The Pursuit of Journalistic News Values through Text Mining Techniques

Evgenia Belyaeva, Aljaž Košmerlj, Dunja Mladenić

Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
Email addresses: firstname.lastname@ijs.si

Abstract

The paper addresses a problem of pursuit of journalistic news values, more specifically *frequency, threshold and proximity* through various text mining methods is presented. We illustrate how text mining can assists journalistic work by finding ideological, often orthodox news values of different international publishers across the world news that contribute to ubiquitous news bias. Our experiments on selected publishers and on news about *Apple's launch of new iPhone 6* and *Apple Watch* confirm that journalists still follow some of the well known journalistic values.

Keywords: News values, newsworthiness, text mining, Apple

1. INTRODUCTION

Every news outlet has a different agenda for selecting which stories to cover and publish. Mass media have traditionally relied on the so-called news values to evaluate the newsworthiness of a story i.e. what to publish and what to leave out, introduced firstly by Galtung and Ruge (1). News values are certain guidelines to follow in producing a news story, so-called ideological factors in understanding decisions of journalists (2). The more news values are present in a story, the more likely that you will see the story featured in different mass media. It is a known and well-studied problem that old system of news values contributes to the ubiquitous media bias.

In the last years there has been a growing interest to work on the intersection of social and computer sciences (3). Text mining is emerging as a vital tool for social sciences and the trend will most likely increase (4). 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 prove old media theories and discover old biased patterns in news across the world.

Some research has been already done to detecting news bias (5, 6), but very little attention paid to automatic detection of news values (7). We argue that in order to understand and detect automatically news bias, it is first important to explain and try to detect news values.

We make a first attempt to automate the detection of several news values by applying various text mining techniques from selected publishers and when reporting about *Apple Corporation* since it has a great impact on our lives and as any technology it is newsworthy by default. Our goal is to distinguish if the theory of newsworthiness by Galtung and Ruge is a valid approach to predict news selection values and to see some interesting recurring patterns in the news.

2. DATA DESCRIPTION

News articles analysed in this paper were first aggregated and processed by the Event Registry[1] - global media monitoring service that collects and processes articles from more than 100.000 news sources globally in more than 10 languages (8).

We extracted news about the Apple Corporation (***iPhone 6*** and ***Watch*** launch) from 16 selected online outlets during the period of 01.09.2014 – 21.10.2014. The time range corresponds to the announcements of the launch of the two above-mentioned products and the start of sales. 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).**

The Publisher	Total Nr. Events	Total Nr. Articles on Apple	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.de	4194	2190	Hannover
Chip online.de	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/Article on Apple

[1] http://eventregistry.org

The Table 1 summarizes the total number of events and the total number of articles reporting on Apple collected and analysed per publisher during the above-mentioned period including the information on the headquarters of each publisher.

Important to note that the websites were selected to cover different geographical places (EU plus USA) in order to identify one of the 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 size of events (i.e. number of articles about them).

3. MINING NEWS VALUES

Galtung and Ruge originally came up with a taxonomy of 13 news values (1), but due to the space limitations, the goal of this work is to identify the first three news values: *frequency, threshold* and *proximity*. Frequency and threshold are impact criteria, calculated through the number of articles per publisher (frequency) and the number of articles per events (threshold), whereas, the proximity criterion is rather about audience identification and geographical distance.

The following Table outlines Galtung and Ruge's theory of news selection and its news values (1).

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

Table 2. News Values by Galtung and Ruge

3.1 Frequency

Frequency as news value refers to the time-span of an event (1). Since Apple has become a new religion of the 21st century, it is newsworthy by default and news about it exists in most outlets around the world. In the experiment, we have analysed the frequency of all articles from the selected publishers mentioning *iPhone* and *Apple 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 mentions related to iPhone 6.

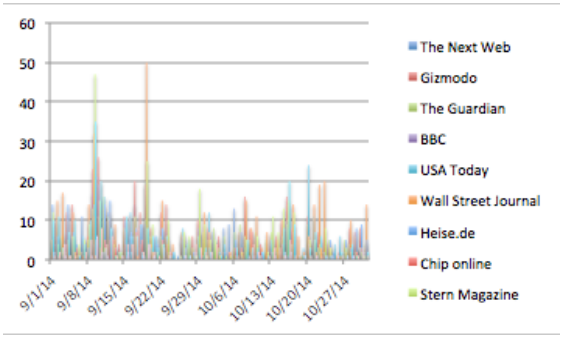

Figure 1. iPhone 6 Frequency Distribution

The frequency measurement experiment indicates that there are two sudden busts in frequency among certain publishers, one corresponding to the announcement of launch *Watch* on the 9th of September 2014 and the second one being the announcement of the *iPhone 6* release on the 19th of September 2014. Both announcements received a much bigger coverage (especially, 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.

The frequency distribution of new *Apple Watch* has a similar to *iPhone 6* trend, having, however, less coverage per publisher, per day. The following Fig. 2 outlines the frequency of *Watch* coverage among the selected publishers during 01.09 – 31.10.2014. The two bursts are also visible in the coverage of *Watch*. It can be explained by journalistic standards to include background information to a story i.e. writing about *Watch* a journalist is likely to mention *iPhone* or simple the Apple Corporation.

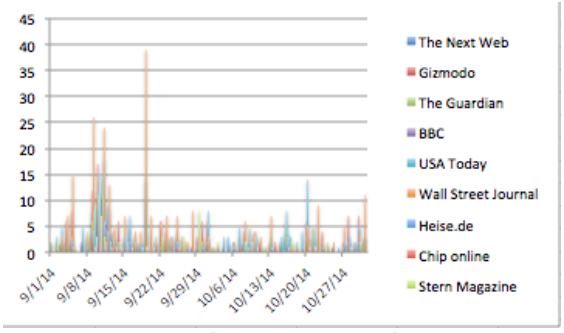

Figure 2. Watch Frequency Distribution

We also measured frequency between specific tech publishers to other international publishers. Our assumption that tech publishers do publish more news with higher frequency on Apple was not confirmed as also seen from Table 1. This partially could be explained by the small size of journalistic teams working for tech publishers in comparison to big media corporations with journalists all over the world like BBC or USA Today.

3.2 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, e.g. thousands of people buying new *iPhone 6* and not just one person buying it in a small local store.

It is indeed difficult to measure something that should have a larger affect on the readers. We understand that events can meet this threshold value either by being large in absolute terms or by 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 infrequent events like new product launches draw more media attention.

Event is understood as a group of articles that are clustered to report on the same issues in the world (9). Our assumption is that a single article might not be always 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 match the threshold value. Note that frequency and threshold values are both impact criteria, threshold is more about the size of an event, whereas frequency should be also understood as events unfolding within the production cycle of a news media and will be reported on repeatedly.

Therefor, 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 JSI Newsfeed[2] – real-time stream of articles from more than 1900 RSS-enabled websites in several major world languages, then we process the articles by a linguistic and semantic analysis pipeline that provides semantic 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 (10).

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), the 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 it up and eventually it will form an event. 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.

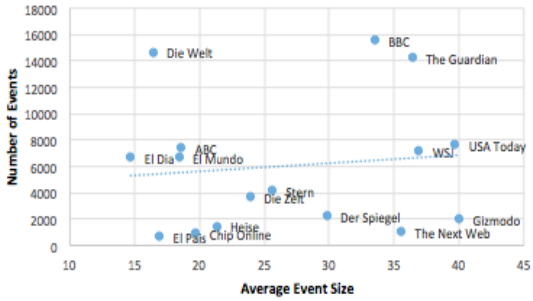

Figure 3. Threshold analysis per publisher

3.3 Proximity

The Proximity value corresponds to physical i.e. geographical or often cultural (in terms of religion or language) closeness of a news story to a media publisher (1). Proximity helps readers to relate to a story on a more personal level. It can change over time and is open to subjective interpretations. However, proximity might also mean an emotional (fear, happiness, pride etc.) trajectory in the audience's eyes, regardless of where it takes place (11).

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 the more intense (higher threshold) the coverage is.

Event location detection is done automatically in 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 try to avoid such cases in our analysis.

The headquarters of each publisher was manually searched for on the official websites of the selected publishers. 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 a story kind.

Since our system was not able to automatically identify location for all events, we use only a sub-selection of our data for each publisher for which we compute the distance in kilometres and calculate how many of them report from the same country and same city where the publisher is. The following Table 4 outlines the proximity experiment results: Total number of sub-Selection of events where Country/City were detected and total

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

number of events where publishers reported either on the same country or the same city where a publisher has headquarters.

Publisher	Total Nr. Country Sub-Selection	Same country	Total Nr. City Sub-Select.	Same city
The Next Web	178	1	174	1
Gizmodo	***371***	***204***	370	15
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	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

Table 4. Geographical proximity analysis per publisher

It has been found that the coverage of most publishers is not local, they do not report on the events close to the headquarters; it can be explained by the fact that the selected publishers are not local publishers and are considered to be the most read outlets in each country and some even in 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 events from Amsterdam and the Netherlands. Whereas, some selected publishers, mainly Spanish and German outlets, (*in italics*) dedicate more or less half of their attention to the news from the same country, which confirms relatively strong proximity news value. Not surprisingly, The Guardian, BBC and the World Street Journal do not support journalistic proximity value since their geographical scope is scattered around the world.

4. DISCUSSIONS AND FUTURE WORK

We made an initial attempt to automate detection of journalistic 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.

Future work will include developing our framework, 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 bias of different publishers. Research designed more specifically and comprising automation of all values could provide more answers to the problems of out-dated and orthodox news 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 evaluation to verify our results.

Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under XLike (ICT-STREP-288342) and XLime (FP7-ICT-611346).

REFERENCES

(1) Galtung, J., Ruge, M. (1965): Structuring and Selecting News. In: Journal of International Piece Studies, In: Journal of International Piece Research, 2 (1), pp. 70-71.

(2) Cotter, C. (2010): News Talk. Investigating the Language of Journalism. Cambridge: Cambridge University Press.

(3) Greening, T. (Ed.) (2000): Computer Science Education in the 21st Century. Springer New York.

(4) Ampofo, L., Collister, S. et al. (2015): Text Mining and social Media: When Quantative Meets Qualitative, and software meets human. In: Halfpenny, P. and Procter, R. (eds.) Innovations in Digital Research Methods. London: Sage.

(5) Ali, O., Flaounas, I., De Bie, T., et al. (2010): "Automating News Content Analysis: An Application to Gender Bias and Readability" JMLR: Workshop and conference Proceedings 11.

(6) Flaounas, I., Turchi, M., et al. (2010) "The Structure of the EU Mediasphere" PLoS ONE. Vol.5, Issue 12.

(7) De Nies, T., D'heer, E., et al. (2012): Bringing Newsworthiness into the 21st Century. In: Proceedings Web of Linked entities Workshop, ISWC. Boston, pp. 106-117.

(8) Leban, G., Fortuna B., Brank J., Grobelnik M., Event Registry – Learning About World Events From News, WWW 2014, pp. 107-111.

(9) Leban, G., Košmerlj, A., Belyaeva, E. et al. (2014): News reporting bias detection prototype. XLike Deliverable D5.3.1.

(10) Carreras, X., Padró, L., et al. (2014): XLike project language analysis services. In: Proceedings of EACL'14: demos, pp. 9-12.

(11) Schults, B. (2005): Broadcast News Producing. Sage Publications, London.