Topic Modelling of the Russian Corpus of Pikabu Posts: Author-Topic Distribution and Topic Labelling

Olga Mitrofanova[1[0000-0002-3008-5514]], Veronika Sampetova[2],
Ivan Mamaev[1[0000-0003-3362-9131]], Anna Moskvina[1], Kirill Sukharev[3]

[1] St. Petersburg State University, Russia,
199034, Saint-Petersburg, Universitetskaya emb. 11,
[2] Speech Technology Center, Russia, 194044, Saint-Petersburg, Vyborgskaya emb, 45E,
[3] Saint-Petersburg Electrotechnical University, Russia,
197376, Saint-Petersburg, ul. Professora Popova, 5
o.mitrofanova@spbu.ru, nikasampetova@gmail.com,
st079541@student.spbu.ru, st017154@student.spbu.ru,
sukharevkirill@gmail.com

Abstract. The paper discusses development of a corpus of Russian posts with hash tags based on Pikabu social network. We developed a balanced and representative corpus as regards the impact of certain authors, the amount and size of their posts. Our study is aimed at the development of probabilistic topic models revealing the authors' interests and preferences, as well as correlation of topics within the corpus. We performed a series of experiments including standard LDA topic modelling and Author-Topic modelling. In course of topic modelling we used algorithms from Python libraries. Experiments allowed to extract groups of authors with similar and related interests. We used topic label assignment based on manually introduced hash tags and labels automatically extracted from the lexical database RuWordNet. That facilitates linguistic interpretation of results.

Keywords: Social Networks, Pikabu, Russian, Topic Modelling, LDA, ATM, Topic Label Assignment

Introduction

Information environment gradually penetrated into our daily life, and the growth of network devices gave rise to a peculiar virtual world with its own rules of digital discourse. Communication within the virtual world is governed by technical equipment of «speakers» and «listeners», the texts created by the digital discourse exhibit the features of various types and forms of speech, the roles of «speakers» and «listeners» turn out to be diversified, communication in itself becomes spectacular, it requires reinforcement by visual content. Therefore, the study of digital discourse should combine methods of cognitive linguistics, content analysis, computational linguistics, sociology and adjacent fields of knowledge.

At present the attention of computational linguists and sociologists is focused on multilevel analysis of social media texts, the core tasks to be solved in empirical stud-

ies being author profiling [1, 2, 3] and topical analysis of online communities [4, 5, 6]. These tasks require corpora collection from web-sources and software elaboration. Proper linguistic processing of social media corpora opens wide opportunities for studies of social opinion and Russian web discourse, cf. recent publications of E. Koltsova and colleagues, LINIS HSE [4] and SCILA [5]; T. Litvinova and colleagues, RusProfiling Lab [6]; S. Bodrunova, I. Blekanov and colleagues, Web-Metrics Research group, SPbSU [7], etc.

Our present study is devoted to the creation and processing of the social media corpus containing posts of various authors from the social network Pikabu [8] which has not been properly investigated. Pikabu is a Russian language community founded in 2009 which is considered as an elaborated analogue to Reddit [9]. The rise of attention to Pikabu was caused by the famous meme Zhdun which flooded social media in 2017. The attractiveness of Pikabu as a source of linguistic data is explained by the medium size of posts (they are not so brief as in Twitter) and by the abundance of the users' hash tags indicating the subject matter of the posts.

Most corpora developed for Russian social networks use Twitter, LiveJournal, Facebook, VKontakte as sources of textual data, e.g., Taiga social network segment [10], GICR [11], Twitter sentiment corpus [12] and the like. Such social media corpora are used for elaboration and evaluation for NLP algorithms, models and tools, cf. Dialogue Evaluation Competition [13]. Social media provide a huge platform for studying topics, opinions, discourse structure of web-communication, that requires corpus-based linguistic resources: lexical databases, sentiment lexicons, formal ontologies, etc. Nowadays predictive distributed word representations are in great demand for text classification, collocation and construction analysis, that's why research community welcomes access to word2vec embedding models pretrained on various text corpora, and social media corpora as well (cf. RusVectōrēs [14]). In the previous work we described and evaluated word2vec models for Pikabu [15] which prove to be useful for further studies of the given source.

In course of experiments we process a newly developed dataset of the Russian Pikabu corpus by means of state-of-the-art algorithms and NLP tools. That gives us the opportunity to form a baseline for further elaboration of our methodology. For the first time we carry out experiments on author-topic modelling of the corpus and obtain data on thematic coherence of posts and on authors' covert clusters. The novelty of our study consists in thorough linguistic interpretation of topic models strengthened by topic label assignment which takes into account hash tags introduced by the authors as well as labels automatically extracted from RuWordNet [16] lexical database. Thereby, our research fills in the gaps existing in contemporary Russian corpus linguistics and social media analysis.

1 Topic Modelling

In recent years we witness the rise of interest in the development and application of topic modelling as a research procedure for data mining and content analysis [17, 18, 19]. In fact, topic modelling is a variety of fuzzy clustering performed for words

and documents, latent semantic relations within a corpus in this case being described in terms of a family of probability distributions over a set of topics [20, 21, 22].

Early versions of topic models were based on algebraic transformations, e.g. classical Vector Space Model (VSM) and Latent Semantic Analysis (LSA), which take into account term-document distribution, term co-occurrence frequency, and may be expanded with dimensionality reduction techniques, e.g. Singular Value Decomposition (SVM). Gradually these models gave way to probabilistic topic models, the most notable of them being Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Expectation-Maximization (EM) Algorithm, etc.

Probabilistic topic models are based on the assumptions that ordering of words within documents and documents within a corpus may be ignored; frequent and rare words do not affect the quality of the topic model; a topic t should be considered as a discrete distribution over a set of words w, and a separate document d as a discrete distribution over a set of topics t; the occurrence of words in a document d is determined by a particular distribution $p(w|t)$.

In our study we use a topic model which is based on Latent Dirichlet Allocation, our choice is explained by the advantages of LDA compared with previously developed methods, as well as its availability in a set of libraries, including gensim [23] and scikit-learn [24] for Python.

Topic modelling allows to build multimodal models which include metadata alongside with intrinsic features of a corpus, e.g., polylingual models for information retrieval, temporal models taking into account the time of document creation, author-topic models which include authorship parameter. The latter type of topic models satisfies the conditions of our experiments.

The Author-Topic model (ATM) [25] may be considered as a refined extension of LDA, combines a topic model reproducing relations between words, documents and topics, and an author model describing relations between documents and authors. In recent years ATM is often used in linguistic and sociological studies, especially in the tasks of user profiling (age and gender detection [26, 27]) and authorship attribution [28, 29, 30].

2 Development of a corpus of Pikabu Russian posts

2.1 Corpus collection

The corpus of Pikabu Russian posts includes texts downloaded from Pikabu social network. Collection of posts was carried out with the help of Pikabu parser adapted from [31]. The parser was developed for Python 3.7 [32] and maintained with lxml [33] and requests [34] libraries.

We improved the original parser by adding the option of arranging posts as regards their authorship. We also added an option of post filtering: deletion of non-Russian texts, images, media-content, punctuation marks, etc. Html-pages parsing was performed by means of BeautifulSoup library [35]. We parsed no less than 100 posts for each author, so preliminary selection of the most productive authors was necessary. We took into account productivity ratings [36] which were published in 2017−2018

but still preserved their actuality in 2019. The parsed posts were ordered from the latest (2019 – end of 2018) to the earliest (middle 2018 and further). Each post was saved in a *.txt file, its name containing the author's ID. Joint data about the authors being saved in a separate document. After preprocessing the corpus size turned out to be 2 161 681 tokens, the total number of texts being 3 059, maximum number of texts for a single author – 100, minimal number of texts for a single author – 16. Some of the authors fell out of the final list of Pikabu users as they posted texts in the image format, e.g. *Oblomoff* (cooking recipes in JPEG) и *IriskaVRF* (comics/pictures).

2.2 Corpus processing

Corpus processing included tokenization, lemmatization and stop-word removal was performed by means of spaCy [37] and pymorphy2 [38] libraries. We modified standard stop-word list by adding Wiktionary data [39] which was parsed by means of BeautifulSoup library [35] in order to extract interjections, pronouns, particles, prepositions, parenthetic expressions, numerals. We added proper names and pejoratives into a stop-word list. All in all, the size of a stop-word list is 1400 items. After stop-word removal the corpus size reduced to 1 144 812 tokens which constitutes about 53% of the initial corpus size.

Before topic modelling we detected bigrams and trigrams in the corpus. In most cases topic models are designed as unigram models which don't take into account regular syntagmatic relations in contexts. At the same time such models may fail to reflect lexical constructions (collocations and idioms) which are broken into separate lemmata, the content of the whole phrase being lost [40, 41]. That's why we came to a conclusion that in the process of topic model development bigrams and trigrams with frequency more than 20 should be added into the dictionary of the model. In our case retrieved n-grams turned out to be frequent functional set expressions, e.g. *любой_случай, всякий_случай 'any_case'*, etc., that's why only a few of them occurred among top 10 topic words in the output.

We also compressed the dictionary by omitting high- and low-frequency items, so that the final size of the corpus turned out to be 8320 tokens in 3059 documents of 39 authors.

2.3 Hash tag analysis

Alongside with posts we parsed users' hash tags which could facilitate linguistic interpretation of the topics generated by our models. We selected three most frequent hash tags for each author, e.g. *видео 'video'*, *мое 'my'*, *длинностекст 'long-text'*, *длиннопост 'long-post'*, etc.; hash tags duplicating users' names: *varlamov, mtd, goodmix*, etc.

While processing frequent hash tags we joined semantically correlated hash tags which could possibly introduce a single topic, e.g.: *авторский рассказ 'author story'* → *рассказ 'story'; строительная история 'building story'* → *строительство 'building'; кулинария, рецепт 'cooking, recipe'*→ *кулинария 'cooking'*, etc. In cases of low interpretability of hash tags we borrowed topic labels from the community titles: e.g. the user *Region89* [42] uses the hash tag *bash im* as the most frequent one,

instead of it we labeled his posts by group names *Истории из жизни 'Life stories'*, *Лига диетологов 'League of nutritionists'*, etc. In some cases hash tags admitted generalization: e.g. country → continent: *США, Канада 'USA, Canada'* → *Северная Америка 'North America'*; *Бразилия, Латинская Америка 'Brazil, Latin America'* → *Латинская Америка 'Latin America'*; *Уганда, Руанда 'Uganda, Rwanda'* → *Африка 'Africa'*, etc.: hyponym → hypernym: *андроид 'android'*, *ios* → *телефон* '*telephone*'; *супергерой, комиксы 'superhero, comics'* → *комиксы 'comics'*, etc. The given hash tag transformations were necessary for evasion of false diversity of topics which could complicate data analysis. Resulting correspondencies are illustrated in Table 1.

Table 1. Examples of authors' hash tags

Author	Number of posts	Hash tags
MadTillDead	100	*Рассказ, офис 'Story, office'*
smile2	100	*История, жизнь 'Story, life'*
CreepyStory	100	*Крипота, страшные истории 'Creep, horror stories'*
AnnrR	96	*Строительство, работа 'Building, work'*
dr. Doctor	22	*Женский бред, форум 'Women's raving, forum'*
Brahmanden	100	*Кулинария, Одесса 'Cooking, Odessa'*
alekseev77	70	*Автосервис 'Car service'*
DoctorLobanov	100	*Медицина, рассказ, война 'Medicine, story, war'*
findeler	95	*Записки строителя, люди 'Builder's notes, people'*
upitko	100	*Фильмы 'Films'*

3 Topic Modelling of a corpus of Pikabu Russian posts

3.1 LDA topic model

LDA topic models were developed for the whole corpus and for its subcorpora. We used gensim library [23] to train the models and pyLDAvis library [43] for Python to visualize topic distributions. The procedure includes 3 stages:

1) LDA parameter choice and model training;
2) evaluation of topic coherence with UMass-measure [44];
3) topic visualization.

For LDA development we split the corpus to segments including representing posts of particular authors. Parameter choice was performed taking into account the variety of author's hash tags which envisages a set of expected topics, and UMass measure as it reflects the level of topic coherence which is treated as a level of human interpretability of the model based on relatedness of words and documents within a topic:

$$\mathrm{UMass}\,(v_i\,,\,v_j) = \log \frac{D(v_i\,,\,v_j) + 1}{D(v_j)}, \qquad (1)$$

where $(v_i\,,\,v_j)$ corresponds to the number of documents containing words v_i and v_j, while $D(v_j)$ shows the number of documents contains v_j [44].

UMass measure was calculated for each author's subcorpus, thus, the most appropriate number of topics in LDA models for different authors was established ad hoc (ranging from 3 to 30). In all cases the topic size was settled as 10 lemmata per topic. Visualization allows to view the most significant lemmata characterizing author's subcorpora and word-topic distributions. Below we present topic distributions and their visualizations for posts of random authors *BadVadim*, *CometovArt*, *FrancoDictator* and *Griffel* (cf. Fig. 1 – 4).

Username ID: BadVadim; hash tags: Оружие 'Weapon'

Topic 1: *оружие, нож, холодное, бабочка, тема, ножевой, видео, ждать, сложный, признать… 'weapon, knife, cold, butterfly, topic, blade, video, wait, complex, recognize…'*

Topic 2: *нож, клинок, оружие, холодное, длина, рукоять, изделие, обух, должный, бабочка… 'knife, blade, weapon, cold, length, handle, product, butt, necessary, butterfly…'*

Topic 3: *нож, пункт, ГОСТ, клинок, рукоять, общий, технический, шкуросъёмный, условие, разделочный… 'knife, point, GOST, blade, handle, general, technical, skinning, condition, cutting…'*

Topic 4: *нож, самооборона, средство, применение, оружие, метр, эффект, случай, дистанция, шокер… 'knife, self-defence, means, application, weapon, metre, effect, case, distance, shocker…'*

Topic 5: *нож, видео, оружие, Россия, сегодня, тема, минута, полиция, сообщество, пневматика … 'knife, video, weapon, Russia, today, topic, minute, police, society, pneumatics…'*

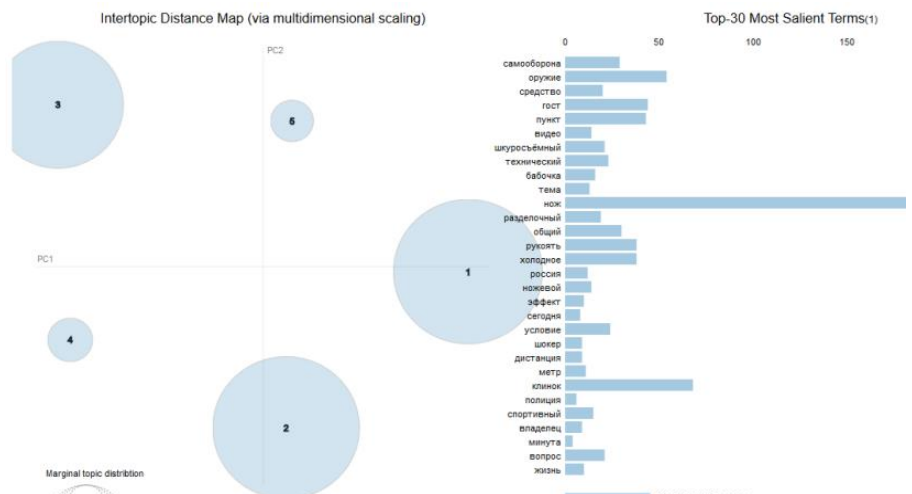

Fig. 1. User *BadVadim*: topic distribution

Username ID: CometovArt; hash tags: Игры 'Plays'

Topic 1: *игра, ролик, фракция, парад, террана, делать, карта, являться, нормальный, хороший… 'play, roller, fraction, parade, terran, make, card, be, normal, good…'*

Topic 2: *проект, скорость, работа, Бог, проблема, древние, Зот, идея, эфир, хороший… 'project, speed, work, God, problem, ancient, Thot, idea, airing, good…'*
Topic 3: *сделать, игра, сезон, серия, момент, играть, трейлер, хороший, увидеть, делать… 'make, play, season, episode, moment, play, trailer, good, see, make…'*
Topic 4: *день, проект, колода, получить, эльф, система, игра, работа, праздник, подборка… 'day, project, block, get, elf, system, play, work, holiday, collection…'*
Topic 5: *орда, альянс, сила, фракция, победить, герой, игра, убить, объединить, зачистить… 'horde, alliance, force, fraction, win, hero, play, kill, join, clean…'*

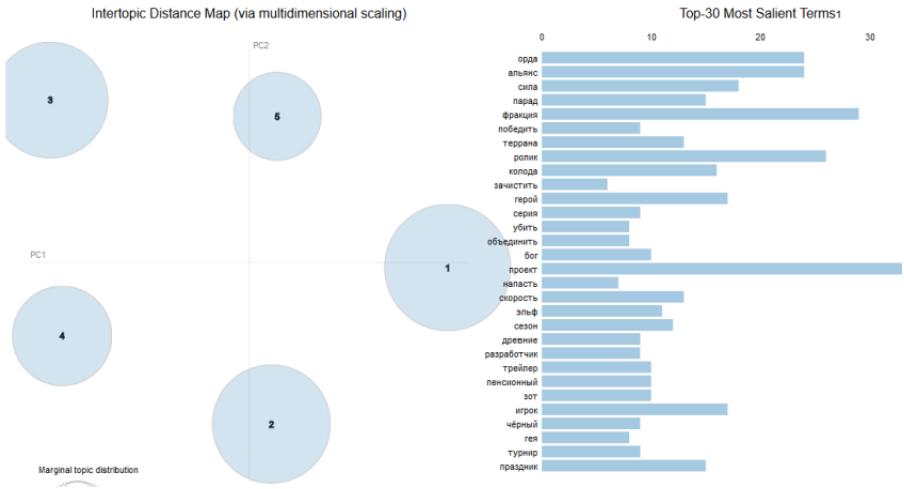

Fig. 2. User *CometovArt*: topic distribution

Username ID: FrancoDictador; hash tags: Политика, Испания 'Politics, Spain'
Topic 1: *страна, беженец, город, евро, Мадрид, Испания, статус, привет, Франко, момент… 'country, refugee, city, Euro, Madrid, Spain, status, salutation, Franco, moment…'*
Topic 2: *страна, Россия, Украина, власть, митинг, посмотреть, действующий, государство, понять, следующий… 'country, Russia, Ukraine, power, meeting, look, active, state, understand, next…'*
Topic 3: *клиент, адвокат, бизнес, суд, процесс, работать, страна, Испания, сделать, право… 'client, lawyer, business, court, process, work, country, Spain, make, law…'*
Topic 4: *деньга, рынок, банка, экономика, банк, государство, кот, дать, валюта, иметь… 'money, market, pot, economy, bank, state, cat, give, currency, have…'*
Topic 5: *страна, дать, Испания, Россия, беженец, деньга, клиент, Швейцария, момент, Франко… 'country, give, Spain, Russia, refugee, money, client, Switzerland, moment, Franco…'*

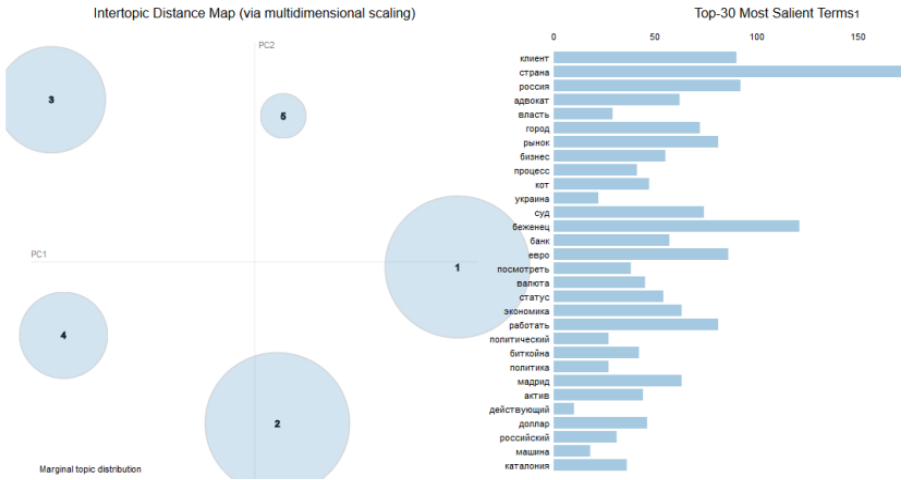

Fig. 3. User *FrancoDictator*: topic distribution

Username ID: Griffel; hash tags: Северная Америка 'North America'

Topic 1: *Канада, страна, химиотрасса, большой, компания, работа, США, находиться, фторид, фтор… 'Canada, country, chemtrail, big, company, work, USA, be, fluoride, fluor…'*

Topic 2: *Канада, США, мир, миллион, два, стать, Торонто, право, арестованный, место… 'Canada, USA, world, million, two, become, Toronto, law, arrested, place…'*

Topic 3: *кофе, Канада, язык, рубль, США, канадский, девушка, слово, Америка, страна… 'coffee, Canada, language, ruble, USA, Canadian, girl, word, America, country…'*

Topic 4: *Канада, русский, канадец, жизнь, жить, страна, уехать, Торонто, гусь, женщина… 'Canada, russian, Canadian, life, live, country, leave, Toronto, goose, woman…'*

Topic 5: *код, валюта, дать, рубль, балл, том, Россия, машина, метр, русский… 'code, currency, give, ruble, score, volume, Russia, car, metre, russian…'*

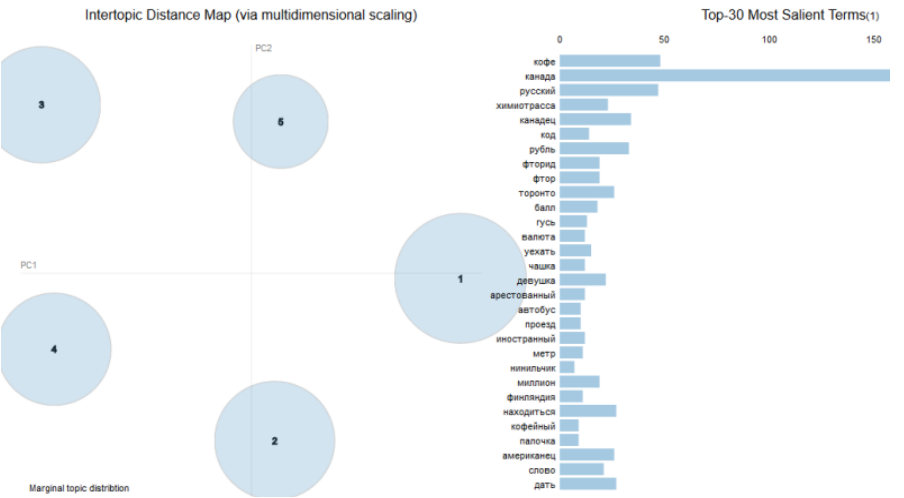

Fig. 4. User *Griffel*: topic distribution

Subcorpora preparation was strengthened by stylometric analysis performed with JGAAP toolkit [45]. Data processing gives evidence in favour of stylometric parameter diversity between subcorpora and their unity within sets of documents written by particular authors.

3.2 ATM Topic Model

ATM was built by means of gensim library [23]. The procedure includes 4 stages:
1) ATM parameter choice and model training;
2) evaluation of topic coherence with UMass-measure;
3) authors' posts similarity estimation by means of Hellinger distance [46];
4) modelling of authors' clusters as regards similarity of their posts.

A series of experiments was carried out to choose the appropriate number of topics for the whole corpus, this number was changed from 10 up to 40 with step 5, the constant size of topics being 10 lemmata. UMass measure was used to define the best experimental settings. The highest UMass values corresponded to topic modelling with 25, 30 and 40 topics. We took into account lowest scores of lemmata repetitions between topics, that were characteristic of the model with 30 topics, this value was selected as the best for the tasks of our study. For each author we selected the most relevant topics generated by ATM and matched them with hash tags. After ATM constructions we evaluated similarity of authors' posts by Hellinger distance [46] which estimates the distance between probability distributions describing topic variety for the authors, thus, author clusters were formed within our model. Examples of such clusters for users *BadVadim*, *CometovArt*, *FrancoDictator* and *Griffel* are given in Table 2.

Author clusters proved to be consistently interpretable with respect to the content of their posts. The user *Griffel* [47] is a participant of the society «Pikabu Users of North America», his associates within a cluster being immigrants and/or travelers, e.g. *Varlamov.ru*, a well-known blogger writing on urbanistics and adventures [48]; the user *goodmix* is a businessman and sauna proprietor, and his associates turn out to be builders and repairmen (*AnnrR*, *alekseev77*, *Scrypto*), etc.

Table 2. Author clusters for users *BadVadim*, *CometovArt*, *FrancoDictator* and *Griffel*

Author	Hash tags	Author cluster	Similarity measure	Hash tags
BadVadim	*Оружие 'Weapon'*	dr.Doctor	0,71	*Женский бред, видеоигры 'Women's raving, video-plays'*
		evilcame	0,64	*Телефон, игры 'Telephone, plays'*
		alekseev77	0,63	*Автосервис 'Car service'*
		AlexGyver	0,60	*Своими руками 'With my hands'*
		goodmix	0,59	*Сауна, истории 'Sauna, stories'*

CometovArt	*Игры 'Plays'*	L4rever	0,75	*Германия, Бонусы'Germany, bonuses'*
		AlexGyver	0,70	*Своими руками 'With my hands'*
		Scrypto	0,69	*Ремонт техники 'Equipment repair'*
		Varlamov.ru	0,68	*Городская среда, архитектура 'Urban environment, architecture'*
		Griffel	0,68	*Северная Америка 'North America'*
Griffel	*Северная Америка 'North America'*	Varlamov.ru	0,99	*Городская среда, архитектура 'Urban environment, architecture'*
		ShamovD	0,98	*Япония 'Japan'*
		FrancoDictador	0,97	*Политика, Испания 'Politics, Spain'*
		L4rever	0,79	*Германия, Бонусы'Germany, bonuses'*
		Esmys	0,78	*Латинская Америка, кулинария 'Latin America, cooking'*
Franco-Dictador	*Политика, Испания 'Politics, Spain'*	Varlamov.ru	0,99	*Городская среда, архитектура 'Urban environment, architecture'*
		ShamovD	0,98	*Япония 'Japan'*
		Griffel	0,97	*Северная Америка 'North America'*
		L4rever	0,79	*Германия, Бонусы'Germany, bonuses'*
		CometovArt	0,68	*Игры 'Plays'*

ATM allows to distribute users over certain groups in accordance with the major topics discussed in the posts. We will mention only a few examples: user group of travelling (*Griffel*, *ShamovD*, *FrancoDictador*, *L4rever*, *Esmys*, etc. describe different cities and countries), user group of narrators (*MadTillDead*, *smile2*, *CreepyStory*, *DoktorLobanov*, *denisslavin*, *ozymandia*, *femme.kira*, *svoemnenie*, *889900*, *Region89*, etc. write short stories), user group of builders and repairmen (*alekseev77*, *AnnrR*, *BadVadim*, *Scrypto*, *AlexGyver*, etc.).

3.3 Topic Label Assignment

There are certain approaches to the improvement of topic models. Traditional topic modelling algorithms, LDA being among them, do not include label assignment as an internal procedure.

At the same time, topic labelling allows to improve informativeness of the models. Labels are considered as single terms or phrases generalizing the topic content. From the semantic point of view, such labels are expected to be either strict hypernyms, holonyms, or at least more abstract lexical items covering the meaning of separate topic words. By default one can choose the first word of a topic as a label, but such labels often turn out to be unconvincing. Although topic words are ranked within a topic and may be related to each other by various syntagmatic and paradigmatic relations [49], such ordering is not obligatorily hierarchical. Labels can be assigned manually in course of human expertise, but in this case they may reflect subjective treatment of topics. Thus, it is necessary to find a tradeoff solution to the problem. Consequently, NLP researches proposed several ways of automatic label assignment [50, 51, 52] which differ in the source of labels (intrinsic data extracted from corpora or extrinsic information from external resources – lexical databases (WordNet, Wikipedia, etc.) or search engines (Google, Yandex, etc.). In our previous studies experiments on automatic label assignment were performed using such external resources as Russian Wikipedia distributional model accessible via ESA (Explicit Semantic Analysis) and Yandex search engine output with morphosyntactic parsing and statistical ranking [53, 54]. Taking into account stylistic peculiarities of social media (Pikabu corpus representing one of them), we expect that labels extracted from encyclopaedic texts (Wikipedia) and news headings (Yandex) may fail to match with the content of Pikabu topics. Therefore, in this study we chose a lexical database RuWordNet [16, 55] suitable for experiments with social media data of mixed stylistic character.

The procedure of label assignment implemented in our project implies a hybrid approach combining human expertise and automatic data processing, involving internal and external sources of candidate labels. On the one hand, we use manually assigned hash tags extracted from Pikabu as topic labels. As hash tags are consciously introduced by the authors, they may be considered more reliable than the first topical words. On the other hand, we extract hypernyms for topic words as candidate labels from RuWordNet. The idea to use lexical hierarchy as a source of topic labels keeps close to the task of automatic extraction of «IS-A» relations and corpus-based taxonomy enrichment. The procedure used in our study implied selection of hypernyms for each topic word which were united in a list and ranked. Both hash tags and hypernyms are ranked in accordance with ipm frequencies from A Frequency Dictionary of Contemporary Russian (based on the Russian National Corpus) by O.N.Lashevskaya and S.A.Sharoff [56].

Samples from our dataset are described in Table 3. As expected, in all four cases the users' hash tags provide the best fit. It should be noted that throughout the corpus hash tags are repeated among top 10 topical words, but not necessarily at the head part. This gives us the reasons to consider them as a solid baseline dataset in further experiments.

As regards the first topical words, in our example they are suitable as topic labels for the topics extracted from the posts of the users *Brahmanden* and *yulianovsemen*, but it is not the case as regards the users *upitko* and *Malfar*: although the words *готовить 'prepare'* и *дело 'case'* have abstract meanings, they turn out to be too ambiguous and vague for being topic labels. As for the corpus in general, the set of top one /

three topical words is rather heterogeneous in meaning, so we may consider them as a tentative – less reliable as hash tags – dataset for evaluation procedure.

Table 3. Examples of topic labels

Author	Topic example	Topic labels = ranked hash tags	Topic labels = ranked RuWordNet hypernyms
Brahmanden	*готовить, добавить, мясо, перец, минута, масло, соус, вкус, лук, бульон… 'cook, add, meat, pepper, sauce, taste, onion, broth…'*	*кулинария, Одесса, рецепт… 'cooking, Odessa, recipe…'*	*продукт, блюдо, кулинария… 'foodstuff, meal, cooking…'*
upitko	*фильм, режиссер, рейтинг, известный, кино, роль, кинолента, известно, история, хороший… 'film, producer, rating, famous, cinema, role, filmstrip, known, story, good…'*	*фильм, рецензия, кино… 'film, review, cinema…'*	*передача, фильм, кино… 'broadcast, film, cinema…'*
yulianovsemen	*дело, случай, сотрудник, следователь, место, райотдел, прокуратура, МВД, орган, розыск... 'case, accident, officer, detective…'*	*милиция, прокуратура, убийство… 'police, procuracy, murder…'*	*политика, право, преступление… 'politics, law, crime…'*
Malfar	*комикс, Бэтмэн, Флэш, фонарь, стать, Икс, вселенная, выпуск, читать, Марвел … 'comics, Batman, Flash, flashlight, become, X, universe, issue, read, Marvel'*	*комикс, супергерой, суперзлодей…. 'comics, superhero, supervillain…'*	*издание, печать, рубрикация… 'edition, print, rubrication…'*

Labels extracted from RuWordNet and assigned to the main topics of the users *Brahmanden* and *upitko* correspond to the content of the topics and partially intersect with hash tags on the lexical level (repetitions are marked in bold: *кулинария*, *фильм*, *рецензия 'cooking, film, review'*). That doesn't hold true for the users *yulianovsemen* and *Malfar*: RuWordNet labels turn out to be rather more general than users' hash tags and topical words (*преступление*, *право*, *издание 'crime, law, edition'*). All in all, according to our observations, RuWordNet topic labels, being semantically correlated with the topics, seem to be rather generalized in comparison with label candidates selected by other methods. In order to improve the results we upgraded the procedure of hypernym selection by using word2vec embeddings extracted from the pretrained corpus model: we assumed that possible label vectors may be similar to averaged topic vectors, but our expectations were partially fulfilled as candidate labels enhanced by word2vec data remained general by meaning. That inspires further

experiments with combination of topic labeling with distributed vector representations.

However, three types of labels constitute a scale of acceptability which is limited by the first topical words as formal labels and RuWordNet labels as the most general ones, the golden mean being the users' hash tags. The combination of expert-based and knowledge-based approaches to topic label assignment requires further quantitative analysis and evaluation, but even in the qualitative aspects it seems to be fruitful as it provides data for topic expansion and may be useful in text rubrication.

Summary

In the given study we managed to create an author-topic model for the corpus of Pikabu Russian posts.

We worked out a procedure for corpora development suitable for processing mixed data from social media: texts and metadata (author's usernames and hash tags). A multilevel analysis of our corpus was performed.

We developed standard LDA models with visualization for subcorpora containing posts of separate authors, that provides data on their interests. A complex Author-Topic model was constructed for the whole corpus, which allows to detect clusters of authors writing on similar topics.

Finally, we carried out experiments on topic label assignment, topic labels being obtained from two sources: manually assigned users' hash tags and hypernyms for topical words automatically extracted from RuWordNet lexical database.

Results achieved by now allow to expand our studies and put forward the next set of tasks: development and processing of various social media corpora, elaboration of the procedure for latent community detection, enhancement of the procedure of hypernym extraction and ranking for automatic label assignment.

Acknowledgements. The authors express their sincere gratitude to Dr. Prof. Natalia Loukachevich (Moscow State University) and colleagues for the access to RuWordNet thesaurus.

References

1. Panicheva, P., Litvinova, O., Litvinova, T.: Author Clustering with and Without Topical Features. In: Speech and Computer 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings. LNAI, vol. 11658, pp. 348-358. Springer (2019).
2. Litvinova, T., Sboev, A., Panicheva, P.: Profiling the Age of Russian Bloggers. In: Artificial Intelligence and Natural Language, 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings, issue 930, pp. 167-177. Switzerland: Springer (2018).
3. Panicheva, P., Mirzagitova, A., Ledovaya, Y.: Semantic Feature Aggregation for Gender Identification in Russian Facebook. In: Artificial Intelligence and Natural Language, 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers, issue 789, pp. 3-15. Switzerland: Springer (2018).

4. LINIS HSE, https://linis.hse.ru/, last accessed 2020/05/08.
5. SCILA, https://scila.hse.ru/, last accessed 2020/05/08.
6. RusProfiling Lab, https://rusprofilinglab.ru/, last accessed 2020/05/08.
7. WebMetrics Research group, SPbSU, http://www.apmath.spbu.ru/ru/structure/depts/tp/webometrics.html, last accessed 2020/05/08.
8. Pikabu, https://pikabu.ru/, last accessed 2020/05/08.
9. Reddit, https://www.reddit.com, last accessed 2020/05/08.
10. Taiga, https://tatianashavrina.github.io/taiga_site/, last accessed 2020/05/08.
11. GICR, http://www.webcorpora.ru/, last accessed 2020/05/08.
12. Twitter sentiment corpus, https://study.mokoron.com/, last accessed 2020/05/08.
13. Dialogue Evaluation Competition, http://www.dialog-21.ru/evaluation/, last accessed 2020/05/08.
14. RusVectōrēs, https://rusvectores.org/ru/models/, last accessed 2020/05/08.
15. Antipenko, A.A., Mitrofanova, O.A.: Comparative Study of Word Associations in Social Networks Corpora by means of Distributional Semantics Models for Russian. International Journal of Open Information Technologies, 8(1) (2020), http://www.injoit.org/index.php/j1/article/view/871/834, last accessed 2020/05/08.
16. RuWordNet, https://ruwordnet.ru/ru, last accessed 2020/05/08.
17. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022. (2003).
18. Blei, J.: Probabilistic topic models. In: Communications of the ACM, vol. 55 (4), pp. 77-84. (2012).
19. Vorontsov, K., Potapenko, A.: Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In: D.I. Ignatov, M.Y. Khachay, A. Panchenko, N. Konstantinova, R. Yavorskiy (Eds.). Analysis of Images, Social Networks and Texts. Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10–12, 2014, Revised Selected Papers, CCIS, vol. 436, pp. 29-46. Springer, Cham (2014).
20. Bodrunova, S, Blekanov, I., Kukarkin, M.: Topic modeling for Twitter discussions: Model selection and quality assessment. In: Proceedings of the 6th SGEM International Multidisciplinary Scientific Conferences on Social Sciences and Arts SGEM2018, Science and Humanities. Sofia, Bulgaria: STEF92 Technology Ltd., pp. 207-214. (2018).
21. Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., & Vorontsov, K.: Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In: G. Sidorov & O. Herrera-Alcántara (eds.), Advances in Computational Intelligence, pp. 169–184. Springer International Publishing (2017).
22. Nikolenko, S., Koltcov, S., & Koltsova, O.: Topic modelling for qualitative studies. Journal of Information Science, 43(1), 88-102. (2017).
23. gensim, https://radimrehurek.com/gensim/, last accessed 2020/05/08.
24. scikit-learn, https://scikit-learn.org/stable/index.html, last accessed 2020/05/08.
25. Rosen-Zvi, M., Thomas Griffiths, Th., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents, http://arXiv:1207.4169, last accessed 2020/05/08.
26. Rangel, F., Rosso, P.: Use of language and author profiling: identification of gender and age. In: NLPCS 2013 10th International Workshop on Natural Language Processing and Cognitive Science CIRM, Marseille, France, pp.177-185. (2013).
27. Panicheva, P., Mirzagitova, A., Ledovaya, Y.: Semantic feature aggregation for gender identification in Russian Facebook. In: A. Filchenkov, L. Pivovarova, J. Žižka (Eds.) Artificial Intelligence and Natural Language, 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers, pp. 3-15. Springer (2017).

28. Argamon, Sh., Koppel, M., Pennebaker, J.W., Schler, A.J.: Automatically profiling the author of an anonymous textю Communications of the ACM – Inspiring Women in Computing, 52 (2), 119–123. (2009).
29. Seroussi, Y., Bohnert, F., Zukerman, I.: Authorship Attribution with Author-aware Topic Models. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, vol. 2, Short Papers, pp. 264-269. Jeju Island, Korea (2012).
30. Zhang, H., Nie, P., Wen, Y., Yuan, X.: Authorship Attribution for Short Texts with Author-Document Topic Model. In: W. Liu, F. Giunchiglia, B. Yang (Eds.) Knowledge Science, Engineering and Management KSEM 2018, LNCS, vol. 11061, pp. 29-41. Springer, Cham (2018).
31. Pikabu parser, https://github.com/silver4one/pikabu_parser, last accessed 2020/05/08.
32. Python 3.7, https://www.python.org/, last accessed 2020/05/08.
33. lxml, https://lxml.de/, last accessed 2020/05/08.
34. requests, https://2.python-requests.org/en/master/, last accessed 2020/05/08.
35. BeautifulSoup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/, last accessed 2020/05/08.
36. Pikabu productivity ratings, https://clck.ru/GK8gP, https://clck.ru/GK8jn, https://clck.ru/GK8kL, last accessed 2020/05/08.
37. spaCy, https://spacy.io/, last accessed 2020/05/08.
38. pymorphy2, https://pymorphy2.readthedocs.io/en/latest/, last accessed 2020/05/08.
39. Wiktionary data, https://ru.wiktionary.org/wiki/Заглавная_страница, last accessed 2020/05/08.
40. Loukachevitch, N., Nokel, M., Ivanov, K.: Combining Thesaurus Knowledge and Probabilistic Topic Models. In: International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2017, pp. 59-71 (2017).
41. Sedova, A., Mitrofanova, O.: Topic Modelling of Russian Texts based on Lemmata and Lexical Constructions. In: Computational Linguistics and Digital Ontologies, vol. 1, pp. 132-144. Saint-Petersburg, ITMO (2017) [in Russian].
42. Region89, https://pikabu.ru/@Region89, last accessed 2020/05/08.
43. pyLDAvis, https://github.com/bmabey/pyLDAvis, last accessed 2020/05/08.
44. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July 12-14, 2012, pp. 952-961 (2012).
45. JGAAP, https://github.com/evllabs/JGAAP, last accessed 2020/05/08.
46. Nikulin, M.S.: Hellinger distance. In: Encyclopedia of Mathematics, Kluwer Academic Publishers (2001).
47. Griffel, https://pikabu.ru/profile/griffel, last accessed 2020/05/08.
48. Varlamov.ru, https://varlamov.ru, last accessed 2020/05/08.
49. Koltsov, S.N., Koltsova, O.Ju., Mitrofanova, O.A., Shimorina, A.S.: Interpretation of Semantic Relations in the texts of the Russian LiveJournal Segment based on LDA Topic Model. In: Proceedings of the XVIIth All-Russia Joint Conference «Internet and Modern Society» IMS–2014, Saint-Petersburg, ITMO, November 19 – 20, 2014, pp. 135-142. Saint-Petersburg, ITMO (2014) [in Russian].
50. Aletras, N., Stevenson, M., Court, R.: Labelling Topics using Unsupervised Graph-based Methods. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2, Short Papers, pp. 631-636. Baltimore, Maryland, ACL (2014).

51. Allahyari, M., Pouriyeh, S., Kochut, K., Arabnia, H.R.: A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling. International Journal of Advanced Computer Science and Applications, 8(9), 335-349 (2017).
52. Bhatia, S., Lau, J.H., Baldwin, T.: Automatic Labelling of Topics with Neural Embeddings. In: 26th COLING International Conference on Computational Linguistics, pp. 953-963 (2016).
53. Erofeeva, A., Mitrofanova, O.: Automatic assignment of topic labels in topic models for Russian text corpora. Structural and Applied Linguistics, 12, 122-147. St. Petersburg University (2019) [in Russian].
54. Kriukova, A., Erofeeva, A., Mitrofanova, O., Sukharev, K.: Explicit Semantic Analysis as a Means for Topic Labelling. In: D. Ustalov, A. Filchenkov, L. Pivovarova, J. Žižka (eds.). Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17-19, 2018, Proceedings, pp. 167-177. Springer, Cham (2018).
55. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., Dobrov, B.V.: Creating Russian WordNet by Conversion. In: Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference «Dialogue», vol. 15, pp.405-415. Moscow: RSUH (2016).
56. Lashevskaya, O.N., Sharoff, S.A.: A Frequency Dictionary of Contemporary Russian (based on the Russian National Corpus), http://dict.ruslang.ru/freq.php, last accessed 2020/05/08.