https://doi.org/10.31449/inf.v42i4.1132 Informatica **42** (2018) 527–533 **527**

Automatic Estimation of News Values Reflecting Importance and Closeness of News Events

Evgenia Belyaeva*, Aljaž Košmerlj, Dunja Mladenić* and Gregor Leban
Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Email addresses: firstname.lastname@ijs.si

*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

Keywords: news values, newsworthiness, text mining, Apple

Received: January 27, 2016

This paper addresses a problem of automatic estimation of three journalistic news values, more specifically frequency, threshold and proximity, by applying various text mining methods. Although theoretical frameworks already exist in social sciences that identify if an event is newsworthy, these manual techniques require enormous amount of time and domain knowledge. Thus, we illustrate how text mining can assist journalistic work by finding news values of different international publishers across the world. Our experiments both on a collection of news articles from different publishers about Apple's launch of new iPhone 6 and Apple Watch and on a wider collection of documents confirm that some journalists still follow some of the well-known journalistic values. Furthermore, we acknowledge that news values are often orthodox and outdated, and no longer apply to all publishers. We also outline possible future implications of our approach to work on interaction between text mining and journalistic domains.

Povzetek: Članek obravnava problem avtomatske ocene novičarskih vrednosti, natančneje: pogostosti, prag pomembnosti ter bližine (oz. relevance), z uporabo različnih metod avtomatske analize teksta. Čeprav v družboslovju obstaja več teoretičnih ogrodij, ki določajo ali je nek dogodek vreden poročanja, te temeljijo na »ročnem« delu in zahtevajo veliko časa ter globoko poznavanje domene. V članku nakažemo, kako lahko avtomatska analiza teksta pomaga pri novinarskem delu z uvidom v novičarske vrednosti na globalnem nivoju in med velikimi, mednarodnimi mediji. Rezultati naših poskusov na zbirki člankov iz različnih virov (spletnih časopisov) o splovitvi izdelkov iPhone 6 in Apple Watch podjetja Apple potrjujejo, da novinarji še sledijo nekaterim uveljavljenim novičarskim vrednostim. Ob tem ugotavljamo, da so nekatere novičarske vrednosti preveč ortodoksne in zastarele ter ne veljajo več za vse novičarske vire. Na koncu orišemo načrtovano nadaljnje delo in možne implikacije uporabe avtomatskih metod analize teksta na novinarsko delo.

1 Introduction

Every news outlet has a different agenda for selecting which news stories to cover. Mass media have traditionally relied on the so-called news values to evaluate newsworthiness of a story i.e. what to publish and what to leave out, introduced firstly by two Norwegian scholars Galtung and Ruge (Galtung & Ruge, 1965). News values are certain guidelines to follow in producing a news story, so-called ideological factors in understanding decisions of journalists (Cotter, 2010). The more news values (12 in total) are present in a piece of information, the more likely that you will see the story featured in different mass media across the globe.

In the last years there has been a growing interest to work on the intersection of social and computer sciences (Greening, 2000). Text mining is emerging as a vital tool for social sciences and the trend will most likely increase (Amolfo & Collister, 2015). Due to the abundance of news information and with the advances in text mining, it is now possible to help journalists to process information in every day job and at the same time to prove old media theories and to discover old, often biased patterns in the news across the world.

Some research has been already done on detecting news bias (Ali & Flaounas, 2010), (Flaounas & Turchi, 2010), less attention has been paid to automatic detection of news values (De Nies & D'heer, 2012), (Al-Rawi, 2017). We argue that in order to understand and automatically detect news bias, it is first important to understand the logic of news selection processes (news values) and try to detect news values on a large-scale.

We make a first attempt to automate the detection of initially three news values by applying several text mining techniques from selected publishers and when reporting about Apple Corporation. Apple has a great impact on our lives and as any technology it has become newsworthy almost by default. Our goal is to distinguish if the theory of newsworthiness by Galtung and Ruge (Galtung & Ruge, 1965) is still a valid approach to predict news selection values. If yes, what are the relevant news criteria we find in our experiments and do we see some interesting recurring patterns in the news?

2 Data description

News articles analyzed in this paper were first aggregated by the Newsfeed and then analyzed by Event Registry[1] – a global media monitoring service that collects and processes articles from more than 100.000 news sources globally in more than 10 languages (Leban & Fortuna, 2014).

We extracted news about the Apple Corporation (iPhone 6 and Apple Watch launch) from 16 selected online outlets during the period of 01.09.2014 – 31.10.2014. We considered the occurrences of the company name Apple and the terms "iPhone" and "Apple Watch". The time range corresponds to the announcements of the launch of the two above-mentioned products and the start of sales. We also considered to include two months of coverage and to check the news about Apple before and after the big events in order to control for variation in media interest in the company. The time range of complete, uninterrupted two months is important because we are also interested in the extent to which editors and journalists write about Apple and how the coverage changes in case if a greater event. The sources under our analysis correspond to the most influential daily news websites, easily accessible, widely read in the following three languages: English (EN), German (DE) and Spanish (ES).

Apple represents an important context for the purpose of this study. The more new and novel ideas and products come from Apple, the more space will be allotted to it by the international media.

The Table 1 summarizes the total number of events and the total number of articles reporting on Apple collected and analyzed per publisher during the above-mentioned period including the information on the headquarters of each publisher. All journals listed in the Table 1 publish daily.

Important to note is that the websites were also selected with the purpose to cover different geographical places (Europe and the USA) in order to identify one of the three news values, i.e. proximity – geographical or cultural proximity of the event to the source. The core available piece of information for each article for our experiment included the date, the location of the event and the location of the publishers' headquarters, as well as the size of the events (i.e. the number of articles about them).

Publisher	Total No. Events	Total No. Articles	Headquarters
The Next Web	1064	1670	Amsterdam
Gizmodo	2007	3911	New York
The Guardian	14299	19997	London
BBC	15582	23852	London
USA Today	7692	13629	Tysons Corner
Wall Street J.	7197	18837	New York
Heise	4194	2190	Hannover
Chip online	907	1212	Munich
Stern	4194	10092	Hamburg
Die Zeit	3722	5600	Hamburg
Die Welt	14683	30359	Berlin
Der Spiegel	2261	2759	Hamburg
El Mundo	6707	8705	Madrid
ABC.es	7431	10388	Madrid
El Pais	686	979	Madrid
El Dia	6700	12752	Barcelona

Table 1: Publishers and Totals of Events/Articles on Apple.

3 Socio-cognitive perspective on news values

One of the most critical questions in the media research field is why certain aspects of reality are selected by journalists and eventually registered to become news. The news selection process is a very complex process influenced by economic, political, organizational and social factors. Over the years the most used explanation of the phenomena prevails to be the so-called theory of newsworthiness, which focuses on explaining the logic of the journalists and media organizations and on predicting what will most likely strike attention of the audience and be selected as news. News selection, i.e. selecting novel pieces of reality, is not purely a journalistic problem; it has its roots in psychology of perception and cognitive science. Looking for new information in order to reach an optimal level of stimulation is fundamental to human behavior (Martindale, 1981). Humans constantly search for arousal or stimulation, often driven by pleasure centers of the brain (Martindale, 1981), (Donohew, Sypher, & Higgins, 1988).

According to Van Dijk, the notion of news values is part of the social cognition since news values are shared by journalists and even by the public of the mass media in an indirect way. They "provide cognitive basis for decisions about selection, attention, understanding, representation, recall and the uses of news information." (Van Dijk, 2009). News audience is not only a vital part of the cognitive processes to label their psychological activation, but the arousal itself is the end product of cognitive process that often occurs automatically. Journalists only need to select an important piece of information and code it in a nice fashionable way so that it "facilitates recognition and heightens the impact of these cognitive processes" (Martindale, 1981). For most journalists, news criteria are something very physical, something that is always present in the back of mind and is an integral part of themselves (Schulz, 2007). These criteria are often unconscious in journalistic practice.

News values are considered to be ground rules or "distillation" of what an identified audience is interested in reading or listening (Richardson, 2007). The most influential contribution came from Galtung and Ruge who underlined a list of 12 news factors, which they divided into eight "culture-free" factors and four "culture-dependent" factors. The following Table 2 outlines Galtung and Ruge's theory of news selection

[1] http://eventregistry.org

News Values	**Short Description**
Frequency	***Time span of an event***
Threshold	***The size of an event***
Meaningfulness/Proximity	***Geographical closeness***
Unambiguity	Clarity of the meaning
Consonance	Conventional expectations of the audience
Continuity	Continuous over time
Unexpectedness	Unplanned/unexpected
Composition	Include other pieces of information
Reference to elite nations	Relate to famous nations
Reference to elite people	Relate to famous people
Negativity	Bad news, conflict-oriented
Personalization	Action of individuals

Table 2: News Values by Galtung and Ruge.

and its news values (Galtung & Ruge, 1965). Most of the news values have a common-sense perception and are simple to understand. The more an event satisfies the bellow-mentioned criteria, the more likely it will be reported in the news.

Later on there have been several attempts to revise the theory of news worthiness (MacShane, 1979), (Harrison, 2006), (Harcup & O'Neill, 2017) despite the existence of several new lists of news values, the theory by Galtung and Ruge has not really been challenged and is still taught at many Journalism schools across the globe. Various scholars from the social sciences field have discussed the theory of news values extensively. However, technologies have now added new "shades" of how we want to test the theory on a larger scale and if and of how the news values have changed over time.

4 Mining news values

Galtung and Ruge originally came up with a taxonomy of 12 factors, but due to the space limitations, the goal of this work is to identify automatically the first three news values: *frequency*, *threshold* and *proximity*. Frequency and threshold are both impact criteria, calculated through the number of articles per publisher (frequency) and a number of articles per events (threshold), whereas, the proximity criterion is considered more about the audience identification and geographical distance. The above-mentioned news values are convenient for the analysis by text mining methods and represent a starting point of a complete framework of automatic detection of 12 news values.

4.1 Frequency

Frequency as news value refers to the time-span of an event (Galtung & Ruge, 1965). For example, a single event on a certain day is more likely to be reported rather than a long process (Fowler, 1991). Since Apple has become a new religion of the 21st century, it is newsworthy by default and news about Apple exists in most outlets around the world on daily basis. In this paper, we understand frequency value as the frequency of all articles from the selected publishers mentioning iPhone 6 and Apple Watch (also known as Watch) respectively. We are interested in finding trends or particular patterns among publishers during the selected period of time. The Figure 1 summarizes the time distribution (i.e. frequency value) of the mentions related to iPhone 6.

The frequency measurement experiment indicates that there are two sudden busts in frequency among certain publishers; one peak corresponds to the announcement of the Apple Watch launch on the 9th of September 2014 and the second peak being the announcement of the iPhone 6 release on the 19th of September 2014. Both announcements received a much bigger coverage (especially, among the following publishers: Wall Street Journal, Stern Magazine and die Welt) in respect to the actual start of sales of the products at the end of October. Applying the logic of the media in our context – announcement of the launches and of the start of sales – announcement is news since the audience did not know when Apple would launch the products and announce the official sales date. The surprise is even larger if the organization is Apple, which is perceived to be the trend maker in technology in general.

The frequency distribution of new Apple Watch has a similar to iPhone 6 trend, having, however, less coverage (number of articles) per publisher, per day. The following Figure 2 outlines the frequency of Watch coverage among the selected publishers during 01.09.2014 – 31.10.2014. The two bursts are also visible in the coverage of the Watch due to the fact that journalists are more likely to complement news pieces with additional, background information (Bell & Garrett, 1998), in our case if a journalist is writing about the Apple Watch he is likely to mention Apple and another Apple product.

We also measured frequency between specific tech publishers in comparison to other international publishers. Our first assumption that technology-oriented outlets do publish more news with higher frequency (number of articles) on Apple was not confirmed as also seen from the Table 1 and Figure 2. This partially could be explained by the small size of editorial and journalistic teams working for tech publishers in comparison to big media corporations with journalists all over the world, for example, BBC, Wall Street Journal or USA Today.

4.2 Threshold

The threshold criterion often refers to the impact of an event and its effect on the readers, i.e. a size needed for an event to become news, for example, thousands of people buying a new iPhone 6 and not just one person buying it in a small local store will get more attention of

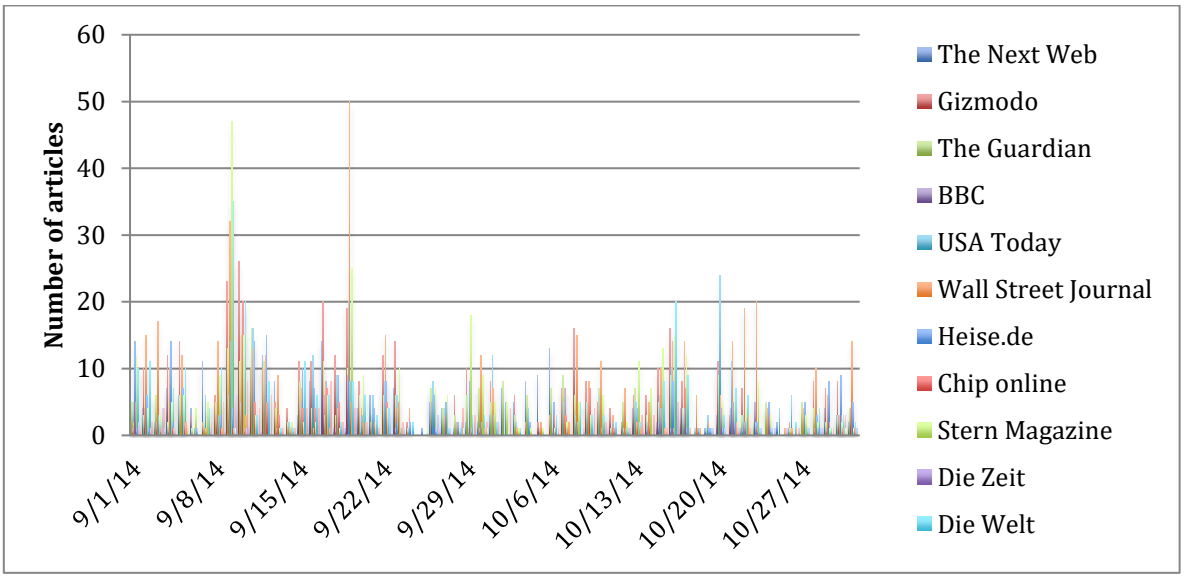

Figure 2: iPhone 6 Frequency distribution.

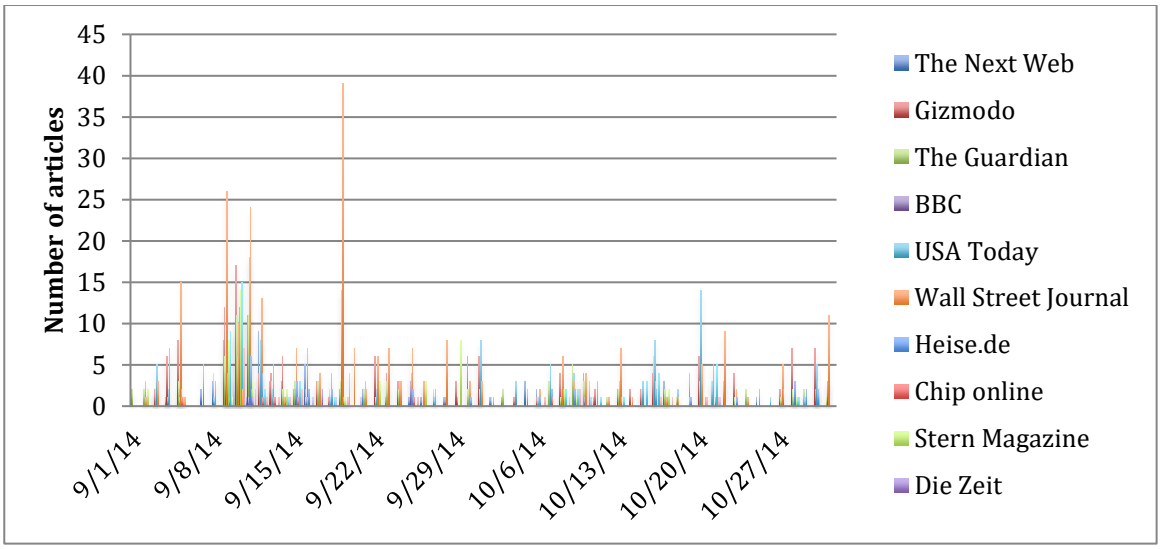

Figure 1: Apple Watch Frequency distribution.

the media. In other words, the bigger, the better, and the cooler a product is, the higher the "amplitude" and the more fuss it will create in the media.

It is indeed difficult to measure something that should have a larger effect on the readers. We understand that events can meet this threshold value either by being large in absolute terms i.e. having a higher frequency or an increase in reporting of a topic. In this experiment, we decided to look at the size of events among the selected publishers without limiting our search to reports about Apple in order to take in more data. The main reason we limit ourselves to Apple related stories in frequency analysis is that we can manually show that remarkable and frequent events like a new product launch draw more media attention.

An event is understood as a group of articles that are clustered to report on the same issues in the world (Leban, Kosmerlj, & Belyaeva, 2014). Our assumption is that a single article might not be very informative, but a group of articles on a certain issue, which is picked up by more publishers can form a part of a bigger story with more impact on the readers and thus match the threshold value. Note that frequency and threshold values are both impact criteria, we see threshold as the size of an event, whereas frequency should also be understood as events unfolding within production cycle of a news media and will be reported on repeatedly.

Therefore, for the threshold analysis we aim at capturing the size of clusters (number of articles of all publishers in event clusters) and assume to witness a greater number of articles that form an event. To note that news articles are first aggregated by the Newsfeed service[2] - a real-time stream of articles from more than 100.000 RSS-enabled websites in several major world languages, then we process the articles by a linguistic and a semantic analysis pipeline that provides semantic

[2] http://newsfeed.ijs.si

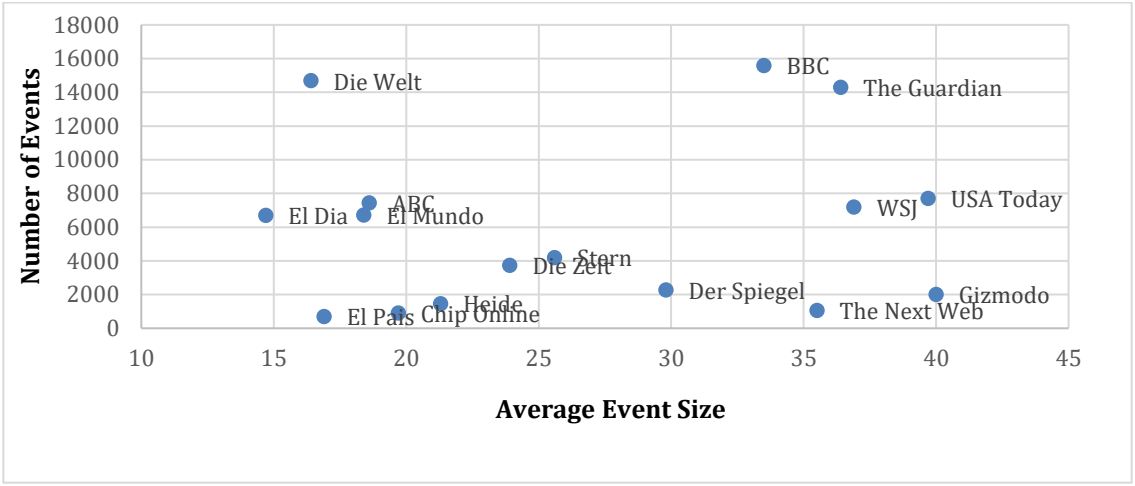

Figure 3: Threshold analysis per publisher

annotations. The semantic annotation tool developed within XLike project comprises three main elements: *named entity recognition* based on corresponding Wikipedia pages, *Wikipedia Miner Wikifier* – detecting similar phrases in any document of the same language as Wikipedia articles and *cross-lingual semantic analysis* that links articles by topics (Carreras & Padro, 2014). The analysis and clustering are then done by the Event Registry. The data in the following scatterplot shows the average event size per publisher during the same period of time and confirms our hypothesis: the higher the threshold (number of articles per event), the greater the impact of a publisher (i.e. The Guardian, BBC), more intense and more frequent the coverage about an event is.

If an article is written by an influential publisher other bigger and smaller publishers will most likely pick the event up and eventually it will form an event (cluster). Interestingly, Spanish and German publishers have a smaller average event size, which could be explained as those publishers are more interested in local events or events within their countries of origin.

4.3 Proximity

The Proximity value (also often referred to as Meaningfulness factor) corresponds to physical i.e. geographical or cultural (in terms of religion or language) closeness of a news story to a listener and thus the media publisher. Proximity helps readers to relate to a story on a more personal, familiar level. It can change over time and is open to subjective interpretations. However, proximity might also mean an emotional (fear, happiness, pride tec.) trajectory in the audience's eyes, regardless of where it takes place (Schults, 2005). An event may happen in a distant place but still be of interest in terms of a certain relevant meaning to the reader: Apple is an American company with headquarters in California, but as the largest technology company its products are praised and sold in every part of the world.

Our assumption when measuring this journalistic value is that the closer the geographical location of the story to the news publisher is, the more frequent and more intense (also higher threshold) the coverage is.

Event location detection is done automatically in the Event Registry in the following way: we first try to identify a dateline in the article (a piece of text at the beginning of every article) that names a location: we assume that this is the location of the event. When the dateline does not appear in the article, we check the Event Registry to use the event's location. A classification algorithm that considers all articles belonging to the same event determines it. In some cases, the Event Registry does not determine the location; we omit such cases in our analysis.

The headquarters of each publisher was manually searched for on the official websites of the selected publishers and Wikipedia[3]. We did not limit our search to the stories reporting about Apple since we assume to see some recurring proximity patterns of the selected publishers in spite of the story topic.

Since our system was not able to automatically identify location of all events, we use only a sub-selection of our data for each publisher for which we compute the distance in kilometers and calculate how many of them report from the same country and same city where the publisher is. The following Table 3 outlines the proximity experiment results: Total number of sub-Selection of Events where Country/City were detected and a Total number of events where publishers reported either on the same country or the same city where a publisher has headquarters.

It has been found that the coverage of most publishers is not local, they do not report on the events close to their headquarters: it can be explained by the fact that the selected publishers are not local publishers and are no longer "national" publishers, but some have become over time "international" outlets whose news is read across the world. Interesting to note that proximity value was not confirmed for, for example, The Next Web – technology oriented website with headquarters in Amsterdam, Netherlands, reported only once on the

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

Publisher	Country sub-Selection	Same country	City sub-selection	Same city
The Next Web	178	1	174	1
Gizmodo	***371***	***204***	370	15
The Guardian	6563	2261	6510	462
BBC	7105	2909	7039	438
USA Today	***4299***	***2842***	4291	0
WSJ	3091	1194	1074	122
Heise	***586***	***262***	585	5
Chip online	211	71	211	1
Stern	***2704***	***1197***	2701	90
Die Zeit	***2505***	***1103***	2504	95
Die Welt	***9185***	***5340***	9182	1248
Der Spiegel	1592	630	1590	55
El Mundo	***4077***	***2269***	4076	861
ABC	***4493***	***2372***	4491	879
El Pais	337	46	337	7
El Dia	***3789***	***2399***	3785	154

Figure 4: Geographical proximity analysis per publisher.

events from their headquarter city. Whereas, some selected publishers, mainly Spanish and German (in italics in Table 3) dedicate more or less half of their attention to the news from the same country, which confirms relatively strong the proximity value. Not surprisingly, the Guardian, the BBC and the Wall Street Journal do not support journalistic proximity value since their geographical scope is scattered around the world. Publishers and some figures in italic letters in the Table 3 represent newspapers that confirm the proximity value that is they report on the events that happen in the same country of their headquarters in respect to the remaining publishers under selection.

5 Discussions and future work

We made an initial attempt to automate detection of journalistic news values, in particular, *frequency* in the context of Apple news, *threshold* and *proximity* in the context of selected publishers. We believe that using text-mining methods is an essential step of interaction between social and computer sciences approaches. This hybrid approach will not only help journalists in their everyday work, but it will also potentially help to identify various ideological patterns or news bias of various global publishers.

The study tested some news values from the theory of newsworthiness formulated by Galtung and Ruge. The assumption that a product launch is likely to get more coverage and thus meet the frequency value was confirmed for most outlets. No assumptions were confirmed on technology-oriented publishers writing more frequently and more intensely on Apple. No influences were found for the proximity news factor in case of international publishers; however, the proximity factor was confirmed for most Spanish and German publishers. These findings suggest continuing exploring the topic in depth.

Future work will include developing our framework further, which will automate the process of assessing newsworthiness of all 12 news values applied to different languages, as well as to different domains like conflicts, natural disasters, political crises etc. By detecting news values through text mining we also aim at confirming still existing ideological patterns i.e. news slant or bias of different publishers. Research designed more specifically and comprising automation of all news values could provide more answers to the problems of outdated and orthodox new values that keep on contributing to the news bias. To our knowledge, there are no automated systems to compare our approach with, thus, in the future we also plan on conducting several evaluations including manual to verify our results.

6 Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under XLike (ICT-STREP-288342) and XLime (FP&-ICT-611346).

7 References

[1] Ali, O., & Flaounas, I. (2010). Automating News Content Analysis: An Application to Gender bias and Readibility. *JMLR: Workshop and conference Proceedings.*

[2] Al-Rawi, A. (2017). News values on social media: News organizations' Facebook use. *Journalism , 18* (7), 871-889. https://doi.org/10.1177/1464884916636142

[3] Amolfo, L., & Collister, S. (2015). Text mining and Social Media: when Quantative Meets Qualitative, and software meets human. In P. Halfpenny, & R. Procter, *Innovations in Digital Research Methods.* London: Sage.

[4] Bell, A., & Garrett, P. (1998). *Approaches to Media discourse.* Oxford: Blackwell Publishers.

[5] Carreras, X., & Padro, L. (2014). XLike project language analysis services. *Proceedings of EACL'14*, (pp. 9-12). https://doi.org/10.3115/v1/E14-2003

[6] Cotter, C. (2010). *News Talk. Investigatingthe Language of journalism.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811975

[7] De Nies, T., & D'heer, E. (2012). Bringing Newsworthiness into the 21st Century. *Web of Linked entities Workshop, ISWC*, (pp. 106-117). Boston.

[8] Donohew, L., Sypher, H., & Higgins, T. (1988). *Communication, Social Cognition, and Affect.* London: Psycology Press.

[9] Flaounas, I., & Turchi, M. (2010). The Structure of the EU Mediasphere. *PLoS ONE , 5* (12). https://doi.org/10.1371/journal.pone.0014243

[10] Fowler, R. (1991). *Language in the News. discourse and Ideology in the Press.* London: Routledge.
[11] Galtung, J., & Ruge, M. (1965). Structuring and Selecting News. *Journal of International Piece Studies , 2* (1), 64-91.
[12] Greening, T. (2000). *Computer Science Education in the 21st Century.* New York: Springer. https://doi.org/10.1007/978-1-4612-1298-0
[13] Harcup, T., & O'Neill, D. (2017). What is news? News values revisited (again). *Journalism studies , 18* (12), 1470-1488. https://doi.org/10.1080/1461670X.2016.1150193
[14] Harrison, J. (2006). *News.* New York: Routledge.
[15] Leban, G., & Fortuna, B. (2014). Event Registry - Learning About World Events from News. *WWW*, (pp. 107-111).
[16] Leban, G., Kosmerlj, A., & Belyaeva, E. (2014). News reporting bias detection prototype. XLike Deliverable D5.3.1.
[17] MacShane, D. (1979). *Using the Media: how to deal with the press, television and radio.* London: Pluto.
[18] Martindale, C. (1981). *Cognition and consciousness.* Homewood, IL: Dorsey.
[19] Richardson, J. (2007). *Analysing Newspapers. An Approach from Critical discourse Analysis.* New York: Palgrave MacMillan. https://doi.org/10.1007/978-0-230-20968-8
[20] Schults, B. (2005). *Broadcast News Producing.* London: Sage Pubblications.
[21] Schulz, I. (2007). The Journalistic Gut Feeling. In I. Schulz, *Journalism Practice* (pp. 190-207). London: Routledge.
[22] Van Dijk, T. (2009). *News as Discourse.* New York: Routledge.