Computer Science • 19(3) 2018 https://doi.org/10.7494/csci.2018.19.3.2753

Joanna Misztal-Radecka

BUILDING SEMANTIC USER PROFILE FOR POLISH WEB NEWS PORTAL

Abstract *The aim of this research is to construct meaningful user profiles that are the most descriptive of user interests in the context of the media content that they browse. We use two distinct state-of-the-art numerical text-representation techniques: LDA topic modeling and Word2Vec word embeddings. We train our models on the collection of news articles in Polish and compare them with a model built on a general language corpus. We compare the performance of these algorithms on two practical tasks. First, we perform a qualitative analysis of the semantic relationships for similar article retrieval, and then we evaluate the predictive performance of distinct feature combinations for user gender classification. We apply the algorithms to the real-world dataset of Polish news service Onet. Our results show that the choice of text representation depends on the task – Word2Vec is more suitable for text comparison, especially for short texts such as titles. In the gender classification task, the best performance is obtained with a combination of features: topics from the article text and word embeddings from the title.*

Keywords user profiling, word embedding, topic modeling, natural language processing, gender prediction

Citation Computer Science 19(3) 2018: 307–332

1. Introduction

1.1. User modeling

Online news services provide a huge number of digital resources, but only a few articles are read by particular users; hence, data sparsity is a critical problem for many applications in this domain. Finding a meaningful representation of user interest is aimed to gather information about a particular user and is helpful in reducing the dimensionality. It also constitutes an important aspect of many tasks such as automatic advertising, personalized recommendations, click prediction, and user segmentation. This has been an active area of research since the 90s [26, 35].

A summary of the solutions used for building user profiles for Internet applications is presented in [10]. The authors grouped different techniques by the methods of information collection (implicit vs. explicit), profile types (long vs. short-term), and user profile representation (such as weighted keywords, semantic networks, concepts, and association rules). Some solutions require explicit user feedback by defining topics of interest or evaluating the articles read by users. In more-recent research, the profile is usually inferred automatically from user behavior (such as content clicks [20] or web searches [4]), and the preferences are defined by the type of content read by them. In [1], the authors performed a user study showing that profile transparency is an important aspect of personalized news systems.

Building a user-preference model from text-extracted features is also a popular technique for automatic recommendations. Content-based recommenders aim at finding items similar to what the user previously liked by building a user-preference model (as in the news recommender proposed by [19,22]). Content-based features are also commonly used to address the cold-start problem in collaborative filtering systems. In [34], the authors introduced a Collaborative Topic Modeling technique that combines the collaborating filtering approach with the content-based features extracted by topic modeling. Another combination of collaborative filtering and content-based features for news recommendation was proposed by [21]. Probabilistic language models for user profiling have also been applied to other domains – [33] proposed a Doc2Vec representation of users and items retrieved from the text descriptions for recommending mobile apps. In [27], the authors applied the Word2Vec algorithm on non-textual features of user check-ins for generating recommendations.

According to [30], Onet is the largest Polish web news portal, visited by more than 21 million real users monthly (as of November 2017). Onet aggregates articles from a wide range of thematic sections such as news, sports, culture, economy, etc. In order to address the diverse interests of the portal's readers, the goal is to understand their preferences in the context of articles that are relevant to them. Based on their behavior in this service, we aim to define the semantic features that are most descriptive of user interests by extracting information from the article contents. Due to the dynamical characteristic of the news content, we focus on short-term profile representations and collect user events over the past 14 days. To evaluate different approaches to user

profiling that incorporate unsupervised techniques, we define two auxiliary tasks to perform a qualitative analysis and quantitative evaluation with supervised methods.

1.2. NLP techniques and related work for Polish language

We apply two distinct approaches to the numerical representation of the text: topic modeling with the LDA algorithm [5] and word embeddings with the Word2Vec algorithm [23]. A comparison of LDA topic modeling and Word2Vec embeddings for the task of document category classification is presented in [15]. We also compare the performances of both approaches in a qualitative way on similar article retrieval as well as in the context of describing user interests on the supervised task of gender prediction.

In the summary of [32], the authors pointed out that the performance of the prediction task depends on the language used: models built for the Spanish language resulted in better performance than those for English. We are interested in articles published in Polish, and we address the problem of representing text models for this language and analyzing the classifier performance depending on the model's parameters.

As numerical text models have gained much attention from researchers recently, many tools and pre-trained models have become available. However, most of these resources are designed for English: there are few public tools available for the Polish language. A 100D vector of Polish words trained with the skip-gram Word2Vec model has been published by [18], while [25] built a model trained for the Polish language from various resources – Polish Wikipedia and National Polish Language Corpora NKJP [31]. The authors evaluated models built with different parameters and vector dimensions on the tasks of analogy and synonym retrieval and concluded that the choice of the model depends on the task. We reuse their models to compare to the models trained on our custom corpus with the same parameters.

1.3. Demography prediction

In this research, we concentrate on user gender prediction as an auxiliary task to evaluate the effectiveness of different approaches to modeling user interests. We motivate this choice by the fact that gender diversifies user behavior; it constitutes a well-defined binary classification problem with practical applications. Moreover, when using interpretable textual features, it is possible to perform a qualitative evaluation by comparing male and female features to the sociological knowledge based on user studies.

A study [11] indicated that a user's browsing history is related to demographic factors such as age, gender, occupation, etc. Among these features, gender is considered to be one of the most important factors that diversify user browsing behavior. A large-scale study on Polish Internet users [29] showed significant differences in the type of content consumed by female and male users – in November of 2017, more than 700,000 women used services related to celebrities as compared to just over 500,000 for men.

A summary of the recent efforts in the context of gender prediction considering various aspects of this problem was presented in [14]. The author found that gender influences the type of content the user browses and that the content of a website (among other factors) is useful information for predicting a user's gender.

Among the diverse solutions to the demography prediction task, two distinct approaches may be distinguished – behavioral (based on clickstream data) [6, 11, 20] and contextual (based on content features) [16, 17, 28]. In [28], the authors compared the results for different feature combinations and found that a combination of features gives significantly better results than using distinct types of features separately, while topic modeling for browsed content gives the best performance among individual features. In our work, we also evaluate topic models as descriptive features for gender-related interests and compare the results with other text representations. For the purpose of this research we focus only on text-extracted features; however, we plan to incorporate other aspects of user behavior in our future work.

For the gender prediction task, several classification algorithms were applied, such as logistic regression [14,17,20], SVMs [11], and random forests [6]. In [14], the authors decided to use logistic regression, as their goal is to find the explanatory variables that are most helpful for prediction rather than to receive the best performing model. We took a similar approach in order to find the topics that diversify the interests for male and female users. Similar to [16], we consider textual features extracted from distinct sections of the content (text and title).

Demography prediction has also been explored in the context of other domains such as e-commerce [9] and authorship profiling (predicting features from user-generated texts). An overview of the Author Profiling Task at PAN 2016 with the objective of gender and age prediction in the cross-genre perspective was presented in [32]. Among the frameworks applied to demography prediction, there are both topic modeling approaches and Word2Vec representations. Moreover, in [2, 3], the authors use Word2Vec-based profiles from social media texts for predicting a user's age and recommendations. We also apply profiles based on Word2Vec word representations and compare their results with the topic modeling approach for predicting user gender. Another study [12] addressed the problem of gender attribution for the CommonCrawl Web corpus for Polish. The authors also analyzed sociological insights from the highest gender-imbalance words. We perform a similar study for a qualitative evaluation of gender-related topics based on user browsing history.

1.4. Paper structure

The paper is organized as follows. In Section 2, we describe various approaches to the numerical representation of text such as word embeddings and topic modeling. Section 3 presents a comparison of the tested approaches in two experimental tasks. First, the methods are compared with respect to the quality of the semantic relationships that they represent for similar article retrieval. Then, the textual features are combined to represent user profiles based on the information extracted from the

browsing history of articles. User profiles are evaluated on the well-defined task of gender prediction, as this demographic feature has been found to be related to user interests. We evaluate the performance of the models and analyze the textual features, describing their interests that are most descriptive of the classification task. Finally, Section 4 summarizes the conclusions and mentions future research directions.

2. Approach

In this section, we give a brief description of the corpus, training data, and models used for the experiments.

2.1. Data description

Our corpus consists of more than 500,000 article texts that cover a wide range of topics (news, sports, culture, cars, economy, etc.) from Polish web news service Onet. We used articles with pageviews for a 14-day period in November of 2017.

As found in [3], for predicting demographic data with Word2Vec-based user profiles, models trained with a custom text dataset yield better results than those using a large general language corpus. Considering also the dynamic nature of news content and changing user interests, we verify this claim and use a Word2Vec model trained on a corpora consisting of Polish Wikipedia and National Corpus of Polish Language NKJP [25] to compare with our models built on a much smaller custom corpus with the same model parameters. As the text preprocessing tools used by [3] are different from ours, we compare models trained on all word forms. The corpora descriptions are presented in Table 1.

Table 1

Text corpora sizes

Corpus	Sentences	Tokens	Unique tokens	Unique lemmas
Polish Wikipedia	12,000,000	184,000,000	3,000,000	2,600,000
NKJP	107,000,000	1,482,000,000	9,200,000	8,400,000
Onet articles	2,000,000	19,000,000	570,000	490,000

For the user profiling task, the training data is their 14-day browsing history (article pageviews) of a sample of more than 100,000 anonymous registered users who provided information about their gender and agreed to the service terms of use (including the processing of the demographic data for marketing and statistical purposes) and accepted our cookie policy. The train labels are encoded gender classes (0 for women and 1 for men). Table 2 shows the class counts. Figure 1 shows the distribution of browsed articles per user in the training set. Since most users read a small fraction of the documents (more than 70% of the users with fewer than 10 article views over 14 days), the click events were very sparse.

Table 2
Class counts for gender classification task

Class	Label	Count	Total [%]
F	0	49,084	47
M	1	54,435	53
Total		103,519	100

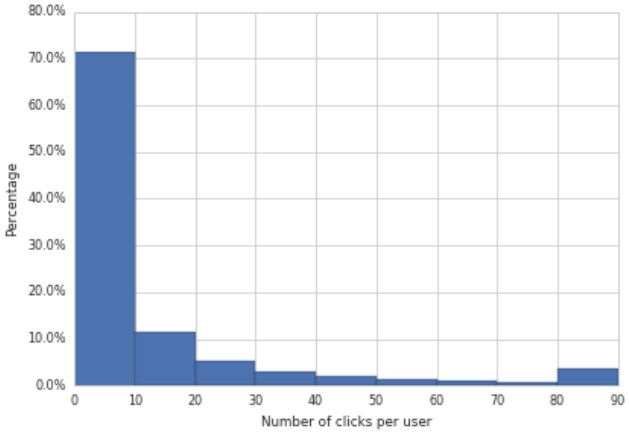

Figure 1. Distribution of clicks per user in training set

2.2. Preprocessing

In the text-preprocessing step, we performed word tokenization with a Polish alpha character regexp tokenizer and stopword filtering (by removing words with more than 10,000 occurrences). Then, in the lemmatization step, we used Polish inflection dictionary SJP [1] and the PyDic python library [8] to retrieve the word base forms. The preprocessing steps are applied to all article texts for training LDA_article, LDA_title, wv_article and wv_title models, while model wv_article_forms is trained on tokenized text with all word forms to be compared with wv_wiki_nkjp_forms.

2.3. Text representation

Here, we describe two distinct algorithms for the numerical text representation used in our experiments – topic modeling with the LDA algorithm [5] and distributional semantics with Word2Vec [23].

Topic modeling aims to represent articles and words as a distribution over a latent set of topics, while distributional semantics is based on the idea that the meaning of a word can be defined by its context. Below, we summarize both approaches.

[1] https://sjp.pl

2.3.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [5] is a generative statistical model of a corpus that defines the representation of documents as a mixture of abstract latent topics that are characterized by the distribution over observed words (assuming Dirichlet priori).

The process of generating word and document topic probability assignments is as follows:

First draw:

- distribution of topics per document $\theta_d \sim Dirichlet(\alpha)$,
- distribution of words per topic $\beta_k \sim Dirichlet(\eta)$.

Then, each word i in document d draws topic index $z_{d_i} \in 1, ..., K$ from topic weights $z_{d_i} \sim \theta_d$ and draws observed word w_{d_i} from selected topic $w_{d_i} \sim \beta_{z_{d_i}}$

We use variational Bayes LDA implementation with online update method [13] and batch size 512 and train it on the article text corpus.

For our experiments, we used the topic distribution for each article as a vector representation of the articles. Hence, article d is represented by vector $\theta_d = z_{d_1}, z_{d_2}, \ldots, z_{d_K}$, where z_{d_i} is the probability of topic i for article d and K is the number of topics. A useful characteristics of this model is its transparency – each topic can be described by a set of words with the highest probabilities. We use these characteristics to analyze the most descriptive features in the context of user interest.

2.3.2. Word2Vec

Word2Vec is a neural network model proposed by [23] that calculates distributed vector representations of words such that similar words are close in the resulting vector space.

We used a skip-gram version of the model where the objective is to maximize the log-linear word classification based on other words in the neighborhood defined with a window size (we use a window with a size of 5). Mathematically, for sequence of words $w_1, w_2, \ldots, w_T$ and window size k, the objective is to maximize the average log-likelihood: $\frac{1}{T} \sum_{t=1}^{T} \sum_{j=-k}^{j=k} \log p(w_{t+j}|w_t)$

We used an implementation with a hierarchical softmax model for computing word probability, which is more efficient than the standard softmax.

We compared our results with a model trained by [25] built on Polish Wikipedia and NKJP [31] corpora. The model is a skip-gram Word2Vec architecture with hierarchical softmax with a vector size of 100.

For calculating embeddings of the article text, we averaged the vector representations of its words. As explained in [24], the representations from the skip-gram model exhibit linear characteristics relevant to its objective function. This allows for the construction of a meaningful phrase vector representation by adding word vectors element-wise. We used an average of word vectors that is independent of the text length (however, it also does not consider word order).

The vector text representation of document d for the Word2Vec model is defined as follows:

$$\bar{v}_d = \frac{1}{|W|} \sum_{w \in w} v_w$$

where $|W|$ is the number of words in a text, w is a word from the model vocabulary, W is a list of the words in the text, and v_w is the vector representation of word w.

We only compared representations built on all tokens from the article content and its title.

2.4. User profile

We constructed user profiles by averaging the vectors of the articles in their browsing history. Thus, the resulting user profile from LDA representations describes the average user's interest in each of the latent topics, while the Word2Vec profile is constructed analogically to the article representation in Section 2.3.2 – assuming the linear characteristics of the skip-gram model [23], the average article vector also approximates the user's semantic interests. A similar approach was proposed for age prediction with Word2Vec-based profiles for social networks in [3].

We used the average of the vectors rather than the sum, as it provides feature normalization in the context of user activity (the vectors represent user interests independently of how many articles they read).

Formally, the user profile is constructed as follows:

$$\bar{v}_u = \frac{1}{|D_u|} \sum_{d \in D_u} v_d$$

,

where D_u is the set of articles that the user read (based on the browsing history), d is an article from the database, and v_d is the vector representation of article d (as defined in 2.3).

3. Experiments

Since the main goal of the presented research is to evaluate the usefulness of different user profile representations in the context of their interest in the news content, we defined two auxiliary tasks described in 3.1 to compare the proposed approaches. Table 3 presents the model configuration that we used for our experiments.

Table 3
Experiment model configuration

Model name	Train data	Preprocessing	Model	Transformed data
LDA_article	article text	stopwords, lemmatization	LDA	article text
LDA_title	article text	stopwords, lemmatization	LDA	article title
wv_article	article text	stopwords, lemmatization	Word2Vec	article text
wv_title	article text	stopwords, lemmatization	Word2Vec	article title
wv_article_forms	article text	stopwords	Word2Vec	article text
wv_wiki_nkjp_-forms	Wiki PL, NKJP	stopwords	Word2Vec	article text

The research questions that are addressed in this work are defined below (Q1–Q4).

(Q1) vector size: *What is the optimum size of the vector representation for the user profiles?* Since the dimensionality of feature space impacts the type of semantic information that it represents (as well as the model performance), the choice of vector size is a task-specific problem. We analyze the results for T1 and T2 for the dimensionsionalities of 10-1000 for both the Word2Vec and LDA representations.

(Q2) feature representation: *Which feature representation is more suitable – Word2Vec or LDA?* Both text representation techniques proved to be effective methods for diverse tasks in the domains of the NLP and recommender systems. LDA provides a self-explanatory topic word distribution, while Word2Vec models the semantic correlations among the texts and may be trained on an external language corpus. However, it is not obvious which representation is more suitable for representing user interest profiles. We evaluate the results of T1 and T2 for both representations.

(Q3) corpus: *For the Word2Vec features, is it better to use a large external corpus or a custom model?* The goal is to verify if using a pre-trained Word2Vec representation on a large external corpus results in a better performance than a much smaller custom text collection. We compare the results for the model published by [25] trained on a large corpus of the Polish language with a custom representation trained on our corpus with the same model parameters.

(Q4) title vs. content: *Is the representation of a title a sufficient source of semantic information for the user profile? Does it improve the model performance when used in combination with article contents?* We compare the results for T1 and T2 for feature representations from both the article and title as well as a concatenation of both.

3.1. Evaluation tasks

We perform a qualitative ad quantitative evaluation of the proposed methods on the following tasks (T1, T2):

(T1) similar article retrieval: First, we perform a qualitative analysis on the task of similar article retrieval and compare the quality of the outcomes on an exemplary set of articles for each of the proposed feature representations.

(T2) gender classification: Secondly, we define a supervised task of gender classification for evaluating the predictive performance of the proposed methods with quantitative classification metrics. We also perform a qualitative analysis of the resulting features comparing to a sociological knowledge related to gender-specific interests.

The configuration details and results for both tasks are described in subsequent sections 3.2 and 3.3.

3.2. Similar article retrieval

We explored the characteristics of the numerical text representation of our corpus by evaluating its performance on similar article retrieval as an auxiliary task for analyzing the impact of particular parameters on the text representations.

We used the nearest neighbor algorithm with cosine similarity metrics for finding the most similar articles.

In the following sections, we present some illustrative examples of retrieved article similarities for different feature configurations.

3.2.1. Impact of vector size on semantic correlations

First, we analyzed the impact of the vector size on the similarity between words and articles. We retrieved similar words and articles for sample items from the corpora for varying vector sizes of the Word2Vec representation. Table 4 presents a sample of the results for word similarity, and Table 5 presents the results for article similarity.

We observed that a smaller vector size implies a more general correlation, while larger vectors work on lower levels of abstraction. As shown in Table 4, the synonyms for the brand name "Toyota" in 10-dimensional Word2Vec space are other brands, while in 100 dimensions, there are particular models of Toyotas (which are hyponyms rather than synonyms). The same observation applies for an article comparison with varying embedding spaces – for short vectors, more general collocations are retrieved, while for larger dimensional spaces, the similarities are very direct. For instance, in the first example in Table 5 for an article about a Peugeot model, a vector size of 10 finds articles about different cars, while a vector size of 1000 yields similar models of Peugeots.

Table 4

Similar words for different vector sizes for Word2Vec model. Smaller vectors yield more-general correlations, while for larger vectors, words are more closely related (hyponym level)

	Vector size			
	10		100	
	word	similarity	word	similarity
Toyota	mazda	0.993	corolla	0.858
	nissan	0.991	avensis	0.858
	vw	0.991	c-hr	0.856
	volkswagen	0.990	prius	0.855
	golf	0.989	auris	0.849

Table 5

Similar articles for different vector sizes for Word2Vec model; smaller vectors yield more-general correlations, while for larger vectors, articles are more closely related

	Peugeot 3008 – nowy SUV z polskimi cenami		
	w2v_article_10	w2v_article_100	w2v_article_1000
1	Tani Opel Crossland X – ale dopiero za rok	Nowa Kia cee'd MY2016 za 59,9 tys. zł	Nowy Peugeot 5008 za 99,9 tys. zł (polskie ceny)
2	SsangYong Tivoli – koreański hit	Tani Opel Crossland X – ale dopiero za rok	Peugeot 308 1.2 — zmiany na lepsze
3	Skoda Karoq – takich cen nie spodziewał się nikt	Ile w Polsce kosztuje Peugeot 508?	Peugeot 3008 2.0 BlueHDi 150 S&S – francuski amant
4	Luksusowa Skoda Superb Combi L&K za 155,6 tys. zł	Peugeot 3008 2.0 BlueHDi 150 S&S – francuski amant	Peugeot 2008 – crossover po liftingu
	Najczęściej kradzione samochody w 2015 roku w Polsce		
	w2v_article_10	w2v_article_100	w2v_article_1000
1	Uwierzysz? To jest nowy Polonez!	10 najczęściej kradzionych samochodów w Polsce	10 najczęściej kradzionych samochodów w Polsce
2	Zwykłe samochody papieża Franciszka	Jakich aut używanych szukają Polacy?	Jakich aut używanych szukają Polacy?
3	Skoda Favorit miała być... Zaporożcem.	Masz taki samochód? Uważaj!	Masz taki samochód? Uważaj!
4	FSO Polonez z serialu "07 zgłoś się"	Sprzedaż aut w 2016 r. – znamy już liderów i najpopularniejsze modele	Samochody roku 1987 – czyli, Voyage, Voyage

3.2.2. Finding similar articles – model comparison

Another task addressed by this experiment was the comparison of retrieved similarities for the diverse corpora and text representations. We compared the results for

models with 100-dimensional vectors. Below, we summarize the observations for the illustrative examples presented in Table 9 in appendix 4 and present general conclusions based on a larger dataset analysis.

Example 1 Article about Blue Lagoon in Iceland

- Interesting consequences of words ambiguity in collocation "Błękitna Laguna" (Blue Lagoon) – some models retrieve correlations to "blue" as a color (wv_wiki_nkjp_forms_titles_100, wv_titles_100) and "Laguna" as model of Renault car (LDA_title_100, wv_wiki_nkjp_forms_titles_100).
- The content retrieved with LDA is not closely related to the queried article since the topic representation is not descriptive of the article's main theme.

Example 2 Article about the ten biggest mistakes in decorating an apartment

- Word2Vec models for title focus mostly on the phrase "ten biggest mistakes" rather than apartment decoration.
- LDA for title retrieves some unrelated content and articles related to renting apartments.
- Other models find articles about apartment decoration.

Example 3 Article about a mountain lake in Kazakhstan

- LDA models retrieve articles related to travel in general.
- Word2Vec models find articles about other lakes around the world.

Table 10 in appendix 4 presents additional results for a corpus of 100,000 articles published over a 3-month period for the Word2Vec model trained with 50-element vectors. Such a text representation yields similar articles that are close semantically, even though they do not have words in common.

3.2.3. Observations

We observed that, for the task of retrieving similarities among articles, the Word2Vec models generally perform better and capture more semantic correlations, while LDA is restricted to comparing only the set of latent topics derived from the corpora (which is not always sufficient). Word2Vec usually succeeds in finding meaningful correlations (even for limited words from the title), while LDA often fails to retrieve relevant similarities when applied only to the title.

Moreover, models that only transform title words are more sensitive to word ambiguity. When using the whole article text, there are more words related to the main theme, and important words may appear several times; hence, the average word vector is more descriptive of the true meaning. This is significant, particularly for the baseline Word2Vec model trained on the external corpora that contains words in other contexts – when using full-article texts, the results are comparable to the same model trained with a custom dataset; but for title words only, it often fails to find the proper word meaning. However, the correlation seems more general than for the models trained with custom corpora.

The feature vector size correlates with the level of abstraction on which the semantic correlations are compared. When using a small vector size (ten elements), the similarities are found on a general level, and related words may be used as synonyms (in an example for a car brand, it retrieves material about other car brands), while for larger vectors (more than 100 dimensions), the retrieved relationships are hyponyms (particular models of this brand). Hence, for the purpose of finding material with very close semantic relationships, one should consider a larger vector size (more than 100 elements), while smaller vectors may find more nontrivial correlations providing the serendipity factor for content-based recommendations.

The Word2Vec model applied to terms from article text and vectors of a size of 50 seems to retrieve meaningful yet non-obvious semantic similarities among the articles (as presented in Table 10).

3.3. Gender classification

In the second experiment, we evaluated the effectiveness of user profiles constructed from different text representations on the gender classification task. We used the user profiles described in 2.4 for the models listed in Table 3. As a simple baseline for comparing features, we used the features constructed from ten manually annotated categories based on article url hosts such as sports, cars, celebrities, tourism, etc. The training data is described in Section 2.1. We used ten-fold cross-validation and the Wilcoxon signed-rank test for validating the models as proposed in [7].

As our primary goal was to evaluate the feature representations of user profiles rather than finding the best performing model, we used a linear classifier – logistic regression with l2 regularization and feature standardization. We compared the accuracy, precision, and recall of the models for both classes.

First, we analyzed the correlation of the text representation vector size on the classification performance for different models built with the Word2Vec and LDA representations.

As the classes were nearly equal (Table 2), we used accuracy as the main classification metric. We also measured the precision and recall for both classes (Table 7).

Next, we analyzed the most important descriptive variables from the logistic regression classifier trained on the LDA features for both the negative (female) and positive (male) feature weights. We explored the most important topics for both classes described by words with the highest probabilities in the LDA representation for their meaningfulness and descriptiveness of user interests compared to the sociological knowledge.

3.3.1. Evaluating number of features in model

We analyzed the correlation of the feature vector size and the predictive performance for vectors of 10, 20, 50, 100, 200, 500, 800, and 1000 elements for the Word2Vec and LDA models. The feature vector is constructed from the article text or the title.

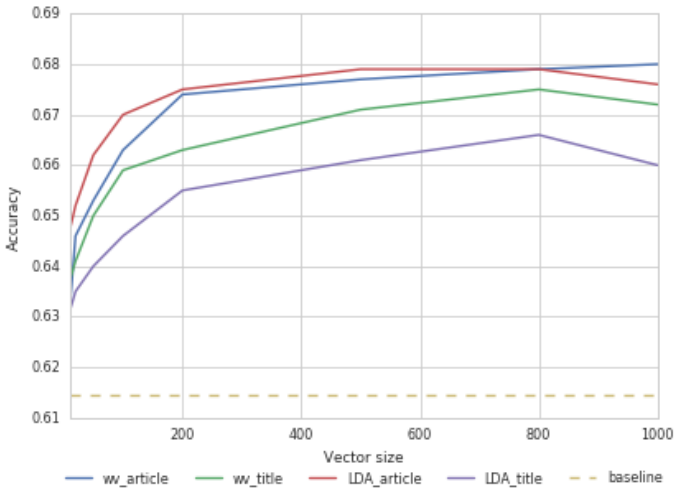

Figure 2. Comparison of model accuracy for varying feature vector sizes

As presented in Figure 2 and Table 6, the classification accuracy is positively correlated with the size of the feature vector for N<=800. However, for vector sizes with more than 200 features, the performance improvement is small. Considering the increased complexity for larger vectors, the use of around 200 vectors seems to be an optimum solution for this problem.

Table 6

Accuracy for different feature vector sizes on gender classification task with logistic regression classifier

	10	20	50	100	200	500	800	1000
wv_article	0.631	0.646	0.653	0.663	0.674	0.677	0.679	0.680
wv_title	0.637	0.641	0.650	0.659	0.663	0.671	0.675	0.672
LDA_article	0.647	0.652	0.662	0.670	0.675	0.679	0.679	0.676
LDA_title	0.631	0.635	0.640	0.646	0.655	0.661	0.666	0.660
baseline	0.614	0.614	0.614	0.614	0.614	0.614	0.614	0.614

3.3.2. Evaluating feature models

Another aspect of the analysis was a comparison of the performance of different models for the classification task. We compared the performance of the logistic regression classifiers for the feature combinations listed in Table 3 with vectors of 100 dimensions. As a simple baseline for feature representation, we used ten manually annotated categories represented as a binary vector for each user (one if the user read an article from the category). Table 7 and Figure 3 present the results for all models.

The results for classifying with different feature representations are presented in Table 7. Figure 3 shows the distribution of the accuracy results for different feature sets for the ten-fold cross-validation.

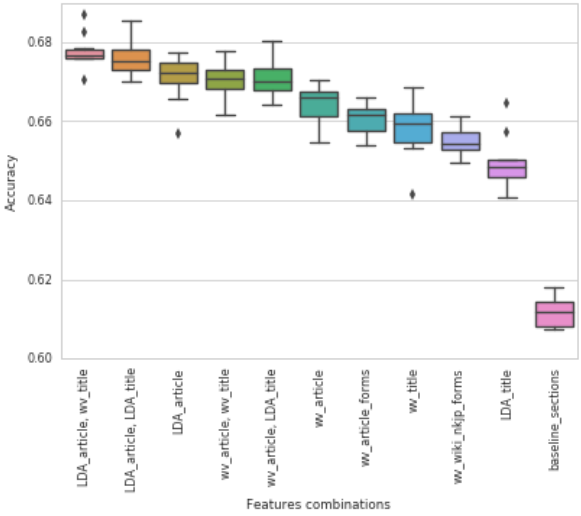

Figure 3. Comparison of feature combinations performance for gender prediction task with ten-fold cross validation

Table 7

Classification metrics for feature combinations on gender classification task with logistic regression

model_name	accuracy	prec0	prec1	rec0	rec1
LDA_article, wv_title	0.678	0.664	0.690	0.653	0.700
LDA_article, LDA_title	0.676	0.660	0.691	0.659	0.692
LDA_article	0.671	0.653	0.687	0.651	0.689
wv_article, LDA_title	0.671	0.656	0.684	0.648	0.691
wv_article, wv_title	0.670	0.653	0.685	0.653	0.685
wv_article	0.664	0.643	0.684	0.654	0.674
wv_article_forms	0.661	0.644	0.675	0.633	0.686
wv_title	0.658	0.643	0.671	0.626	0.688
wv_wiki_nkjp_forms	0.655	0.633	0.676	0.649	0.661
LDA_title	0.650	0.630	0.667	0.630	0.668
baseline_sections	0.612	0.585	0.640	0.632	0.594

All of the feature combinations based on text representations yield a significantly higher performance than the baseline model.

The LDA representation gives a significantly higher accuracy as compared to Word2Vec when using the whole article text (an accuracy of 0.671 for LDA and 0.664 for Word2Vec), but Word2Vec gives a better result for titles only (Word2Vec – 0.658, and LDA – 0.65).

The models transforming the article texts significantly outperform the title-only models (an accuracy of 0.658 title vs. 0.664 article for Word2Vec; 0.65 title vs. 0.671 article for LDA), but combining both the article and title features yields a statistically significant improvement over both models.

The best performing feature combination is LDA for the article text combined with Word2Vec from the title (0.678), showing that these two approaches capture distinct characteristics of user semantic preferences. The model resulted in a 0.066-higher accuracy than the baseline.

The Word2Vec representation with all word forms yields a significantly lower performance than the the one based on lemmas, but it gives a better result than the external corpus (0.661 vs. 0.655).

Another observation is that, for all of the feature representations, the precision and recall for the positive class (men) are higher that for the negative class (women), meaning that the male features are more characteristic. Weight distributions for the LDA features presented in Figure 4 support this hypothesis.

The conclusion is that, when long texts are available for building user profiles, LDA gives a slightly better performance, and concatenating the information from the titles improves the results. However, when only short texts are provided (such as titles), using Word2Vec for profile construction is a better solution and provides satisfactory results. Moreover, the experimental results demonstrate that feature representations based on the custom corpus outperform the ones trained on a much larger external corpus. However, if the custom dataset is too small, Word2Vec trained on a larger external corpus gives a significantly better performance than the baseline.

3.3.3. Descriptive feature analysis

In this section, we explore the features that are most descriptive of user gender prediction for differing vector sizes.

An advantage of using Latent Dirichlet Allocation as a representation of the semantics of user preferences is its transparency for inspecting words assigned to particular topics. We used the logistic regression model for gender classification, as it provides a straightforward interpretation of explanatory variables with associated weights.

We observe in Figure 4 that, for a lower number of topics, the distribution of gender-related topics is skewed – there are fewer male topics (positive weights), and female interests (negative weights) are more widely spread. For $N = 20$, all of the sports-related topics are in one cluster; for a higher N, they are separated into sub-categories.

Table 8 shows that the LDA model resulted in some high-quality textual descriptions of the topics that capture user interests. The most important female topics

(among 500 features) relate to children, fashion, women's rights, diets, and cosmetics, while male topics concern mostly sports and the military. Figure 4 shows that, for more-general topics (N = 20), the most distinctive features for women relate to celebrities (which correspond to the results from the Polish Internet user study in [29]).

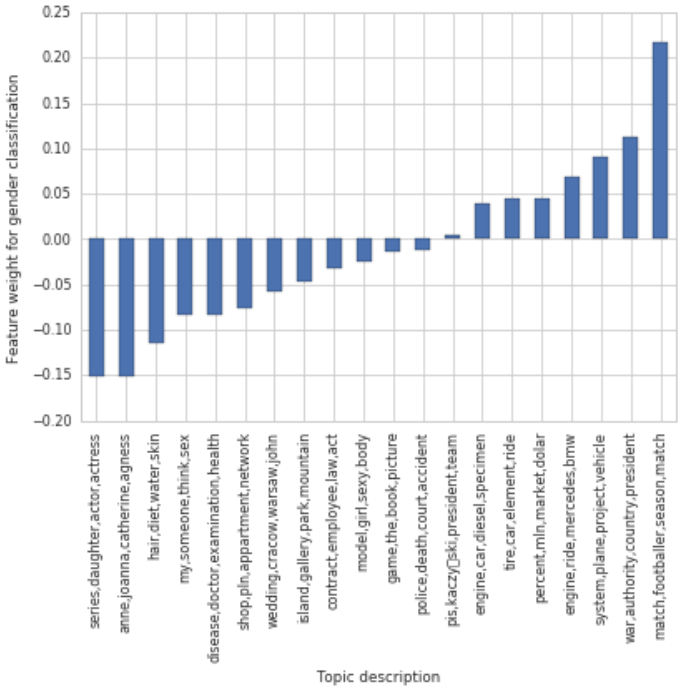

Figure 4. Model weights for LDA features (positive weights for male, negative for female features). Words in topics descriptions are translated from Polish

Table 8

Top 15 important features for gender classification (positive weights for male, negative for female features); Polish words from the model and English translations

Female topics PL	Female topics ENG	Weight	Male topics PL	Male topics ENG	Weight
dziecko maluch zabawa urodzić pociecha	child baby play birth kid	–0.041	Polak czerwoni biały Dawid Szwecja	Pole red white David Sweden	0.031
nosić spodni ubranie spódnica modny	wear trousers cloth skirt trendy	–0.032	żołnierz wojskowy wojska siła armia NATO	soldier military army power army NATO	0.031

Table 8 (cont.)

Female topics PL	Female topics ENG	Weight	Male topics PL	Male topics ENG	Weight
kobieta kobiecy siła prawa aborcja młoda	woman female power rights abortion young	–0.032	klub kontrakt umowa podpisać klub transfer	club contract sign club transfer	0.0304
jeść jedzenie posiłek śniadanie zdrowa kaloria	eat food meal breakfast healthy calories	–0.029	najlepszy wykonanie interwencja zagranie	best execution intervention play	0.026
miłość Tomek prawdziwy kochać romantyczny	affection Tom real love romantic	–0.027	bramki zdobyć trafienie poprzeczka	goal win hit crossbar	0.026
makijaż usta kosmetyk krok rzęsa powieka podkład	makeup lips cosmetic step eyelash eyelid	–0.026	piłka siatka futbolówka nożny odbić uderzyć	ball net football foot bounce hit	0.026
dom mieszkać pokój wrócić domowy mieszko biały	house live room come_back domestic live white	–0.025	gola strzelić trafienie napastnik asysta	goal hit attacker assist	0.0250
Katarzyna Janowski siła gość Zielińsk	Catherine Janowski power guest zielińsk	–0.025	macierewicz MON antoni szef obrona BBN	Macierewicz Ministry of Defense Antoni boss defense National Security Bureau	0.025
Paulina Młynarski dziennikarka Agata Polsat	Paulina Młynarski journalist Agata Polsat	–0.025	Kubica Williams Renault Bolid Formuła powrót	Kubica Williams Renault Bolid Formula come_back	0.024
fryzura warkocz krótki galeria bob długi trend	hairstyle braid short gallery bob long trend	–0.024	walka pojedynek ring waga walczyć pięściarz	fight duel ring weight fight boxer	0.023
córka ojciec tata córeczka moja spacer urodzić	daughter father dad daughter my walk give_birth	–0.024	zwycięstwo pokonać wygrana trzeci odnieść	victory defeat win third achieve	0.023

Table 8 (cont.)

sklep zakup klient ikea internetowy promocja	shop buy client Ikea Internet sale	–0.023	Barcelona Hiszpański Messi Camp Nou Katalonia	Barcelona Spanish Messi Camp Nou Catalonia	0.019
gwiazda taniec Cejrowski znany największy	star dance Cejrowski popular biggest	–0.023	Legia Warszawa Magiera malarz Astana Jacek	Legia Warsaw Magiera painter Astana Jacek	0.018
rodzic szczepienie noworodek tata rodzicielski	give_birth vaccination new_born dad parental	–0.022	euro milion strefa zapłacić transfer Neymara	euro million zone pay transfer Neymar	0.018
ciąża poród urodzić ciężarna brzuch ciążowy	pregnancy childbirth give_birth pregnant belly	–0.022	KSW Dublin gala Lipski MMA Mariusz Norman	KSW Dublin gala Lipski MMA Mariusz Norman	0.018

4. Conclusions and future research

We compared two distinct approaches to numerical text representation and constructing user profiles based on the articles they read. We used Latent Dirichlet Allocation for topic modeling and a Word2Vec neural network for defining the embedding of words based on their semantics and averaged word vectors as a representation of the text from the whole article. We trained both representations on our dataset consisting of more than 500,000 news articles in Polish and compared their performance with the Word2Vec model trained on an external corpus and a simple baseline with a binary vector of ten manually assigned categories.

We tested the performance of each model on two different tasks – the retrieval of similar documents, and the prediction of user gender based on their browsing history. We evaluated the results for varying sizes of numerical text representation vectors, considering features extracted from the whole article texts and only from their titles.

In both experiments, the Word2Vec models outperformed LDA for short or incomplete texts such as article titles. We also observed that the size of the feature vector has a strong impact on the final results.

For the gender classification task, all of the proposed models outperformed the baseline; the best results were obtained for a feature combination from the LDA article terms combined with Word2Vec on the article titles: we noticed that the title introduces some extra semantic information. The optimum size of the feature vectors for both models is close to 200 dimensions: larger feature vectors have a small impact on the performance.

We also explored the explanatory variables for the gender classification task and discovered that the most important features from the LDA topics provide high-quality descriptions of user segments.

In future research, we plan to explore other representations of document semantics as well as different approaches to aggregating the user interest profiles that are more descriptive than average vectors. We also plan to evaluate the user profiles in other experimental tasks such as content personalization and automatic recommendations.

Appendix A: Results for retrieving similar articles

Table 9

Top four similar articles retrieved by distinct text representations for three exemplary articles

Example 1	Example 2	Example 3
Islandia – Błękitna Laguna (Blue Lagoon)	10 największych błędów, jakie popełniamy, urządzając mieszkanie	Niezwykłe jezioro Kaindy (Kajyngdy) w Kazachstanie – podwodny las w górach Tien-Szan
LDA_article_100		
Kosmos	Dlaczego hotele używają białej pościeli?	USA – Hawaje – Molokini
Farma	Jak dobrać dywan do wnętrza?	Nieodkryty klejnot Alaski
Australia	Styl rustykalny od podstaw	Jemen – Sokotra – wyspy dziwolągów
Zagrożenie powodziowe w Polsce	Na jaki kolor pomalować sypialnię?	Najpiękniejsze miejsca na Ziemi
LDA_title_100		
Oto 8 najmniejszych miast Polski. Ich historie są ciekawsze niż ci się wydaje	Likus, znaczy luksus	10 cudownych górskich miasteczek w Europie
Największa, sztuczna laguna na świecie powstanie w Dubaju	Nikt nie chce drewna z wiatrołomów	Dinant – jedno z najpiękniejszych miast Belgii przyklejone do skał nad Mozą
Renault Koleos – od teraz także w Europie	Chcesz wynająć komuś mieszkanie? Nie daj się oszukać	Huacachina – oaza na pustyni Atacama niesamowitą atrakcją Peru
Mentawaje – szamani z indonezyjskiej dżungli	10 przykazań wynajmującego. Musisz to wiedzieć zanim wynajmiesz komuś mieszkanie	Belgrad na weekend: atrakcje i przewodnik po stolicy Serbii

Table 9 (cont.)

Example 1	Example 2	Example 3
w2v_article_100		
10 najpiękniejszych lagun na świecie	Sypialnia	Klejnot Gór Kaskadowych – jedno z najpiękniejszych jezior w USA
Laguna Colorada – niezwykłe jezioro w Boliwii	Co zamiast płytek w kuchni – najnowsze trendy	23 mało znane miejsca na niezapomniane wakacje
Różowa laguna w Las Coloradas – rajski zakątek Meksyku	Pięknie urządzone 65-metrowe mieszkanie w Kielcach – idealne dla rodziny	Najpiękniejsze miejsca na świecie – skarby natury wg Lonely Planet
Najlepsze naturalne SPA na świecie	Fajnie urządzone mieszkanie w domu z lat 70-tych. To wnętrze do mieszkania, a nie tylko do oglądania!	Jezioro Guerlédan – największe jezioro Bretanii odkrywa swoje tajemnice
w2v_title_100		
Laguna Colorada – niezwykłe jezioro w Boliwii	1 na 6 Polaków popełnia ten błąd. Tak można wysadzić dom!	Grüner See w Austrii – lazurowe jezioro z fantastycznym podwodnym krajobrazem
Różowa laguna w Las Coloradas – rajski zakątek Meksyku	5 największych błędów, które mężczyźni popełniają w swoim ubiorze	Jeziora Plitwickie, wodospady Krka i góry Welebit czyli inne oblicze Chorwacji
Blue Valentine	Jeśli chcesz być bogaty, nie popełniaj tych 7 błędów	Curon – wioska zatopiona w alpejskim jeziorze
Damajagua na Dominikanie – 27 wodospadów, z których można zjeżdżać do lagun	Cztery najczęstsze błędy w urządzaniu łazienek	Jaskinia Melissani na wyspie Kefalonia w Grecji – jaskinia z jeziorem, po którym pływają łodzie
wv_wiki_njp_articles_100		
10 najpiękniejszych lagun na świecie	Polacy starzeją się najszybciej w Europie. Zainteresowanie architekturą dla seniorów rośnie	Gdzie na żagle i kajaki w Śląskim – aktywnie nad jeziorem
Najlepsze naturalne SPA na świecie	Siła argumentu w sprawie czystości	Najpopularniejsze kąpieliska w polskich miastach
Jachty bogaczy. Tak się pływa w luksusie.	Sypialnia zgodna z nauką feng shui	Krajobrazy jak z Chorwacji. W Polsce mamy Lazurowe Jezioro
Najlepsze plaże nudystów. Top 10 najpiękniejszych plaż	Jak twoja mała kuchnia może stać się miejscem o dużych możliwościach, które pokochasz?	Najpopularniejsze kąpieliska w polskich miastach

Table 9 (cont.)

Example 1	Example 2	Example 3
wiki_njp_forms_titles_100		
Czarno to widzę #2. Recenzje: Big K.R.I.T., Igorilla, Moses Sumney, Sonar i Nai Palm	5 największych błędów, które mężczyźni popełniają w swoim ubiorze	Laguna Colorada – niezwykłe jezioro w Boliwii
Nowy Seat Ibiza 1.0 TSI – Leon w miniaturze	Jeśli chcesz być bogaty, nie popełniaj tych 7 błędów	Lokalna egzotyka: Jezioro Płaskie i Jezioro Krejwelanek
Ford F-350 Super Duty King Ranch: amerykański sen	Chcesz utrzymać się w pracy? Nigdy nie popełniaj tych błędów	Klejnot Gór Kaskadowych – jedno z najpiękniejszych jezior w USA
Bikini Blue	1 na 6 Polaków popełnia ten błąd. Tak można wysadzić dom!	Grüner See w Austrii – lazurowe jezioro z fantastycznym podwodnym krajobrazem

Table 10

Exemplary results for retrieving similar articles with Word2Vec model (vector size 50) with TF-IDF article term filtering

title	similar_1	similar_2	similar_3	similar_4
Twórca „Moonlight" szykuje nowy film	„Egzorcysta" od 19 czerwca tylko na FOX	Nowe plotki o Star Wars od studia Visceral	Nie żyje reżyser i zdobywca Oscara kultowego filmu „Rocky"	„Detroit": nowy spot promujący film Kathryn Bigelow
Sezon na kabriolety. Przegląd rynku	Duże kombi za dychę	Autobusy na wakacje w cenie około 30 tys. zł	Pomysł na tanie ściganie – 5 aut dobrych do KJS-ów	Mini hatch 5D – miejska zabawka za nierozsądne pieniądze
Jak fajnie wykorzystać sklejkę w aranżacji wnętrza? Zobaczcie mieszkanie Beaty!	Dwupokojowe mieszkanie w Krakowie – tak w stylu skandynawskim można urządzić 50 metrów	Otoczony brzozami 160--metrowy dom w Rembertowie. Piękny!	Trzy najczęstsze błędy w urządzaniu przedpokoju i pięć wzorowych projektów z polskich domów	Jak szybko i tanio odmienić mieszkanie – pięć pomysłów od profesjonalistów

Table 10 (cont.)

Dani Ceballos skazany na Real Madryt, Barcelonę skreślił wiele lat temu	Dani Alves prosi o transfer. Juventus potwierdza	Oficjalnie: Sandro Ramirez piłkarzem Evertonu	Higuain pospieszył się z pożegnaniem Alvesa	Wielki klub chce Teodorczyka! Potężna oferta
Balans praca – życie jest możliwy. I opłacalny	Życie jest elastyczne, ale czy twoja praca też?	10 sposobów by zadbać o pracownika	Kariera na miarę osobowości	7 cech najgorszego pracodawcy

References

[1] Ahn J., Brusilovsky P., Grady J., He D., Syn S.Y.: Open User Profiles for Adaptive News Systems: Help or Harm? In: *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 11–20, ACM, New York, USA, 2007. http://dx.doi.org/10.1145/1242572.1242575.

[2] Alekseev A., Nikolenko S.I.: Predicting the age of social network users from user-generated texts with word embeddings. In: *Proceedings of 5th IEEE Artificial Intelligence and Natural Language Conference (AINL)*, St. Petersburg, pp. 3–13, 2016.

[3] Alekseev A., Nikolenko S.I.: Word Embeddings of User Profiling in Online Social Networks, *Computación y Sistemas*, vol. 21(2), pp. 203–226, 2017.

[4] Bai X., Barla Cambazoglu B., Gullo F., Mantrach A., Silvestri F.: Exploiting Search History of Users for News Personalization, *Information Sciences*, vol. 385–386, pp. 125–137, 2017. http://dx.doi.org/10.1016/j.ins.2016.12.038.

[5] Blei D.M., Ng A.Y., Jordan M.I.: Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, vol. 3(1), pp. 993–1022, 2003. http://dl.acm.org/citation.cfm?id=944919.944937.

[6] De Bock K., Van den Poel D.: Predicting Website Audience Demographics for-Web Advertising Targeting Using Multi-Website Clickstream Data, *Fundamenta Informaticae*, vol. 98(1), pp. 49–70, 2010. http://dx.doi.org/10.3233/FI-2010-216.

[7] Demšar J.: Statistical Comparisons of Classifiers over Multiple Data Sets, *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. http://dl.acm.org/citation.cfm?id=1248547.1248548.

[8] Dorosz K.: PyDic – Single dictionary API, 2013. https://pydic.readthedocs.io/en/latest/index.html.

[9] Duc D.T., Son P.B., Hanh T., Thien L.T.: A Resamping Approach for Customer Gender Prediction Based on E-Commerce Data, *Journal of Science and Technology: Issue on Information and Communications Technology*, vol. 3(1), pp. 76–81, 2017. http://jst.udn.vn/ict/index.php/jst/article/view/40.

[10] Gauch S., Speretta M., Chandramouli A., Micarelli A.: *User Profiles for Personalized Information Access*, pp. 54–89. Springer, Berlin–Heidelberg, 2007. http://dx.doi.org/10.1007/978-3-540-72079-9_2.

[11] Goel S., Hofman J.M., Sirer M.I.: Who Does What on the Web: A Large-Scale Study of Browsing Behavior. In: *ICWSM*. 2012.

[12] Graliński F., Borchmann Ł., Wierzchoń P.: "He Said She Said" – a Male/Female Corpus of Polish. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France, 2016.

[13] Hoffman M., Bach F.R., Blei D.M.: Online Learning for Latent Dirichlet Allocation. In: Lafferty J.D., Williams C.K.I., Shawe-Taylor J., Zemel R.S., Culotta A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 856–864. Curran Associates, Inc., 2010. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

[14] Ivanova E.: *Predicting website audience demographics based on browsing history*. G2 pro gradu, diplomityö, Aalto University School of Business, 2013. http://urn.fi/URN:NBN:fi:aalto-201403171565.

[15] Jędrzejowicz J., Zakrzewska M.: *Word Embeddings Versus LDA for Topic Assignment in Documents*, pp. 357–366. Springer International Publishing, Cham, 2017. http://dx.doi.org/10.1007/978-3-319-67077-5_34.

[16] Kabbur S., Han E.H., Karypis G.: Content-Based Methods for Predicting Web-Site Demographic Attributes. In: *2010 IEEE International Conference on Data Mining*, pp. 863–868. 2010. http://dx.doi.org/10.1109/ICDM.2010.97.

[17] Kim I.: *Predicting Audience Demographics of Web Sites Using Local Cues*. David Eccles School of Business, University of Utah, 2011. https://books.google.pl/books?id=jxxxMwEACAAJ.

[18] Kędzia P., Czachor G., Piasecki M., Kocoń J.: Vector representations of polish words (Word2Vec method), 2016. http://hdl.handle.net/11321/327. CLARIN-PL digital repository.

[19] Kompan M., Bieliková M.: *Content-Based News Recommendation*, pp. 61–72, Springer, Berlin–Heidelberg, 2010. http://dx.doi.org/10.1007/978-3-642-15208-5_6.

[20] Liu J., Dolan P., Pedersen E.R.: Personalized News Recommendation Based on Click Behavior. In: *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pp. 31–40. ACM, New York, NY, USA, 2010. http://dx.doi.org/10.1145/1719970.1719976.

[21] Lu Z., Dou Z., Lian J., Xie X., Yang Q.: Content-based collaborative filtering for news topic recommendation. In: *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 217–223, 2015.

[22] Luostarinen T., Kohonen O.: Using Topic Models in Content-Based News Recommender Systems. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings*, No. 085, pp. 239–251. Linköping University Electronic Press, 2013.

[23] Mikolov T., Chen K., Corrado G., Dean J.: Efficient Estimation of Word Representations in Vector Space. In: *CoRR*, vol. abs/1301.3781, 2013. http://arxiv.org/abs/1301.3781.

[24] Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.: Distributed Representations of Words and Phrases and their Compositionality. In: Burges C.J.C., Bottou L., M. Welling, Ghahramani Z., Weinberger K.Q. (eds.), *Advances in Neural Information Processing Systems*, 26, pp. 3111–3119, Curran Associates, Inc., 2013. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[25] Mykowiecka A., Marciniak M., Rychlik P.: Testing word embeddings for Polish. In: *Cognitive Studies | Études cognitives*, vol. 17, 2017. https://doi.org/10.11649/cs.1468.

[26] Özgöbek Ö., Gulla J.A., Erdur R.C.: A Survey on Challenges and Methods in News Recommendation. In: *WEBIST (2)*, pp. 278–285. 2014.

[27] Ozsoy M.G.: From Word Embeddings to Item Recommendation. In: *CoRR*, vol. abs/1601.01356, 2016. http://arxiv.org/abs/1601.01356.

[28] Phuong D.V., Phuong T.M.: *Gender Prediction Using Browsing History*, pp. 271–283, Springer International Publishing, Cham, 2014. http://dx.doi.org/10.1007/978-3-319-02741-8_24.

[29] Polscy internauci w listopadzie 2017. http://pbi.org.pl/raporty/polscy-internauci-listopadzie-2017.

[30] Polski internet w listopadzie 2017. http://pbi.org.pl/badanie-gemius-pbi/polscy-internauci-listopadzie-2017.

[31] Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B.: *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa, 2012.

[32] Rangel F., Rosso P., Verhoeven B., Daelemans W., Potthast M., Stein B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings*, CLEF and CEUR-WS.org, Évora, Portugal, 2016.

[33] Stiebellehner S., Wang J., Yuan S.: Learning Continuous User Representations through Hybrid Filtering with doc2vec. In: *CoRR*, vol. abs/1801.00215, 2018. http://arxiv.org/abs/1801.00215.

[34] Wang C., Blei D.M.: Collaborative topic modeling for recommending scientific articles. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 448–456. ACM, New York, NY, USA, 2011. http://dx.doi.org/10.1145/2020408.2020480.

[35] Webb G.I., Pazzani M.J., Billsus D.: Machine Learning for User Modeling, *User Modeling and User-Adapted Interaction*, vol. 11(1-2), pp. 19–29, 2001. http://dx.doi.org/10.1023/A:1011117102175.

Affiliations

Joanna Misztal-Radecka
Grupa Onet-RAS Polska, ORCID ID: https://orcid.org/0000-0002-2959-4004

Received: 15.01.2018
Revised: 18.05.2018
Accepted: 20.05.2018