**An Approach to Detect and Analyze the Impact of Biased Information Sources in the Social Media**

Jarkko Paavola, Harri Jalonen

Turku University of Applied Sciences, Turku, Finland

jarkko.paavola@turkuamk.fi

harri.jalonen@turkuamk.fi

**Abstract:** The paper presumes that social media is an environment where local and small events may escalate into bigger and even global ones in a very short period of time. This is because social media offers opportunities for discussion of shared interest in way which cannot be controlled: everything that can be exposed will be exposed – for all intents and purposes. This possibility has also changed the landscape of discussions of controversial issue, such as foreign and security policy. Compared to traditional mass media, social media enable disclosing opinions without censorship. Nowadays people have access to online discussions, blogs and even websites entirely devoted to sharing negative information. It has been seen that, during crisis situations social media has become a major way of affecting people’s opinions. Consequently we are witnessing the rise of trolls – individual who shares inflammatory, extraneous or off-topic messages in social media, with the primary intent of provoking readers into an emotional response or of otherwise disrupting normal on-topic discussion. Based on the lack of censorship, on the one hand, and trolling behaviour, on the other, the paper aims to understand the rise and diffusion of extreme opinions in Twitter. This is a case study paper, where the analysed case is Twitter messages on Ukrainian crisis during 2014 written in Finnish language. The aim is to utilize sentiment analysis for the automatic detection of trolling behaviour. Sentiment analysis provides tools for strategic communications for the automatic analysis of social media discussions and to recognize opportunities for participating in the discussion at the most effective stage.

Keywords: Opinion mining, Social media, Emotion, Twitter

**1. Introduction**

Popular social media sites magnetise hundreds of users nowadays. A sophisticated guess is that in the OECD countries, two thirds of the population use social media more or less regularly. Worth noting is that social media is not confined to technology, but involves cultural, societal and economic consequences. This paper defines social media as a constellation of shared technologies that allow "activities, practices, and behaviors among communities of people who gather online to share information, knowledge, and opinions using conversational media" (Safko and Brake, 2009). Social media is not an alternative to real life, but an irremovable part of it. Through billions of users, social media constitutes a transformative and collective force, which has the ability to initiate revolutionary processes, as seen in the Arab Spring.

The Arab Spring has shown that the significance of the change entailed with the emergence of “public spheres” comes concrete in countries which have prohibited or hampered the use of social media. Turkey, for example, blocked access to Twitter in March 2014. Prime Minister Recep Tayyip Erdogan had vowed to "wipe out Twitter" after users spread allegations of corruption. Although the ban has been condemned as violating freedom of expression and individual rights by the country's top court, the case expresses the concern the ruling groups may have on social media.

Seemingly social media has become a place where people discuss and debate controversial topics and try to affect public opinion. Studies have shown that social media discussions are prone to polarise in two opposite directions and end up in contradiction. Studies also pinpoint that negatively spoken contents are different from positive one. According to one sentiment analysis of the affective nature of online political comments, positive comments exceeded negative ones, but that positive comments decreased over time while negative comments increased over time (Robertson et al., 2013). One possible explanation is provided by Sobkowicz and Sobkowicz (2012), who have argued that political online discussion turns into negative due to the need to attract attention. To grab attention in social media, it seems that users are obliged to rely on expressing provocative opinions. Once negative sentiment takes over it is difficult to stop. Negative statements tend to follow negative statements (Chmiel et al., 2011).

Theoretically the paper is based on three interconnected premises that is seen playing an important role in inducing conflict seeking discussions and expressing negative contents. Firstly, there exists a freedom of speech in social media. This allows people to express contradictory and sensitive material without any censorship. In other words, there are no gatekeepers, who decide which material can be published and what not. Secondly, the lack of gatekeeping enables also people who have extreme opinions and/or suspicious aims to join social media discussions and intentionally provoke others. Therefore, social media has provided new opportunities for trolling. Thirdly, peer information differs for official announcements. Especially during crisis situations social media has become a major way of affecting people’s opinions. Although, people are aware of propaganda behind official announcements, social media has a more personal feeling, which makes is it seem to be more trustworthy source when people are directly communicating about what is happening during an event. This motivates in some quarters to utilize trolling systematically to affect public opinion.

For example, recently, a Finnish newspaper reported that there are approximately twenty active persons posting constantly pro-Russia messages in Finnish language regarding the situation in Ukraine to social media. Their aim seems to be opinions on Russia to be divided among Finnish people: they offer “another side of the story”.

Empirically this case-study paper focuses on Twitter discussions related to Ukraine crisis in Finnish language. Previously developed sentiment analysis tool will be tested how these kind of biased actors in Twitter microblog could be detected. Parameters affecting the influence for general polarity on the subject will be also discussed. The paper provides motivation for strategic communications to utilize automatic analysis of social media discussions and to recognize opportunities for participating in the discussion at the most effective stage.

This paper is structured as follows: in the second section, the theoretical premises of the paper are provided. The third section presents the previously developed social media analysis tool. In the fourth section, the results obtained from Twitter discussions is presented. Finally, conclusions are drawn in the fifth section.

**2. Social media – the lack of gatekeepers, trolling behaviour and loaded with emotions**

Social media have brought with it ‘media life’, which Deuze (2011) calls “the state where media has become so inseparable from us that we do not live with media, but in it” (Karppi, 2014). In a hyperconnected, network society, posts on Twitter cause stock market crashes and overthrow governments (Pentland, 2014).

Unsurprisingly, life in social media is as messy as it is in the real world. However, social media is still a relatively new phenomenon whose consequences cannot be fully predicted. However, some sophisticated guesses can be made. One is that behaviour in social media contradicts with the theory of gatekeeping. According to the theory, which was originally developed by Kurt Lewin (1943), gatekeeping is the process through which information is filtered for dissemination. Every medium has gatekeepers, who select and confine the information flows. Reporters, for example, decide which sources are chosen to be included in a story, whereas editors decide whether stories are printed or covered. In contrast to traditional mass communication, social media is an unregulated context allowing ordinary people to publish almost anything that comes to their minds. There is no need/room for gatekeepers in social media. Social media offers opportunities for discussion of shared interest in way which cannot be controlled: everything that can be exposed will be exposed – for all intents and purposes. Interaction produces emergent behaviour which cannot be understood on the basis of what is known about the actors involved in that interaction.

A great deal of social media behaviour is affected by negativity bias. It means that people are much more likely to recognise and be influenced by negative information shared in social media. There are studies that have addressed social media sites which are dedicated to allowing people to vent. Rant-sites, as they are called, provide people a forum to rant, for example, about firms and their products and services. Rant-sites particularly attract people who feel anger. Martin et al. (2013) have studied how anger is expressed in these sites. As a main finding they suggested that “reading and writing online rants are likely unhealthy practices as those who do them often are angry and have more maladaptive expressions than others”. As peculiar it may sound, some individuals energise themselves by sharing negative and detrimental information. Noble et al. (2012) have labelled these individual as trolls. Troll is an individual who shares “inflammatory, extraneous or off-topic messages […] in social media, with the primary intent of provoking readers into an emotional response or of otherwise disrupting normal on-topic discussion”. Contrary to a dissatisfied customer, it is in the troll’s deliberate intention to damage an organisation or a community. Studies also show that if users are allowed to post comments anonymously, it lowers the risk of losing face, and therefore increases the odds of showing negativity such as anger and disgust (Derks et al., 2007, Yun and Park, 2011).

Social media is loaded with emotions. Several studies have shown that emotionally spoken contents spread effectively in social media (Bae and Lee, 2012, Verhagen et al., 2013). As a result local and small events escalate into bigger and even global ones in a very short period of time. Adapting Rubin (2011), social media is seen transforming into “one big global amplifier through which emotional experience is transmitted and strengthened”. The essential thing is that, although emotions are felt on an individual level, in social media, they can simultaneously be shared to and by the others. This has a wide range of “real-life” consequences. It has been shown, for example, that social, political, cultural and economic events are correlated with Twitter mood levels (Gilbert and Karahalios, 2010). Emotions speak to people more effectively than pure facts. Unfortunately, this has opened the door also for discussions motivated by biased emotions and without link to reality.

The only way to comprehend the consequences of lack of gatekeeping, trolling behaviour and arguments based on flawed emotions is to look at the inception and evolution of discussions. In the following, this will be done by using the sentiment analysis approach.

**3. The description of analysis tool**

Turku University of Applied Sciences has developed a social media analysis tool in Tekes (Finnish funding agency for technology and innovation) funded NEMO project, which analyzes business value of negative emotions. So far, the tool has been used to detect the polarity (positive-neutral-negative) of the messages, and to visualize dynamics between messages.

The tool consists of following parts implemented with open source software:

* Crawler, which collects to local database all Tweets and Facebook messages posted in Finnish language. The crawler uses interfaces that allow collecting messages several years backwards. Naturally, messages written in any language can be mined. Language detection filter is utilized is used to limit the amount of collected message.
* Visualization tool, which shows keyword appearance geographically and as a time-series. Visualization tool allows to analyze several parameters e.g. by the sender of the message. In this paper, the visualization tool is the main tool for detecting trolls. The tool structures discussion threads to the form, where it is easily seen actors who are commenting Russia related tweets systematically.
* Interface to produce training sets for automated analysis.
* Sentiment analysis. This part involves also complex operation of changing words in Finnish language to the basic form as Finnish language does not contain prepositions, but words have inflectional endings.

Sentiment analysis, or opinion mining, refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. The aim is to determine the attitude of a speaker or a writer. Sentic computing is a multi-disciplinary approach to sentiment analysis that exploits both computer and social sciences to better recognise, interpret, and process opinions and sentiments over the Web (Cambria and Hussain, 2012). The reference shows that sentiment analysis can be also successfully used to detect also trolls in the social media. The troll in the social media refers to a person having the deliberate intent of provoking readers into an emotional response.

Technically the analysis here is performed with R-language (R Project, 2015). The R-code divides the text material to be analyzed to training and analysis parts. With the training material different data mining models can be constructed. After the training, testing materials are used to verify classification ability. Tweets are modified before analysis so that all capital letters are changed to lower case, extra spaces and numbers are removed. Text material is analyzed for pre-defined positive and negative words, and their percentage share of the length of message. Emoticons are calculated to different variable. Based on created variables decision-tree and logic regression analyses are run. These models are then integrated to tool to analyze future messages.

This paper should be considered as work in progress. For example, the accuracy of analysis can be improved with vocabulary sets specifically designed for the purpose. In this paper, general vocabulary was utilized.

**4. Analyzed case in Twitter (messages written in Finnish)**

The following keyword list (translated here to English for clarity) was utilized to locate and store Twitter messages:

* Crimea, Eastern Ukraine
* Kiev
* Independence square
* Gas, oil
* economic sanctions
* freezing assets
* currency, ruble
* aid convoy
* invasion
* pro-Russia
* separatist, separatism
* flight MH17
* state debt
* riot police
* sniper
* Odessa fire
* Tymoshenko
* Yushchenko
* Yanukovych

For each located tweet also the whole discussion thread was stored. Thus, even though the keyword search was limited to the year 2014, the final time span of collected messages was up from 2009 to 2015.

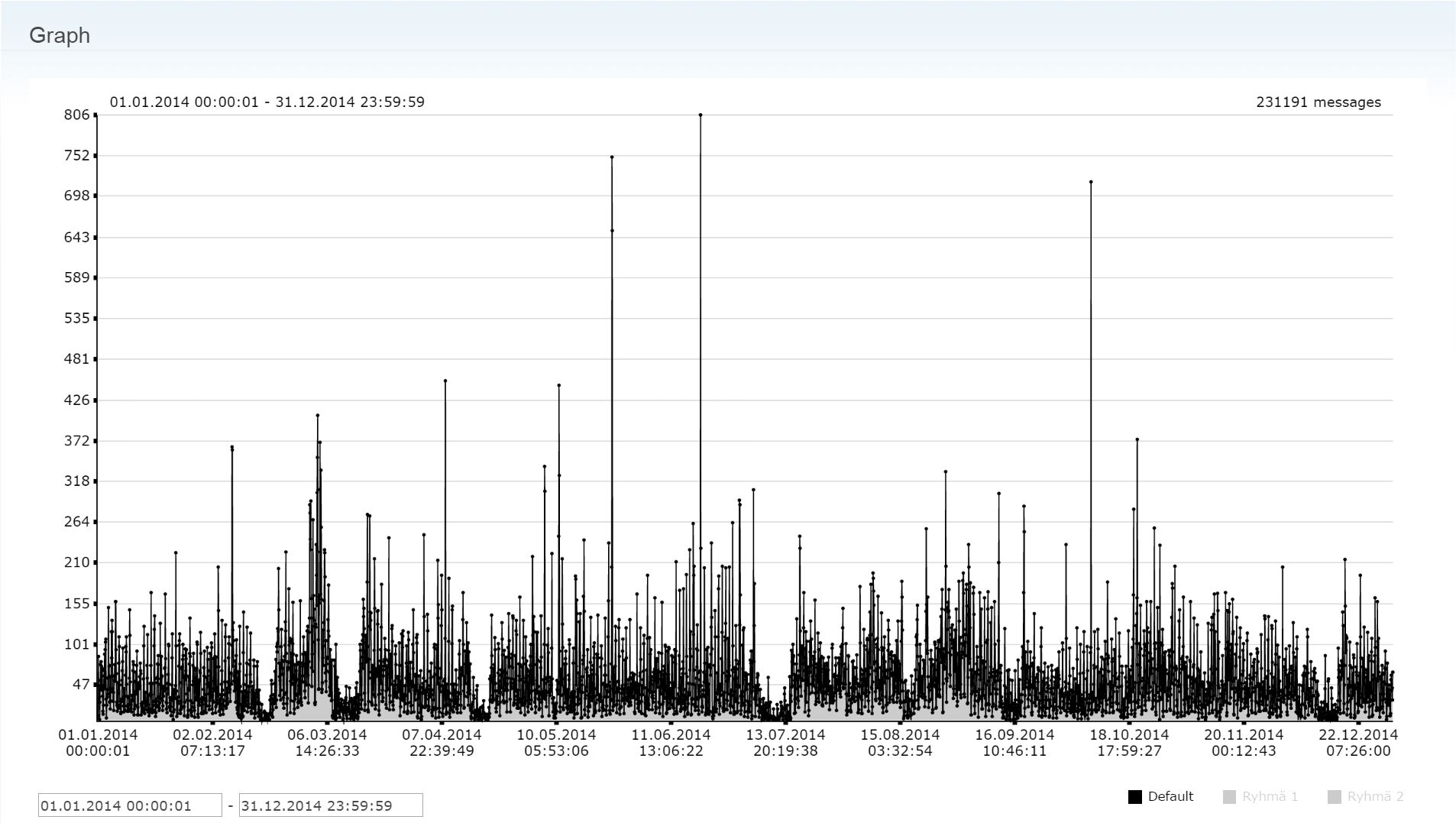


Figure 1: Timeline of Twitter messages. One measurement point represents the accumulated number of messages for two-hour period. Time span shown is 1.1.2014-31.12.2014. Total number of messages shown in the figure is 231191.

Figure 1 shows constant message activity for the whole year 2014. It was assumed that some events would be clearly visible as spikes in tweeting activity (such as 20th of February when Kiev saw the day of violence, when at least 88 people were killed in 48 hours; 25th of May when Ukraine elected Petro Poroshenko as president; or 17th of July when Malaysia Airlines flight MH17 from Amsterdam was shot down near the village of Grabove in rebel-held territory, with the loss of 298 lives), but actually that was not case. Further work in the future is required whether the word list was not sufficient or if there are other factors affecting the result.

When looking at the most active tweeters during the analysis period, the following list is produced:



When trying to locate trolls we can discard first actors 1, 6, 18, 22, 24, and 28 as they are news broadcasts using Twitter as one source of distribution. However, interesting behavior could be found by analyzing replies to the news.

In this case study, the sentiment analysis could not be applied to detect trolls as such. With quick manual browsing of some active participants to the discussion and replies to their messages it was easy to find users who systematically try to neutralize all criticism against Russia. Thus, to be effective in detecting trolling the data mining algorithm should be tuned to detect opposite opinion posted systematically to Russia related negative messages – or equivalently positive messages related to the western countries or actions taken against Russia.

The efficiency of the troll could be considered to be dependent on two main factors: activity and the amount of people exposed for the biased communication. Also, the success of triggering response should be considered. Analysis tool allows to collect following information from Twitter messages

* The amount of followers – This parameter defines how many people will see the post directly
* The amount of retweets – this parameter could considered as the “acceptance and endorsement” of given point-of-view
* The amount of favourites – This parameter indicates how the post is perceived by followers
* The amount of replies – this can be seen as the effect of gaining interest
* The amount of replies written to other users’ posts – this shows the activity of communicating user

The implementation of the analysis of these parameters is in the roadmap for the sentiment analysis tool.

**5. Conclusions**

The paper showed that social media is an environment where local and small events may escalate into bigger and even global ones in a very short period of time. This is because social media offers opportunities for discussion of shared interest in way which cannot be controlled: everything that can be exposed will be exposed – for all intents and purposes. This possibility has also changed the landscape of discussions of controversial issue, such as foreign and security policy. Compared to traditional mass media, social media enable disclosing opinions without censorship. Nowadays people have access to online discussions, blogs and even websites entirely devoted to sharing negative information. It has been seen that, during crisis situations social media has become a major way of affecting people’s opinions. Consequently we are witnessing the rise of trolls – individual who shares inflammatory, extraneous or off-topic messages in social media, with the primary intent of provoking readers into an emotional response or of otherwise disrupting normal on-topic discussion. Trolling has become a form of information war.

Finnish newspaper articles suggest that the information war is active also with personal emails to journalists and politicians. Trolls try to affect also directly persons that report news and participate in political discussion in the Parliament. This activity has been so strong that Foreign Affairs Committee of the Parliament had a discussion event about the trolling phenomenon late January 2015. The main conclusion of the event was that in information war the democracy principle and the freedom of speech is used to justify the delivery of dis-information. Trolls try to mitigate opposite opinions for example by labelling counterparts as racists. For example, the member of Parliament Pertti Salolainen gave very strong comments that during his long political career this kind of lobbying of political stance is new and disturbing phenomenon. He emphasizes that there is not enough knowledge on this kind of information war available. People should be educated to understand and evaluate the trustworthiness of information presented in the media – both traditional moderated media and chaotic social media.

This kind of very aggressive trolling has raised up the discussion of freedom of speech. Worried participants in the discussion see that the trolls are abusing the freedom of speech. In principal, as a consequence for the fear that trolls will be successful in blurring the truth, people are hoping for gatekeeper to arrive also to social media. This will be unlikely, thus strategic communications must ready for automated detection of trolls and careful planning when to participate in social media discussions as negative sentiment has been shown to be dominant in social media.

**References**

Bae, Y. and Lee, H. (2012) “Sentiment analysis of twitter audiences: Measuring the positive or negative

influence of popular twitterers”, *Journal of the American Society for Information Science and Technology*,

Vol 63, No. 12, pp 2521–2535.

Cambria, E. and Hussain, A. (2012) *Sentic Computing - Techniques, Tools, and Applications*, Springer.

Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A. and Hołyst, J. A. (2011) “Collective Emotions Online and Their Influence on Community Life”, PLoS ONE, Vol 6, No. 7, pp 1–8.

Derks, D., Arjan, E. R. and von Grumbkow, J. (2007) “Emoticons and social interaction on the Internet: the importance of social context”, *Computers in Human Behavior*, Vol 23, pp 842–849.

Deuze, M. (2011) “Media Life”, *Media, Culture and Society*, Vol 33, No. 1, pp 137–148.

Gilbert, E. and Karahalios, K. (2010) “Widespread worry and the stock market”, *The Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington, DC, USA.

Karppi, T. (2014) *Disconnect.Me – User Engagement and Facebook*, Doctoral Dissertations, Annales Universitatis Turkuensis, Ser. B Tom. 376, Humaniora, University of Turku.

Lewin, K. (1943) “Psychology and the process of group living” *Journal of Social Psychology*, Vol 17, pp 113–131.

Martin, R. C., Coyier, K. R., VanSistine, L. M. and Schroeder, K. L. (2013) “Anger on the Internet: The Perceived Value of Rant-Sites”, *CyberPsychology, Behavior and Social Networking*, Vol 16, No. 2, pp 119–122.

Noble, C. H., Noble., S. M. and Adjei, M. T. (2012) “Let them talk! Managing primary and extended online brand communities for success”, *Business Horizons*, Vol 55, pp 475–483.

Pentland, Al. (2014) *Social Physics. How Good Ideas Spread – The Lessons from a New Science*, The Penguin Press, New York, NY.

Robertson, S. P., Douglas, S., Maruyma, M. and Semaan, B. (2013) “Political discourse on social networking sites: Sentiment, in-group/out-group orientation and rationality*”, Information Policy: The International Journal of Government and Democracy in the Information Age*, Vol 18, No. 2, pp 107–126.

R Project (2015) http://www.r-project.org/

Rubin, A. (2011) “Living in the age of emotional rationality: Wendell Bell, social media and the challenges of value change”, *Futures*, Vol 43, pp 583–589.

Safko, L. and Brake, D. K. (2009) *The Social Media Bible. Tactics, Tools and Strategies for Business Success*, Wiley.

Sobkowicz, P. Sobkowicz, A. (2012) “Properties of social network in an internet political discussion forum”, *Advances in Complex Systems*, Vol 15, No. 6.

Verhagen, T., Nauta, A. and Feldberg, F. (2013) “Negative online word-of-mouth: Behavioral indicator or emotional release?”, *Computers in Human Behavior*, Vol 29, pp 1430–1440.

Yun, G. W. and Park, S.-Y. (2011) “Selective posting: Willingness to post a message online”, *Journal of Computer-Mediated Communication*, Vol 16, pp 201–277.