Position Papers of the of the 17th Conference on Computer Science and Intelligence Systems pp. 35–39
DOI: 10.15439/2022F285
ISSN 2300-5963 ACSIS, Vol. 31

Detecting Uninformative Research Job Titles via Classifier Failures – Zero Shot Approach

Martin Víta
Faculty of Informatics and Statistics, Prague University of Economics and Business
Ekonomická 957, 148 00 Praha 4–Kunratice, Czech Republic
Email: info@martinvita.eu

***Abstract*—The aim of this paper is to introduce a novel approach to detecting "uninformative" job titles in research domain, i.e., detecting titles that convey little or no information about the focus and/or content of a particular job – like "Academic staff member AP/2", "PhD student position" etc. Such job titles decrease the success rate of job advertisements. The proposed approach belongs to zero shot approaches – it exploits only existing, easy accessible classification of jobs to research fields and it does not require any additional (manual) annotations. This work introduces an experimental corpus and provides preliminary results of our approach.**

I. Introduction

BASED on an internal survey of *ResearchJobs.cz*[1], job advertisements with well prepared, informative titles gain more attention from potential candidates in terms of (unique) users visits than vacancies having only general titles like "Postdoc position", "Academic staff member" etc. Moreover, as shown in [1], a job title is a suitable feature in predicting CTR[2] of job advertisements. Hence, a question of automated detecting of inappropriate job titles naturally arises in this setting.

Obviously, the task of detection uninformative titles can be straightforwardly addressed by common supervised ML techniques requiring an annotated corpus labeled in a binary way (informative/uninformative). However, a preparation of such a corpus is resource-extensive activity.

Our approach is based on the assumption that *an appropriate job title provides us enough information to classify the job advertisement to a correct research field*. Moreover, we assume that a classification of job offers to predefined fields is commonly available (usually selected from predefined categories by the user when submitting the advertisement).

Roughly said, *if a correctly working classifier of research fields assigns an incorrect label to a job title – since the label is known – then the job has an inappropriate title*, i.e., the failure of the classifier indicates an uninformative or even incorrect title. For example, if a classifier assigns "Computer science" label to a job entitled only "PhD student" submitted by the user within "Medical sciences" field, than we can conclude that the title is not appropriate, since it did not provide enough information to predict the research field correctly. In contrast, "Postdoc in therapy for neuromuscular diseases" labeled as "Medical sciences" by the user (who submitted the offer) and also by the classifier, than it indicates *sufficiently informative* title ensuring correct (automatic) classification.

Unlike ordinary classification task where *the text is the only input*, in our task, the input consists of the text (job title) as the first part and also of the human selected/submitted class as the second part. The result then depends on the *difference or identity of predicted and submitted class*.

Such a tool for detecting uninformative job titles can be directly used for an automatic feedback to users when submitting their advertisements and/or together with an automatic recommendation of a more suitable title.

The paper organization follows the standard IMRAD structure: Section 2 provides an overview about methods, i.e., models and data in our case. Section 3 contains results, Section 4 then the corresponding discussion. Since this paper has a "proof-of-concept character", the paper is completed with the overview of further work research directions.

II. Models and Data Involved

In this section, we provide a description of ML models and data used for training and zero shot task testing.

A. Core Idea of Zero Shot Approach

As already mentioned in Introduction, the keystone of the proposed approach is a classifier assigning a research field to a job title.

The trained classifier will be subsequently used to classify job titles from the (zero shot) test set where research field labels are known and these items are also equipped with a binary (informative–uninformative) human labels which serves as a gold-standard. If there is a mismatch of "real" research field label and the output of the classifier, than the title is marked as `uninformative`, otherwise marked as `informative`. The evaluation w.r.t. these two labels is performed further in a standard way.

The basic dataset to be used in this work (see Subsection 2.3) contains 5,341 positions from Euraxess portal, thus the same number of job titles. The median length (number of characters) of job titles is 57, whereas the the average is 65.76, 1st quartile 36 and 3rd quartile 85 characters. Thus we are dealing with classification of short texts.

[1]Czech job portal focused on research and academic vacancies

[2]Click-Through Rate, CTR is defined as the number of clicks that a given advertisement receives divided by the number of times it is shown.

Classification of short texts belongs to one of the traditional tasks of ML/NLP with a long history [2]. This direction of research was often driven by motivation for sentiment analysis of tweets [3] and other social media content.

B. Models for Classification

The general task of classification of short text can be tackled by several ML methods. However, the aim of this work is not to focus on the classification itself – but later on the zero shot [4] part. This section provides an overview of models as well as corresponding features involved. In this work we will deal only with neural networks based models, namely:

- Character-based 1D-convolutional neural network (Char-CNNs),
- Convolutional Word2Vec-based model (CNNs),
- Universal sentence embeddings (USE).

CharCNN Character-based convolutional networks are a frequent choice for processing of short texts [5]. Their advantages are – among others – that they can be employed in language agnostic setting [3], they are robust to misspellings and they can easily deal with special character combinations such as emoticons etc.

In this model, a job title is represented as a fixed length sequence (255 characters, since it is the maximal length of the job title) of one-hot encoded character vectors – in this case we deal with 128 characters, shorter titles are padded with zero vectors. Therefore, the corresponding matrix has dimension 128×255. This matrix is subsequently processed by 1D-convolutional layer with 25 filters of kernel size equal to 3 (i.e., we are processing "character-trigrams") and the result of this convolutional layer is fed to 1d-max pooling layer with pool-size again equal to 3. The decision is made by standard softmax dense layer with 9 output neurons (i.e., number of output classes). The final model contains 28,759 trainable parameters, the hyperparameters of the model (number of filters etc. were set using grid search).

Diagram of the architecture is shown on Figure 1[3].

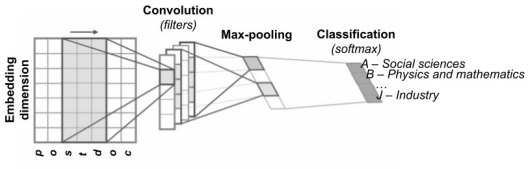

Fig. 1. Character-based CNN architecture

CNNs Convolutional word2vec-based models belong to traditional architectures for text classification [2]. Convolutional neural networks were successfully used in short text classification (tweets in particular) in many branches, including biomedical domain [6].

[3]Diagrams are modified versions of one from: https://towardsdatascience.com/convolutional-neural-network-in-natural-language-processing-96d67f91275c

In this setting, a job title is represented of a fixed length (30 words in our case) sequence of word2vec [7] embeddings. Shorter titles are padded by zero vectors. Since we deal with texts of research domain, we did not used general word2vec embeddings but pretrained embeddings of dimension 200 learned on texts of scientific (biomed) domain, that were used in [8]. Representation of words that occur in a job title but are not contained among words with pretrained embeddings is uniformly set to zero vectors.

This sentence matrix (200×30) is fed to a 1D-convolutional layers with 22 filters and kernel size 3 followed by 1d-max-pooling layer with pool-size of 3. Hyperparameters were set again using grid search. Finally, softmax classification output layer is used. This model has 15,211 trainable parameters in total, the overall architecture is depicted on Figure 2 and it is formally similar to the previous case, however, here we do not use one hot encoding.

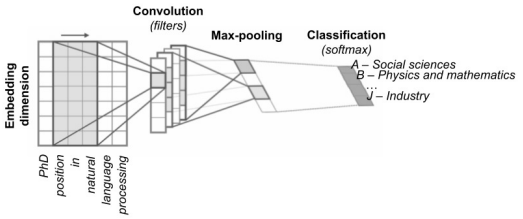

Fig. 2. Word2vec-based CNN architecture

USE As an example of more advanced methods – transformer-based representations, we used *Universal Sentence Encoder (USE)* [9], successfully applied in many areas such as Semantic Textual Similarity (STS), [10]. The trained model implementation was obtained from TensorFlow Hub[4]. It provides a 512-dim sentence (text snippet) representations – in our work, the pretrained network was used straightforwardly for feature extraction. These representations were subsequently fed into a dense layer (dim: 32; the number was obtained again by grid search), followed by the output softmax layer as in the previous cases. The model has 16,713 trainable parameters in total.

Implementation Details: The complete implementation of this work was elaborated in R + Keras library[5]. As optimizer, RMSprop [11] was used in all training scenarios. The number of epochs varies from 12 to 16 depending on the model and data involved.

C. Data Involved

The data used in this work can be basically divided into two groups: data used for training the "research field classifier" and data used for testing the zero shot approach (informative/uninformative classification).

[4]https://tfhub.dev/google/universal-sentence-encoder/4
[5]https://cran.r-project.org/web/packages/keras/index.html

TABLE I
DISTRIBUTION OF LABELS IN BASIC JOB TITLES DATASET

Research Field (class)	Instances
A – Social sciences	754
B – Physics and Mathematics	541
C – Chemistry	224
D – Geosciences	117
E – Biosciences	309
F – Medical Sciences	515
G – Agriculture	461
I – Informatics/computer science	262
J – Industry	588

1) Data for Learning the "Research Field Classifier": The key dataset is a database dump of Euraxess portal[6] from March, 2021. It was provided in the form of one large XML file. Each position has several attributes, however, from our point of view, only a few of them are relevant: job title and research field, in particular.

The raw dataset contains 5,341 positions. Each position is assigned to at least one research field (for example: *Psychological sciences*, *Physics* etc.) and may be assigned also to research subfields (for example: *Psychology*, *Applied physics* etc.). The total number of research fields is 41, including two special labels *All* and *Other*.

To ensure to deal with a single-class classification task, we filtered out only positions that have just only one research field label, and moreover, we did not take into account positions having *All* or *Other* labels. This resulted in a reduced dataset of 3,771 positions whereas each position is labeled by one of 39 labels. However, this labeling is strongly unbalanced – top 3 classes are *Engineering*, *Agricultural sciences* and *Medical sciences* containing 465, 461 and 446 positions respectively. On the other hand, the least numerous labels in this reduced dataset are *Criminology*, *Ethics in social sciences* and *Ethics in physical sciences* with 1, 1 and 2 occurrences respectively.

In order to deal with more balanced classification and to reduce the number of classes, we used a coarser Czech classification system of research branches[7] having 9 classes (more precisely, it deals with 10 classes, the last one is K – Defense, but this field is not taken into account). Simple handcrafted transformation rules were prepared. The utilization of this classification has also other reasons that will be obvious later in this chapter. The distribution of labels in the dataset of position titles is provided in Table I.

Subsequently, we randomly selected 3,000 of items (positions) to be the training set for research filed classification. The rest of 771 positions were left for further preparation of test set of zero shot ("informative/uninformative" classification task).

To achieve better classification accuracy, we also prepared an auxiliary annotated dataset of a bigger volume – research project titles together with their research branch classification. These data were taken from open data section of Czech R&D Information System[8] which gathers (meta)data about all R&D projects in the Czech Republic funded by public sources. This dataset contains 43,694 items (project name–classification pairs). The aim of exploiting this dataset was to extend the original training data by this easily obtainable stuff. Each model is trained both with the original basic dataset and this enriched one.

To provide a better idea of items in this auxiliary dataset, we randomly select three examples of project titles with corresponding classification labels.

- *Phospholipid metabolizing enzymes as new components of salicylic acid signalling pathway:* C – Chemistry
- *Communities and resources in late prehistory of Jebel Sabaloka, central Sudan: from analysis to synthesis:* A – Social sciences
- *Optimization of hunted species management in relation to the sustainable forest management:* G – Agriculture

2) Test Data for Zero Shot Classification: The second part of the data involved is the test dataset for zero shot classification, i.e., job titles manually labeled as informative or uninformative.

There were 771 remaining jobs (job titles) from Euraxess dataset that were not intended for training, whereas 102 (!) of them were manually marked as uninformative; the rest (i.e., 669 items) is considered as informative. This dataset of 102 uninformative job titles was subsequently enriched by another set of 48 uninformative titles (annotated manually again) which were obtained from a randomly shuffled collection of jobs from ResearchJobs.cz portal (this portal uses also the "A–J research branches" classification). Hence the number of uninformative examples in the test dataset reached 150. To obtain a balanced test set, 150 items with *informative* titles were randomly selected from already mentioned list of 669 items. This dataset can be provided upon a (mail) request. The content of uninformative subset of job titles is illustrated using a wordcloud, see Figure 3. Inter-annotator agreement was not investigated in this context.

Obviously, typical, i.e., most frequent, words in uninformative part of job titles are general names of academic/research positions (professor, PhD), words linked to hiring process (call, applications), general duties (teaching, research).

In addition to general words common for both classes, the informative job titles contain bigger amount of relatively infrequent words denoting particular research fields (physics, biology) and corresponding specific words (quantum, molecular).

Examples of informative and uninformative job titles randomly selected from this zero shot test dataset are provided in the following list (job title – user selected research field classification – informative/uninformative label):

- Doctoral student in Economic History (A – Social sciences): informative

[6]Euraxess portal is one of the most important European job portals focused on research and academic position. It publishes positions solely in English – https://euraxess.ec.europa.eu/jobs/.

[7]IS VaVaI: https://www.isvavai.cz/

[8]https://www.isvavai.cz/open-data

Fig. 3. Wordcloud generated from uninformative job titles

TABLE II
RESULTS IN DIFFERENT SCENARIOS (MODEL–DATA)

Architecture	Dataset	10-fold	UninfTask
CharCNN	basic	0.4413	**0.6900**
	ext	0.4100	0.6167
CNN	basic	0.5717	0.7600
	ext	0.5720	**0.7833**
USE	basic	0.5480	***0.7900***
	ext	0.5897	0.7867

- PhD scholarship in 6G Wireless Communications (J – Industry): informative
- Assistant Professor FSI UJEP (J – industry): uninformative
- Fellowship for Postdoctoral Researcher (F – Medical sciences): uninformative

III. RESULTS

Since our work relies on two-way classification, the evaluation is based mainly on accuracies. The evaluation has basically two levels: evaluation of research field classifier and evaluation of uninformative/informative classifier.

The evaluation of research field classifier is done as an average of accuracies in 10-fold cross validation, the evaluation of uninformative/informative classifier as standard accuracy.

The results in different scenarios (model-data) are summarized in Table II

Confusion Matrices: More detailed view on bold-marked results (i.e., best results for each model) are available via confusion matrices: Table III, Table IV and Table V.

IV. DISCUSSION

The best results were achieved using Universal Sentence Encoder. As can be seen from the confusion matrix, our algorithm based on research field classifier failures was able to detect 79 % of uninformative job titles. It should be mentioned that this proposed approach inherently implies certain error arising from the fact that in some cases the classifier can predict the correct class of uninformative title by chance.

TABLE III
CONFUSION TABLE FOR USE MODEL

		Predicted label	
		Uninformative	Informative
True label	Uninformative	126	24
	Informative	39	111

TABLE IV
CONFUSION TABLE FOR CNN MODEL

		Predicted label	
		Uninformative	Informative
True label	Uninformative	122	28
	Informative	37	113

On the other hand, 26 % of informative job titles were marked as uninformative – i.e., predicted and true label were *not equal* in the case of informative title ("informativeness" was labeled manually). Preliminary human conducted analysis indicates that most of these cases were borderline items with respect to output classes (classification system) and the the assumption of dealing with jobs that are *assigned just to one class* in our setting. A position "PhD in robotics" can serve as an example: both *Computer science* and *Industry* labels are relevant in this case, analogous situation is often between medical and biological sciences – the correct "real-world" assignment is the subject of the *whole text of job detail* which is not taken into account due to the main aim of this work. Hence an alert when predicted research field label and label selected by the user are not identical as a side-effect points out possibly confusing title.

According to observations of confusion matrix of CNN approach, we see that both USE and CNN are comparable. In the number of false negatives, CNN approach slightly outperforms USE, in true positives, the situation is reversed. The results of CharCNN are strongly below expectations.

In both successful approaches (USE and CNN) the effect of additional training items (research project titles) is marginal, moreover, in CNN approach training without additional data lead to better performance. Notable effect of enriching the training dataset was observed only in convolutional character-based approach.

The cases of *trully informative job titles labeled as uninformative* will be a subject of further investigations on a larger

TABLE V
CONFUSION TABLE FOR CHARCNN MODEL

		Predicted label	
		Uninformative	Informative
True label	Uninformative	123	27
	Informative	66	84

dataset. Generally, the sources of this misclassification belong to the following two groups:

1) wrong category assignment by user during submission of the job advertisement,
2) incorrect work of the research field classifier.

Relatively poor performance of the research field classifier in 10-fold cross validation at first sight is caused by the following reasons:

1) High proportion of uninformative job titles in the Euraxess dataset: according to our preliminary human experiments it consists approximately 1/8 of the dataset.
2) Frequent presence of borderline case (as described above).
3) Relatively high number of output classes (thus also a trivial majority vote classifier achieves very low accuracy).
4) Occasional occurrence of job titles in languages other than English, misspellings etc.

V. Conclusion and Further Work

We have introduced a novel zero shot approach to detection of uninformative job titles in research domain based on exploiting incorrect predictions of a job field classifier. We prepared corresponding experimental corpora and provide some preliminary results.

Further Work

Our results of this zero shot approach indicate that this chosen direction is promising. Nevertheless, there is a large room for improvement, mainly in the sense of exploiting fine-tuned variants of BERT [12] and its variants like SentenceBERT [13] and others [14].

For our preliminary experiments, the single label setting was chosen due its simplicity. However, the nature of the task is rather multi-label, thus we will adopt our approach for multi-label classification. The side effect is that we can immediately use larger datasets (without filtering jobs that are assigned just to one class).

Further generalization may lead also to fuzzy point of view: rather than speaking about crisp "text–class" membership function we can deal with a fuzzy membership: each job advertisement (and job title so) may belong to more classes with different degrees of membership – as an example we can consider positions like "Postdoc in cancer research" which are spanned between biological and medical sciences. Fuzzy approach to sentiment analysis of tweets [15] can serve as an inspiration.

As already mentioned, this work is restricted only to job titles in English. Another direction of further investigations can be naturally focused on language agnostic as well as multilingual approaches (analogous to [16] for instance) which will be able to detect uninformative titles also in other languages.

A separate chapter in further research is a language generation for improving job titles – given a text, i.e., content of the job advertisement, the task is to create an appropriate, *informative* job title. As a promising direction seems to be application of GPT transformers [17] for language generation as well as summarization techniques [18] – extreme summarization of particular parts of job detail (e.g., requirements) may be a suitable addition to an uninformative prefix (like "Postdoc", "PhD student" or "Assistant Professor").

References

[1] M. Jiang, Y. Fang, H. Xie, J. Chong, and M. Meng, "User click prediction for personalized job recommendation," *World Wide Web*, vol. 22, no. 1, pp. 325–345, 2019. doi: 10.1007/s11280-018-0568-z

[2] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019. doi: 10.3390/info10040150

[3] J. Wehrmann, W. Becker, H. E. Cagnini, and R. C. Barros, "A character-based convolutional neural network for language-agnostic twitter sentiment analysis," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966145 pp. 2384–2391.

[4] P. K. Pushp and M. M. Srivastava, "Train once, test anywhere: Zero-shot learning for text classification," *arXiv preprint arXiv:1712.05972*, 2017. doi: 10.48550/arXiv.1712.05972

[5] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[6] L. Akhtyamova, M. Alexandrov, and J. Cardiff, "Adverse drug extraction in twitter data using convolutional neural network," in *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*. IEEE, 2017. doi: 10.1109/DEXA.2017.34 pp. 88–92.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[8] G.-I. Brokos, P. Malakasiotis, and I. Androutsopoulos, "Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering," in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 2016. doi: 10.18653/v1/W16-2915 pp. 114–118.

[9] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018. doi: 10.48550/arXiv.1803.11175

[10] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:1708.00055*, 2017. doi: 10.48550/arXiv.1708.00055

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. doi: 10.48550/arXiv.1810.04805

[13] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. doi: 10.48550/arXiv.1908.10084

[14] A. Gillioz, J. Casas, E. Mugellini, and O. Abou Khaled, "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020. doi: 10.15439/2020F20 pp. 179–183.

[15] C. Jefferson, H. Liu, and M. Cocea, "Fuzzy approach for sentiment analysis," in *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE, 2017. doi: 10.1109/FUZZ-IEEE.2017.8015577 pp. 1–6.

[16] A. F. M. de Paula, R. F. da Silva, and I. B. Schlicht, "Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models," *arXiv preprint arXiv:2111.04551*, 2021. doi: doi.org/10.48550/arXiv.2111.04551

[17] B. Ghojogh and A. Ghodsi, "Attention mechanism, transformers, bert, and gpt: Tutorial and survey," 2020.

[18] I. Cachola, K. Lo, A. Cohan, and D. S. Weld, "Tldr: Extreme summarization of scientific documents," *arXiv preprint arXiv:2004.15011*, 2020. doi: 10.48550/arXiv.2004.15011