Bringing Newsworthiness into the 21st Century

Tom De Nies[1], Evelien D'heer[2], Sam Coppens[1], Davy Van Deursen[1], Erik Mannens[1], Steve Paulussen[2], and Rik Van de Walle[1]

[1] Ghent University - IBBT - ELIS - Multimedia Lab
{tom.denies,sam.coppens,davy.vandeursen,erik.mannens,rik.vandewalle}@ugent.be
[2] Ghent University - IBBT - MICT
{evelien.dheer,steve.paulussen}@ugent.be

Abstract. Due to the ever increasing flow of (digital) information in today's society, journalists struggle to manage and assess this abundance of data in terms of their newsworthiness. Although theoretical analyses and mechanisms exist to determine if something is worth publishing, applying these techniques requires substantial domain knowledge and know-how. Furthermore, the consumer's need for near-immediate reporting significantly limits the time for journalists to select and produce content. In this paper, we propose an approach to automate the process of newsworthiness assessment, by applying recent Semantic Web technologies to verify theoretically determined news criteria. We implemented a proof-of-concept application that generates a newsworthiness profile for a user-specified news item. A preliminary evaluation by manual verification was performed, resulting in 83.93% precision and 66.07% recall. However, further evaluation by domain experts is needed. To conclude, we outline the possible implications of our approach on both the technical and journalistic domains.

Keywords: News, Newsworthiness, Semantic Web, Relevance, Named Entity Recognition

1 Introduction

Every day, journalists have to choose from numerous items to decide which events qualify for inclusion in the media, as they cannot all fit the available space. The media act as 'gatekeepers', who control and moderate the access to news and information. In today's society, connected by the Internet, the amount and flow of information and communication has increased dramatically. In addition, users become involved as creators and distributors of content. The abundance of (digital) information makes it difficult for journalists to manage and assess the events in terms of their alleged newsworthiness.

Within the scholarly field of journalism studies, the selection of news by the media is said to be influenced by a range of factors. As specified in Sect. 2, several criteria have been determined over time to assess news value. However, most of these are based on purely human analysis and reasoning. At the time they were created, large-scale empirical evaluation of these criteria would have been extremely labor intensive, if not impossible. Now, recently developed techniques and services in information technology enable us to revisit newsworthiness research, and construct an automated approach to help journalists in assessing news value. Our goal is to provide a tool that will help journalists in both news selection and reader targeting, and additionally, will aid sociological researchers in verifying and determining news selection criteria.

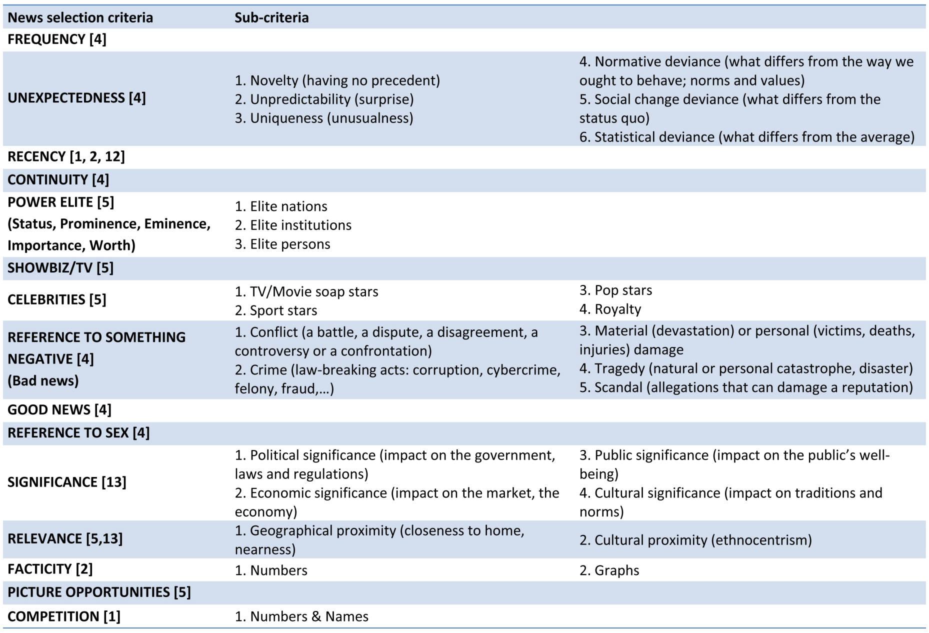
The remainder of the paper is structured as follows: first, we provide a sociological perspective on news value assessment. Next, an overview is given of our proposed approach to revive this perspective using state-of-the-art information processing techniques. We break down the approach into its components, and provide details about each type of analysis, including how it was realized in the proof-of-concept implementation. Finally, a first evaluation of the automatic approach is performed, and the possible improvements and future research opportunities are discussed.

2 A Sociological Perspective

Sociological research on the process of news selection in newsrooms has resulted in various overviews of 'news values' or 'news selection criteria'. The concept of newsworthiness is built on the assumption that certain events get selected by media above others based on the attributes or 'news values' they possess. The more of these news values are satisfied, the more likely an event will be selected. If an event lacks one news value, it can compensate by possessing another. Hence, journalists' criteria for selecting the news are cumulative, making stories significant based on their overall level of newsworthiness. The earliest attempt for a systematic approach of determining newsworthiness by news values, is a taxonomy by Galtung & Ruge [1], that triggered both scholars and practitioners in examining aspects of events that make them more likely to receive coverage [2, 3]. For analytical purposes, the concept of news values is valuable to understand that news selection is more than just the outcome of journalists' 'gut feeling'. To decide what is news and what is not, journalists consciously and unconsciously use a set of selection criteria, that help them assess the newsworthiness of a story or an event [4, 5].

In Table 1, an overview is given of the different news values that are available in the literature of journalism studies, categorized in news values and their subfactors. Here, we will give a brief review of the most important studies in this domain.

In the 1960's, Galtung & Ruge [1] published a theory of news selection, which provided a taxonomy of 12 news values that define how events become news. More specifically, they discerned (1) *frequency*: the time-span of the event to unfold itself, (2) *threshold*: the impact or intensity of an event, (3) *unambiguity*: the clarity of the event (4) *meaningfulness*: the relevance of the event, often in terms of geographical proximity and cultural similarity, (5) *consonance*: the way the event fits with the expectations about the state of the world, (6) *unexpectedness*: the unusualness of the event, (7) *continuity*: further development of a



News selection criteria	Sub-criteria	
FREQUENCY [4]		
UNEXPECTEDNESS [4]	1. Novelty (having no precedent) 2. Unpredictability (surprise) 3. Uniqueness (unusualness)	4. Normative deviance (what differs from the way we ought to behave; norms and values) 5. Social change deviance (what differs from the status quo) 6. Statistical deviance (what differs from the average)
RECENCY [1, 2, 12]		
CONTINUITY [4]		
POWER ELITE [5] (Status, Prominence, Eminence, Importance, Worth)	1. Elite nations 2. Elite institutions 3. Elite persons	
SHOWBIZ/TV [5]		
CELEBRITIES [5]	1. TV/Movie soap stars 2. Sport stars	3. Pop stars 4. Royalty
REFERENCE TO SOMETHING NEGATIVE [4] (Bad news)	1. Conflict (a battle, a dispute, a disagreement, a controversy or a confrontation) 2. Crime (law-breaking acts: corruption, cybercrime, felony, fraud,...)	3. Material (devastation) or personal (victims, deaths, injuries) damage 4. Tragedy (natural or personal catastrophe, disaster) 5. Scandal (allegations that can damage a reputation)
GOOD NEWS [4]		
REFERENCE TO SEX [4]		
SIGNIFICANCE [13]	1. Political significance (impact on the government, laws and regulations) 2. Economic significance (impact on the market, the economy)	3. Public significance (impact on the public's well-being) 4. Cultural significance (impact on traditions and norms)
RELEVANCE [5,13]	1. Geographical proximity (closeness to home, nearness)	2. Cultural proximity (ethnocentrism)
FACTICITY [2]	1. Numbers	2. Graphs
PICTURE OPPORTUNITIES [5]		
COMPETITION [1]	1. Numbers & Names	

Table 1. News selection criteria as identified and specified in literature.

previous newsworthy story, (8) *composition*: a mixture of different kind of news, (9) *reference to elite nations*, (10) *reference to elite persons*, (11) *reference to persons*: events that can be made personal, (12) *reference to something negative*. Bell [6] used Galtung and Ruge's list as a starting point, but redefined some of them and added the value of facticity: a good story needs facts (e.g. names, locations, numbers and figures). Harcup and O'Neill [2] also made a revision of Galtung and Ruge's criteria for the contemporary news offer, with special attention for the entertainment offer available in newspapers. More specifically, the following values were added: (1) Events with *picture opportunities*, (2) *Reference to sex*, (3) *Reference to animals*, (4) *humor*, (5) *showbiz/TV* related events and (6) *good news* (e.g. acts of heroism). In addition, concerning elite persons they made a distinction between *power elite* (e.g. Prime minister) and *celebrities* (e.g. Pop Stars). McGregor [7] also proposed news values reflecting the modern way of news selection, with a focus on TV news. He referred to events that are *visually* accessible and recordable. In addition, when events involve tragedy, victims or children, it is likely to appeal to the *emotions* of the audience. Concerning negative events, *conflict*, *scandal* and *crime* are highlighted. Bekius [8] added *competition* as a value, which explains the extensive coverage of sports. Finally, Shoemaker and Cohen [3] extended the notion of significance, by demarcating four sub-dimensions: *political* significance, *economical* significance, *cultural* significance and *public* significance.

3 A Technical Perspective

While the aspects of news as described in Sect. 2 provide a valuable guideline for the theoretical analysis of news, this remains a task for specialists, labor intensive and prone to human error. However, recent advances in Semantic Web technologies have made new tools and services available on the Web to automatically analyze the content of an article. In Figure 1, these technologies are mapped to the news values from Table 1 they can be used to detect. Each of these analyses will be discussed in detail in Sect. 4. The mapping is structured as follows: each straight, full line indicates that an analysis or detected entity or topic contributes to the connecting news criteria in some way. The dotted, curved lines represent the subdivision of news criteria into several sub-criteria, which are then connected to one or more analyses by full lines. The attentive reader will notice that there are two news criteria in the mapping that were not present in the literature overview in Table 1: *social significance* and *technological significance*. This is because when implementing our approach, we noticed a significant number of articles that complied with these criteria, and had no corresponding news value mapping. Therefore, we added these news values ourselves. Whether they can actually be considered news selection criteria in the newsrooms, remains to be investigated by extensive social research. However, our approach will provide the ideal means to this end.

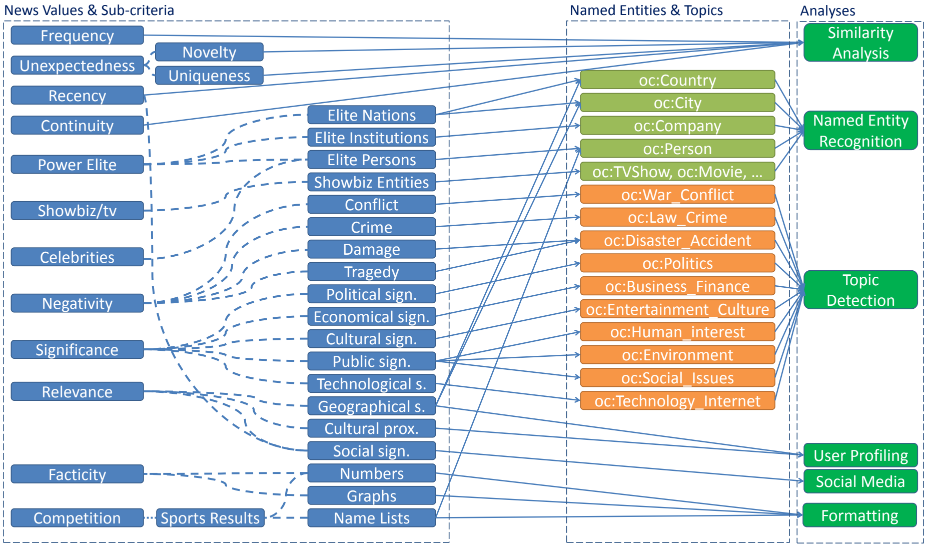


Fig. 1. News values and sub-criteria mapped to Named Entities, Topics and analyses using Semantic Web technologies

4 Determining Newsworthiness

The aim of our research is to provide users (news professionals, such as journalists and media researchers) with information about the newsworthiness of an arbitrary news article. However, it would not be feasible, nor desirable for our system to output a single newsworthiness score. Instead, a score between 0 and 1 will be generated for each news value discussed in Sect. 2. This way, a *newsworthiness profile* is generated, based on the content of the article. To achieve this, we take the content input (plain text), and analyze it in different ways. The results of this analysis are then traced back to the news values, using the mapping illustrated in Figure 1. This mapping leads to six analysis 'clusters', namely: similarity analysis, Named Entity Recognition (NER), topic detection, social media, reader targeting and content format analysis. Each analysis adds a score to one or more corresponding values. After the final analysis, a normalization is performed before returning the final result.

4.1 Named Entity Recognition

A crucial step in our analysis is gathering as much descriptive metadata about the news articles as possible. However, manual addition of metadata by the authors is often neglected. Therefore, we need to automatically generate these metadata ourselves. To do this, we use publicly available tools known as Named Entity Recognition(NER) services. These services accept plain text as input, and output a list of linked Named Entities (NEs), detected in the text. For our implementation, our NER service of choice is OpenCalais[3], a well-established, thoroughly tested [9] and freely available NER service. However, it is our goal to include additional NER services over time, thereby linking to more nodes in the Linked Data cloud, and creating a more complete mapping of Linked Data resources to news values. For a thorough overview of NER services and their performance, we refer to the NERD framework, by Rizzo & Troncy [10]. At the time of writing, OpenCalais is able to detect 21 types of entities from plain text, of which a selection is shown in Figure 1. A complete list of these entity types is provided in the online OC documentation[4]. Conforming to the mapping in Figure 1, these entity types contribute to the *eliteness*, *geographical significance*, *showbiz/TV* and *celebrities* criteria. The amount at which a detected entity contributes to its respective news value(s) is determined by the *relevance score* assigned by OpenCalais. This relevance score represents the importance of an entity in the text it occurs in. For each entity detected in the text, the relevance score is accumulated in the corresponding news value(s).

[3] http://www.opencalais.com

[4] http://www.opencalais.com/documentation/linked-data-entities

4.2 Topic Detection

As an addition to the NER, OpenCalais also categorizes the article into one of 17 topics[5]. We associate each of these topics to one or more news criteria, as in Figure 1. The topics of the article contribute to the *negativity* and *meaningfulness* criteria. Because these are very broad criteria, we subdivide negativity into *conflict*, *crime*, *damage* and *tragedy*. Similarly, meaningfulness is subdivided into *political*, *economical*, *cultural*, *public* and *geographical significance*. One of the predefined topics OpenCalais detects, technology, occurred frequently during the testing phase of development, and had no straightforward mapping to one of the values. Therefore, a new news value, termed *technological significance*, was added. Whether this new news value is also used by media practitioners, remains to be investigated in future work. Analogous to the relevance scores in Sect. 4.1, OpenCalais assigns a *confidence score* to each topic. This score is added to the corresponding news criteria for each detected topic.

4.3 Similarity Analysis

Next to NER and topic detection, another important component of automated news analysis is identification of any prior mentions of an article or its content. Comparing the content of an article to strategically chosen reference sets gives us information about the *frequency*, *novelty*, *uniqueness* and *recency* of that kind of content. Here, the challenge is to find a good method for retrieving similar articles, as well as suitable reference sets.

For the retrieval of similar articles in a reference set, we have developed a Named Entity based similarity measure. The approach uses the NEs detected in an article to compute a similarity measure. When comparing two articles A and B, we create two vectors representations a and b of their NEs, where a_i is the weight of NE i in document A (analogous for B), as determined during the NER step (in our case, the relevance score assigned by OpenCalais). The similarity between the documents is then calculated as the *cosine similarity* of the vectors, given by Formula 1.

$$Sim(A,B) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}} \tag{1}$$

When no NEs were detected, we revert to the classic “bag of words” approach, using *Term Frequency - Inverse Document Frequency (TF-IDF)* weights for every word in the text. For a more extensive description and evaluation of this NE based similarity measure, we refer to [11], where the same method is used for the clustering of a large set of news articles. Here, the NE based similarity is used to select all news articles from a reference set that are more similar to the input article than a certain threshold.

For the selection of the reference sets, it is important to keep in mind which news values the similarity analysis will influence. As seen in Figure 1, the criteria

[5] http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/document-categorization

related to similarity are all relative to recent publications. In other words, for this application, it is better to use a smaller and dynamic reference set of recent news items (e.g. one or two weeks), rather than a large news database, gathered over a long period of time. In our implementation, we use the New York Times "Most Popular" API[6] to acquire the the most viewed articles of the last 7 and 30 days to be used as our reference sets.

After retrieving all articles with similarity to the input article above a certain threshold T (in our case, $T = 0.7$), from both reference sets, we use the size of these retrieved sets to make a decision about the related news values. If the article has more than one similar article, using the reference set of the last 7 days, a positive score is added to the *recency* value. If it has no similar articles in the 7-day set, but it has more than one in the 30-day set, the scores of the *frequency* and *continuity* criteria are increased. Finally, if no similar news items were found, a score is added to the *unexpectedness*. The most suitable amount at which the scores are increased will have to be determined by further testing. In the first implementation, a straightforward "+1" is performed.

4.4 Social Media

An important group of actors in today's media landscape that cannot be ignored, are the social media. Social networks such as Twitter and Facebook play a leading role in the spreading of news stories. These media act as a barometer for news trends, and provide the necessary services and APIs to be utilized as such. These APIs allow us to measure the relevance and public significance of a certain news item, as a form of crowdsourcing. The social media are important indicators of the newsworthiness of an article. Once a topic is trending on Twitter and/or Facebook, the information related to this topic is spread widely in a very short timespan. Therefore, we incorporate a social media API in our approach. More specifically, we cross-reference the Named Entities and topics extracted from the content of the input news item, to trending topics and keywords on Twitter, using the Twitter API[7]. Each trending word or sentence is compared to the string labels of the detected NEs and topics. Additionally, these labels are expanded with their synonyms and semantically similar words, found using a publicly available lexical tool, such as DISCO[8]. If a NE or topic (or one of their synonyms) occurs in the list of currently trending topics on Twitter, the *social significance* and the *recency* score of the news item are increased by one.

4.5 Reader Targeting

While we are aiming towards a general and inherent model of newsworthiness, the influence of the targeted reader's whereabouts and preferences cannot be ignored. Local news might have entirely different news values contributing to

[6] http://developer.nytimes.com/docs/read/most_popular_api

[7] http://dev.twitter.com/

[8] http://www.linguatools.de/disco/

its newsworthiness than national or global news. Therefore, we will take these two aspects of the targeted news consumers into consideration. The *cultural proximity* of a news article to a group of users is determined by adding 1 for each entity or topic relating to culture detected in the article, such as TV shows, plays, etc. Similarly, the *geographical proximity* of an article to a user is increased by 1 when specific locations such as cities or regions appear in the article. These criteria give a journalist an indication of which audience his/her article can be targeted towards.

4.6 Content Format Analysis

Research shows that *facticity* plays an important role when it comes to newsworthiness [6]. Although no off-the-shelf techniques are in place for detecting facticity, there are some constructs we can look for. Although far from the only contributing factors to facticity (such as *correctness*, which is beyond the scope of this work), lists and tables are good indicators that facts are being summed up. Lists and tables are elements that are fairly easily recognizable in digital content, due to their notation using a markup language. For example, lists that are included in a news article on a web page will be represented using the HTML <*ul*> tag. However, these HTML lists and tables are often used as a means to create the layout of a web page. To make the distinction from this type of usage, only lists and tables that contain numerical values will contribute to the facticity. Additionally, we look for repetition of certain structural patterns using regular expressions, such as "*:*$" and "-*$" (where '*' represents a wild card, and '$' a new line). Other indicators of facticity are frequent occurrences of currencies (detectable by many NER services), or other numerical values. Each of these detected indicators will contribute to the facticity value in the resulting newsworthiness profile. The amount at which the facticity score is increased, is to be determined empirically.

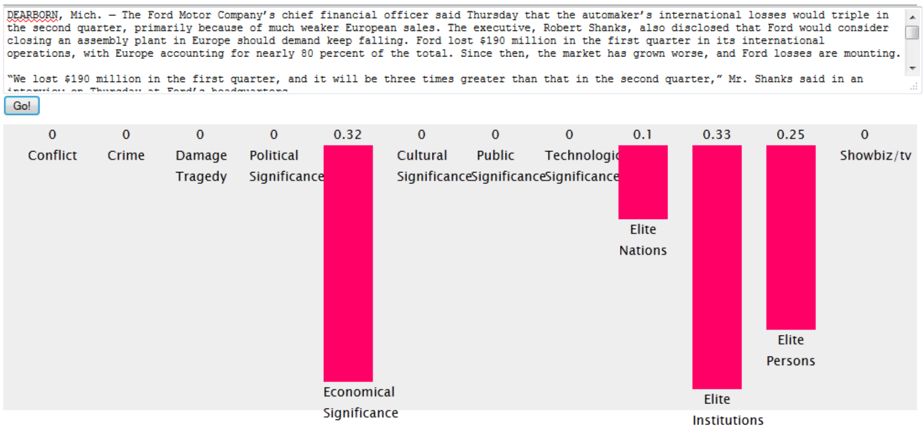
4.7 Normalization and Presentation

When all analyses have been completed, we obtain an array of scores, each linked to a news value. However, these scores have no well-defined bounds, as each type of analysis uses different calculations, precision and weights to add a score to a news value. Therefore, we normalize each news value score s to a normalized score s_n as in Formula 2, where S_0 is the set of non-zero scores in the array.

$$\forall s \in S_0 : s_n = \frac{s}{\sum_{s_i \in S_0} s_i} \tag{2}$$

This implies that the total of all non-zero scores is always 1, and each score indicates the contribution of its news value to the newsworthiness of the article.

Now, the score array is visualized to the user, in the form of a bar chart or similar graphical representation. This way, the user is instantly presented with an overview of the newsworthy aspects of the input article. Figure 2 shows a screen shot of how this is achieved in the proof-of-concept implementation.



DEARBORN, Mich. — The Ford Motor Company's chief financial officer said Thursday that the automaker's international losses would triple in the second quarter, primarily because of much weaker European sales. The executive, Robert Shanks, also disclosed that Ford would consider closing an assembly plant in Europe should demand keep falling. Ford lost $190 million in the first quarter in its international operations, with Europe accounting for nearly 80 percent of the total. Since then, the market has grown worse, and Ford losses are mounting.

"We lost $190 million in the first quarter, and it will be three times greater than that in the second quarter," Mr. Shanks said in an interview on Thursday at Ford's headquarters.

Fig. 2. Example of visualization of newsworthiness scores in our proof-of-concept application. The analyzed article contains information on the stock of a major automotive company. The software correctly determined the economical significance of the article, and indicates the presence of some important persons, nations and organizations.

5 Evaluation

Until now, analysis of news and assessment of its newsworthiness was performed by domain experts, such as journalists and media researchers. Unfortunately, this means that there are no automated systems to compare to our approach. Therefore, as a first evaluation of our approach, we constructed a gold test set, using publicly available news that is assumed to be newsworthy, and verified the results of our approach manually using the latest version of the proof-of-concept software. Note that even though not all features were implemented yet at the time of writing, we chose to evaluate the output of all implemented features, rather than provide no evaluation at all.

Our test set was created using the New York Times Developer APIs[9]. The New York Times offers a large selection of its data openly for non-commercial use. To obtain a set of articles that can safely be assumed to be newsworthy, we made use of the "Most Popular API". As an initial set, we requested the most viewed articles from the 30 days prior to the submission of this paper (thus ranging from July 8th until August 6th). From this set, we then retained only those URLs which were annotated with more than five concepts, using the "Semantic API". This extra selection step ensures us of suitable data, in which a NER service should stand a good chance of finding Named Entities. Finally, using the "Article Search API", we obtained the summary[10] of each article, and processed it using our approach. The similarity analysis (see Sect. 4.3) was

[9] http://developer.nytimes.com/docs/read/{most_popular_api, semantic_API, article_search_api}

[10] Note that the New York Times API does not release the full text of articles.

skipped, since we use the same Most Popular API as a reference set, which would result in each newsworthiness profile containing the "recency" value.

After applying our approach, we obtained a set of 224 articles, each annotated with extracted entities, topics, and a newsworthiness profile. We assessed these newsworthiness profiles manually for correctness ourselves. For each profile, we assessed two things:

1. Are there any news criteria missing in the profile, that are perceived in the article? (*recall*)
2. Are there any excess news criteria in the profile, that are not perceived in the article? (*precision*)

When performed for all newsworthiness profiles, the first assessment gives us the number of profiles without any missing news criteria, or the *recall* of the newsworthiness profiles. The seconds assessment results in the number of profiles without any falsely detected news criteria, or the *precision* of the newsworthiness profiles. After careful evaluation of the 224 newsworthiness profiles, a **precision of 83.93%**, and a **recall of 66.07%** was obtained. The most common mistakes occurred when two very distinct news criteria occurred in the same article. For example, one of the articles describes a football coach harrassing one of his players. This will trigger both the Crime and Competition criteria, while in this case, only Crime is relevant. This could perhaps be solved by defining disjointness rules between the news criteria.

Overall, these preliminary results indicate that we have found the basis for a viable, automated method to extract the contributing news criteria that determine an article's newsworthiness. However, as discussed in the next section, further evaluation by domain experts is still needed when the approach is fully implemented. The processed set of articles, complete with newsworthiness profiles and their evaluations, can be found in machine-understandable JSON format at the following url: `http://users.ugent.be/~tdenies/AVALON/ISWC/evaluation.json`. A graphical interface to browse this data is available at `http://users.ugent.be/~tdenies/AVALON/ISWC/`.

6 Discussion and Future Work

While the initial impression of our approach is positive, a more extensive evaluation will be performed soon. The first evaluation is slightly biased, as it relied on verification of the generated results, because it is a relatively fast and straightforward process. A more thorough, unbiased approach is to ask a panel of domain experts to manually compose a newsworthiness profile, based on the same news values as our approach, and compare it with a profile generated by our approach, using more than one NER service (and corresponding news value mapping). However, before we initiate this extensive, time-consuming evaluation, further optimization of our proof-of-concept software is required, to avoid excess repetition of the process. Once this technological optimization is complete, we will also assess usability with media practitioners in order to discern improvements in terms of ease of use (e.g. the visualization of the newsworthiness scores) and user-perceived reliability (e.g. the professionals' trust in the tool's accuracy). In addition, future research will focus on the question to what extent the selection of news of particular media outlets can be adequately predicted by automatically generated newsworthiness assessments. Such a study has the potential to provide valuable insights as to what extent the inherent subjectivity of human news selection can be approximated by objective parameters and algorithms for newsworthiness assessment.

Our approach also provides an opportunity to gain a new perspective on the research towards news value assessment. The automated approach allows for a large-scale analysis of published works. This way, it becomes possible to verify whether the theoretically determined news criteria still hold for today's media landscape, or even to identify new criteria. As experienced during development, comparing automated and human news selection mechanisms might reveal news values not yet identified in the literature of journalism studies. One specific use case that has recently emerged, is that of citizen journalism. As users become more informed, they often participate actively in the production of news. Examples of this sort of participatory journalism are often found in blogs, opinion pieces and local newspapers. It is to be expected that these citizen journalists will use a distinctly different set of criteria to determine whether something is newsworthy to them. Our approach allows to investigate this objectively.

7 Conclusions

We succeeded in building an application to automatically assess the newsworthiness of a news article, starting from only the content. We used a unique combination of sociological research and technological advances, to create a mapping of long established and recently emerged news values to a set of automated analysis clusters. Through Named Entity Recognition, topic detection, similarity analysis, social media, reader targeting and text format matching, our approach assigns a score to a set of news criteria, to create a newsworthiness profile for an arbitrary piece of text. The generated newsworthiness profiles of 224 news items were manually verified, and exhibited a precision of 83.93% and recall of 66.07%. Of course, further evaluation is needed, in collaboration with domain experts. This approach provides many opportunities for future work, including large-scale analysis of news content by domain experts, potentially uncovering new criteria that contribute to newsworthiness in the current media landscape.

Acknowledgments. The research activities in this paper were funded by Ghent University, IBBT (a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union, in the context of the IBBT project Smarter Media in Flanders (SMIF). Companies involved are Belga, Concentra, VRT and Roularta, with project support of IWT. We also extend thanks to The New York Times, for their comprehensive open news API.

References

[1] Galtung, J., Ruge, M.: The structure of foreign news. Journal of peace research **2** (1965) 64–90

[2] Harcup, T., O'neill, D.: What is news? Galtung and Ruge revisited. Journalism studies **2** (2001) 261–280

[3] Shoemaker, P., Cohen, A.: News around the world. Content, practitioners, and the public. Recherche **67** (2006) 02

[4] O'sullivan, T., Hartley, J., Saunders, D., Montgomery, M., Fiske, J.: Key concepts in communication and cultural studies. Routledge London (1994)

[5] Schultz, I.: The journalistic gut feeling. Journalism practice **1** (2007) 190–207

[6] Bell, A.: The language of news media. Blackwell Oxford (1991)

[7] McGregor, J.: Terrorism, war, lions and sex symbols: Restating news values. What's news (2002) 111–125

[8] Bekius, W.: Werkboek journalistieke genres. Coutinho (2003)

[9] Iacobelli, F., Nichols, N., Birnbaum, L., Hammond, K.: Finding new information via robust entity detection. In: Proactive Assistant Agents (PAA2010) AAAI 2010 Fall Symposium. (2010)

[10] Rizzo, G., Troncy, R.: NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In: (ISWC'11) Workshop on Web Scale Knowledge Extraction (WEKEX'11). (2011)

[11] De Nies, T., Coppens, S., Van Deursen, D., Mannens, E., Van de Walle, R.: Automatic discovery of high-level provenance using semantic similarity. In: Proceedings of the 4th International Provenance and Annotation Workshop IPAW 2012, LNCS 7525, Springer, Heidelberg. (2012) 97–110