Multidimensional Analysis of Trust in News Articles (Student Abstract)

Avneet Kaur,[1*†] **Maitree Leekha,**[2*] **Utkarsh Chawla,**[2*] **Ayush Agarwal,**[2*] **Mudit Saxena,**[3*]
Nishtha Madaan,[4] **Kalapriya Kannan,**[4] **Sameep Mehta**[4]
[1]Indraprastha Institute of Information Technology, New Delhi, India
[2]Delhi Technological University, New Delhi, India
[3]Shiv Nadar University, New Delhi, India
[4]IBM Research AI- INDIA
avneet14027@iiitd.ac.in, {maitreeleekha, utkarshchawla}_bt2k16@dtu.ac.in, {ayush286, mudit10saxena}@gmail.com,
{nishthamadaan, kalapriya.kannan, sameepmeehta}@in.ibm.com

Abstract

The advancements in the field of Information Communication Technology have engendered revolutionary changes in the journalism industry, not only on the part of the journalists and the media personnel, but also on the people consuming these news stories, who today, are only a click away from all the updates they need. However, these advances have also exposed the prevailing venality, wearying off the trust of the public in news media. How then, does an individual discern that which, out of the countless news stories for an incident, should be trusted? This work introduces a system that presents the user a multidimensional analysis for trust in news from various media sources based on the textual content of the articles, assessment of the journalists' perspectives and the temporal diversity of the issues being covered by the media houses publishing the news articles. Our experiments on a self-collected dataset confirm that the system aids in a comprehensive analysis of trust.

1 Introduction

The media and journalism industry has seen a massive transition in the way it operates in the past few years. Unlike conventional times, the consumer of this digitized generation has quick access to almost everything happening around the globe, thanks to the internet. While this shift towards the digital media has made the channel between the consumers and the journalists less obscure, the public's trust in the news media has declined. The increasing political polarization of the media houses and the backing monetary profits has created a sense of skepticism amongst the public about the credibility of the news disseminated by them. In such a situation, how does a person discern that which, out of the many available sources, offers the news which is closest to the facts, is authentic and therefore, can be trusted? Given the sheer number of media houses and news sources, it is tedious for a person to analyze manually the articles delivered by each one of them for trust. Furthermore, oftentimes people only consider the textual content of the articles, whereas other factors such as the record of the journalist who wrote it and the media house that published it, also play an important role. Inspired by the theoretical models formulated by Kohring and Matthes, we propose a novel system to help the user analyze the articles from various news sources for trust at three different levels, for examining:

Trust in the selectivity of facts- How articles about the same event but from different media houses have varying textual content, leading to a biased representation of facts.

Trust in journalistic assessment- How an article writer may tend to collude different perspectives to misrepresent a news entity.

Trust in topic selectivity- How diverse are the issues covered by a media house over a period of time, to ensure that the news coverage does not reflect a topic selectivity bias.

2 Methodology

The system can be decoupled into three modules- Content Based Analysis, Journalistic Assessment and Temporal Diversity Analysis, each of which analyzes the articles from various news sources for trust at different levels.

2.1 Content Based Analysis

This module contrasts the articles from different media houses based on their linguistic content and examines if the representation by any media house is biased in comparison to the trusted representation. Therefore, as the first step, the user selects the news source offering the article in a form that she believes is the most authentic as the trusted source.The system then analyzes how the articles from other sources (**A**) differ from the trusted source article (**T**), based on the following lexical bias analysis metrics:

Centrality shift: In graph theory, centrality indicates the importance of a node in the graph. Measuring the centrality shift helps in examining how the centrality (*i.e.*, importance) of an entity differs in **A** and **T**. We find the betweenness centrality shift of the common entities in the semantic role graph of the two articles, **T** and **A**, constructed using the IBM Watson's NLU API.

Sentiment and Emotion Shift: We use IBM Watson's NLU API to find the difference in the sentiments (positive, neutral and negative) and emotions (sadness, joy, fear, anger and

*All the authors contributed equally, and wish that they be regarded as joint First Authors.

†Mobile No.: +91 8143911339

disgust) associated with the common entities in the entity graph constructed in the previous step.

Over-representation: To analyze if the article **A** has over-represented (*i.e,* exaggerated) the content, we find the entities which are present **A** but not in **T**, and contrast their centrality scores.

Furthermore, we combine the above lexical scores using the TF-IDF scores of the entities to compute the corresponding article-level bias analysis metrics, and further average them to form a **Bias Score**, which reflects the overall extent of change in representation of the content between **A** and **T**.

2.2 Journalistic Assessment

This module analyzes how a particular journalist represents different perspectives about a news article, by using Metro-Maps (Shahaf, Guestrin, and Horvitz 2012), to provide a comprehensive view of all the stories written by a person, and linked to the current article (**A**), in the form of a storyline. For constructing a metro-map, we follow the steps briefly described below:

Data Pre-processing: First, we use NLTK library to extract all the news entities, *i.e.*, all the Noun Phrases (NPs) from the articles. Then, we extract the relations, which are essentially phrases present between two succeeding entities in a sentence.

Entity Graph Construction: Now, we construct a weighted, directed entity graph, $G(V, E)$, where the nodes (V) are the entities (NPs) and the edges (E) represent the relations (phrases). The weights of the edges are the cosine similarity between the word-embeddings (pre-trained) of the entities corresponding to the nodes. Finally, we make this initial graph news-centric by retaining only the nodes which are reachable from the nodes corresponding to the entities present in **A**, *i.e.*, N_A.

Relationship centres & Dense sub-graphs: We extract the sub-graphs which depict stories pertaining to the news entities present in **A**, *i.e.*, the nodes in N_A. We extend the CePS algorithm (Tong and Faloutsos 2006) to locate relationship centres around a particular news entity $q \in N_A$ i.e., the entities which are the representatives of stories around q. Having found k relationship centers, we use them to sample k dense sub-graphs from G, each depicting a different story around q. We execute probabilistic random walks, where the nodes corresponding to the edges having larger edge weights and higher in-degrees have a greater probability of getting selected. Therefore, all the dense sub-graphs are *coherent* and maximize *coverage*. Random walks are performed for all the k relationship centers to produce k dense sub-graphs depicting context around q. This is done for all the news entities $q \in N_A$.

Relation Canonicalization: To effectively present the relations, we use relation canonicalization to standardize the different relations that exist between two entities. Using the pre-trained embeddings of the words present in the relations (phrases) between two entities, we cluster them into c groups. Using the TF-IDF scores of the words present in each cluster, we assign the top-scoring word as the cluster label. The label of the cluster to which an edge (relation) belongs determines it's representation.

2.3 Temporal Diversity Analysis

The distribution of the topics from a particular news genre selected by a media house for publishing may not be neutral. To analyze this temporal bias in topic selection, we use Latent Dirichlet Allocation (LDA) on the news articles of a particular media house and genre. More precisely, given a media house M, and a genre g, we segregate all the articles published by the media house which belong to the genre g month-wise. For all the articles published in a month, we extract the key topics discussed in them using LDA.

3 Experimental Results and Conclusions

We also analyze the system's working on a self-collected dataset. We crawled the websites of 3 news sources and performed article matching using keyword similarity to extract corresponding articles from 3 genres (number)- National (1905), Sports (704) and Entertainment (68). We analyzed the distribution of the Bias Score on the dataset, with Source-1 as the trusted source. Table 1 shows the number of samples with *high* scores, potentially the ones which have changed the representation to a large extent, with Source-3 having more such articles than Source-2.

	Entertainment	Sports	National
Source-2	17	175	594
Source-3	25	261	853

Table 1: The number of articles having bias score greater than the respective distribution's third quartile.

We analyzed the temporal topic diversity using LDA for the articles in the dataset and constructed metro-maps by combining together the articles from a particular journalist. To further examine the under- and over-represented topics in the dataset, we used attention modeling to analyze the difference in the amount of *attention* being given to a topic by various sources, which differs from the LDA analysis since it also incorporates context. Essentially, we train an attention based text summarization model and extract the attention weights of the top attention words in the trusted article and compare their weights in other source articles.

The work embarks the starting point for the analysis of trust from a linguistic point of view. Overall, the system improves the understanding of bias in representation present in articles from different sources and therefore, helps in analyzing them for trust. We also deployed the interface for the system to aid comprehensibility through visualizations.

References

Kohring, M., and Matthes, J. 2007. Trust in news media. *Communication Research* 34:231–252.

Shahaf, D.; Guestrin, C.; and Horvitz, E. 2012. Trains of thought: Generating information maps. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, 899–908. New York, NY, USA: ACM.

Tong, H., and Faloutsos, C. 2006. Center-piece subgraphs: problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 404–413. ACM.