Propaganda Barometer – a Supportive Tool to Improve Media Literacy Towards Building a Critically Thinking Society

O. Khriyenko

Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FIN-40014 Jyväskylä, Finland
oleksiy.khriyenko@jyu.fi

Abstract— **To smartly consume a huge and constantly growing volume of information, to identify fake news and resist propaganda in the context of Information Warfare, to improve personal critical thinking capabilities and increase media literacy, people require supportive environment with sophisticated technology facilitated tools. With rapid development of media, widespread popularity of social networks and fast growing amount of information distribution channels, propaganda and information warfare enter an absolutely new digital technology supported cyber era. Propaganda mining is not a trivial and very time consuming process for human. And, as with any new technology, human need certain time to understand its actual purpose, learn and adapt own behavior making consumption of a technology more valuable, beneficial and enjoyable. To make adoption faster and minimize possible harmful influence, we need to find a proper way to apply currently available technologies and knowledge for elaboration of supportive tool that helps information consumers to become more independent, insightful, and critical.**

*Keywords - **supportive learning environment; skills development tool; critical thinking; media literacy; propaganda mining; information warfare; fake detection; artificial intelligence; cognitive computing; IBM Watson; NLP.***

I. INTRODUCTION

> *"… the digital universe is also a locale for competition and confrontation. Cyberspace has become a new domain for unfair competition and espionage, disinformation and propaganda, terrorism and criminality.",*
> ***Manuel Valls,** Prime minister of France,2014*

Nowadays, in the context of widespread occurrence of terrorism and hybrid wars, when relations between the countries all over the world become unstable, when politicians try to control the public and dictate their will, when some countries are trying to dominate over others, among other aspect of National Cyber Security, protection strategy against information war becomes very crucial. To make citizens aware of the risks of manipulation and propaganda techniques used by malicious players is an important responsibility of every government.

"Propaganda needs to be clever, smart and efficient," said Russian Defense Minister Sergei Shoigu referencing to the fact that the Russian Defense Ministry has formalized its information-warfare efforts with a dedicated propaganda division. One of the goals of information war is to create chaos not only in the information sphere, but also within society itself. As an example, some Finns had started to spread aggressive pro-Kremlin disinformation without fact checking after being exposed to the propaganda [1]. As a result, some people had protested outside Yle [1] (Media Company) headquarters in Finland after being agitated by disinformation on social media. In 2015, Finnish cybersecurity expert Jarno Limnéll stated[2] that the phenomenon of pro-Russia information influencing will continue to grow and, being not a member of NATO, Finland has to pay attention to propaganda and information war issues even more than Baltic countries (who are members of coalition) and other neighboring countries of Russia (e.g. Ukraine, Belarus, Georgia, etc.) do. Year later, CEPA's Information Warfare Initiative[3] has addressed impact of Russian disinformation in their "Winning the Information War" report [2], where has presented dozens of case studies and stated that the Kremlin's use of information as a weapon is not new, but its sophistication and intensity are increasing. Moreover, modern Russian propaganda is cleverly targeted, technically adept and cynically fact-free.

When his army annexed Crimea, Vladimir Putin went on TV and informed the world there were no Russian soldiers in Ukraine. At that moment he wasn't lying so much as saying that the truth doesn't matter anymore. "Those miserable, scared liars", with these words Alexander Scherba (ambassador of Ukraine in Austria) has commented[4] work of totalitarian fake news with reference to the translation made by interpreter during the Putin's and Steinmeier's joint statement made in Moscow 25th October 2017. When Steinmeier said in German

[1] https://yle.fi/

[2] http://kioski.yle.fi/omat/this-is-what-pro-russia-internet-propaganda-feels-like

[3] http://infowar.cepa.org/

[4] https://www.facebook.com/o.scherba/posts/10213092008147661

DOI: 10.5176/2251-3043_6.1.113

- "Annexion" (with reference to Russian annexation of Crimea), interpreter chokes and then translates - "Unification with Russia". When fact-checking agencies rate 78% of Donald Trump's statements untrue but he still becomes a US Presidential candidate – then it appears that facts no longer matter much in the land of the freedom. When the Brexit campaign announces some claims and then they are shrugged off as a 'mistake' by some of Brexit leaders - then it is clear that we are living in a "post-fact" or "post-truth" world. Politicians and media have always lied – but now, we live in the world where they don't care whether they tell the truth or not.

"We are all the defenders. So, all those who receive information, we are the defenders of Finland."[5],
Sauli Niinistö, *President of Finland, 2016*

Democracy can be considered as an effective form of government if the public, that supposed to rule it, is well-informed about and able to independently and critically think about national and international events. Very often politicians manipulate with minds of the public. If citizens do not recognize propaganda in the news when exposed to it, they cannot reasonably determine what media messages have to be supplemented, counter-balanced, or thrown out entirely. Among the countermeasures needed is a proper information defense mechanism that protects people and societies from troll attacks and disinformation, mechanism that educate people, make them critical thinking and able to resist manipulation of their mind. If an information defense mechanism is not developed, the propagandists will gain new victories, they will oppress and confuse even more people, and gain the ability to mobilize people to commit serious actions outside the information sphere. Therefore, one of the targets for government is to facilitate media literacy of citizens and build critically thinking society capable to defend itself. The same goal should be aimed by any citizen who willing to build actual democracy in the country ruled by dictator or corrupted ruling elites.

With rapid development of media, widespread popularity of social networks and fast growing amount of information distribution channels, propaganda and information warfare enter an absolutely new digital technology supported cyber era. Digital era gave us a possibility to generate and spread over huge amount of information in a very short period of time. While one fact of a lie is detected, thousands more have been created same time. Such a constant growth of disinformation volume makes unreality inevitable. It does not mean that world of technology is the cause of the problem. Technology is just a tool for those who use it for particular purpose. Trying to take control of social media, Russian leadership has mobilized a new information warfare tool known as 'trolls' – a virtual army of fake social media Putin-fans [3].

Any new technology or service offered nowadays requires certain time for people to understand its real purpose and influence on their lives. It takes time to adopt a technology and optimize a use of it. Thus, it is not obvious anymore that checking of Facebook (or some other social network) is the first thing people do when wake up in the morning and continue to do almost every hour during the day. However, almost everyone did it when just joined the network. So, people learn and adapt their behavior making consumption of a technology more valuable, beneficial and enjoyable. The same should happen with modern media consumption and all associated with it challenge. Propaganda mining is not a trivial and very time consuming process for human, especially when we are dealing with huge amount of information. Being a facilitator for the problem, technology should become an enabler of corresponding solution for it. To make adoption faster and minimize harmful influence, we just need to find a proper way to apply currently available technologies and knowledge to develop appropriate tools that will guide people for effective management of a huge amount of information. We need to elaborate supportive tool that helps information consumers to become more independent, insightful, and critical in responding to the news media messages.

The main objective of the paper is to revise current efforts regarding propaganda and fake news detection and make feasibility study of their applicability for development of supportive environment capable to improve media literacy towards development of critically thinking society. Thus, relying on average information consumer, authors are focusing this work on technologically facilitated solution that helps people to develop their own ability of critical thinking and become the most powerful tool against the Information Warfare. Next section introduces challenges of modern journalism and fact-checking initiatives in the context of propaganda and Information War. Chapter 3 covers currently available tools to improve media literacy. Further, in Chapter 4 authors introduce Propaganda Barometer - a supportive tool for critical thinking skills development. Addressing requirements for such tool functionality, chapter introduces possible practical solutions with respect to automated support for critical thinking skills development and propaganda techniques detection. Conclusions with future work directions finally wrap-up the work.

II. Challenges Of Modern Journalism And Fact-Checking Initiatives

One of significant principles of true journalism is journalistic objectivity. This principle of journalistic professionalism refers to fairness, disinterestedness, factuality, nonpartisanship, and usually encompasses all of these qualities. Objectiveness, accurate and fact based investigative information presentation to the public, enable the audience to make up their own mind about a story and decide what they believe and what they do not. Journalists need to present the facts whether or not they like or agree with those, remaining neutral and unbiased regardless of own personal opinion and beliefs. Unfortunately, it is very difficult to meet really true journalism nowadays. Concepts of the appropriate role for journalism[6] vary between countries. In some countries, the news media is controlled by a government intervention that makes it dependent body. If not controlled by the government,

[5] http://yle.fi/uutiset/presidentti_niinisto_infosodasta_me_kaikki_olemme_maanpuolustajia/8388624

[6] https://en.wikipedia.org/wiki/Journalism

the profit motive becomes an issue. Many journalists are just employees who earn money performing any orders. Unavailability of true journalism in countries where almost all the news media belong to ruling government and oligarchs (who are controlling or are controlled by it) is a perfect basis for influence on people opinions, making them absolutely loyal to any decisions or actions of the government. It becomes a perfect tool for information war against own citizens with a purpose to constantly keep a control over them, to stay in power as long as possible protecting own corruption and never be punished for that. Similarly, heavily adopting various propaganda techniques and producing an avalanche of fake news, it become possible to influence audience in other countries bringing an information war on international level.

Nowadays, traditional media are no longer the only source of news and information. The Web, and social media (social network sites) in particular, has revolutionized the way in which information is disseminated. The platforms where content can be freely shared, enabling users to actively participate and influence to information diffusion, made mind manipulation process even easier than ever before. Social networks became channels for spam dissemination and intentionally crafted fakes, making our current times the age of misinformation [4][5].

As an example of a research platform that fights the battle against fake news and is focused on the potential of AI (in particular machine learning and natural language processing) to identify fake news stories, we may highlight the Fake News Challenge[7] initiative. However, making parts of the job much easier and more efficient for human fact-checkers with support of automated systems, according to Fake News Challenge, "It won't be possible to fact check automatically until we've achieved human-level artificial intelligence capable of understanding complex human interactions, and conducting investigative journalism." Rapid spreading of fake information made the fake news sites the object of intent attention of not only experts and journalists, but also management of those online services that have contributed to their spread. For example, Facebook introduces the function of user driven tagging of the news veracity, making Facebook staff willing to pay attention for further analysis. Google tries to combat fake news introducing a Fact-Check Feature on both the news.google.com website and in the Google News and Weather applications. It enables publishers to show a "Fact Check" tag in Google News for news stories identifying articles that include information fact checked by news publishers and fact-checking organizations. In turn, all this requires publisher to meet the corresponding criteria[8] and follow certain procedures. However, due to the low (close to zero) cost of sharing information, there are too many parties involved in spreading news, making it nearly impossible to check and regulate all false news sources. There are also some commercial services (e.g. TrustServista[9]) that use Artificial Intelligence algorithms to determine the trustworthiness of news articles, tackle misinformation and fake news propagation in a more efficient way, and find the original source of information. However, these tools do not target average reader, but aim to shorten investigation times for media professionals instead.

Became a very hot topic, fact checking and fake detection attract not only research communities but also startups to run new projects aimed at these challenges. Been facilitated by media and business accelerators (e,g. Matter[10]), as well as general growing demand, some of them (e.g. Rootclaim[11]) base their solutions on combination of math and crowdsourcing. Others try to overcome the anonymity introducing constraints to information distribution environments (e.g. Authenticated Reality[12]). Applying principles of Data Journalism[13] and Open Data[14] (and open-data government initiatives such as Data.gov and Data.gov.uk. in particular), Vigilant[15] and Grafiti[16] help fact-checkers to easily access and use open information about activities and decisions of the government, various financial documents and registers of property rights, etc. to report confirmation or refutation of materials in more attractive for readers form via interactive visual representations.

Among the tools to support fact-checking, we may highlight a universal handbook - Wolfram Alpha[17]. This computational knowledge engine allows user get actual answer to the question instead of set of links to relevant documents. Using Semantic Web [6] and Linked Data [7] technologies, it supports user with access to semantically relevant information, as well as helps to process an image and recognize allocated there objects using Computer Vision techniques. There are some other examples of services that perform search based on statistical data (e.g. Statista[18] and Zanran[19]).

One more initiative that tries to apply collective intelligence against the fakes via crowdfunding and crowdsources is WikiTribune[20]. Been organized by Jimmy Wales (founder of Wikipedia[21]), it similarly applies the same principles and models as Wikipedia does, and supposes to provide ad-free news media platform implementing evidence-based journalism through collective contribution and responsibility of professional journalists and a community of volunteers. Most probably this initiative will face a lot of challenges that caused WikiNews[22] (one more initiative of Jimmy Wales) failure, as well as should deal with such negative aspect of Wikipedia as "edit warring". The most resent one (in the context of Russian propaganda) occurs with respect to Anne of Kiev, when following Vladimir Putin's statements, her place of birth Kievan Rus' has been changed to Rus' and her name to Anne of

[7] http://www.fakenewschallenge.org/

[8] https://developers.google.com/search/docs/data-types/factcheck

[9] https://www.trustservista.com/

[10] https://matter.vc/

[11] https://www.rootclaim.com/

[12] http://thenewinternet.com/

[13] https://en.wikipedia.org/wiki/Data_journalism

[14] https://en.wikipedia.org/wiki/Open_data

[15] https://vigilant.cc/

[16] https://grafiti.io/

[17] https://www.wolframalpha.com/

[18] https://www.statista.com/

[19] http://www.zanran.com/

[20] https://www.wikitribune.com/

[21] https://www.wikipedia.org/

[22] https://en.wikinews.org/wiki/Main_Page

Rus. Let's hope that team behind WikiTribune has analyzed previous mistakes and is ready to meet new challenges.

In addition to content-based fake detection, there is a source-based approach towards quality classification of materials based on a source of origin and a distribution chain of it. Being able to associate a material with authors or distributors who belong to sources of fakes, materials could be classified as fake with calculated confidence level. In this context, Blockchain[23] technology could be considered as the most promising facilitation technology nowadays. Userfeeds[24] is one of the first startups who recently announced their planes to develop a blockchain based fake verification tool. However, using this approach, true stories generated or distributed via such unreliable nodes could be also classified as fakes. Thus, tools, which are based on this approach alone, would not be that much useful in addressing our goal of media literacy improvement. But, smart combination of both approaches in conjunction with intelligent automated techniques based on Deep Learning, NLP, Cognitive Computing and other AI related technologies, could lead towards valuable results for real time reader guidance and development critical thinking skills.

To support human in assessment of huge amount of surrounding information, there are research achievements with respect to automatic deception detection using logistic regression [8], distance-based methods [9], neural network and advanced text processing [10], evolutionary algorithms [11], etc. Trust and reputation issues, which are closely correlated with fake information, are also addressed for this purpose [12][13]. More recently, automatic fake detection has gained increasing interest [14][15][16]. There are also attempts to detect fake news indirectly, regardless of actual content, based on the users that interact with them ("liked" them) [17]. Fake news on social media has been occurring for several years making it a powerful source for fake news dissemination. A review on existing fake news detection methods under social media scenarios [18] provides a basic understanding on the state-of-the-art fake news detection methods. However, there are still many challenging issues to be further investigated, since fake news detection is still in the early age of development.

While a lot of research towards automated AI based fake news detection is going on nowadays, making this topic very popular [19][20][21]; reasonable criticism still exists. With respect to the interview given by Paul Shomo (a Senior Technical Manager at security firm Guidance Software) to Fox News[25], fake news producers could figure out how to get around the AI algorithms. He says it's "a little scary" to think an AI might mislabel a real news story as fake (known as a false positive). Recent studies by Google Brain have shown that any machine learning classifier can be tricked to give incorrect predictions, and it is possible to get them to give pretty much any result you want. Examples that support this point of view have been present by Dave Gershgorn in his article "Fooling The Machine"[26], where it was shown that certain manipulation with test samples may lead to wrong image classification/recognition by well-trained neural network model. For example, applying an adversarial attack, someone may print a "noisy" ATM check written for $100—and cash it for $1000000; or swap a road sign with a slightly modified one that would set the speed limit to 200, making it pretty dangerous for a world of self-driving cars; or redraw a car's license plate to fool the road cameras; etc. In this case, how can we expect a person unconditionally believe a decision made by machine instead of simply believe that news is not faked? In the mentioned above Fox News article, Darren Campo (adjunct professor at the NYU Stern School of Business) says that fake news is primarily about an emotional response and people won't care if an AI has identified news as fake, unless the news matches up with their own worldview. He tells "Fake news protects itself by embedding a 'fact' in terms that can be defended… While artificial intelligence can identify a fact as incorrect, the AI cannot comprehend the context in which people enjoy believing a lie." Therefore, we need a learning tool that will help to improve media literacy of information consumers enabling them to make own decision; tool that not only automatically detects fakes and recognizes propaganda techniques used in the news, but provides corresponding evidences and explanations; tool that presents alternative point of views, and helps to elaborate personal trust rating of information sources.

III. Media Literacy Facilitation Tools

Among attempts to help information consumers to increase their media literacy and awareness about fake news, we may admit valuable contribution of various initiatives and projects formed as an effort to prevent propaganda (e.g. EUvsDisinfo[27], Polygraph[28], StopFake[29], PropOrNot[30], Bellingcat[31], Politifact[32], etc.). These are good sources of processed and fact-checked by experts articles, as well as learning materials for those willing to spend time by reading analytics to be familiar with propaganda and disinformation cases. Talking about technology facilitated tools that aim the same target of media literacy improvement, we may highlight various gamified learning applications and browser plugins.

Gamification in education process is widely used approach. However, gamified approach towards fact checking is something that has appeared recently. As an example, browser-based game "Factitous"[33] simply allows user to check his/her ability to guess whether given article is fake or real. As soon as used makes his/her chose, application tells correct answer and provide some short explanation with a link to original materials. Another similar fact checking learning game "Post

[23] https://en.wikipedia.org/wiki/Blockchain

[24] https://userfeeds.io/

[25] http://www.foxnews.com/tech/2017/02/21/how-ai-fights-war-against-fake-news.html

[26] https://www.popsci.com/byzantine-science-deceiving-artificial-intelligence

[27] https://euvsdisinfo.eu/

[28] https://www.polygraph.info/

[29] https://www.stopfake.org/en/news/

[30] http://www.propornot.com/

[31] https://www.bellingcat.com/

[32] http://www.politifact.com/

[33] http://factitious.augamestudio.com/

Facto"[34] additionally asks user to define his/her feelings after reading the article and provides explanation why exactly such feelings are caused by the material (actually explains feelings that are predefined for the article in advance). Further, Post Facto presents excerptions from article and asks user to select suspicious ones where most probably some fact checking should be done. If user guessed correctly, game provides corresponding explanations. It also points out some element of real materials such as existence of actual author, logic of content delivery; allows user to directly access a map to check any location mentioned in the article, or search for images used in materials to possibly find an original source of it. Probably the most impressive game related to face news is "Fake It To Make It"[35]. Being inspired by the way how people have earned money creating fake news sited during US president election in 2016, Amanda Warner have created this strategy type game that models the process of the fake sites promotion. Appearance of such learning games in context of media literacy facilitation is a good trend. However, these are only the first steps and current tools are mainly based on predefined examples with manually processed by human guidance and explanations. By applying more sophisticated techniques for automated fake detection, such solutions could be turned into real-time supportive tools for ordinary consumer of information that help on-the-fly analyze any content constantly improving own critical thinking skills.

Talking about tools capable of on-the-fly content analysis, there are some fake detection solutions implemented as plugins for web browsers. The "Fib"[36] chrome-extension goes through Facebook feed in real time and alerts user by verifying the authenticity of posts (status updates, images or links). Similarly, "B.S. Detector"[37] and "Fake News Alert"[38] perform not only with Facebook, but also with Twitter and any other sites and warn users about unreliable news sources simply checking the links against the predefined list of the links to suspicious/fake sites. The "Fact Checker"[39] is community-driven fact-checking platform that flags incorrect or fake news articles and provides direct links to evidence documents and data that either support or contradict assertions. Similarly, "PropOrNot Propaganda Flagger"[40] marks sites and search engine results with "YYY" marker if propaganda elements were recognized in the source. However, both mentioned solutions are limited by a database of the original evidence documents and a list of propaganda sites (associated with PropOrNot) respectively, which are manually filled by users via crowdsourcing approach.

Thus, analogically to gamified solutions mentioned before, browser extensions work on a similar principle. They rely on a manual list of sites likely to contain propaganda and fake content. The emergence of such tools is a great step in the fight against propaganda and false information in the web. Attract the attention to, warn and make information consumers aware of propaganda and fakes could be seen as the first steps in media literacy facilitation. The "human" fact-checking approach that is also adopted by Facebook and Google to identify, validate and assess the reliability of the material, does not resolve the problem, since there is much larger number of people are writing fakes than those who checks them. Moreover, the time needed to verify the facts and to refute is much longer than to write some fake story.

IV. Propaganda Barometer - A Supportive Tool For Critical Thinking Skills Development

The fake news phenomenon is closely connected to a filter bubble problem caused by personalized search of news feeds when readers only encounter stories that they are likely to "like" (to click or comment on). In the context of social network based fake news epidemic, there is a criticism towards technology companies like Facebook, Twitter, and Google, whose algorithms influence who sees which stories. Thus, been hooked by certain worldview once, readers usually encounter stories that confirm pre-existing beliefs. Average information consumer very often does not want (or is not capable) to recognize propaganda and distinguish fake news. The problem concerns not only social network platforms, but also any news search and recommendation systems. Therefore, we have to help readers to build own critical thinking capabilities and be able to unhook themselves via awareness and real-time supportive guidance.

Being able to capture falsity, detect patterns of propaganda and mind manipulation methods, offer alternative points of view by present conflicting information and sources; such "Propaganda Barometer" tool may become a useful learning environment to improve media literacy towards development of critically thinking society. In this chapter we will address possible practical solutions with respect to the functionality requirements of the tool. To find practical solution and elaborate intelligent tool we have to apply different technologies including (but not limiting to): text analysis and Natural Language Processing, Semantic Web and Linked Data, Data Mining, information and service integration, image and video data processing and object recognition, emotion and sentiment analysis, human-computer interaction, etc. Within our project we have been focused on IBM Watson[41] cognitive computing[42] capabilities offered via IBM Bluemix[43] cloud. However, there are a lot of other cognitive computing services offered by IT giants like Google, Microsoft, Intel, Facebook, and other smaller service and application providers.

[34] http://www.postfactogame.com/

[35] http://www.fakeittomakeitgame.com/

[36] https://devpost.com/software/fib

[37] https://chrome.google.com/webstore/detail/bs-detector/dlcgkekjiopopabcifhebmphmfmdbjod?ref=producthunt

[38] https://chrome.google.com/webstore/detail/fake-news-alert/aickfmgnhocegpdbfnpfnedpeionfkbh/related

[39] https://chrome.google.com/webstore/detail/fact-checker/cokfgekpmhapkgfieefhfjicphlollje

[40] https://chrome.google.com/webstore/detail/propornot-propaganda-flag/ogmjlhmfnmhhcllijlbaomamgfaiflai

[41] https://www.ibm.com/watson/

[42] https://en.wikipedia.org/wiki/Cognitive_computing

[43] https://www.ibm.com/cloud-computing/bluemix/

A. Automated support for critical thinking skills development

The Foundation for Critical Thinking[44] defines critical thinking as "the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action". Therefore critical thinking requires person to apply various intellectual tools to deliberately and systematically process diverse information so that (s)he can make better decisions and generally understand things better. Among principles of critical thinking we may distinguish three main ones: *awareness of biases in own thinking, reversing things, evaluation of evidences.*

All of us have biases in our thinking. Critical thinkers should be aware of their cognitive biases and personal prejudices, as well as their influence on thinker's seemingly "objective" decisions and solutions. To make reader aware of own biases, the supportive tool should provide possibility to compare initial reader's attitude and feelings (like/support or dislike/disagree with author's point of view) regarding the news against attitude and feelings assessed based on provided by the toll explanations, evidences and possible alternative points of view. By keeping log of reader's personal self-evaluations and recognized mind manipulation evidences, the tool tracks dynamics and makes reader's bias level visible for him/her.

Very often propagandists try to spoof reality and swap actual causality. It may seem obvious that X causes Y, but what if Y caused X? Reversing things might be a good approach to mine the truth. To develop reader's ability of reverse thinking, the tool should suggest materials with alternative causality. For this purpose, at the first stage, the tool searches for materials similar to the target document. It could be done through third party search engine(s). Optionally, filtering/sorting of search results could be done through entity-based text similarity measure (e.g. Jaccard, Tf-Idf, Cosine, etc.) and domain ontology based entity google semantic similarity measure [22]. Having relevant set of candidates, the tool performs extraction of logical chains (causality) from the text and selects materials with revers chains. It could be done by extracting "entities", "semantic roles" (RDF[45] triples of information in a form of subject-predicate-object) and "relations" (predicates that link two entities) from output of IBM Watson Natural Language Understanding[46] (NLU) cognitive computing service. Merging corresponding equal or semantically close objects and subjects of extracted RDF triples (statements), we build logical chain(s) of analyzed document. The next challenging step is to recognize (sub)chain(s) that represents actual implication and states that A causes B, since not all extracted from text chains represent that. For this purpose several approaches could be used, for example, two ontology-based ones. The most straightforward relies on semantic similarity of chain predicate(s) to the property "*cause/imply*". The more sophisticated approach is based on knowledge-based inference. Based of available fact statements, OWL Property Chain(s)[47] could be inferred by ontology reasoners. It is also would be possible to apply Neural Network based approach. However, to train a model, we need to collect sophisticated training set of human processed text samples with corresponding labels stating "A causes/implies B". Such labeled set could be afforded by collective intelligence applying crowdsourcing model.

And of course, essential part of critical thinking is ability of evidence evaluation. As we highlighted before, automation of actual fact-checking is probably the most challenging task. It is not clear in which way to automatically analyze author's conflict of interests and how shown evidences were gathered, by whom, and why. However, there are a lot of examples when, for example, politicians change their points of view depending on context change. Therefore, keeping recording of their statements formalized in RDF format, it would be possible to apply semantic reasoning and identify possible contradictions between them. By warning the reader about such cares, the tool should support him/her to keeping personal ranking of trust for such politicians, as well as for any other entities (e.g. people, organizations, parties, information sources, news media companies, etc.) that has been caught in lie. And every time, when the entity is recognized by the tool, reader should be informed about it. They will, however, say something fair from time to time. This is due to the fact that if they were biased every time they spoke, they would soon run out of credibility. Reader should be careful about do trust them twice. Regarding analysis of image (video) based evidences, there is not that many things could be done automatically. However, taking into account that one of the usually used tactics is to produce as much as possible fakes within limited period of time neglecting of their quality, it might be still possible to detect some of them automatically applying image processing techniques. For example, in the beginning of the war in Donbass (eastern Ukraine), Russian propaganda news have shown a lot of video interviews with so named "local people", however, in many of them the same guest performers "gastrolery" been involved playing different roles. Therefore, applying face recognition techniques and associating those people with subject entities from corresponding RDF-formalized documents, it would be possible to detect a conflict of "roles". Another example is related to fake images[48] that were used as "evidence" to "proof" that Ukrainian fighter shot the Flight MH17 (the Malaysia Airlines plane crashed after being hit by a Russian-made Buk missile over eastern Ukraine). After analysis of the images (particularly size comparison of plane and other objects on the field) it has been proven that the images are fakes. Of course, automation of such analysis of objects relations is not trivial task, but it could be considered as possible next step.

[44] http://www.criticalthinking.org/

[45] https://www.w3.org/RDF/

[46] https://www.ibm.com/watson/services/natural-language-understanding/

[47] https://www.w3.org/TR/owl2-primer/#Property_Chains

[48] https://www.kompravda.eu/daily/26307/3186146/

B. Propaganda in context of information warfare and its automated detection

Propaganda[49] uses emotional appeals instead of presenting solid evidence to support a point. Propaganda techniques are widely applied in variety of application domains. Advertisers, salespeople, and politicians often lack adequate factual support for their points, so they appeal to our emotions by using propaganda techniques. Similarly, propaganda is a powerful weapon in information wars.

Created in 1937 to educate the American public about the widespread nature of political propaganda, the Institute for Propaganda Analysis (IPA) is best-known for identifying the seven basic propaganda techniques: Name-Calling, Glittering Generality, Transfer, Testimonial, Plain Folks, Card Stacking, and Bandwagon. According to the authors of a book on propaganda, "these seven devices have been repeated so frequently in lectures, articles, and textbooks ever since that they have become virtually synonymous with the practice and analysis of propaganda in all of its aspects." [23]. However, there are more than 50 various propaganda techniques[50] and more than 20 of them could be considered as the most common, and successful. Following we address some of them and present possible approaches for their automated detection using available cognitive competing tools.

The first group of techniques is aggregated under umbrella of “association”. It covers: *Transfer* technique, where target (subject of the article) is associated with something positive that people admire, desire, or love; *Guilt By Association*, where someone's reputation is damaged by associating them with negative event or activity, an unattractive person or organization, etc.; as well as *Name-Calling*, where emotionally loaded language is used to turn people against a target (e.g. product, person, movement, etc.). And, it doesn't matter if there is an actual association or not. Author does not tell that it is the target that is (or does) good or bad. The goal is to build an emotional context around the target. To recognize this type of propaganda technique, we have to recognize such emotional context in certain part(s) of the material and detect that actual target is not a direct subject in there. For sure, emotional either positive or negative attitude to particular thing (event, person, etc.) depends on many factors including cultural, historical, social, etc., and is very difficult to be identified automatically without subject domain knowledge and statistical data on actual people’s attitude. As soon as a base of such facts will be collected, it will be possible to perform more precise and sophisticated classification. Otherwise, we may apply basic general approach for text-based emotion analysis using IBM Watson NLU and Tone Analyzer[51] cognitive computing service. Services allow classification of text inputs within five emotional categories (sadness, joy, fear, disgust, and anger) and tones (e.g. polite, frustrated, sad, sympathetic, etc.) with corresponding confidence levels. Since input data is not always a text, in case of image based materials, we may apply IBM Watson Visual Recognition[52] service or other object recognition and image analysis services. Unfortunately, so far, image based emotion recognition is done on basis of detected faces and has nothing to do with other detected object. Therefore, we see elaboration of more sophisticated image emotion recognition based on detected objects and their relations as a next challenging opportunity for further research. Nevertheless, in addition to IBM Watson, there is a variety of other services[53] capable to detect emotion form text or images. Regarding the second part, we are able to recognize unavailability of explicit direct linkage of the target with the emotion-analyzed part of the material by applying RDF triple extraction with IBM Watson NLU service (as has been mentioned in the section 4a). Having the text formalized into a set of RDF statements (triples), we may assess a role of the target as an RDF subject there. Again, it becomes more challenging when we deal with images trying to formalize them in a form of RDF statements. Applying object recognition techniques we may extract entities that might be associated with RDF subjects and RDF objects of a resulting RDF document. Being able to also retrieve actions, activity and other contextual information for the image, it might be possible to combine this information with knowledge from domain ontologies and infer possible RDF predicates to finally RDF subjects and RDF objects in our RDF document. Moreover, in those cases when image contains a text, there are services capable to recognize a text in image, allowing us to perform further emotion analysis and RDF-based formalization.

Another group of propaganda techniques is associated with citation/quotation. The *Misinformation* technique involves reporting information in such a way that the final message of the story is not true, it's what the propagandist wants you to believe. Within citation, propagandist gives a half truth about someone's position, usually flops/twists it or takes it out of context it has been originally present. As a result, information it is presented in a misleading fashion. Another *Unproven "Facts"* technique is used by a writer to "prove" a position by starting to quote "studies", "reports", and "experts" as "proving" this or that, but they never mention the study's name, location (where copies can be found), or the conditions specific to the experiments. By applying *He Said She Said* technique, the authors can say something they know isn't true, or isn't fair, but they want to say it anyway. Be careful and do not mix it up with association-based technique discussed before. Here, the target is not someone else, but the propagandists themselves. And they do not want to be associated with a negative or false statement, but would like to “speak” and deliver it to the reader. Sometimes, propagandists anonymize and abstract the originator of the statement (no matter whether it is or does not exist at all). That's why they say "some people say", rather than "I say". Currently, with our tool we do not aim at fact-checking, rather we would like to increase visibility and awareness of the reader of used propaganda techniques. Therefore, initial goal is to identify the fact of citation/quotation to warn reader to pay attention to proper citing (reference to the originals); and then detect, whether

[49] https://en.wikipedia.org/wiki/Propaganda

[50] https://en.wikipedia.org/wiki/Propaganda_techniques

[51] https://www.ibm.com/watson/services/tone-analyzer/

[52] https://www.ibm.com/watson/services/visual-recognition/

[53] https://nordicapis.com/20-emotion-recognition-apis-that-will-leave-you-impressed-and-concerned

object of quotation is an abstract entity (e.g. "studies", "reports", "experts", “people”, etc.), and whether citation is emotionally negative. So, applying the same semantic formalization method and transform the document into a set of RDF statement, we are able to recognize the triples with RDF predicates semantically relevant to “citation”. For this purpose we build a corresponding set of properties including “say”, “tell”, “state”, “mention”, “write”, etc. At the same time, we may recognize “citation” triples with abstract RDF subjects – abstract entities semantically similar to "studies", "reports", "experts", “people”, etc. Alternatively, having big enough set of labeled training sample of texts classified by availability of citation, it would be also possible to apply Neural Network based classification approach incorporated in IBM Watson Natural Language Classifier[54] service.

The *Backstroke* technique assumes that propagandist systematically belittling the goals of the subject of the article (target) as the goals are being listed. For every step forward for the target, the propagandist pulls the reader back. Using sentiment analysis feature of IBM Watson NLU service, we may identify whether sentiment of certain piece of text is positive, neutral, or negative. Therefore, analyzing the patterns of sentiment modification (difference between sentiment classifications of sequential parts of the text with corresponding confidences) in the document, we can make a probabilistic conclusion about the use of *Backstroke* technique.

Another propaganda technique where we may apply emotion-oriented approach for its detection is *Over Humanization*. It is a technique to tell a story by focusing on the real people who the story impacts. At the same time it could be used to manipulate when a propagandist tries to mask an issue by making anyone who has a valid disagreement look evil due to all the human suffering talked about in the story. This technique can be used with any potential tearjerker topic. Therefore, using NLU service, we may try to detect tearjerker emotion that has high enough confidence.

Very often, when Russian government uses information warfare against own citizens, it differs from traditional forms of propaganda. In this case, aim of propaganda is not to convince or persuade, but rather to undermine. Instead of agitating audiences into some action, it seeks to keep them hooked and distracted, passive and paranoid. Very often instead of addressing internal political, economic, etc. problems of the country, all the news and talk-shows are about other countries, other nations, some opponents. It is a very popular technique among populists, when been asked a concrete question, they start to tell how the issue is going somewhere else and what others are (or are not) doing regarding the issue. Such technique could be recognized via detecting difference between subjects of a question and corresponding answer.

Another popular technique among populists when they do not want or do not know what to answer to the question is to start to talk about some general well known and accepted things making reader confused. Of course, applying simple text analytics it might not be possible to figure it out whether speaker has answered the question or not. However, we may try to recognize whether parts of the answer belong to the same topic or not. For such classification purpose we can use IBM Watson Natural Language Classifier service. If “answer” contains parts which are classified to different topics it might mean that speaker tried to confuse reader with generalized blurred “answer”. However, before we would be able to actual classification, we have to train the classifier based on labeled clusters of text samples associated with different topics.

The *Bandwagon* technique tells us to support a certain issue because, in effect, “everybody else is doing it”. In such a way author convince reader to jump on the bandwagon and share certain opinion since everybody else is there already and reader should not be left behind. It means that message will contain many similar RDF statements with different RDF subjects, as well as such generalized subject as “everybody”, “everyone”, “all”, etc. As soon as density of such statements is happened to be higher than certain set threshold, we may suspect use of this propaganda technique. Somewhat similar from the use purpose point of view to the *Bandwagon* and at the same time citation based technique is *Testimonial*. The idea behind this technique is that testimony of famous/respected people influences the viewers that admire those people. Thus, as soon as the tool detects “citation” statement with RDF subject recognized as famous person (celebrity, politician, cultural/public activist, etc.), it could be considered as an indicator for possible alert.

So far, discussed above propaganda technique detection approaches were based on internal features of the material. However, there are techniques detection of which requires comparison with other materials. Being doing (or even being just panning to do) some negative and unwelcome things, propagandist starts to blame someone in doing so. With no matter whether it is true, it is enough to tell the story with many details to make it sound realistic. As a result, there are fewer chances that people start to even think that the propagandist can do such things being sharing all the details of it. To warn reader about possibility of this technique been applied, the tool should process other alternative articled that cover the same things (activities, event, etc.) which are directly associated with the propagandist for the initial document. As soon as such alternative articled are found, reader will be warned about it. Similar approach could be applied to tackle *Not Talking at all about Something* technique. To look and sound positive for target auditory, propagandist may speak about well accepted by target majority things dropping out and hiding some other relevant to the topic things that might be harmful for his/her reputation. Therefore, as soon as the tool finds other documents where the same topic is discussed, it would be possible to compare sets of things mentioned in different document. Thus, extra things that have been discussed in other document (as well as corresponding links to the sources) could be introduced to the reader for further assessment.

C. Prototype implementation challenges

The mentioned techniques could be applied not only to text based information sources. Similarly, video and audio materials could be processed been converted to text via use of corresponding cognitive computing services (e.g. IBM Watson

[54] https://www.ibm.com/watson/services/natural-language-classifier/

Speech-to-Text[55]). Aggregating mentioned detection features, the Propaganda Barometer could be implemented in a forma of web browser plugin and inform user about suspicions highlighting corresponding part of a document. Support for further analysis of related sources with relevant explanations could be performed via corresponding widgets. Moreover, be more cognitive and user friendly to information consumer, the Propaganda Barometer can support natural language conversation implemented with IBM Watson Conversation[56], Speech-to-Text and Text-to-Speech[57] services.

In our project we were focused on feasibility study of applicability of current IBM Watson Cognitive Computing services, and on this stage did not target actual plugin implementation. Therefore, only some basic prototypes of mentioned above features (mainly associated with IBM Watson NLU service) were implemented. To analyze unstructured text for categories, concepts, keywords, semantic roles, entities, relations, as well as emotion and sentiments, NLU service uses a default general domain language model which can categorize documents into 1 083 categories and recognize up to 24 entity types, 433 entity subtypes and 53 relation types. However, the main limitations and bottlenecks we faced concern text formalization into RDF format. Based on experimental results we may conclude that default general model is not fine-tuned for some specific domains, the semantic roles analyses are not accurate enough for reliable triple extractions yet. It is still possible to improve the extraction performance by connecting customized domain oriented model in conjunction with the default model for domain-specific analyses.

In order to create a custom model for Watson NLU, IBM offers Watson Knowledge Studio[58] (WKS) - a stand-alone product that aims to better involve field-experts in the training of supervised machine learning models in order to process unstructured data. Nevertheless, creation of sophisticated domain specific language model with WKS requires comprehensive analysis of problem domain from knowledge management expert, as well as time consuming affords from domain experts to improve the model with extra supervised machine learning based facilitation.

V. Conclusions

The problem of fake content has become so global that it has turned its attention to variety of researchers, entrepreneurs and big companies aimed at combating fake content. They use different approaches to solve the problem. Some of them provide a platform for placement of guaranteed-tested content. Others try to automate the manual fact-checking. Additionally, attempt to solve this problem leads to the creation of projects aimed at global changes in the Web. Today, amount of researchers (research groups) and startups aimed at fake news resolving problem is evolving, and their solutions are rather the results of the first search, a kind of alpha version of future full-fledged products. Nevertheless, they deserve attention and study.

Based on the review of current approaches, it is still not clear that machine learning offers the best hope for near-term solutions for fact-checking in contrast to crowdsourcing that perhaps may offer greater one. Wikipedia may be the most prolific assembler of facts the world. Currently, crowdsourcing has demonstrated more short-run potential for performing accurate and flexible fact-checking. However, the nature of online news publication has changed, such that traditional vetting from potential deception is impossible against the flood arising from content generators. Therefore, elaboration of more sophisticated intelligent algorithms for automated fake detections are needed, and perhaps, the next reasonable step is to combine/merge the approaches and heavily utilize human processed crowdsourcing achievements for new generation of automated solutions.

Thus, so far, we cannot much rely on automated solutions, as well as manually processed results cannot be produced with sufficient speed and volume. Fake news and propaganda present a serious problem for democracy that relies upon an informed citizenry. And likely, there is more that today's technology giants can do to combat explicit propaganda from propagating through social networks and dominating search results. However, each citizen, every person should itself be responsible for own behavior and decisions made. Therefore, in this paper authors are focused on automated supportive learning environment that helps an average information consumer to improve own media literacy being warned of possible own biases and manipulation techniques (with some probability been applied to influence his/her mind), tool that via corresponding explanations and guidance facilitates reader's critical thinking skills development and makes him/her the most powerful shield against the Information Warfare.

In contrast to currently available fake detection plugins, which are based on manually human-processed fact-checking; Propaganda Barometer is mainly aimed at on-the-fly content analysis with automated detection of propaganda techniques and reader's critical thinking development, addressing its main principles: *biases in own thinking, reversing things, evaluation of evidences.*

Our future work will be mostly focused on improvement of unstructured data transformation into structured machine-processable form of RDF triples. For this purpose we are planning to facilitate the process by applying domain specific language model, as well as move towards automated image-based content formalization into RDF form.

References

[1] J. Aro, "This is what pro-Russia Internet propaganda feels like—Finns have been tricked into believing in lies". Kioski.yle.fi, 24 June 2015.

[2] E. Lucas and P. Pomeranzev, "Winning the Information War: Techniques and Counter-strategies to Russian Propaganda in Central and Eastern Europe". A Report by CEPA's Information Warfare Project in Partnership with the Legatum Institute, August 2016. URL: http://infowar.cepa.org/Winning-the-Information-War

[3] J. Aro, "The cyberspace war: propaganda and trolling as warfare tools". European View 15, 1 (2016), pp. 121-132.

[55] https://www.ibm.com/watson/services/speech-to-text/

[56] https://www.ibm.com/watson/services/conversation/

[57] https://www.ibm.com/watson/services/text-to-speech/

[58] https://www.ibm.com/us-en/marketplace/supervised-machine-learning

[4] P. Heymann, G. Koutrika and H. Garcia-Molina, “Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges”. IEEE Internet Computing, 11(6):36–45, November 2007

[5] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli and W. Quattrociocchi, “Science vs Conspiracy: Collective Narratives in the Age of Misinformation”. PLOS ONE, 10(2):e0118093, February 2015.

[6] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, Scientific American 284(5), pp.34-43.

[7] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

[8] M. Sharifi, E. Fink and J. G. Carbonell, “Detection of Internet scam using logistic regression”. In IEEE International Conference on Systems, Man, and Cybernetics, pages 2168–2172, October 2011.

[9] A. Ishak, Y. Y. Chen and S-P. Yong, “Distance-based hoax detection system”. In Proceedings of the International Conference on Computer Information Science (ICCIS), volume 1, pages 215–220, June 2012.

[10] M. Vukovic, K. Pripuzic and H. Belani, “An Intelligent Automatic Hoax Detection System”. In Knowledge-Based and Intelligent Information and Engineering Systems, pages 318–325. Springer, Berlin, Heidelberg, September 2009.

[11] I. Yevseyeva, V. Basto-Fernandes, D. Ruano-Ordas and J.R. Mendez, “Optimising anti-spam filters with evolutionary algorithms”. Expert Systems with Applications, 40(10):4010–4021, August 2013.

[12] B.T. Adler and L. Alfaro, “A Content-driven Reputation System for the Wikipedia”. In Proceedings of the 16th International Conference on World Wide Web, WWW ’07, pages 261–270, Banff, Alberta, Canada, 2007. ACM.

[13] J. Golbeck and J. Hendler, “Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks”. In Engineering Knowledge in the Age of the Semantic Web, pages 116–131. Springer, Berlin, Heidelberg, October 2004.

[14] X. Chen, R. Chandramouli and K.P. Subbalakshmi, “Scam Detection in Twitter”. In Katsutoshi Yada, editor, Data Mining for Service, number 3 in Studies in Big Data, pages 133–150. Springer Berlin Heidelberg, 2014

[15] J. Ito, J. Song, H. Toda, Y. Koike and S. Oyama, “Assessment of Tweet Credibility with LDA Features”. In Proceedings of the 24th International Conference on World Wide Web, WWW ’15 Companion, pages 953–958, New York, NY, USA, 2015. ACM

[16] N. Conroy, V. Rubin and Y. Chen, “Automatic Deception Detection: Methods for Finding Fake News”. In Proceedings of ASIS&T2015, At St. Louis, MO, USA, 2015.

[17] E. Tacchini, G. Ballarin, V.M. Della, S. Moret and L. Alfaro, “Some Like it Hoax: Automated Fake News Detection in Social Networks”. Technical Report UCSC-SOE-17-05, School of Engineering, University of California, Santa Cruz, 2017. arXiv:1704.07506 [cs.LG]

[18] K. Shu, S. Wang, A. Sliva, J. Tang and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective”. ACM SIGKDD Explorations Newsletter. 19. . 10.1145/3137597.3137600.

[19] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan and J. Han, “A survey on truth discovery”. ACM Sigkdd Explorations Newsletter,17(2):1–16, 2016.

[20] S. Mukherjee and G. Weikum, “Lever-aging joint interactions for credibility analysis in news communities”. In CIKM’15

[21] G. Weikum, 2017. “What computers should know, shouldn’t know, and shouldn’t believe”. In WWW’17.

[22] R.L. Cilibrasi and P. M.B. Vitanyi, The Google Similarity Distance. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, March 2007, 370–383.

[23] J. Combs and D. Nimmo, “The New Propaganda: The Dictatorship of Palavar in Contemporary Politics”. New York: Longman Publishing Group, 1993.

Oleksiy Khriyenko (1981) obtained Engineer’s degree in Computer Science (Intelligent Decision Support Systems) in 2003 from the Kharkov National University of Radioelectronics, Ukraine. Later, Oleksiy Khriyenko obtained a Master’s degree in Mobile Computing from MIT department (University of Jyväskylä, Finland). Since 2008 he is Ph.D. from the same department. His research interests include: Artificial Intelligence, Deep Learning and Cognitive Computing, Semantic Web and knowledge engineering, multi-agent systems, Web of Things and ubiquitous services, context-sensitive adaptive environments, etc. Currently, Oleksiy Khriyenko does research, lecturing and is involved in management of international master programs (WISE, COIN) at IT faculty, University of Jyväskylä. (http://users.jyu.fi/~olkhriye)