Extended Abstract: Towards Automated Contextualization of News Articles

Stephan C. Escher Jan L. Reubold Richard Kwasnicki Joachim Scharloth
Lutz M. Hagen Thorsten Strufe
TU Dresden, Dresden, Germany
<firstname>.<lastname>@tu-dresden.de

1 TRUSTWORTHINESS OF NEWS PROVIDERS

The World Wide Web is a dominant medium for news and information exchange. Together with TV it enjoys far larger regular audiences in the U.S. than print and radio, and recent studies suggest that also the gap to TV consumption is closing fast[1]. The reasons are diverse, ranging from ease of (commonly: free) access to the democratization of publishing, as the cost for production and publication essentially disappears.

Democratization, however, raises new challenges. Where consumers needed to learn and judge on the order of a few to a dozen news papers, TV stations, or other media outlets in the past, media literacy has become much more challenging in the face of hundreds of news sites, blogs, and even seemingly legitimate satire publishers. Social media has further amplified the potential reach of an individual's post. Confronted with the amount of posts fed into social media, the users are overwhelmed and unable to judge the trustworthiness and credibility of posts [2, 4]. Publishing fictitious but upsetting posts has even become it's own business model[2], which is alarming considering that recent studies suggest that over 60% of U.S. adults use social media as their primary news source[1].

This development has caused a surge in misinformation. Parties trying to further their agendas have gained equal access to powerful tools of mass publication. Manipulation using the main information channels is nothing new in itself. The combination of cheap and fast production lines and global accessibility lift this threat to an entirely new level. To encounter this situation, the users may need support in judging the news in their feeds. In a study of Schwarz and Morris [3] users were supported by augmenting posts with specific, corresponding features. While their results suggest that contextualization of articles allowed users to make more accurate judgments, the selection of features used for the augmentation is critical. For one, subjective features weaken the trustworthiness of such an approach, similar to manual rating. Especially fringe audiences may feel a manual choice to be an attempt of elitist censorship. Secondly, crowd-based features provide unreliable information [2]. Thus, a suitable solution requires: (*i*) features to be identified by an automatic, data-driven approach, (*ii*) features to be based on 'objective' information, and (*iii*) to relate to the content of an article.

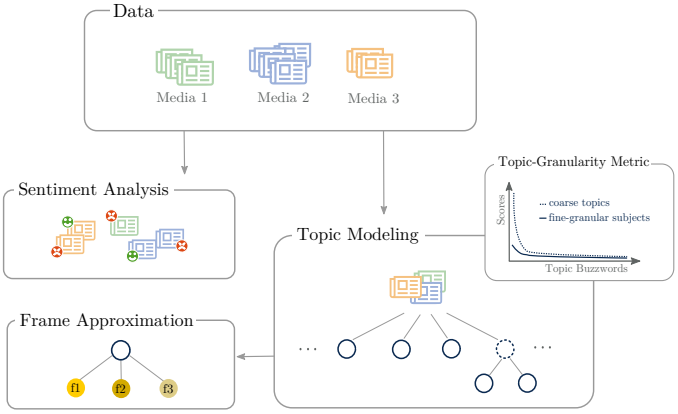

Figure 1: Workflow of the proposed approach: articles are processed by an hierarchical topic model using a topic-granularity metric and a sentiment analysis; framings are extracted from the subjects.

In this abstract, we report on our ongoing work to provide users with an automated contextualization of news articles. By presenting fairly abstract characteristics about the publisher (e.g. approximated agenda or topic spectrum) and bringing articles into context with documents covering a similar subject, the user is provided with information for assessments. While this does not represent an absolute score of trustworthiness, it shall help to indicate tendencies.

An example from our preliminary results demonstrates that news portals such as *Russia Today* and *Sputnik News*, which in part are deemed by the U.S. as a principal international propaganda outlet[3], focus significantly more on U.S. and EU specific topics than other publishers.

2 BRINGING LIGHT TO THE DARKNESS

In this work, we apply NLP techniques to extract latent information about publishers and their articles, such as their topic spectrum and potential agenda. A topic spectrum is defined as a summary of the topics identified in the articles of a publisher. Additionally, articles are augmented by approximated framings publishers use to present information. Frames are crucial cornerstones for understanding an authors intentions. Given some information, frames can alter the consumers perception without altering the information itself. Cumulative information on used framings and corresponding sentiments assists users to better understand the agendas of publishers.

[1] e.g. https://goo.gl/48qCJX
[2] http://money.cnn.com/interactive/media/the-macedonia-story/
[3] https://www.dni.gov/files/documents/ICA_2017_01.pdf

MIS2, 2018, Marina Del Rey, CA, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

Given a set of news articles, we identify topics on a subject level. Articles corresponding to a specific subject (topic) are then used to approximate framings. The idea is that corresponding articles can be grouped based on their content. Combined with a simple sentiment analysis on each article, we provide the following features:

I Publisher contextualization
  i Top N subjects of a news provider combined with corresponding sentiment trends
  ii Global topic tree (hierarchical topic model) with selectable views for each news provider

II Article contexualization
  i Viewed article compared to corresponding articles in a latent subject space
  ii Viewed article compared to corresponding media landscape within the subject (sentiment and frames used)

While $(I.i)$ and $(I.ii)$ show general information about a publisher and do not change on an article basis, $(II.i)$ and $(II.ii)$ change depending on the subject discussed in the article. The latter information allows us to display an article in the context of corresponding observed framings with distance information and related articles. Combined with the former, the provided information allows users to get a better picture of the background of an article. Note, that we do not interpret the results on a credibility basis, thereby, preventing to introduce biases into the analysis.
In order to develop and test our algorithm, we collected articles from German news papers and online blogs (2013-2017). In preliminary evaluations, we focused on a subset of 500,000 articles from the year 2015.

2.1 Hierarchical Topic Model

The hierarchical topic model is the most crucial component of the proposed approach. While models exist to extract the underlying hierarchical topic structure of a set of articles, we use an existing approach to model a flat topic structure and extend it to the hierarchical case. The reason is that models such as the hierarchical LDA [1] (hLDA) do not suite our case. For example, applying hLDA resulted in serious RAM issues. Therefore, we utilize a topic extraction algorithm that applies a non-negative matrix factorization and LDA.[4]

In a first step we learn a hierarchical topic model. It splits topics in a recursive fashion until a topic describes a specific subject. To find the subject level, we make use of a metric based on the word scores within topics as the recursion anchor. Given an identified topic, we process it further until it reaches the recursion anchor. The anchor is defined by a metric that signals the algorithm to stop if its value is below a certain threshold. The metric m is defined as the average distance between the normalized top T word weights characterizing a topic t,

$$m(t_i) = \sum_{j=1,k=2}^{k=T} \frac{t_i[j] - t_i[k]}{T-1}, \tag{1}$$

where t_i denotes the i-th topic and the metric m is computed by taking the top $T = 10$ words of t_i into account.

[4]https://goo.gl/WzBd56
[5]https://www.philosophersmag.com/essays/26-the-fact-opinion-distinction

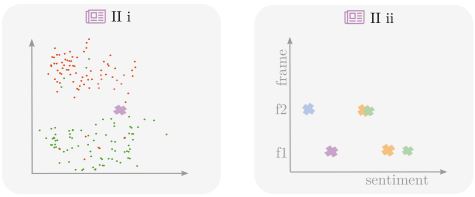

Figure 2: Article contextualization: (left) 2-dimensional latent subject space (PCA on vector representations of articles) with sentiments represented by the hue of marks from negative (red) to positive (green); lilac mark represents the viewed article; (right) frames used by publishers with the corresponding average sentiments.

2.2 Identification of Frames

In a second step, we coarsely approximate framings. Based on a subject the algorithm groups articles that report in a similar fashion about the subject. We test two approaches for finding these. For one, we apply the topic extraction algorithm used for the identification of topics in our hierarchical model. Here, we only take articles into account that belong to a certain topic. As a second solution, we test a clustering based on the vector representations of the articles within a topic using word embeddings. As depicted in Fig. 2 (left), the idea is that articles belonging to specific framings form point clouds in the latent subject space.

3 DISCUSSION

Approaching the problem of trustworthiness of news providers is challenging. Merely thinking about the distinction between fact and opinion presents difficulties[5].

Therefore, we approach the problem from a different point of view. Instead of defining features to compute a trust-score, we focus on providing information on the frame of an article and general trends in the articles published by the corresponding news provider. Understanding how an author wants his audience to perceive provided information and if his intention is a general trend of the publisher, allows users to better assess the content of an article.

Additionally, the results obtained by our analysis of articles (based on the topic hierarchy, framings, and sentiments) creates follow-up questions. Given the detailed vector space representation of articles one could potentially track articles back to their sources.

This clearly is no mature solution to fake news and misinformation. It merely serves to demonstrate a first step towards automatic, data-driven contextualization, on the way to support users in assessing trustworthiness and credibility of publishers and articles.

REFERENCES

[1] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. 2004. Hierarchical topic models and the nested chinese restaurant process. In Advances in neural information processing systems. 17–24.
[2] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. Journal of communication 60, 3 (2010), 413–439.
[3] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 1245–1254.
[4] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. 2016. Evaluating Information: The Cornerstone of Civic Online Reasoning. (2016).