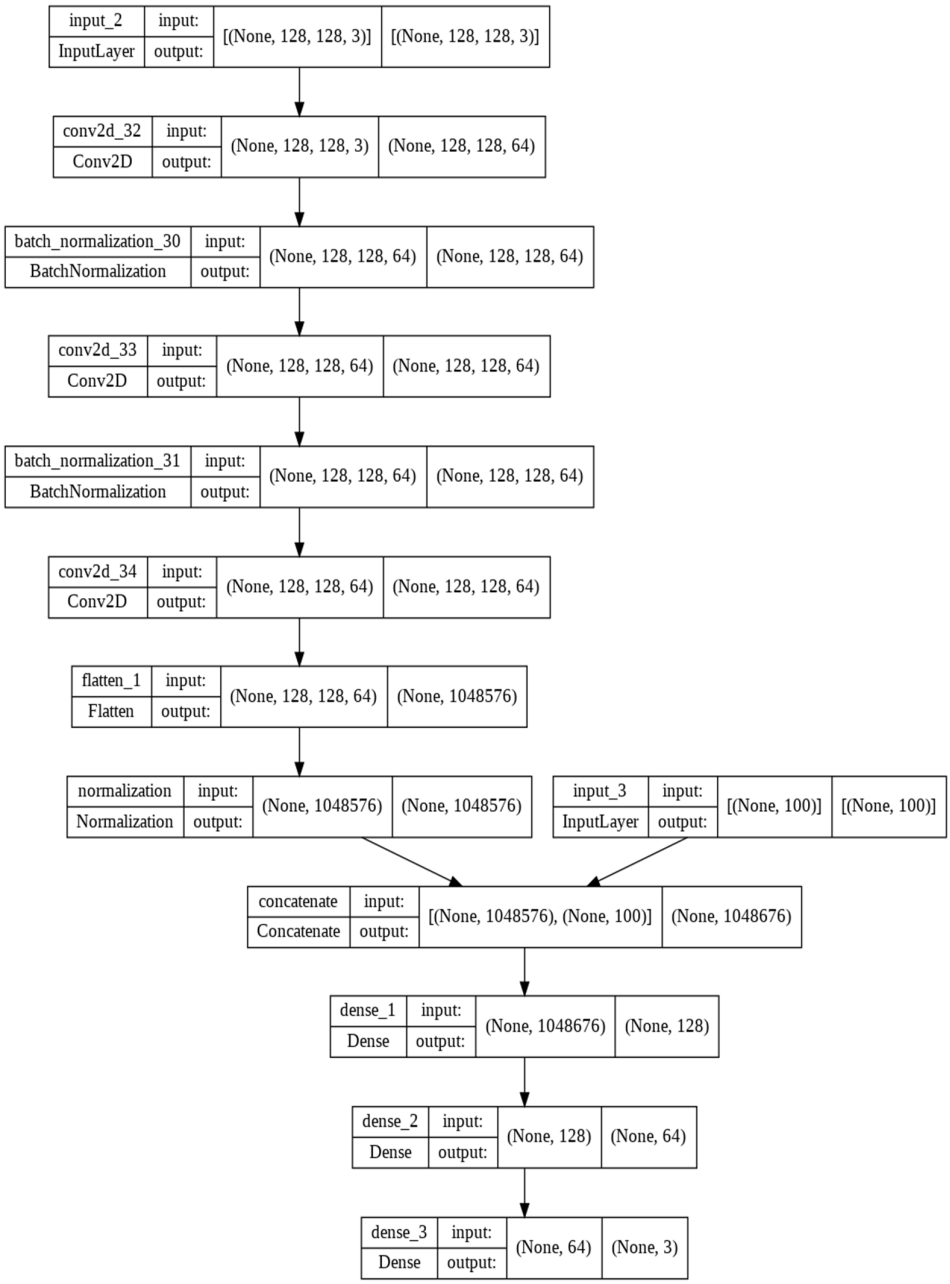

Figure 1. The architecture of the proposed deep neural network.

Using a more formal notation, we may assume that each document D_i is a collection of words w_{ij} and images g_{ik}. Each word w (irrelevant of the document it appears in) is assigned a vector representation v R^{100} by examining the context words of w in a large text corpus. We consequently compute the probability that w-m is in the context of w:

$$p(w_j) = \frac{expexp\left(v_{w_j} \cdot v_{w_{-m}}^T\right)}{\sum_{n=-m}^{m} expexp\left(v_{w_j} \cdot v_{w_n}^T\right)}, \quad (1)$$

and the objective of the skip-gram model is to predict the context of central words, thus finding the vectors v that minimize the loss function which corresponds to the cross-entropy averaged over the corpus. Consequently, the vector representation of a document D_i v_{D_i} is the average of all its word vectors v_{ij} and thus $v_{text,Di}$ R^{100}. The respective image

input for each g_{ik} image is a tensor in $R^{128 \times 128 \times 3}$, since we employ 3 color channels and resize all images to 128 × 128 bits. After passing the tensor to the CNN and flattening, the resulting vector from the visual content of D_i is $v_{image,Di}$ in $R^{1048576}$. The two vectors (i.e. $v_{text,Di}$, $v_{image,Di}$) are concatenated and fed to the dense layers of the neural network.

In CNN architecture, neural network had two inputs: one for images, one for text. Then, they were concatenated. To represent text of paper in numeric form, Word2Vec Continuous Skipgram was applied. Word2Vec Continuous Skipgram embeddings for the Kazakh language were made available by the University of Oslo. The embeddings have been trained on the Kazakh CoNLL17 corpus. Words that do not have embeddings are ignored. The embeddings are of size 100. We took the embeddings for each word in each document and aggregated the information for the whole document using the average of the word embeddings.

To classify papers only by text, Naïve Bayes algorithm was used. Naïve Bayes is a probabilistic classification algorithm based on the Bayes theorem. The algorithm demonstrates good results when the words among the different document classes are not repeated. The proposed system combines the two worlds (i.e., text and images) in a neural network architecture that takes two inputs: the text representation and the image representation. Each branch of the network handles its own type of content and extracts an internal representation. The two representations are concatenated and then fed to three consecutive dense layers as shown in Figure 1 that demonstrates the overall neural network architecture.

4. Experimental Evaluation

4.1. Dataset

The dataset employed for the evaluation of the models consisted of 140 documents from the different series: 40 papers of mathematics, mechanics and computer science, 50 papers of biology, and 50 papers of geography. Each paper in the dataset contained at least one image and at most 22 images. The dataset was split into training (110 documents) and test (30 documents) using stratified random sampling.

4.2. Results

In order to evaluate the performance of the three classification algorithms, we measured the prediction accuracy, which for the three-class task corresponds to the number of correct predictions divided by the total number of classified documents. The first step of the evaluation process is to fine-tune the models' hyperparameters using the training dataset. For this purpose, we employ accuracy for evaluating the performance of the model during training. The first experiment examines how the proposed CNN model accuracy evolves with the number of training epochs. As demonstrated in Figure 2, the model achieves a very high accuracy in the 18th epoch, and after this it is more or less stable until the 20th epoch. The maximum value is in the 18th epoch, when the prediction accuracy in the training set reached 99.09%.

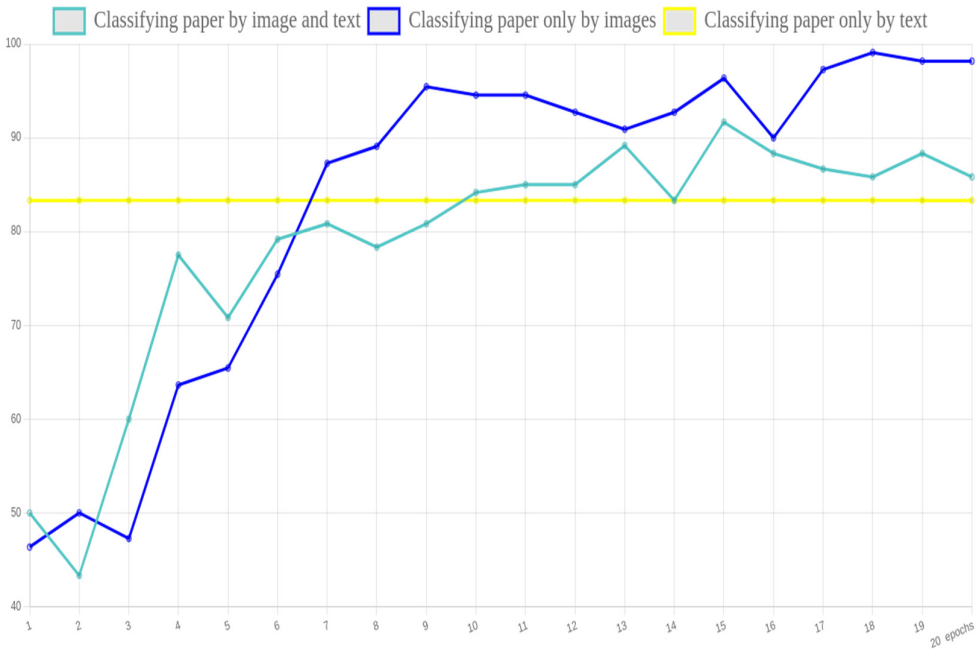

Figure 2. Trend of accuracy in different epochs.

The second experiment examines the effect of batch size to the prediction accuracy, keeping the number of training epoch stable (i.e., 10 epochs) and evaluating on the training data. As shown in Figure 3, the accuracy has an increase when the number of batch size increases, reaching the maximum accuracy (99.106% of the maximum accuracy) for batch size equal to 8.

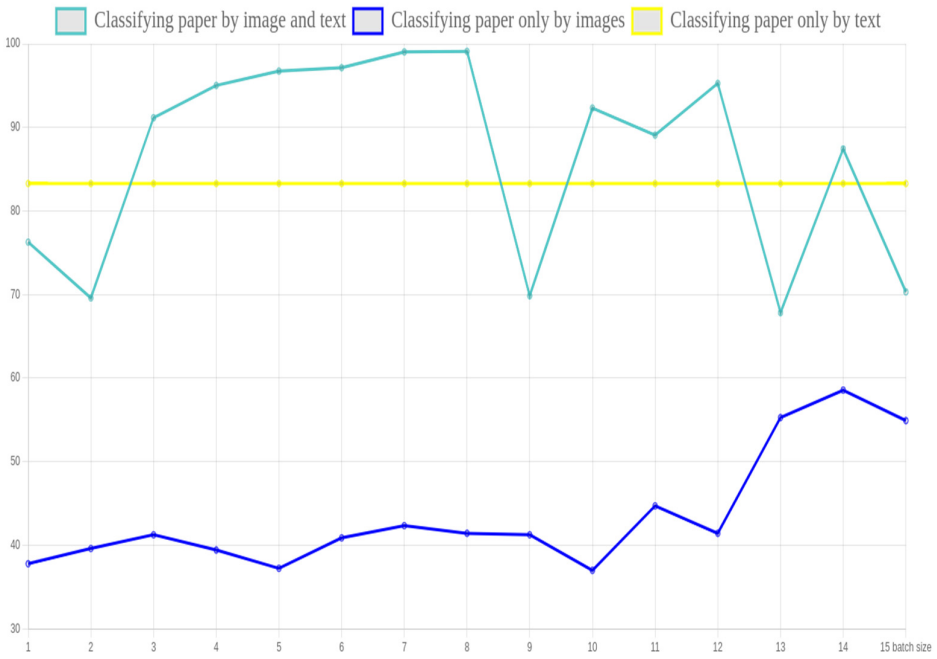

Figure 3. Trend of accuracy in different batch sizes.

With these two parameters fixed, we train our data fusion model and evaluate its performance in the test data. In order to test the statistical significance of our results, we

reshuffled and split the dataset into training and test 10 times and we report the average accuracy. The average prediction accuracy of the three classifiers in the ten runs is depicted in Table 2. The statistical significance analysis of the results shows that the deep neural network model that combines image and text information is significantly better than the other two models at 87.68%, whereas the other two models are comparable with each other.

Table 2. Accuracy of classifier.

Classification	Accuracy (%)
Naïve Bayes	83.33
CNN model with pure images	83.23
Deep NN model with images and text	**87.68**

4.3. Discussion of Results

The results in Table 2 show that using the textual content of documents or their images only can give good prediction for the scientific document topic and their results are comparable. This answers our first research question (i.e., RQ1) concerning the comparison between text and visual content contribution to the classification task. This is quite promising for the task, since in the lack of linguistic resources for a language we can simply rely on the visual content (i.e., images) of the documents in order to detect their thematic category.

With respect to the second research question (i.e., RQ2), it is clear from the results in Table 2 that the prediction performance increases significantly when the two sources of information are fused together. This is even more important for the Kazakh language or other East or North European languages since it allows to bridge the gap created by the lack of rich linguistic resources with the visual content of the documents.

By analyzing the results of the 10 runs, using T-test, we are able to evaluate the statistical significance of the improvement we achieved using both images and text. The statistical significance analysis of the results shows that the deep neural network model that combines image and text information is significantly better (at the 95% confidence interval) than the other two models at 87.68%, whereas the other two models are comparable with each other. This means that the achieved results are constantly better and not only on a random split of the dataset into training and test. Of course, more experiments on larger datasets will allow to further test our hypothesis, which is so far validated. A further analysis of the confusion matrices in our experiments revealed that when the documents were classified using text only by the Naïve Bayes classifier, all documents of the mechanics class were properly detected. There were misclassifications for the documents of the biology and geography classes. This probably happened because the documents in the two classes share many words in common or use related terminology. The number of classification errors decreased when the image content was employed, which is another indication that in classes that use overlapping or related terminology the visual content can be beneficial for the classification task.

5. Conclusions

In the current work, we evaluated the performance of three document classification models in classifying scientific papers written in the Kazakh language. The two models took advantage of the text and image content, respectively, whereas the third model that we propose took advantage of the fusion of textual and image content that existed in all documents. The results were very promising for the fusion model. This experiment demonstrated the ability of information fusion to improve the document classification task for scientific documents and opened new possibilities to use more information such as structural information captured by the visual analysis of documents, special parts of the textual content such as the citations, author names and affiliations, the abstract or the

author-defined keywords, etc. Following our paradigm, in order to take advantage of additional information, more complex deep neural network architectures can be employed that codify the textual and visual information items separately and then fuse them in the layers close to the network's output layer. The work in this field is part of our future work on developing a search engine for scientific publications in the Kazakh language, and document classification is a critical step since it will narrow down the search space at query processing time.

Author Contributions: Conceptualization, A.B., I.V. and D.A.; methodology, A.B., I.V. and D.A.; software, I.V. and D.A.; validation, A.B., I.V. and D.A.; formal analysis, A.B., I.V. and D.A.; investigation, A.B., I.V. and D.A.; resources, A.B., I.V. and D.A.; data curation, A.B., I.V. and D.A.; writing—original draft preparation, D.A.; writing—review and editing, I.V.; visualization, I.V. and D.A.; supervision, A.B. and I.V.; project administration, A.B.; funding acquisition, this research received no external funding. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Embeddia project had site at: https://www.clarin.si/repository/xmlui/handle/11356/1277 (accessed on 1 August 2022). Site of FastText embeddings was available at: https://fasttext.cc/docs/en/crawl-vectors.html (accessed on 1 August 2022). Word2Vec Continuous Skipgram was available through http://vectors.nlpl.eu/repository/ (accessed on 1 August 2022).

Acknowledgments: Authors would like thank Magzhan Kairanbay for helping doing experiment of current paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Halkidi, M.; Nguyen, B.; Varlamis, I.; Vazirgiannis, M. THESUS: Organizing Web document collections based on link semantics. *VLDB J.* **2003**, *12*, 320–332.
2. Bharathi, G.; Venkatesan, D. Improving information retrieval using document clusters and semantic synonym extraction. *J. Theor. Appl. Inf. Technol.* **2012**, *36*, 167–173.
3. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. (IJDWM)* **2007**, *3*, 1–13.
4. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150.
5. Kastrati, Z.; Imran, A.S.; Yayilgan, S.Y. The impact of deep learning on document classification using semantically rich representations. *Inf. Process. Manag.* **2019**, *56*, 1618–1632.
6. Osman, A.H.; Barukub, O.M. Graph-based text representation and matching: A review of the state of the art and future challenges. *IEEE Access* **2020**, *8*, 87562–87583.
7. Babić, K.; Martinčić-Ipšić, S.; Meštrović, A. Survey of neural text representation models. *Information* **2020**, *11*, 511.
8. Mikolov, T.; Deoras, A.; Povey, D.; Burget, L.; Černocký; J Strategies for training large scale neural network language models. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 196–201.
9. Pollak, S.; Pelicon, A. EMBEDDIA project: Cross-Lingual Embeddings for Less-Represented Languages in European News Media. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, Ghent, Belgium, 1–3 June 2022; pp. 291–292.
10. Ulčar, M.; Robnik-Šikonja, M. High quality ELMo embeddings for seven less-resourced languages. In Proceedings of the 12th Conference on Language Resources and Evaluation, Marseille, France, 13–15 May 2020; pp. 4731–4738. *arXiv* **2019**, arXiv:1911.10049.
11. Khusainova, A.; Khan, A.; Rivera, A.R. Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation. *arXiv* **2019**, arXiv:1904.00365.
12. Yessenbayev, Z.; Kozhirbayev, Z.; Makazhanov, A. KazNLP: A pipeline for automated processing of texts written in Kazakh language. In Proceedings of the International Conference on Speech and Computer, St. Petersburg, Russia, 7–8 October 2020; Springer: Cham, Switzerland, 2020; pp. 657–666.
13. Makhambetov, O.; Makazhanov, A.; Yessenbayev, Z.; Matkarimov, B.; Sabyrgaliyev, I.; Sharafudinov, A. Assembling the kazakh language corpus. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, 18–21 October 2013; pp. 1022–1031.

14. Makazhanov, A.; Sultangazina, A.; Makhambetov, O.; Yessenbayev, Z. Syntactic annotation of kazakh: Following the universal dependencies guidelines. In Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages, Kazan, Russia, 17–19 September 2015; pp. 338–350.
15. Yelibayeva, G.; Sharipbay, A.; Bekmanova, G.; Omarbekova, A. Ontology-Based Extraction of Kazakh Language Word Combinations in Natural Language Processing. In Proceedings of the International Conference on Data Science, E-learning and Information Systems 2021, Petra, Jordan, 5–7 April 2021; pp. 58–59.
16. Haisa, G.; Altenbek, G. Deep Learning with Word Embedding Improves Kazakh Named-Entity Recognition. *Information* **2022**, *13*, 180.
17. Cai, Y.L.; Ji, D.; Cai, D. A KNN Research Paper Classification Method Based on Shared Nearest Neighbor. In Proceedings of the NTCIR-8 Workshop Meeting, Tokyo, Japan, 15–18 June 2010; pp. 336–340.
18. Zhang, M.; Gao, X.; Cao, M.D.; Ma, Y. Neural networks for scientific paper classification. In Proceedings of the First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06), Beijing, China, 30 August–1 September 2006; pp. 51–54.
19. Jaya, I.; Aulia, I.; Hardi, S.M.; Tarigan, J.T.; Lydia, M.S. Scientific documents classification using support vector machine algorithm. *J. Phys. Conf. Ser.* **2019**, *1235*, 12082.
20. Kim, S.W.; Gil, J.M. Research paper classification systems based on TF-IDF and LDA schemes. *Hum.-Cent. Comput. Inf. Sci.* **2019**, *9*, 30.
21. Zhang, Y.; Zhao, F.; Lu, J. P2V: Large-scale academic paper embedding. *Scientometrics* **2019**, *121*, 399–432.
22. Risch, J.; Krestel, R. Domain-specific word embeddings for patent classification. *Data Technol. Appl.* **2019**, *53*, 108–122.
23. Lv, Y.; Xie, Z.; Zuo, X.; Song, Y. A multi-view method of scientific paper classification via heterogeneous graph embeddings. *Scientometrics* **2022**, *127*, 30.
24. Mondal, T.; Das, A.; Ming, Z. Exploring multi-tasking learning in document attribute classification. *Pattern Recognit. Lett.* **2022**, *157*, 49–59.
25. Harisinghaney, A.; Dixit, A.; Gupta, S.; Arora, A. Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. In Proceedings of the 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), Faridabad, India, 6–8 February 2014; pp. 153–155.
26. Audebert, N.; Herold, C.; Slimani, K.; Vidal, C. Multimodal deep networks for text and image-based document classification. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Würzburg, Germany, 16–20 September 2019; pp. 427–443. *arXiv* **2019**, arXiv:1907.06370v1.
27. Li, P.; Jiang, X.; Zhang, G.; Trabucco, J.T.; Raciti, D.; Smith, C.; Ringwald, M.; Marai, G.E.; Arighi, C.; Shatkay, H. Utilizing image and caption information for biomedical document classification. *Bioinformatics* **2021**, *37* (Suppl. 1), i468–i476.
28. Jiang, X.; Li, P.; Kadin, J.; Blake, J.A.; Ringwald, M.; Shatkay, H. Integrating image caption information into biomedical document classification in support of biocuration. *Database* **2020**, *2020*, baaa024, https://doi.org/10.1093/database/baaa024.
29. Kaur, G.; Kaushik, A.; Sharma, S. Cooking is creating emotion: A study on hinglish sentiments of youtube cookery channels using semi-supervised approach. *Big Data Cogn. Comput.* **2019**, *3*, 37, https://doi.org/10.3390/bdcc3030037.
30. Shah, S.R.; Kaushik, A.; Sharma, S.; Shah, J. Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models. *Big Data Cogn. Comput.* **2020**, *4*, 3, https://doi.org/10.3390/bdcc4010003.