SPATIAL PLANNING TEXT INFORMATION PROCESSING WITH USE OF MACHINE LEARNING METHODS

I. Kaczmarek[1, *], A. Iwaniak[2,6], A. Świetlicka[3,6], M. Piwowarczyk[4,6], F. Harvey[5]

[1] Institute of Spatial Economy, Wroclaw University of Environmental and Life Sciences, Poland - iwona.kaczmarek@upwr.edu.pl
[2] Institute of Geodesy and Geoinformatics, Wroclaw University of Environmental and Life Sciences, Poland
[3] Institute of Automatic Control and Robotics, Poznan University of Technology, Poland - aleksandra.swietlicka@put.poznan.pl
[4] Faculty of Computer Science and Management, Wrocław University of Science and Technology, Poland - mateusz.piwowarczyk@pwr.edu.pl
[5] Department of Cartography and Visual Communication, Leibniz Institute for Regional Geography, Leipzig, Germany - f_harvey@leibniz-ifl.de
[6] Wroclaw Institute of Spatial Information and Artificial Intelligence, Poland – adam.iwaniak@wizipisi.pl

Commission IV

KEY WORDS: spatial planning documents, zoning plan, unsupervised machine learning, LSTM, neural networks, NLP

ABSTRACT:

Spatial development plans provide an important information on future land development capabilities. Unfortunately, at the moment access to planning information in Poland is limited. Despite many initiatives taken to standardize planning documents, the standard for recording plans has not yet been developed. Each of the planning areas has a symbol and a category of land use, which is different in each of the plans. For this reason, it is very difficult to carry out an analysis enabling aggregation of all areas with a specific, the same development function.
The authors in the article conduct experiments aimed at using machine learning methods for the needs of processing the text part of plans and their classification. The main aim was to find the best method for grouping texts of zones with the same land use. The experiment consists in an attempt to automatically classify the texts of findings for individual areas into the 10 defined categories of land use. Thanks to this, it is possible to predict the future land use function for a specific zone text regulation and aggregate all zones with specific land use type.
In the proposed solution for the classification problem of heterogeneous planning information authors used *k*-means algorithm and artificial neural networks. The main challenge for this solution, however, was not the design of the classification tool but rather the preprocessing of the text. In this paper an approach for text preprocessing as well as selected methods of text classification is presented. The results of the work indicate greater use of CNN's usability to solve the problem presented. *K*-means clustering produces clusters, in which texts are not grouped according to land use function, which is not useful in the context of zones aggregation.

1. INTRODUCTION

Spatial planning information is information that, as the name suggests, apply to a planned and regulated state. In the context of spatial planning, it is, therefore, information that concerns desirable future changes that are planned in and as administrative space. The sources of this information are spatial planning documents. Depending on the level of detail they may contain plans, directions or spatial policy implemented in a given area. These documents can also contain more detailed rules and arrangements regarding future land development.

The basic planning document in Poland in communities is the local spatial development plan, which is drawn up at the local (municipal) level. It is the most important document shaping spatial planning in the commune. This is primarily due to the fact that the local plan has the status of an act of local law and its provisions directly determine the rights of property owners. They have direct binding effects for property owners. The plan is also a document based on which other decisions are made, e.g. measures for economic development or decisions on environmental conditions. Planning information from the plan is of key importance for monitoring planned changes in land use.

Poland has a hierarchical planning system. Plans are prepared at national, regional and local level. In Poland, municipalities are not obliged to create local plans (Böhm A., 2008). However, a lack of a local plan causes spatial development that is often chaotic and uncontrolled, resulting in dysfunctional spatial organization. Work has been undertaken for many years to standardize local plans in Poland. In some European countries, i.e. the Netherlands (IMRO standard) or in Germany there are standards for recording plans, thanks to which it is possible to compare and analyse planned changes in space across the whole country. In Poland each region/city has its own rules for developing plans. Although there are legal conditions determining the minimum scope of the local plan (Regulation, 2003), this scope is extended and invariably modified. The lack of standards for writing and publishing plans makes obtaining comprehensive information on planned changes on a scale larger than a city or region very difficult. Performing even a simple analysis, for example, finding all areas for multi-family housing in a region, is currently very difficult. It can be an easy task on a city scale that uses a certain standard. At present, at the scale of the entire region, it is impossible. Analysis of this kind of

* Corresponding author

information is important from the point of view of monitoring planned investment processes. Lack of monitoring can cause serious effects in space, such as urban sprawl and escalating spatial conflicts (Kazak J., 2013).

The authors in the article undertake a study involving the processing of the text parts of the local spatial development plan and its classification into specific categories of land types for analysis and comparisons. For this purpose, they use data that describes detailed regulations for individual zones in a plan, available in the form of HTML files.

For the solution of the classification of heterogeneous planning information problem we propose the use of well-known methods of supervised and unsupervised learning methods. Both of them have some advantages and disadvantages. The main advantage of unsupervised learning method is that the data does not need to be labelled, what minimize human efforts to obtain data manually.

However, supervised learning models thanks to direct feedback and evaluation based on it are able to produce more accurate results. In general we can use supervised learning to predict exact known output and unsupervised learning for tasks where we don`t know what exact output should be and we want to find hidden patterns in our data. Indeed unsupervised learning methods are more adaptive as they infer patterns straight from the data, not from their labels. The target variables may change over time and lead to Concept Drift or Concept Evolution. This characteristic could be useful in tasks where all sets of labels are not completely known or they are evolving. One of disadvantages of unsupervised methods is that they are harder to evaluate. If we know the labels and want to compare accuracy of classification task with the clusters produced by selected unsupervised algorithm we need to take care of things like labels permutations among clusters, homogeneity and completeness of labels in particular clusters. We could also combine supervised and unsupervised learning methods and use semi-supervised learning where some of the training data are missing their labels.

From the unsupervised learning methods we chose one of the simplest ones - the *k*-means method (MacQueen, 1967). It is a well-known technique for data cluster analysis and together with its simplicity it constitutes a great tool which quickly provides training results. From the supervised learning methods we decided to focus on artificial neural networks, which are well known from their advantages like: generalization ability, prediction of output data based on input data without the need to explicitly define the relationship between them or that they can learn any dynamic, non-linear input-output relationship arbitrarily well (Świetlicka et al., 2019). It was also shown that artificial neural networks can solve many of the natural language processing (NLP), like classification, named entity recognition (NER), image captioning, language translation and many others (Brownlee, 2019). Supervised machine learning (e.g. artificial neural networks) models require a lot of labeled data to reach good accuracy – but in many cases we know all set of labels but we don`t have resources to label our whole dataset. In such cases we could use techniques called Few-Shot Learning to produce classifiers trained with a very small amount of labelled training data.

Using machine learning methods in classification of texts from zoning plans requires several steps, as presented in Fig. 1.

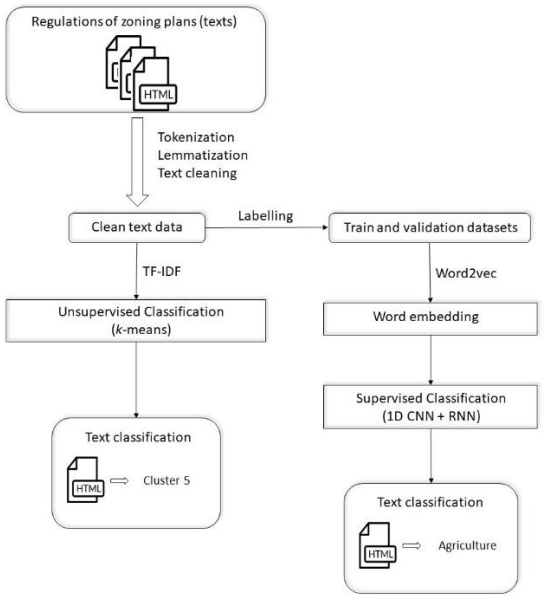

Figure 1. Workflow of conducted works for texts classification.

The paper is organised as follows. In section 2 we are providing a short description of the context of our problem. In section 3 we describe methods of preprocessing and supervised and unsupervised methods of classification, while in section 4 we provide results of the proposed methods. Finally in section 5 we provide a short conclusion with prospective future works.

2. BACKGROUND AND CONTEXT OF A PROBLEM

The local spatial development plan (zoning plan) is an act of law in which the purpose of a given area is determined, the location of public purpose investments is determined and the development methods and development conditions are determined. It consists of text - plan text and graphic - plan drawing (Figure 2). Both parts are integrally connected and legally binding. The zoning plan in Poland, like in many European countries, sets out the so-called zones where specific types of arrangements apply (regulations, restrictions, rules). The plan defines the general regulations that apply to the entire area covered by the plan and specific regulations that apply to specific zones in the plan.

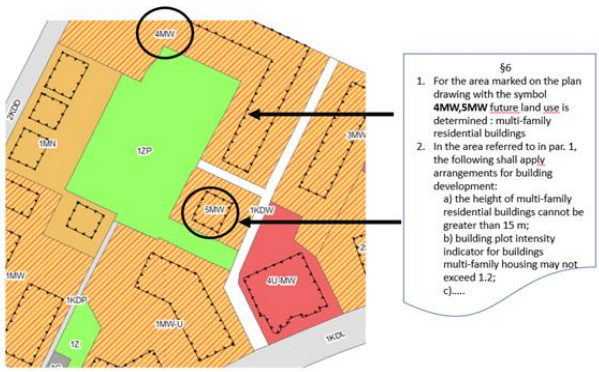

Figure 2. Fragment of a drawing of a plan with the text related to the drawing.

The representation of zoning plans varies greatly. The differences appear in both the textual and graphic parts. In the textual part they concern both the substantive content and various ways of

editing the plan's text. In the graphic part, the differences are mainly associated with the use of various symbols and markings defining the future land use of the area. Sharing local plans also varies. Most municipalities provide plans in the form of a georeferenced raster image. A small percentage of plans are available in vector form. Usually, larger cities that have a standard, choose this form of sharing plans (Kaczmarek et al., 2014).

The service of National Integration of Local Spatial Development Plans has been available in Poland, following the INSPIRE Directive, since 2017 (National Integration of Local Spatial Development Plans, 2020) It provides current integrated local spatial development plans from 1296 local government units from 2477 municipalities in the country using the WMS (Web Map Service) standard. Currently 303 units are vector plans, while the remaining 993 are in the form of a raster with geo-reference (as of 30/03/2020). Detailed plans for specific areas have been made available for the plans in vector form using the GetFeatureInfo method (Fig.3). This form of the integrated document is particularly desirable because it provides the user not only with information about what is on the plan drawing, but also detailed land development regulations recorded in the text part of the plan. The interpretation of the plan based on the graphic part (plan drawing) allows obtaining information only about the intended purpose of the area. Obtaining information about other arrangements, i.e. land development rules, building lines in force, etc. is only possible after consulting the plan text.

Figure 3. Fragment of the planning document available at geoportal.gov.pl together with the corresponding fragment of detailed regulations.

Thanks to this form of integration, it is possible to obtain information directly from the geoportal not only about the textual regulations related to a given zone, but also it is possible to download the entire content of the plan (usually in the form of PDF files) and the legend of the plan drawing.

Due to the lack of a uniform classification of future land development, it is not possible to integrate all areas for which single-family housing is intended. According to the current ordinance and currently prevailing planning practice, the symbol reserved for this type of area is MN. However, due to the occurrence of often mixed functions, such areas can be represented using many different symbols, e.g. 1-10MN_RM, MNu, MN/U, 30dMN_U, MNU_B1, MNMT3, and many others. It is not possible to uniquely identify the purpose of the area by its symbol and the designation.

The authors carry out research aimed at assigning individual texts of the plans' findings together with their symbols to 10 defined categories of land use, i.e.:

1. communication areas
2. agriculture
3. single-family housing
4. multi-family residential development
5. residential areas
6. areas of technical and production buildings
7. service development areas
8. green areas and water
9. areas of technical infrastructure
10. other

The above classification is based on the regulation regarding the required scope of the local plan (Regulation, 2003). The regulation specifies basic land use categories as well as graphic symbols used while preparing spatial development plans in Poland (Jaroszewicz, 2016).

The research question is if machine learning methods can help to solve problems in spatial planning related to the integration of zoning plans. Are there methods that allow for automatic classification of the areas with the same land use and then allow their aggregation in the wider scale (e.g. country)?

3. METHODS

3.1 Preprocessing of data

The main challenge in the task of classification of spatial planning documents is the problem of preprocessing of the text - which is obviously in Polish - with complex declensions of nouns, adjectives, adverbs, and counting words.

According to many available on-line articles, Polish is in the top 10 of the most difficult languages in the world (e.g. Macedo, 2015). It is indeed very complex, not only from the pronunciation point of view but mostly because of its grammar. Seven cases of nouns together with the conjugation of verbs creates a wide variety of a single word's forms. As a simple example in Table 1 we are showing a comparison of the word "run" with its conjugation in English and in Polish. An additional problem is related to the complexity of a single sentence, which in Polish can be built with many dependent clauses.

ENGLISH: RUN
run, ran, runs, running
POLISH: BIEGAĆ
biegam, biegasz, biega, biegamy, biegacie, biegają, biegałem, biegałeś, biegał, biegaliśmy, biegaliście, biegali, biegałam, biegałaś, biegała, biegałyśmy, biegałyście, biegały, biegało, biegaj, biegajmy, biegajcie, biegałbym, biegałabym, biegałbyś, biegałabyś, biegałaby, biegalibyśmy, biegalibyście, biegałybyście, biegaliby, biegałyby, biegający, niebiegający, biegająca, niebiegająca, biegające, niebiegające, biegając, bieganie, niebieganie, …

Tab. 1. Conjugation of the word "run" in English and Polish languages.

The preprocessing of text includes preparation of data for further text mining. The main steps of preprocessing text are: tokenizing, stemming and lemmatization. Tokenization is the process of dividing text into meaningful pieces, which are called tokens. The text is divided into words, which then are grouped into

sequences. The best grouping is with respect to sentences. Stemming allows to extract the base form of words (eg. playing - play). Lemmatization is similar to stemming but takes into consideration the morphological analysis of the words.

Text representation can be done in different ways. One can be a bag-of-words where words appear independently and the order is not taken into account (Huang, 2008). The bag-of-words model allow us to represent text as numerical feature vectors. In such vector only the occurrences of given word is counted in text. In aim to reduce the weight of unimportant words in a text *tf-idf* method (term-frequency - inverse document-frequency) is often used (Aizawa, 2003, Robertson, 2004). It allows to describe the relevance of words in the text. The relevance is high if the word is often found in a specific text and rarely in others. *Tf-idf* can be calculated by multiplying the value of term frequency with inverse document frequency. There are many formulas to calculate term frequency (*tf*), the most common is:

$$tf(t,d) = log(\frac{N}{df(t)}) + 1, \quad (1)$$

where *t* denotes single term, *d* is a single document, *df* is a document frequency of t, and *N* is a number of considered documents. Inverse Document Frequency (*idf*) value can be then calculated from the following formula:

$$idf_j = log(\frac{N}{df_j}) \quad (2)$$

Different methods require different preparation of data. While for the unsupervised methods we chose the *tf-idf* method, for the supervised algorithms we decided to use standard tokenization methods.

To perform tokenization of text we decided to focus only on the basic forms of words. For this purpose we reached for the so called *Morfeusz* - inflectional analyzer and generator for Polish language morphology (Woliński, 2014). This public tool enables analysis of the text in order to extract the basic meaning of the word together with its features (lemma; tag; entity type: name, surname, geographical name, etc.). The basic forms of words were used to create the Tokenizer, which allows to vectorize a text corpus, by turning each text into a sequence of integers (Fig. 4).

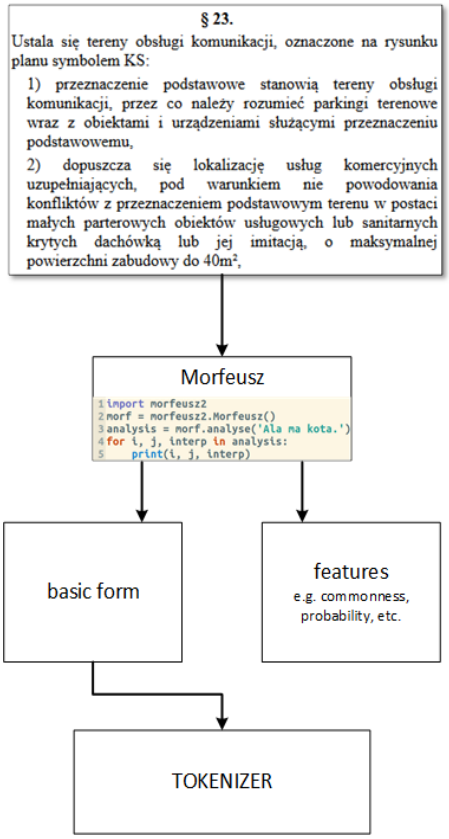

Figure 4. Preliminary language „cleansing".

3.2 K-means (unsupervised learning)

As a part of the first stage of research, the classification was performed by the unsupervised method using the *k*-means algorithm. This algorithm is one of the simplest and best known algorithms for clustering (MacQueen, 1967). The purpose of the *k*-means algorithm is to divide a given data set into *k* clusters, in which the number of clusters is defined by the user. The main task is to find centroids in each of the clusters. These centroids are created on the basis of mutual similarity, measured by a specific measure of distance (usually Euclidean distance). The algorithm classifies objects in such a way that the variance inside a single cluster is the smallest and the highest between all the clusters at the same time.

3.3 Artificial neural networks (supervised learning)

For the classification problem we have considered a few different artificial neural network structures and finally chose one for further analysis. Structure of this neural network is based on convolutional and GRU (Gated Recurrent Unit) layers arranged into a multi-channel structure (see Fig. 5). Each thread uses the embedding layer, which enables representing words as *n*-dimensional vectors (word2vec). By using the embedding layer we are able to convey the relative meaning of words for the neural network. Embedding layer is a matrix of coefficients, where each row represents a single word. Polish word2vec is available in many versions thanks to the Polish Academy of Science. This great collection of representations of Polish words was created with Python Gensim package (Řehůřek, et. al., 2010) based on two text corpuses: National Corpus of Polish (*pol. Narodowy Korpus Języka Polskiego - NKJP*) and Polish Wikipedia (see Fig. 6).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-95-2020 | © Authors 2020. CC BY 4.0 License.

Computational simplicity of convolutional layers enables making more sophisticated, extensive structures. In the presented solution we used three parallel threads as shown in Figure 5. Each thread was implemented with a different kernel size of the convolution, which can be understood as a grouping factor of words in the input sequence. The feed-back connections in the recurrent layer result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus, the network can maintain the previous state, allowing it to perform such tasks as sequence-prediction that are beyond the power of a standard multilayer perceptron. The last fully-connected layer (Dense) is performing the classification. As an activation function in this layer we used *softmax* function, while in the remaining layers we used *relu* (rectified linear unit) function. The model was compiled with *categorical crossentropy* loss function and *Adam* optimizer. To prevent neural network from overfitting we added *dropout* layers, which were set to drop 20% of neurons in the preceding convolutional and fully-connected layers.

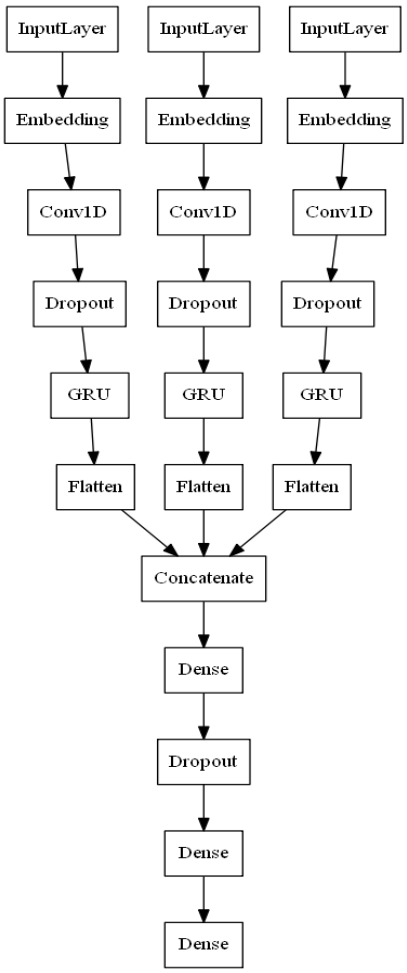

Figure 5. Structure of the multi-channel convolutional neural network in the new architecture.

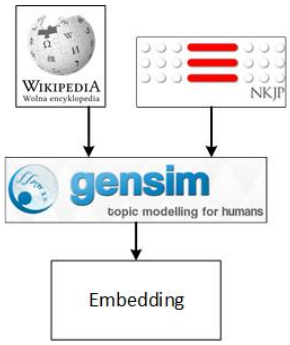

Figure 6. Preparation of the Embedding layer.

4. RESULTS

4.1 K-means clustering

Ten classes have been defined in the input parameters to the model. One of the main issue in *k*-means clustering is determining the number of clusters which has an effect on the clustering results. A number of methods for estimating the optimal number of clusters have been proposed e.g. elbow method (Coates et al.,2012), cross-validation (Kaufman et al.,1990), gap statistic method (Tibshirani et al., 2001). In this case the number of clusters was determined by number of land use categories. The aim was to group texts into 10 defined categories, therefore none of the statistical methods was used.

Entry to the model were fragments of plan regulations (text) without assigned categories. The Frequency-Inverse Document Frequency (*tf-idf*) algorithm was used to convert the text to numerical form. The *scikit learn* package was used to convert the texts to numeric form. The *tf-idf* algorithm is implemented in *TfidfVectorizer* module.

The results of clustering are presented in Fig. 7.

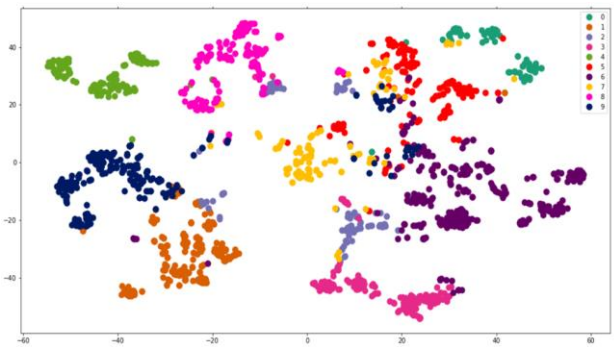

Figure 7. Results of k-means clustering.

Figure 7 shows the separate clusters grouping according to the *k*-means clusters. In order to check the usefulness of clustering for the needs of spatial planning, the analysis and verification of texts in individual clusters were carried out. The analysis of the most commonly used words in the separated clusters indicates the low utility of the *k*-means algorithm for the solution to the problem presented in the article, i.e. the grouping of texts according to the category of future land use. In order to check whether there is a correlation between the created clusters and the categories that were assigned to each of the zones, a combination of both information was presented in Fig.8 and Table 2. The figure and table present the number of texts that have been classified for each of the resulting clusters together with the category of land use adopted for each text. For example, in cluster No. 1 the texts

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-95-2020 | © Authors 2020. CC BY 4.0 License.

of regulations were grouped in the following categories: agriculture (0 texts), communication areas (80), areas of technical and production buildings (0), areas of technical infrastructure (8), single-family housing (34), other (0), residential areas (0), green areas and water (38), multi-family residential development (30), service development areas (12). It is noticeable that in cluster #2 there are only text labelled with communication areas category (517). In other clusters there are texts assigned to many categories of future land use. Results of unsupervised clustering with *k*-means clustering indicate that they do not meet or are sufficient in the context of automatic classification of zones based on their land use. With CNN the results are far more insightful.

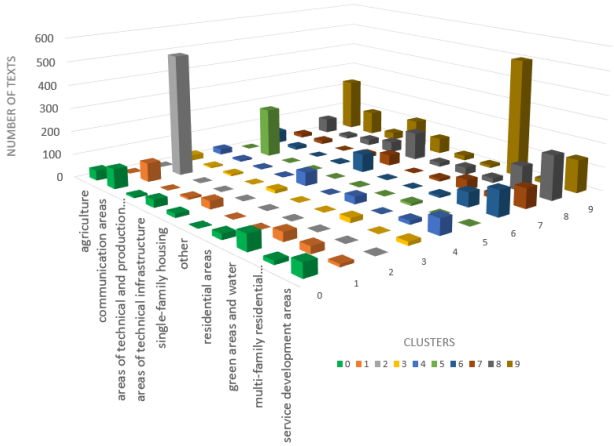

Figure 8. Comparison of number of texts in clusters with assigned category for each text.

category \ cluster	0	1	2	3	4	5	6	7	8	9
agriculture	43	0	0	24	24	0	54	16	64	228
communication areas	85	80	517	6	8	212	14	16	13	102
areas of technical and production buildings	8	0	0	8	4	0	0	1	26	32
areas of technical infrastructure	33	8	0	0	2	0	4	18	45	118
single-family housing	17	34	0	14	58	0	78	47	119	67
other	2	0	0	0	0	2	2	0	17	21
residential areas	24	0	0	0	32	0	0	12	31	12
green areas and water	69	38	0	20	0	10	10	44	20	501
multi-family residential development	16	30	0	0	16	0	62	22	107	18
service development areas	57	12	0	16	68	0	110	85	188	139

Tab.2 Evaluation of *k*-means clustering.

4.2 Artificial neural networks

Artificial neural networks were implemented in Python with use of Tensorflow 2 and Keras packages (Gulli et al., 2019). To carry out the experiment we had only about labeled 4 100 texts (fragments of plan regulations). Training data includes symbol of a zone, textual regulations of a zone and label: category of land use (Fig.9).

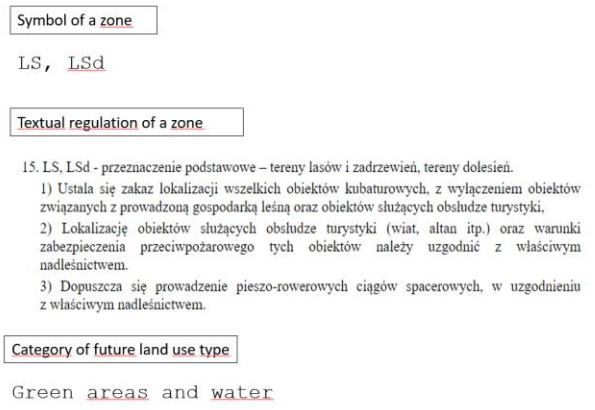

Figure 9. Training data: symbol of a zone, textual regulations of a zone, category of land use

In order to train and check the effectiveness of prediction, the entire set was divided into a train (75%) and validation (25%) datasets. The training and validation samples were generated randomly.

In Figures 10 and 11, the model loss and accuracy are presented. From these performance plots we can observe, that despite a discrepancy between waveforms for train and validation sets, that occur between 5th and 20th epoch, after around 25 epochs they start to converge. This observation leads us to a conclusion, that the neural network after around 30 epochs is already trained, as the value of loss does not go any lower, and at the same time the value of accuracy does not go any higher. Further training was not necessary, however it did not make the neural network overtrain. This allows us to conclude that the architecture was properly chosen. After 60 epochs, loss of train and validation data reached 9.7960e-04 and 0.1033, while accuracy reached 1 and 0.9823, respectively.

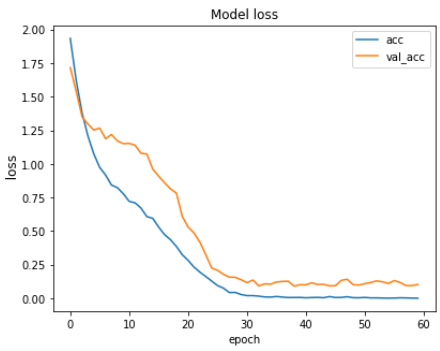

Figure 10. The plot of model loss on the training and validation datasets over training epochs.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-VI-4-W2-2020-95-2020 | © Authors 2020. CC BY 4.0 License.

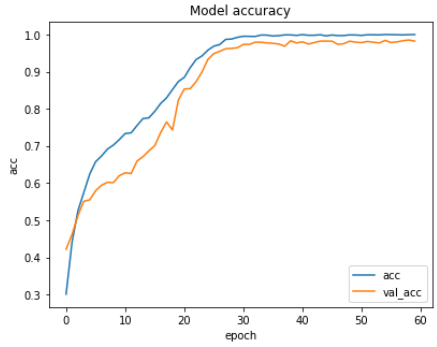

Figure 11. The plot of model accuracy on the training and validation datasets.

In any classification problem (binary or multiclass) accuracy seems to be insufficient, as classes can be unbalanced. That is why we reached for precision, recall and F_1 scores together with a plot of the ROC curve (Receiver Operating Characteristic). As we considered a multiclass problem the scores were calculated as a weighted average of the scores, where the support was chosen as the number of true instances for each label. The precision, recall and F_1 scores reached 0.9817, 0.9823 and 0.9816, respectively, where we used a weighted average. Additionally in Figure 12 the ROC curves are presented, each ROC curve was calculated for a single label versus all the remaining ones. Analysing this waveform we can observe that even with a small threshold we can obtain a high TPR (True Positive Ratio) for a small FPR (False Positive Ratio), which leads us to a conclusion, that our model of neural network classifies correctly most of the texts.

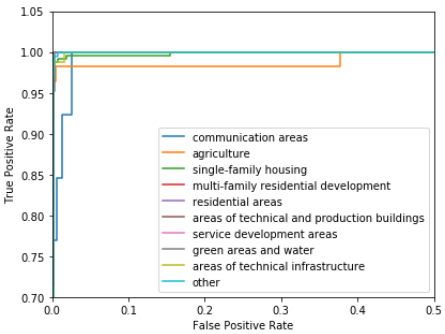

Figure 12. ROC curve

5. CONCLUSIONS AND FUTURE WORKS

While this analysis shows the potential of classification approach using select machine learning methods, it already demonstrates that it provides helpful analytical ways to assess and synthesize Polish planning activities in regions and possibly even larger areas. The results of the experiments show that machine learning methods can be used with success in helping solving problems in the spatial planning domain. The classification of textual regulations of zones can be information enriching the spatial data (e.g. published in GML). Thanks to this, it is possible to get an answer to the question about the total area of zones with a specific type of land use in the area where vector data is available.

Each of the methods which was used in the aim of text classification has its own drawbacks. Using the CNN requires a lot of labeled data, while *k*-means results did not provide expected results in the context of grouping areas with the same land use. Working on NLP problems it is almost impossible not to consider Transformers by Hugging Face, which at this moment offers over 30 pre-trained models, including BERT (Vaswani, et. al. 2017, Devlin, et al. 2018), RoBERTa (Liu, et al. 2019) and XLNet (Yang, et al. 2019). The biggest struggle is again with the morphology of the Polish language, while these models are trained on full sentences, which in Polish, as we mentioned, can be very long, with many dependent clauses. Adapting these models or preparing text to make it possible to use them is one of our next milestones.

Future work involves at first developing the CNN and connecting the spatial features in text to the graphic planning documents to be able to better assess the accuracy of the classifications, also given the noted grammatical issues in the Polish language.

In future work, experiments are also planned using other unsupervised methods, i.e. DBSCAN and Affinity propagation. It is also planned to extend and modify the classification of zones in order to extract more detailed information on the future land use of the zone. Future work also includes experiments involving the extraction from the text of the spatial plan document of other relevant information, i.e. indicators related to land development (e.g. floor area ratio, maximum building height).

ACKNOWLEDGEMENTS

The research is co-financed by the Wroclaw University of Environmental and Life Sciences in Poland under the B030/0004/20 support project „An innovative scientist”.
The research is co-financed by the The National Centre for Research and Development in Poland under the POIR 01.01.01-00-1274/17-00 project „Building a knowledge base on real estate”.

REFERENCES

Aizawa A., An information-theoretic perspective of tf–idf measures, 2003, *Information Processing & Management*, 39 (1), pp. 45-65, doi: 10.1016/S0306-4573(02)00021-3.

Böhm A., Effectiveness of the legal instruments for landscape protection and shaping existing in Poland, *Czasopismo techniczne Architektura*, Wydawnictwo Politechniki Krakowskiej, 2008, R. 105, z. 1-A, pp.137-146.

Brownlee J., Deep Learning for Natural Language Processing, 2019.

Coates A., Ng A., 2012, Learning Feature Representations with K-Means In *Neural Networks: Tricks of the Trade*, eds. Montavon, et al. , Springer, Berlin Heidelberg, 561–80

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

Gulli A., Kapoor A., Pal S., 2019, Deep Learning with Tensorflow 2 and Keras. Regression, Convnets, GANs, RNNs. NLP, and more With Tensor FOV 2 and the Keras ADI, Packt, Birmingham - Mumbai.

Huang A., Similarity measures for text document clustering, 2008, In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference* (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56.

Jaroszewicz J., Piotrowska L., 2016, *Implementation of the inspire directive in Poland in the scope of spatial data 'land use' theme*. Geomatics, Landmanagement and Landscape, 4, 125–157. DOI:http://doi.org/10.15576/GLL/2016.4.125.

Kaufman L., Rousseeuw P.J.,1990, Finding Groups in Data: an Introduction to Cluster Analysis. *John Wiley & Sons*, New York.

Kaczmarek I., Iwaniak A., Łukowicz J., 2014, New spatial planning data access methods through the implementation of the INSPIRE Directive, *Real Estate Management and Valuation*, vol. 22, no. 1, pp. 12-24, doi: 10.2478/remav-2014-0002

Kazak J., Szewrański Sz., 2013, Indicator-based environmental assessment of spatial planning with the use of Community Viz. In I. Ivan, P. Longley, J. Horak, D. Fritsch, J.Cheshire, T. Inspektor (Eds.) GIS OSTRAVA 2013 - Geoinformatics for City Transformation. Ostrava: Vsb-Tech Univ Ostrava

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V., RoBERTa, 2019, A Robustly Optimized BERT Pretraining Approach, *arXiv 1907.11692.*

Macedo H., Japanese, Finnish or Chinese? The 10 Hardest Languages for English Speakers to Learn, 2015, url: https://unbabel.com/blog/japanese-finnish-or-chinese-the-10-hardest-languages-for-english-speakers-to-learn/, last visit: 04/19/2020.

MacQueen J., 1967, Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: Statistics, 281--297, University of California Press, Berkeley, Calif.

National Integration of Local Spatial Development Plans, URL: http://integracja.gugik.gov.pl/cgi-bin/KrajowaIntegracjaMiejscowychPlanowZagospodarowaniaPrzestrzennego, last visit: 30/04/2020

Regulation of the Minister of Infrastructure of 26 August 2003 regarding the required scope of the project of a local spatial development plan (pol. Rozporządzenie Ministra Infrastruktury z dnia 26 sierpnia 2003 r. w sprawie wymaganego zakresu projektu miejscowego planu zagospodarowania przestrzennego), Dz.U. 2003 nr 164 poz. 1587.

Řehůřek R., Sojka P., 2010, Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45-50, Valletta, Malta, doi: 10.13140/2.1.2393.1847.

Robertson S., Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation* 60 (5), 2004, pp. 503-520, doi: 10.1108/00220410410560582.

Świetlicka A., Kolanowski K., Kapela R., 2019, Training the Stochastic Kinetic Model of Neuron for Calculation of an Object's Position in Space, *Journal of Intelligent & Robotic Systems*, doi: 10.1007/s10846-019-01068-0.

Tibshirani R., Walther G., Hastie T., 2001, Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411-423

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin. I., 2017, Attention Is All You Need, *arXiv 1706.03762.*

Woliński M.,2014, Morfeusz reloaded, In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, pages 1106–1111, Reykjavík, Iceland, ELRA.

Yang Z., Dai Z., Yang Y., Carbonell J. G., Salakhutdinov R., Le Q. V., 2019, XLNet: Generalized Autoregressive Pretraining for Language Understanding, *arXiv 1906.08237*.