NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

*Oleg S. Nagornyy, Olessia Y. Koltsova*

# MINING MEDIA TOPICS PERCEIVED AS SOCIAL PROBLEMS BY ONLINE AUDIENCES: USE OF A DATA MINING APPROACH IN SOCIOLOGY

BASIC RESEARCH PROGRAM

WORKING PAPERS
SERIES: SOCIOLOGY
WP BRP 74/SOC/2017

*Oleg S. Nagornyy*[1], *Olessia Y. Koltsova*[2]

# MINING MEDIA TOPICS PERCEIVED AS SOCIAL PROBLEMS BY ONLINE AUDIENCES: USE OF A DATA MINING APPROACH IN SOCIOLOGY

Media audiences that represent a significant part of a county's public may hold opinions on media-generated definitions of social problems different from those of media professionals. The proliferation of user-generated content makes such opinions available, but simultaneously demands new automatic methods of analysis that media scholars still have to master. In this paper, we show how topics regarded as problematic by media consumers may be revealed and analyzed by social scientists with a combination of data mining methods. Our dataset consists of 33,877 news items and 258,121 comments from a sample of regional newspapers. With a number of new, but simple indices we find that issue salience in media texts and its popularity with audience diverge. We conclude that our approach can help communication scholars effectively detect both popular and negatively perceived topics as good proxies of social problems.

Keywords: social problem, online media, topic modeling, sentiment analysis, Russia.

JEL Classification: Z

[1]National Research University Higher School of Economics. Internet Studies Lab, Research Assistant; E-mail: onagorny@hse.ru

[2]National Research University Higher School of Economics. Internet Studies Lab, Laboratory Head. E-mail: ekoltsova@hse.ru.

# Introduction

Analysis of user-generated content has been increasingly seen as potentially useful for revealing audiences' perceptions of agendas discussed in media, in particular those perceived as social problems, and, consequently, for tracing existing or emerging social tensions. Methods of this research are still in their cradle. Not only various algorithms and software tools still need improvement, but, more importantly, they need to be adapted to the goals of media and communication studies and integrated into full-cycle "production chains" of social science knowledge. This integration is a separate goal, different both from mathematical algorithm creation, and from software development. In this paper we show how communication scholars can benefit from using data mining methods combined with qualitative manual analysis of texts for the goal of detecting social problems via user content.

We start from viewing social problem as a complex discursive phenomenon that emerges and develops in public arenas in an interplay of efforts of interest groups, media professionals, lay media and internet users and other actors. We proceed with the assumption stating that readers' comments to professional media content are the spaces where professional definitions of problems may be altered or even subverted. Our empirical analysis allows us to argue that salience of topics in professional media content and their popularity with commenting audience do diverge. We also show that topics which are both the most popular with readers and the most negatively commented are good indicators of social problems.

The rest of the paper is structured as follows. In the section below we define the concept of social problem and explain why and how it is related to readers' comments. Next, we describe how two relevant data mining approaches – topic modeling and sentiment analysis – can be used in communication research, notably for our goals. We then outline our research hypotheses and describe our data in two respective sections. We devote the subsequent section to the procedure of finding media agendas, or topics. The main empirical section presents our results concerning relative topic salience in news, their popularity with the audience, and the prevailing polarity of comments for each topic. It also contains qualitative analysis of texts associated with the most popular and most negatively commented topics. In conclusion, we summarize the results, describe limitations of our approach and outline directions for further research.

## Theories of social problems formation by the public and the media

Application of automatic big data analysis to the goals of communication research faces a difficulty connected to the nature of issues social scientists aim to capture, such as social problems or topics, which are inherently ambiguous.

Social problem can be most broadly defined as an undesirable situation or condition that characterizes a society on the whole or some of its parts (and that can be eliminated only if a collective effort is made). The major difficulty with this concept can be formulated as a question: undesirable to whom? Early approaches to social problems define them as social pathology that objectively undermines social health, and that should be diagnosed and treated (Smith, 1911). These approaches refer to medical metaphors and thus presuppose that the agents of problem detection and treating are experts competent in social science. This approach has been widely criticized on various grounds, including subjectivity: experts may be prone to mistakes, biases (Yasaveyev, 2004) and simply frauds.

Development of democratic institutions and a more relativistic vision of social issues has lead researchers to regard the public as the source of problem definition. This vision gave rise to public opinion polls aimed at detecting problems most important for the public (Lauer, 1976). Still, it was not clear how widely an issue was to be recognized as a problem to be considered as such by sociologists (Ibid). Lauer observes that most of the time the number of people was to be "considerable". Merton and Nisbet (1971) defined social problem as "a way of behavior that is regarded by a substantial part of a social order as being in violation of one or more generally accepted or approved norms". Other social scientists have tried to narrow down the public responsible for defining social problems to certain groups, such as: "issue specific groups", "strategically placed groups" or just interest groups. Spector and Kitsuse (1977: 67) give a more radical definition of social problems regarding them as "activities of individuals or groups making assertions of grievances and claims with respect to some putative conditions…"

This later approach can also be addressed with some criticism. In societies where the institution of interest groups does not exist or is suppressed, social problems have to be considered non-existent. Even in Western societies, definition of social problems based on public opinion can be problematic since public opinion itself may be manipulated, notably by the media, that can choose which issues to (de)problematize. The ever-lasting academic debate on the scale of media influence on public opinion ranges from early theories of total control (Lasswell, 1927) to modern modifications of minimal effects theory (Bennet & Iyengar 2008). From this stream of thought, of special relevance to us are several ideas: first, the claim from agenda setting theory about ability of

media to influence rather audience's attention than opinion (Cohen 1963). Second, Hall's ideas about variability of human "decoding" of media messages defined by him as dominant/hegemonic, oppositional and negotiated (Hall, 1980), as well as his reflections on the reverse dependence of media on externally set agendas. And third, of special relevance is the most radical approach that reduces the media to filters or public arenas (Hilgartner & Bosk, 1988) where social problems compete for public attention as a scare resource. This theory views social problems as results of efforts of those who promote them to stay on the public agenda reinforced by inter-media linkages and constrained by the limited "carrying capacity" of public arenas. As empirical evidence has been mixed, by now it has become generally accepted that the relation of the media and public opinion is complex, and the process of social problem definition of which both the media and the public are a part is a constantly developing societal negotiation.

The advent of the Internet has added new layers of complexity to this picture (Zhou & Moy, 2007). Nowadays, public opinion is no longer locked within the datasets collected by pollsters. It has flooded into the blogs, social network sites (SNSs) and, most importantly, to the forums of the online media, where much of the discussion of public affairs takes place. The media on their part can no longer ignore this new reality (Lee et al., 2014). They seem to be still holding the leading positions in the sphere of agenda setting since this activity demands the biggest professional effort, although even here the evidence is mixed (Wallsten, 2007; Sayre et al., 2010). Further, by commenting media messages, readers may apply oppositional decoding, thus altering news framing, and influence opinions of others (Lee & Yoon 2010; Kim 2015). As the process of social problems definition is now publicly available and documented, one can directly observe and discern problems of different "scale" that become salient in certain times at certain levels of the society: local, regional, national etc. This provides a possible answer to the question of how much public is needed for defining a social problem: *if an issue attracts highest attention from the audience of the media of the respective level and is problematized, it may be considered a social problem*. We argue that the volume and polarity of reader comments might be a good proxy for this concept. That is, we assume that if a topic or an issue receives a large number of reader comments that find it problematic, it may be defined as a social problem.

By suggesting this, we do not mean that audiences' opinions, particularly expressed in comments, necessarily mirror distribution of opinions over the general population. In fact, there is lack of studies on this matter. It is known that commenters may differ from non-commenters by their psychological (Wu & Atkin 2017) and demographic features (Stroud et al 2016). However, various earlier research of letters to the editor that had dominated reader feedback before the Internet finds that those letters roughly reflect public opinion, especially on salient issues

(McCluskey & Hmielowski 2012: 306). Simultaneously, the authors find that comments present a wider range of opinions than traditional letters, and thus should be a better proxy for public opinion. But the most important point is that even though comments, just like regular media content, may be not representative, they do play an important role in social problems definition and general public opinion formation (Henrich & Holmes 2013). In particular, critical comments have been found to shift opinions of readers away from the opinion expressed in media messages (Lee & Yoon 2010) and more generally to alter readers' opinions via influencing their perceptions of others' opinions about the messages (Kim 2015).

One potential limitation of this approach is that the volume of comments may not rigorously reflect audience's interest in an issue; instead, it may be influenced by structural features, such as news release time, style, genre, polarity and other content-independent parameters (Weber 2014; Liu et a; 2015). Polarity of comments may also depend on this, although there is no relevant research available. Therefore, it makes sense to look for the volume and polarity of comments generated directly by topics, not by individual news items. For this, a method of topic extraction from news collections is needed that we develop further below.

## Topic modeling and sentiment analysis for social problems research

A traditional way to detect so defined social problems would be to perform a standard manual content analysis (Krippendorf, 2004). In this case assessors (coders) would have to answer the following questions: (1) what social issue is a given media text devoted to? (2) is this issue problematized? (3) does this issue attract much public attention in the form of comments (4) does a given comment problematize this issue?

A severe limitation of this approach is that it does not scale to the volume of internet content. We therefore develop an approach based on a number of methods of automatic text analysis. First, we detect agendas set by the media by revealing their topical structure; second, we measure the volume of attention to each topic as expressed in the number of comments, and third, we measure the prevailing sentiment in the respective comments considering negative sentiment as the best proxy to the issue being problematized.

Recently, a lot of effort has been made to introduce automatic analysis of texts in different languages into media, journalism and communication studies (Grimmer & Stewart, 2013; Günther & Quandt, 2016). Beyond these initial methodological overviews, media scholars have started to use automatic approaches to detect topics (Flaounas et al., 2013; Scharkow, 2013) and related concepts, such as media frames in news texts (David et al., 2011; Burscher et al., 2014). However,

all these studies have been based on supervised machine learning approach where topics, frames or other categories are known beforehand. In our case, we seek to extract salient topics from the media texts, assuming we have no prior knowledge of media agendas, that is, we seek to find newly emerged and perhaps unexpected topics. Unsupervised methods, akin to cluster and factor analysis, are usually applied to such tasks. An unsupervised approach most suitable for high-dimensional data, such as texts, is a group of algorithms known as topic modeling (TM) (Blei, 2003; Steyvers & Griffiths, 2007). In social science, it has been used mostly outside media studies – e.g. for mining topics in legislative speeches (Quinn et al., 2010) or abstracts of scientific papers (Diesner & Carley, 2010). However, more recently it has been successfully applied for mining topical structure in the Russian-language blogs (Koltsova & Koltcov 2013; Koltsova & Shcherbak, 2015) and group pages in the Russian social networking site VKontakte (Voskresenskiy et al., 2015; Alexandrov et al., 2016). These algorithms ascribe each word and each text to a number of latent variables (topics) with varying probabilities and thus may be viewed as methods of fuzzy co-clustering of words and texts over topics. What follows from this is that words order in such models is largely ignored, as is polysemy, metaphorism and other text features; however, despite those simplifications TM fairly well imitates human understanding of what a topic is. Within TM, a topic is defined algorithmically as a set of most often co-occurring words, and only when such set is interpretable by humans, it indicates an issue addressed in the most probable texts. A topic is thus not a social problem since situations covered in media are not necessarily undesirable.

As mentioned above, an indicator of a topic being a social problem is its negative perception by the readers or framing by the text authors. The most immediate way to detect polarity of such perception is to perform sentiment analysis (SA). This group of methods varies in its complexity (Pang & Lee, 2008). Here we leave out approaches based on supervised machine learning since such instruments are not available for Russian language political texts, at least publicly. We also leave out the entire family of joint topic-sentiment models that mostly use TM to extract aspects of objects being rated by customers and dominant sentiment toward those aspects within customer reviews (Jadhav, 2014; Dermouche et al., 2015), while our goal is to find sentiment outside the texts in which these topics have been touched.

The SA approach we use is based on sentiment lexicons — lists of words with pre-defined sentiment polarity that are being searched for in texts, after which the texts are ascribed a sentiment score as some integral grade derived from the individual word scores. Sentiment polarity may be understood differently: most generally, as positive vs negative, or in a more nuanced way, e.g. support vs criticism.

It should be noted that general negative terms, of course, do indicate something undesirable, but they do not necessarily indicate that the situation is seen as social and demanding a collective effort to be eliminated – that is, a social problem. For instance, negative words may indicate grief or fear expressed in relation to an accident or a disaster (Thelwall et al., 2011, p. 413). However, it is virtually impossible to discuss social problems without negative terms. Therefore, general negative sentiment can be used as a first signal for a social scientist or a policy maker to get acquainted with the respective user content in more detail, and to find out if a social problem is emerging.

## Hypotheses

In addition to revealing which topics are perceived as social problems by media audiences, our approach allows testing hypotheses about relation between topics' in-text salience, popularity with audiences and problematization. Existing research finds that media professionals, including journalists, often have vague ideas about their audiences' demographic features, interests (Atkin et al 1981) and the content of their comments (Ürper & Çevikel 2014). Our first hypothesis may thus be formulated as follows:

*H1:* topic salience in media texts will not be related to topic popularity among the commenting audience.

Previous research (Weber 2014; Liu et al 2016) also finds that negative news items (e.g. those addressing damage rather than success) get more comments than positive ones. To our knowledge, there is no research exploring relation between the volume of comments and their polarity which would reveal whether readers are more inclined to leave negative comments. From the revealed inclination to comment on negative news, our assumption is that they would. Therefore our second hypothesis sounds as follows:

*H2:* The more negatively the topics are perceived by the commenting audience, the more popular they will be with this audience.

## Data collection and preprocessing

In this research we choose to study social problems at the regional level as we hypothesize that at this level more specific problems may emerge (McCombs & Funk 2011), while the national Russian media are, first, more controlled and, second, overwhelmed with international news, notably the Ukrainian crisis. We use the data from the media of the Omsk region, an area in the Southern Siberia. By its population and GRP it is a bit above the median among other regions,

while it is exactly at the median by the population density, the share of urban population and the share of Russians. It thus can be considered a typical Russian region.

We define an Omsk media outlet as a website registered as such in the region, targeting only Omsk audience and having a certain level of penetration – no less than 10,000 unique users per month. According to the Agency of Regional Research, the leading Omsk marketing organization, 18 such online outlets we registered in Omsk as of June 2014, the time closely preceding data collection[3]. Four of them embraced 65% of all the visits, and we decided that only those four outlets were worth of the effort of laborious data scraping. Since Omsk regional media are not polarized politically, we believe that audience size is a sufficient criterion for making our sample representative. On the whole, we follow a procedure typical for sampling media organizations in media studies.

Our sample thus included all news items and respective comments from *Gorod55* (http://gorod55.ru/, 6,302 news items), *BK55* (http://www.bk55.ru/, 14,078 items), *NGS Omsk* (http://ngs55.ru/, 4,780 items) and *Omsk-inform* (http://www.omskinform.ru/, 8,727 items) for the entire year from September 1, 2013 to September 1, 2014. The collection comprised 33,887 news items and 258,121 comments. On average, the four sources published 116 news per working day, and 33 news per holiday. Distribution of comments per news item is, as expected, uneven, but not much skewed. Around 80% of news items (26,783) got at least one comment. Average number of comments per news item is 7.6, with *Gorod55* taking the lead with 10.7 and *Omsk-Inform* lagging back with only 2.7. It thus can be seen that the production leader and the feedback leader do not coincide.

After collecting our data, we performed a number of standard text preprocessing procedures. We first eliminated repetitive structural elements, such as section headings that could have skewed TM results. Next, we performed tokenization (transformation of texts from sequences of symbols into sequences of words) with a tokenizer from Pattern[4] software that had been selected from among five competing instruments. To obtain correct word frequencies, we performed lemmatization which reduces all forms of the same word to a single form, with *pymorphy2*[5] lemmatizer. Finally, we cleared the collection from stop-words – items that carry no or nearly no meaning but can deteriorate the quality of further analysis. We combined a lexicon, manual selection from among

[3] Agency of Regional Research. Rating Omsk online media. Retrieved August 25, 2014 from http://omsk-journal.ru/publ/9-1-0-116
[4] http://www.clips.ua.ac.be/pattern
[5] https://github.com/kmike/pymorphy2

100 most frequent words and all words that occurred in the collection only once; we thus reduced the number of unique words from 118,718 to 69,447.

# Finding agendas in news with topic modeling

## Learning the topic model

The most widely used algorithm of topic modeling is known as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). It assumes that topic distribution over words and texts has a Dirichlet prior – that is, it follows a certain pre-defined pattern. The distribution of individual words and texts over topics is, however, unknown, and the aim of the algorithm is to restore it – i.e. to determine the "weight" of each word and each text in each topic. There are two main methods to do it: variational method (Blei et al., 2003) and the one using Gibbs sampling (Steyvers and Griffiths, 2007); in addition, there is a large number of software implementations of both of them. Here we use the standard Gensim software[6] implementing variational method because it has been widely accepted in the LDA community. Some examples of topics (in the form of the lists of the most probable words and their probabilities) revealed by this algorithm in our collection are presented in Table 1.

**Tab. 1. Examples of topics**

| Theaters | Ivan Klimov's murder | Russia-Ukraine-US relations |
|---|---|---|
| 0.032*Omsk | 0.017*Ivan | 0.044*Russia |
| 0.027*theatre | 0.017*Omsk | 0.021*Ukraine |
| 0.013*play | 0.014*Klimov | 0.014*president |
| 0.010*culture | 0.014*murder | 0.013*country |
| 0.009* festival | 0.011*Lebedovoy | 0.013*Putin |
| 0.008*Russia | 0.011*boxer | 0.009*USA |
| 0.007*actor | 0.009*conflict | 0.007*Ukrainian |
| 0.006* trouper | 0.009*Yan | 0.006*Vladimir |
| 0.005* troupe | 0.008*version | 0.006*state |
| 0.005* theatrical | 0.006*police | 0.005*declare |
| 0.005* viewer | 0.006*case | 0.005* authority |
| 0.005*Vladimir | 0.006*Russia | 0.005*military |
| 0.005* hall | 0.006*wound | 0.004*territory |
| 0.005*scene | 0.006*investigation | 0.004*American |
| 0.005*name | 0.005*crime | 0.004*side |
| 0.005*producer | 0.005*man | 0.004*head |
| 0.005*director | 0.005*shooting | 0.004*Kiev |
| 0.004*art | 0.005*information | 0.003*against |
| 0.004*creative | 0.005*region | 0.003*power |
| 0.004*Sergei | 0.005*November | 0.003*sanction |

[6] https://radimrehurek.com/gensim/

As it can be seen, the topics are easily interpretable; however it is important to be aware of the method's limitations one of which is the problem of selection of the "right" number of topics. As most other LDA problems, it has not yet received a final solution as the concept of LDA quality has not yet been formally defined. The most commonly used measure of perplexity (Jelinek et al 1977) monotonously decreases with the growth of the number of topics and thus gives no clues on their "right" number. Therefore, we use a metric offered by Arun et al (2010). Having obtained 19 topic solutions with step = 5 in the range between 5 and 100, we find several minima of Arun's measure and end up in choosing one of them (which corresponds to 50 topics) based on manual topic assessment.

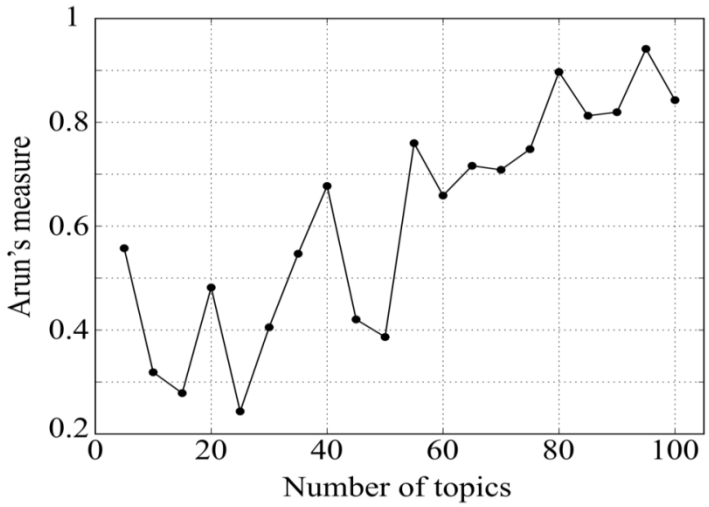


Fig. 1. Dependence of Arun's measure on the number of topics.

## Detecting topics and their salience

The obtained topics were labeled based on top words and top texts, by two researchers who then agreed on the labels. LDA always yields a certain proportion of uninterpretable topics (e.g. those crystallized around pejoratives), but in our case we obtained only one completely uninterpretable topic and four topics whose interpretation aroused some difficulties (marked *). By summing probabilities of all texts in a given topic, one can obtain an index of the topic's salience in the collection, that is assess how widely this topic is covered by journalists compared to other issues. The list topics sorted by this index is presented in Table 2.

**Tab. 2. Topics ranked by salience**

| Topic | Index of the topic's salience |
|---|---|
| Car accidents | 0.0478 |
| Criminal news | 0.0448 |
| Fires | 0.0389 |
| Local authorities: appointments, resignations & statements | 0.0383 |

| | |
|---|---|
| Regional taxes & fuel prices | 0.0378 |
| Russia, Ukraine & US international relations | 0.0369 |
| Trials on economic crime | 0.0347 |
| Sport, hockey | 0.0333 |
| *Urban development (misc) | 0.0319 |
| Omsk region industrial developemnt | 0.0313 |
| Public transport and traffic | 0.0306 |
| Urban landscaping & greening | 0.0286 |
| Regional parliament activities | 0.0255 |
| Accidents with children | 0.0251 |
| Olympic Games 2014 & Omsk athletes | 0.024 |
| Police actions drug, alcohol & counterfeit money crimes | 0.0237 |
| Local authorities: appointments, resignations & statements | 0.0233 |
| Stray dogs & doghunters | 0.0228 |
| Weather | 0.0227 |
| Abridgments of traffic law | 0.0215 |
| Arbitration court and the Mostovik case | 0.021 |
| Urban demography & housing payments | 0.0179 |
| Education | 0.0179 |
| IT & military high tech | 0.0176 |
| Schools, orphanages & child charity | 0.0161 |
| Urban events & openings | 0.0157 |
| Theaters & festivals | 0.0157 |
| Beauty contests & their winners | 0.0152 |
| Real estate: contruction | 0.0148 |
| Control & regulation of enterprises | 0.0143 |
| Macroeconomic events: currency rates & oil prices | 0.0143 |
| Holidays & VIP weddings | 0.014 |
| Movies and Movie stars | 0.0129 |
| Housing: heating | 0.0128 |
| Ads of banking services | 0.0111 |
| Libraries, literature & art | 0.011 |
| *Regional elections and misc. | 0.011 |
| Yury Gamburg resignation | 0.0108 |
| Missing persons announcements | 0.0107 |
| Street & bridge reconsturction & maintenance | 0.0106 |
| Concerts | 0.0106 |
| Hockey | 0.0103 |
| Ivan Klimov’s murder | 0.0097 |
| Housing & the case of disabled Akhmetov | 0.0093 |

| | |
|---|---|
| Car sales | 0.0092 |
| *Military holidays | 0.0082 |
| *Uninterpretable | 0.0072 |
| *Omsk media, plants & animals | 0.0063 |
| Crimea accession | 0.0062 |
| *NATO warships in Black sea, Russian rocket launch & contests | 0.0032 |

As expected, local and regional topics outnumber national and international topics. The most salient topics are also regional, however, Ukrainian crisis stands particularly high in this list. Not surprisingly, also, accidents and disasters occupy the three top positions. The topics can be divided into event-driven and issue-driven (including problem-driven). Since our data is news, issue-driven topics do not necessarily contain texts that discuss general trends, causes or consequences of social phenomena as such; most of the time they group together events related to a certain issue or a problem, for instance, schooling or economic crime. We can see that issue-driven topics outnumber those centered around a single event, which is natural because a single event has to be as salient as an entire "issue" to compete for being detected as a separate topic. The Sochi Olympic Games is the absolute winner among event-driven topics, however, it is the only non-regional topic in this category. Apart from the traditionally well-covered military holidays, including Victory Day, two resonant local events appear in this list: resignation and subsequent arrest of the Omsk vice-mayor Yury Gamburg, and the murder of the locally famous boxer Ivan Klimov. Although the sportsman's murderers have not been found, many news items and comments share a belief in that his murder was connected to his conflict with the local "Gipsy baron" Yan Lebedovoy and perhaps was committed or ordered by someone from the local Roma community. Both events seem to be not very important when judged solely by their salience in the collection. However, further analysis of comments brings them back to our attention.

# Finding popular and problematic topics with comment analysis

## Detecting topic popularity with readers

News item popularity among readers is usually measured through the number of views or clicks, however, clicks contain no data for further analysis of polarity of readers' feedback. This can be most easily assessed through the number of comments (Shoemaker et al., 2010), however, a comment is an attribute of a news item, and no news item belongs to any topic entirely, which is why we have to develop an index of topic popularity. Here we, first, multiply the probability of a topic in each text by the number of comments received by the respective text, and then calculate the sum of those products. However, according to such index more salient topics will always look more

commented because high values of salience are usually obtained when a topic is moderately present in multiple texts, rather than when it is highly salient in a limited number of texts. Therefore, we normalize our index via dividing it by the mean probability of the given topic in all texts. We obtain the following formula:

$$Y_t = \frac{D \sum_{d=0}^{D} prob_{dt} \times qcomments_d}{\sum_{d=0}^{D} prob_{dt}},$$

(1)

where $Y_t$ — popularity index of topic *t*, *D* — number of documents (texts), *d* — document number, $prob_{dt}$ — probability of topic *t* in document *d*, $qcomments_d$ — number of comments on document *d*.

Finally, for better representation we normalize all the obtained popularity scores to the range 0-100%. Ten most commented and ten least commented topics are presented in Table 3.

**Tab. 3. Topic popularity among readers (based on the volume of comments)**

| Topic | Topic popularity among readers |
|---|---|
| 1. Russia, Ukraine and the USA international relations | 100.0% |
| 2. Stray dogs & doghunters | 94.0% |
| 3. Yury Gamburg resignation | 86.8% |
| 4. Local authorities: appointments, resignations & statements | 79.9% |
| 5. *Urban development (misc) | 76.2% |
| 6. Abridgments of traffic law | 74.9% |
| 7. Libraries, literature & art | 70.3% |
| 8.Crimea accession | 69.7% |
| 9. Ivan Klimov's murder | 67.3% |
| 10. Regional taxes & fuel prices | 65.7% |
| … | … |
| 41. Housing: heating | 39.8% |
| 42. Movies and Movie stars | 39.6% |
| 43. Urban events & openings | 39.2% |
| 44. Olympic Games 2014 & Omsk athletes | 38.7% |
| 45. Education | 37.0% |
| 46. Concerts | 37.0% |
| 47. *Regional elections and misc. | 35.4% |
| 48. Fires | 34.6% |
| 49. Ads of banking services | 34.2% |
| 50. Car sales | 20.7% |

It would be logical to expect that readers turn to regional and local media for regional or local topics. However, the leading topic in our popularity list is centered around the relations between Russia, Ukraine and the USA in the context of the Ukrainian crisis. As we saw before, this

topic is also one of the most salient. The latter fact alone could have been explained by the bias of local journalists who could be copying the propagandistic agenda of the national media. Nevertheless, this topic is much discussed. Furthermore, accession of Crimea is also particularly popular. This can be explained, first, by the fact that Ukraine's "departure" from Russia, accession of Crimea and the international sanctions that followed have deeply affected all aspects of people's lives in Russia. Second, a shift of audience's preferences from local to national to international news was noticed already a few decades ago (Bogart, 1989), especially among younger, urban, male and more educated consumers. A more recent study has found that online audiences consume more national and international news than traditional audiences (Tewksberry, 2006), which means that our findings are broadly consistent with the existing trends detected by other methods.

Another observation concerns the two local topic-forming events – Gamburg's arrest and Klimov's murder – both of which are much discussed, despite being moderately covered. This illustrates a possible gap between editorial policies on and customers' perceptions of particular issues. Overall, there is no correlation between topic salience in texts and topic popularity among readers – that is, H1 is confirmed.

## Detecting negatively perceived topics with sentiment analysis

Although sentiment analysis instruments are quite developed for English language (González-Bailón & Paltoglou, 2015), for Russian language, until very recently, no lexicons were publicly available. In this research we use *LINIS Crowd* (http://www.linis-crowd.org/) released in early 2016 (Koltsova et al., 2016) and aimed at SA in user generated content related to politics and public affairs. *SentiRuLex*, a more general sentiment lexicon (Lukachevitch & Levchik, 2016), was released after the data analysis for this paper had been completed. Some other lexicons whose quality is unknown were rejected; *LINIS Crowd*, as compared to reported quality of publicly unavailable lexicons, has performed fairly well (Koltsova et al., 2016).

*LINIS Crowd* had been made for the well-known SA freeware *SentiStrength* originally aimed at English language[7]. This software ascribes two scores to each text: negative and positive. We used the default version that calculates the two text scores as the maximum of the grades of all occurring words of the respective class (positive or negative). This approach has shown optimal results for short texts (tweets) (Thelwall et al., 2010), that are close in length to news comments we

[7] http://sentistrength.wlv.ac.uk/

seek to analyze. The overall sentiment score of each comment was calculated as the sum of the negative and the positive scores.

The aggregated polarity of all comments related to a given topic was computed similarly to the topic's popularity score, but instead of the number of comments the formula contains the mean sentiment score of all comments on a given news item:

$$Y_t = \frac{D\sum_{d=0}^{D} prob_{dt} \times sent_d}{\sum_{d=0}^{D} prob_{dt}}, \quad (2)$$

where $P_t$ — polarity index of topic $t$, $D$ — number of documents (texts), $d$ — document number, $prob_{dt}$ — probability of topic $t$ in document $d$, $sent_d$ — mean sentiment score of comments on document $d$.

As with the popularity score, the polarity scores were normalized to the range 0-100%. Ten most positively commented topics and ten most negatively commented topics are presented in Table 4.

**Tab. 4. Prevailing polarity of comments on various topics**

| Topic | Positivity score |
|---|---|
| 1. Olympic Games 2014 & Omsk athletes | 100.00% |
| 2. Hockey | 90.33% |
| 3. Beauty contests & their winners | 89.82% |
| 4. Theaters & festivals | 89.36% |
| 5. Street & bridge reconsturction & maintenance | 88.75% |
| 6. Sport, hockey | 88.67% |
| 7. Holidays & VIP weddings | 87.51% |
| 8. Public transport and traffic | 86.86% |
| 9. Weather | 86.33% |
| 10. Local authorities: appointments, resignations & statements | 86.02% |
| … | … |
| 41. Abridgments of traffic law | 57.01% |
| 42. Ads of banking services | 54.26% |
| 43. Police actions drug, alcohol & counterfeit money crimes | 54.23% |
| 44. Accidents with children | 50.62% |
| 45. Missing persons announcements | 49.85% |
| 46. Car sales | 47.42% |
| 47. Car accidents | 47.28% |
| 48. Fires | 42.19% |
| 49. Criminal news | 23.59% |
| 50. Ivan Klimov's murder | 22.74% |

It is not surprising that the most positive emotions are related to holidays, entertainment, sports and culture. Crimea accession is also an obvious case for any Russian, since Crimea-driven rise of the national pride and the feeling of victory have been overwhelming. At the negative end, all topics except Klimov's murder are related to disasters and crimes framed as separate cases. They arouse general negative emotions, such as grief, fear and anger, without bringing audiences to problem definitions. The only topic within which problem definition does occur is Klimov's murder although at the first glance it, too, may seem an event-driven topic. However, the main negativism of the comments is not related to grief, but is directed towards the inefficient performance of the police unable or unwilling to find the murderers. Thus, a social problem as it is defined by the readers in relation to this case may be formulated as police (investigation) inefficiency, corruption and overall absence of justice.

An important observation is that popularity scores and polarity scores do not correlate (r=0.08); that is, *H2* is not supported. We find all types of topics: popular positive (Crimea accession), popular negative (Ivan Klimov's murder), unpopular positive (weather), and unpopular negative (fires). As it can be concluded from Shoemaker et al. (2010), this is not always the case, and there exist audiences inclined to comment either mostly on negative or mostly on positive news items.

Since we defined a social problem as both popular and negatively commented topic, we need an index that captures both of those features, e.g. an additive index. Table 5 shows topics that take top ten positions when sorted by such index. Some of them are still event-driven (accidents and crimes), but quite a number of others, marked with italics, easily reveal their problem-driven nature after just a brief look at the comments to the most relevant texts. Stray dog topic reflects tensions between defenders of animal rights and those who try to "clear" the city by killing dogs. Macroeconomic topic is related to people's economic pessimism and mistrust to governmental economic policy. NATO warships topic is overwhelmed with hate speech towards "Americans" and is driven by Russia-US political tensions. Finally, the topic of Russia-Ukraine relations produces the most polarized and diverse comments that can be grouped into the following types: (1) hostile to Ukrainians (prevailing); (2) hostile to Putin because of the war with Ukrainians; (3) hostile to Putin because of insufficient war with Ukrainians; (4) supportive of Putin; (5) hostile to separatists in the Eastern Ukraine; (6) supportive of separatists. While in the NATO warships topic the source of the problem is defined unanimously, in Ukrainian topic one can see competing definitions of social problems, but the aggressive polarization itself may be regarded a social problem (although such understanding of social problem goes beyond the operational definition adopted here). Thus, although not all popular and negatively commented topics are devoted to social problems,

popularity-negativity score can serve as an alarm signal for performing a more detailed manual analysis of the respective texts.

**Tab. 5. Top ten topics ranked jointly by negative polarity and popularity**

| Topic | Negativity-popularity index |
|---|---|
| *1. Ivan Klimov's murder* | 144.56% |
| *2. Russia, Ukraine & US international relations* | 137.62% |
| *3. Stray dogs & doghunters* | 130.30% |
| 4. Criminal news | 124.61% |
| 5. Abridgments of traffic law | 117.89% |
| 6. Yury Gamburg resignation | 113.70% |
| 7. Car accidents | 108.62% |
| *8.* NATO warships in Black sea, Russian rocket launch & contests* | 107.74% |
| 9. Accidents with children | 100.98% |
| *10. Macroeconomic events: currency rates & oil prices* | 99.03% |

# Conclusion and future research

In this paper we have demonstrated that although introduction of data mining techniques into communication research does not liberate a social scientist from manual labor, it nevertheless helps channel human effort to narrower text subsets where the phenomena of interest are most likely to occur. We have shown how a combination of basic topic modeling and sentiment analysis coupled with a number of simple indices can be used to reveal social problems as they are defined by online news readers in thousands of their comments. By doing so, we have obtained sociological conclusions without reading the entire collection.

First, we have shown that the volume of attention to certain topics demonstrated by the media and by their audiences diverge. This can indicate either censorship or erroneous editorial policies, an both findings may be valuable social science or marketing outcomes. Second, we have seen that prevalence of negative emotions always accompanies obviously negative events, such as disasters, crimes and conflicts, and is never observed for such topics as culture or entertainment. This suggests that the method correctly detects the prevailing sentiment. Third, prevalence of negative emotions combined with topic popularity among the readers can indicate that the audience perceives the given topic as a social problem, and we have seen that such topics are not necessarily most covered by the media; in other words, they are not necessarily those that are perceived as

important by media professionals. This means that the approach may be capturing the complex interplay of the roles of the media and the public in the process of problem definition. Fourth, prevalence of negative emotions sometimes indicates polarization of the readers that use negative lexicon as a tool in an aggressive discussion. Lack of social consensus online can also be viewed as a social problem, and it might be used to predict the offline clashes in future.

We must also name some more directions for future research that stem from the current limitations of the proposed approach. Nearly all its steps need developing quality tests. The quality of the proposed indices heavily depends on the quality of the methods that are at their core – topic modeling and sentiment analysis. As it was noted, quality assessment of the former is a serious and unresolved problem, while Russian-language sentiment analysis is underdeveloped just due to the lack of resources. Thus, much will depend on the computer science and computational linguistics communities, however, communication scholars can and should contribute to formulation of the end tasks and to elaborating concepts of quality relevant to communication studies and a broader social science. Thus, understanding what is "true" topic salience or a "truly" problematized topic, against which any methods could be tested, is still waiting for further methodological effort.

The second direction of inquiry closely related to the former is to define and detect social problem as such in a more precise way. Once we have learned to find truly popular and truly problematized topics, we still need to understand how much popular and how much problematized they should be to qualify as social problems. One way is to compare topics to other topics or to previous time periods, but a more fundamental solution is to find out which level of anxiety in comments corresponds to the concept of social problem, for instance, which level may result in visible social consequences. Here, we point at an obvious need for theoretical work to be done within communication studies and broader social science.

# REFERENCES


Alexandrov, D., Gorgadze, A., & Musabirov, I. (2016). Virtual Caucasus on VK Social Networking Site. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 215–217). New York, NY, USA: ACM.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 391–402). Springer Berlin Heidelberg.

Atkin C.K., Burgoon J.K. & Burgoon M. (1983). How Journalists Perceive the Reading Audience. *Newspaper Research Journal* 4(2): 51-63.

Bennett L. & Iyengar S. (2008) A New Era of Minimal Effects? The Changing Foundations of Political Communication. *Journal of Communication*, 58(4): 707-731.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bogart, L. (1989). *Press and Public: Who Reads What, When, Where, and why in American Newspapers*. Psychology Press.

Burscher, B., Odijk, D., Vliegenthart, R., Rijke, M. de, & Vreese, C. H. de. (2014). Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures*, *8*(3), 190–206.

Cohen, B. (1963). *The press and foreign policy*. New York: Harcourt.

David, C. C., Atun, J. M., Fille, E., & Monterola, C. (2011). Finding Frames: Comparing Two Methods of Frame Analysis. *Communication Methods and Measures*, *5*(4), 329–351.

Dermouche, M., Kouas, L., Velcin, J., & Loudcher, S. (2015). A Joint Model for Topic-sentiment Modeling from Text. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (pp. 819–824). New York, NY, USA: ACM.

Diesner, J., & Carley, K. M. (2010). A Methodology for Integrating Network Theory and Topic Modeling and Its Application to Innovation Diffusion. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing* (pp. 687–692). Washington, DC, USA: IEEE Computer Society.

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2013). Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content: topics, style and gender. *Digital Journalism*, *1*(1), 102–116.

Gonzalez-Bailon, S., & Paltoglou, G. (2015). *Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources* (SSRN Scholarly Paper No. ID 2558788). Rochester, NY: Social Science Research Network.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297.

Günther, E., & Quandt, T. (2016). Word Counts and Topic Models. *Digital Journalism*, *4*(1), 75–88.

Hall, S. (1980). Encoding/decoding. In *Culture, media, language* (pp. 128–138). London.

Henrich N. & Holmes B. (2013). Web news readers' comments: Towards developing a methodology for using on-line comments in social inquiry. *Journal of Media and Communication Studies*, 5(1): 1-4.

Hilgartner, S., & Bosk, C. L. (1988). The Rise and Fall of Social Problems: A Public Arenas Model. *American Journal of Sociology*, *94*(1), 53–78.

Jadhav, N. (n.d.). *Topic models for Sentiment analysis: A Literature Survey. Unpublished manuscript*. Retrieved from http://www.cfilt.iitb.ac.in/resources/surveys/Topic-Models-For-Sentiment-Analysis-2014-Nikhilkumar-Jadhav.pdf

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63–S63.

Kim, Y. (2015) Exploring the Effects of Source Credibility and Others' Comments on Online News Evaluation. *Electronic News*, 9(3), 160-176.

Koltcov, S. N., Koltsova, O. Y., & Alexeeva, S. B. (2016). An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. In *Computational Linguistics and Intellectual Technologies* (pp. 277–287). Moscow.

Koltsova, O., & Koltcov, S. (2013). Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet*, *5*(2), 207–227.

Koltsova, O., & Shcherbak, A. (2014). "LiveJournal Libra!": The political blogosphere and voting preferences in Russia in 2011–2012. *New Media & Society*, 1461444814531875.

Krippendorff, K. H. (2003). *Content Analysis: An Introduction to Its Methodology* (2nd edition). Thousand Oaks, Calif: Sage Publications, Inc.

Lasswell, H. D. (1938). *Propaganda technique in the World War*. New York, P.Smith (reprinted in 1938).

Lauer, R. H. (1976). Defining Social Problems: Public and Professional Perspectives. *Social Problems*, *24*(1), 122–130.

Lee, E.-J. & Yoon, J.J. (2010) What Do Others' Reactions to News on Internet Portal Sites Tell Us? Effects of Presentation Format and Readers' Need for Cognition on Reality. *Communication Research*, 37(6), 825–846.

Liu Q., Zhou M., & Zhao X. (2015). Understanding News 2.0: A framework for explaining the number of comments from readers on online news. *Information & Management.* 52(7): 764-776.

Loukachevitch, N., & Levchik, A. (2016). Creating a General Russian Sentiment Lexicon. Presented at the Language Resources and Evaluation Conference.

McCluskey M. & Hmielowski J. (2012). Opinion expression during social conflict: Comparing online reader comments and letters to the editor. *Journalism*, 13(3): 303-319.

McCombs M. & Funk M. (2011) Shaping the Agenda of Local Daily Newspapers: A Methodology Merging the Agenda Setting and Community Structure Perspectives. *Mass Communication and Society*, 14:905–919.

Merton, R. K. and N., R. (1971). *Contemporary Social Problems* (Third Edition edition). Harcourt Brace Jovanovich.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, *2*(1–2), 1–135.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228.

Sayre, B., Bode, L., Shah, D., Wilcox, D., & Shah, C. (2010). Agenda Setting in a Digital Age: Tracking Attention to California Proposition 8 in Social Media, Online News and Conventional News. *Policy & Internet*, *2*(2), 7–32.

Scharkow, M. (2011). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761–773.

Shoemaker, P. J., Johnson, P. R., Seo, H., & Wang, X. (2010). Readers as gatekeepers of online news: Brazil, China, and the United States. *Brazilian Journalism Research*, *6*(1).

Smith, S. G. (1991). *Social Pathology*. New York: Macmillan.

Spector, M., & Kitsuse, J. (1977). *Constructing Social Problems*. Menlo Park: CA: Cummings.

Steyvers, M., & Griffiths, T. (n.d.). Probabilistic topic models. In T. Landauer, D. C. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. NJ: Erlbaum.

Stroud, N.J., Van Duyn, E. & Peacock, C. (2016) *News Commenters and News Comment Readers.* Engaging News Project Report. Available at: https://engagingnewsproject.org/wp-content/uploads/2016/03/ENP-News-Commenters-and-Comment-Readers1.pdf

Tewksbury, D. (2003). What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet. *Journal of Communication*, *53*(4), 694–710.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter Events. *J. Am. Soc. Inf. Sci. Technol.*, *62*(2), 406–418.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Ürper, D. Ç. & Çevikel T. (2014) Reader comments on mainstream online newspapers in Turkey: Perceptions of web editors and moderators. *Communications*, 39(4): 483-504.

Voskresenskiy, V., Sukharev, K., Musabirov, I., & Alexandrov, D. (2014). Online Communication in Apartment Buildings. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics* (pp. 52–55). Springer International Publishing.

Wallsten, K. (2007). Agenda Setting and the Blogosphere: An Analysis of the Relationship between Mainstream Media and Political Blogs. *Review of Policy Research*, *24*(6), 567–587.

Weber P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New Media and Society*, 16(6): 941 –957.

Wu, T.-Y. & Atkin D. (2017). Online News Discussions. Exploring the Role of User Personality and Motivations for Posting Comments on News. *Journalism and Mass Communication Quarterly. 94*(1): 61-80.

Yasaveyev, I. G. (n.d.). *The construction of social problems by means of mass communication media*. Kazan University Press. [in Russian]

Zhou, Y., & Moy, P. (2007). Parsing Framing Processes: The Interplay Between Online Public Opinion and Media Coverage. *Journal of Communication*, *57*(1), 79–98.

Oleg S. Nagornyy
National Research University Higher School of Economics. Internet Studies Lab,
Research Assistant; E-mail: onagorny@hse.ru

Olessia Y. Koltsova
National Research University Higher School of Economics. Internet Studies Lab,
Laboratory Head. E-mail: ekoltsova@hse.ru