Content Analysis of some Social Media of the Occupied Territories of Ukraine

Volodymyr Lytvynenko¹, Iryna Lurie², Svitlana Radetska³, Mariia Voronenko⁴, Natalia Kornilovska⁵, Daria Partenjucha⁶

Informatic and Computer Science Department, Kherson National Technical University, Bereslav Shosse, 24, Kherson, Ukraine, 73008

immun56@gmail.com¹, iil@rambler.ru², rad_svete@rambler.ru³, mary_voronenko@i.ua⁴, knv06061971@rambler.ru⁵, partenjukha@rambler.ru⁶

Abstract. The paper analyzes the activities of the Internet publications in the temporarily occupied territories. The analysis of the tools for the detection and monitoring of free and paid services, data analysis of social media to monitor references and comments on social networks are presented in the paper. This paper includes an example of a practical text analysis and extraction of information online. It is shown that the use of computing environment KNIME opens new pathways in a variety of social media and determine the user behavior. This technique improves the process of analytical studies.

Keywords: Social Media, KNIME, Content Analysis, Network Analytics, Text Mining, Occupation

1 Introduction

Analysis of the texts is one of the most common types of the scientific and the scientific and practical analyses. Content analysis relates to the researches affecting the text, images, audio and video and allows to write content in a social context. Understanding the social meaning of the document involves the process of identifying of the responses the document may receive in public life, the degree of originality of the document, its difference from other documents of a particular kind. Today there is a popular database based on the archives of the media. Despite of the valuable role played by the Internet in the media, the reasons for its development and the role of market information has been studied insufficiently.

To determine the general orientation of the electronic media source vectors it is needed to consider their similarities and differences in order to identify social factors that affect this orientation, to investigate them from a scientific point of view.

The essence of this research is due to an attempt to analyze the specific functions of the electronic media in the temporarily occupied territories of Ukraine. Taking into
consideration the fact that Ukraine is currently in terms of the information war, it is necessary to obtain full and objective view of the events, which take place there.

The aim of this paper is to consider the role of the information technology use for the analysis of the electronic media in the temporarily occupied territories of Ukraine.

Monitoring and analysis of the electronic media is the systematic tracking of the news reports from the Internet resources, which enables timely to identify and predict the trends in the emergence of competitive situations using quantitative and qualitative analyses. Quantitative analysis is used for the primary assessment of the reports on the objective activities of analysis in the mass media. It gives the possibility to estimate the overall amount of time to cover the subjects and themes selected for analysis in the electronic media.

Qualitative analysis is used to examine the main characteristics of the information field, forms of its presentation, its emotional orientation or absence of infringement of the legislation and to evaluate the information concerning the object of analysis as positive, negative or neutral according to its content.

When it is impossible to conduct qualitative and quantitative analyses we are able to use media effects, such as plots designed to manipulate public opinion. The main criterion for the use of this assessment is the lack of relevance, accuracy, transparency, facts, balance, diversity, timeliness, clarity.

2 State of modern informational on-line publications

There is no clear definition of the Internet media. It is believed that the Internet media (Internet edition online newspaper) is the regularly updated information site that serves as the media and has its permanent audience. It differs from the traditional media only in the field of activity whereas the functions and purposes are identical [1, 2]. Most online publications, which work as the news agencies are not registered due to the lack of the legal regulations of the Internet publications. This fact leads to misunderstanding of the situation by journalists who are beginning to think that there is a complete lack of permissiveness and responsibility for propagating information use [3, 4].

Media often faces the shortcomings of the national legal framework. This makes the media vulnerable to the pressure from various institutions (public, private and criminal), manipulation and intimidation of the political elite, low-culture audience of regional media and, conversely, very high level of self censorship by the media representatives in relation to coverage and analysis of the political sphere [5, 6]. Massive social and political actions against the Ukrainian authorities to protect the Russian language, during anti-government, federalist, pro-Russian and separatist slogans which were shared at the end of February – beginning of March 2014 in the cities of the south-eastern Ukraine after the power change of government, exacerbated the conflict between the West and East Ukraine and led to the instability and separating of the society [7].

As a result the Crimean Peninsula was annexed by the Russian Federation, and the unrecognized states, the so called Donetsk and Luhansk Republics were proclaimed. Citizens involved in the activities of the republics were called separatists and terrorists. They were engaged in extremist propaganda and military action. Due to this the network media representing a radical resources aimed at inciting ethnic hatred, the implementation of pro-Russian views, opinions and ideas emerged.
List of the electronic media in the temporarily occupied territories of Ukraine, is given in Table 1.

**Table 1. SWOT-analysis of the electronic media**

<table>
<thead>
<tr>
<th>Name</th>
<th>Options</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>“RIA News of the Crimea” <a href="http://crimea.ria.ru/">http://crimea.ria.ru/</a></td>
<td>Providing of timely information and news about events in the Crimea, Russia, world</td>
<td>The speed of information. International distribution. Easy navigation of the system</td>
<td>Trying to present impartial news, but the problem is the contradiction between global and national interests.</td>
</tr>
<tr>
<td>“News of the Crimea” <a href="http://news.allcrimea.net/">http://news.allcrimea.net/</a></td>
<td>Independent online publication that covers events that happen in the Crimea</td>
<td>Support news archive, functioning of keyword search, the opportunity to discuss the news on social networks.</td>
<td>Many embarrassing and news layouts, making it difficult to view.</td>
</tr>
<tr>
<td>News media “Russian Spring” <a href="http://rusvesna.su/">http://rusvesna.su/</a></td>
<td>Informative and analytical portal that provides quick information about events in New Russia, in Ukraine and in the world</td>
<td>Represent the views of the experts, politicians, reports from the places of military events, information is presented in 5 languages, multimedia information.</td>
<td>Submission of false information. Too much advertising. At this website in section &quot;Help&quot; fundraising services for the sustenance of the resource is held. But, it is not clear who gets the money</td>
</tr>
<tr>
<td>Internet publication “DNR24” <a href="http://dnr24.su/">http://dnr24.su/</a></td>
<td>Information about events in eastern Ukraine, Donetsk and Lugansk regions.</td>
<td>The speed of information. Simple and easy navigation of the system.</td>
<td>The lack of censorship. The radical nature of news. Images to the users. Worsening political and armed conflict</td>
</tr>
<tr>
<td>News Agency ”Antifa-shyst” <a href="http://antifashist.com/">http://antifashist.com/</a></td>
<td>Anti-Fascist Forum of Ukraine, which aims at presenting the news on political and military action.</td>
<td>Multimedia. A large number of articles.</td>
<td>Inconvenient online design of the resource. The subjectivity of the information provided. Shocking news content.</td>
</tr>
<tr>
<td>News Agency ”The news front” <a href="http://news-front.info/">http://news-front.info/</a></td>
<td>An independent agency network. Truthful information about the events in the New Russia, Ukraine and the world.</td>
<td>The ability to view the site in the 5 languages. Apps for mobile platforms IOS and Android.</td>
<td>Conducting surveys from the front line and exaggeration of the events taking place in the temporarily occupied territories, distorted information.</td>
</tr>
</tbody>
</table>
In today's world of mass media, there is a huge amount of available data. The problem is to convert them into useful and relevant form. Under these conditions, the actual content is the personification of the global system WWW, which is impossible without the development and implementation of the appropriate techniques and tools [8]. A large number of messages, comments which reflect the mood of the users in terms of escalation and information confrontation have been considered [9].

KNIME is easy to use as a graphical tool for all types of the process analyses - data access, data conversion, initial study, the powerful mining, visualization and report's generating. This open integrated platform offers more than 1,000 modules (nodes), developed by the community KNIME. The procedures implemented through the workflows and workflow consists of nodes. The Nodes (Units) are responsible for the implementation of various procedures in the working process, can be found in the "repository nodes." Data View are the components which serve to visualize data (graph, charts). Each node has a configuration setting window. These tools can give an initial overview. They are not suitable for a deeper understanding of the behavior, needs, problems, desires or tendencies of the individuals as these tools and services do not actually provide any data for presenting of the summarized data [10]. Combining of text mining and network analysis have been conducted for the Information Agency "anti-fascist" - a news resource that aims to present reports about political, economic and military action in the temporarily occupied territories of Ukraine. In general, the community contains about 7,000 comments, 10,930 articles on politics and about 85,000 users. Community of members is very active with more than 100 responses to the topic. The majority of people leave their comments and are registered under their nickname, but some comments are anonymous. In particular, the Twitter account of the online edition "anti-fascist", which has 1,389 readers, followers and 28,839 tweets has been analyzed. With the help of the Follower wonk the data about users, followers of the pages were obtained. Predictive text analysis, network analysis methods and inverse transformation of the raw data into the usual information were applied basing on the information that is the methods of clustering and modeling were used. Twitter- appearance of the account is presented in Fig. 1.

![Twitter account of the anti-fascist news agency](image)

**Fig. 1.** Twitter account of the anti-fascist news agency

To understand the mood of the user it is necessary to determine the level of his/her relations up the nature of his/her comments, accordingly they can be positive or negative. The level of relationship can also be used to classify users in future. In order to classify the mood of the lexicon containing words (tips) in addition to other information, their polarity is used to indicate the words as positive or negative. There polarity: positive, neutral or negative has been used. For each user-follower the
frequency of positive and negative words use has been determined and calculated in accordance with the participation in the anti-fascist page retweets. The difference between these frequencies determining the attitude of the user is the following:

$$\lambda (u) = f_{\text{pos}}(u) - f_{\text{neg}}(u),$$

where $f_{\text{pos}}(u)$ – the frequency of the positive words use; $f_{\text{neg}}(u)$ - the frequency of the negative words use. Positive $\lambda$ – defines positive users and negative $\lambda$ – negative users.

In the first stage of the process, several traditional anti-fascist processing reading units for data mining are used. Then each post is converted into the data type of the document for further text analysis operations. Finally, the node "dictionary Tagger" tag links polarity of each word in the document column (Fig. 2).

Now all the words in tweets and retweets are marked as positive or negative. It is possible to begin the estimation of the relationship of each user. There are two types of important nodes in a category Text Processing KNIME:

- Node "BoW- creator" accumulating the words (Bow) for a set of documents consists of two columns: one containing the document and another containing the terms occurring in certain tweets.
- Node "TF" calculates a relative term frequency (TF) of each part of the documents and adds a column containing the term frequencies, calculated by the division of the absolute frequency of the term usage, which is found in the post, the number of terms of the post.

The frequency of negative and positive terms is aggregated above the user's ID to obtain the general frequency of negative and positive words use. Then the level of each user relationship is calculated as the difference between the frequencies of terms (Fig. 3).
Members were divided into two categories namely "positive" and "negative." Suppose that a custom of the level relations are Gaussian distribution around the mean with variance $\mu_\lambda$ and $\sigma_\lambda$ and the most users around $\mu_\lambda$ are neutral. Therefore, it can be foreseen that the users with the level of relationships $\lambda$ within $\mu_\lambda \pm \sigma_\lambda$ are neutral, while the users with $\lambda$ in the left turn of Gauss distribution ($\lambda \leq \mu_\lambda - \sigma_\lambda$) are negative and the users with $\lambda$ in the right turn of Gauss distribution ($\lambda \leq \mu_\lambda + \sigma_\lambda$) are positive. Based on the calculated values for $\mu_\lambda$ and $\sigma_\lambda$, the results of binning process is 807 negative users and 582 positive users.

In Fig. 4 a graph distribution of all known users and followers is shown. The X-axis represents the frequency of the positive words use, Y represents the frequency of the negative words use by the users. Negative users are painted red, positive users are painted green.

The user who uses a great amount of different words (positive and negative) is «Turmalay», it can be seen in the upper right corner in Fig. 5. However, it is not a user with the highest level of relations, as the level of its social authority is 60.
The average frequency of words (positive and negative) used by positive users is 418, which is almost twice as bigger than the negative users – 217. Thus, the negative users often do not write.

The final workflow for processing and retweets records is shown in Fig. 6.

The main purpose of the network analysis is to identify the leaders and followers based on the status and structural state of the users in the current network. After filtering of all articles and comments of the users, 26 non-interconnected components were created. 25 contained only three smaller tops and one - 24,055 tops from 98,150 countries. The network, created on the basis of anti-fascist data page is extremely complex (Fig. 7). The importance of networking visualization becomes clear, the focus on specific areas with identified leaders and followers is considered to be of great importance. Leaders are the users who create their own tweet or comment, which becomes the topics for discussion, for example tweets concerning vital political issues. These users may be of interest to those individuals who are involved in the formulation of the public opinion, as they attract a lot of attention to their
publications. Followers are the users who creates retweets of the leaders’ entries but they do not get the comments by themselves.

**Fig. 7. Full network of the Internet publication (интернет-издание) "Antifascist"**

In order to identify the leaders and the followers let us use centrality index from web analytics. This figure is based on assigning each vertex of two different values, combining the weight of the authority and the weight of the node. Peak meets the high weight of the hub when it refers to large numbers and high level of credibility. Therefore, the high weight is assigned by the hub users who often react to the articles published by others.

Fig. 8 shows the diagram of leaders scattering against the weight for all users belonging to the major components. X-axis represents the follower account in the calculation of the mass of the hub, while the Y axis represents the account of the leader per weight of authority.

**Fig. 8. Plot band of the leaders against the followers among the total number of users**

User «Vadim Stolicin» attracts immediate attention and the highest level of the authority equals to 1, and high concentration estimates 0.9. So he gets a lot of comments from other users to his / her message (high score leader) and at the same time he/she often comments the articles and comments of other users. This person really is one of the most active at discussions of political issues in the anti-fascist page. Another user «neokomm» also can be of great interest for the investigation. The user has the greatest weight of authority 4, but a very small concentration of estimates 0, meaning that it has a very large number of followers, but never responds to their
records.

On the opposite side of the scattering diagram, you can see user's «nyepx» data, who has the highest concentration of 1 ratings and moderate the authority of 0.2, which means that he leaves a lot of comments on the reports of others, but rarely writes his/her own posts, and even if he / she does write posts, they rarely get comments. He/she is one of the best followers.

Fig. 9 shows the workflow KNIME, which filters out all anonymous posts and users, creating a network of people based on the anti-fascist data set, it draws the largest component, calculates the weight of the authority for each user using R Networks of obtaining 7 plugins, and visualizes the weight and the credibility into the dispersion chart.

![Fig. 9. Hybrid KNIME and R workflow to extract information from the network](image)

Information about the actual author of the text, sentiment expressed in it, as well as graphs and numbers of readers and defendants can not reveal the position of the person in relation to all others in the community, and can not detect the interaction between this user and others.

The network analysis is suitable for detection of anomalies, such as "I will vote for you, and you will vote for me," which is a classic problem to extract text used in the analysis of sentiment. The workflow which combines the text mining results with the results of the network analysis can be seen on Fig. 10.

![Fig. 10. Workflow KNIME combining the analysis of connections and text mining](image)

Based on the analysis of the attitudes and the extracted from the texts information it is possible to place each user to schedule scattering with level of orientation on the X axis and the pusher or the account leader on the axis Y.

Not all the positive attitudes of thinkers on the right of the diagonal will matter for our marketing campaigns. In fact, despite the positive thoughts in the information they mostly react to someone else's original thought (positive or negative). On the other hand, there are some friendly-oriented users, whose attitude is above the
diagonal, they are definitely the leaders. These users can be considered as those whose positive contributions are determined by the community attitude.

3 Conclusions

The main trends and online publications in the temporarily occupied territories were analyzed. There is a wide range of tools for identifying, monitoring and analysis, there are both paid and free services for analyzing social media data. All these tools allow to track the comments and references in the social networks.

The combination of intelligent text analysis and extraction of information from various offers new possibility to penetrate into social media and determine the user’s behavior that would be impossible using each approach separately.

This approach can also be improved by including additional sources of relevant data relating to specific priority areas, such as company and product names, political parties, well-known users. Additional data will further strengthen the capabilities of the method of identification, segmentation and development of interesting groups. This technique introduces additional features into the data intended for the user to improve the analytical research process.

References

4. Domarev, V.V.: Zashchita informatissi i bezopasnosti' kom'pyuternykh sistem.K., DiaSoft, 1999480 s (In Ukrainian)
11. Killian Thiel et al, "Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining" KNIME 2012. (In English)
