NewsDeps: Visualizing the Origin of Information in News Articles

Authors: Felix Hamborg, Philipp Meschenmoser, Moritz Schubotz, Philipp Scharpf, Bela Gipp

News dependency detection, content borrowing, news plagiarism detection, bias by commission and omission of information.

Abstract

In scientific publications, citations allow readers to assess the authenticity of the presented information and verify it in the original context. News articles, however, for various reasons do not contain citations and only rarely refer readers to further sources. As a result, readers often cannot assess the authenticity of the presented information as its origin is unclear. In times of “fake news,” echo chambers, and centralization of media ownership, the lack of transparency regarding origin, trustworthiness, and authenticity has become a pressing societal issue. We present NewsDeps, the first approach that analyzes and visualizes where information in news articles stems from. NewsDeps employs methods from natural language processing and plagiarism detection to measure article similarity. We devise a temporal-force-directed graph that places articles as nodes chronologically. The graph connects articles by edges varying in width depending on the articles’ similarity. We demonstrate our approach in a case study with two real-world scenarios. We find that NewsDeps increases efficiency and transparency in news consumption by revealing which previously published articles are the primary sources of each given article.

1.1 Introduction

The rise of online news publishing and consumption has made information from various sources and even other countries easily accessible [13], but has also led to a decrease of reporting quality [20, 38]. The increasing pace of the publish-consume cycle leaves publishers with less time for journalistic investigation and information verification. At the same time, the pressure to publish a story soon after the event has happened rises, as competing outlets will do so likewise. Also, journalists routinely copy-edit or reuse information from previously published articles, which increases the chance of spreading incorrect or unverified information even more [12]. In 2010, a study showed that over 80% of articles reporting on the same topic did not add any new information, but merely reused information contained in articles published previously by other outlets [44].

Currently, regular news consumers cannot effectively assess the authenticity of information conveyed in articles. Understanding where information stems from, and how the information differs from the used sources could help readers to assess the authenticity of such information and ease further verification. For the same reasons, documentation of the origin of information is a fundamental standard in scientific writing. News, however, do not contain citations and only rarely refer readers to further articles [5].

The objective of our research is to identify and visualize information reuse (or content borrowing) in articles. These news dependencies manifest themselves as text snippets reused from prior articles. A field that aims at finding instances of text and information reuse is plagiarism detection. The relatedness of plagiarism detection (PD) to our project can be seen directly in the definition of plagiarism: “[…] the use of ideas and/or words from sources […]” [22]. These news dependencies should be found between articles, e.g., how similar are two articles content-wise, and within articles, e.g., which piece of information in one article was used from another article. We call this task news information reuse detection (NIRD). By showing how articles reuse information from other related articles, NIRD helps users to assess the articles’ authenticity. Additionally, NIRD increases the efficiency of news consumption as users can see if an article contains novel information or is a mere copy of other articles.

In Section 2, we give a brief overview of related work. In Section 3, we propose NewsDeps, a NIRD approach that integrates analysis and visualization of news dependencies. Our main contribution is a visualization that reveals which articles reuse information from other articles. We demonstrate this functionality in a case study in Section 4. Our study shows that NewsDeps also provides an overview of current topics, a common use case in regular news consumption. The article concludes with a discussion of future work (Section 5) and a summary (Section 6).

1.2 Background and related work

This section first describes fundamental forms of information reuse in the context of our research objective. Second, we discuss previous methods used to measure semantic textual similarity and approaches using them to identify information reuse.

1.2.1 Forms of information reuse

Information reuse in news is common and often necessary, e.g., due to resource limitations of publishers [9]. In other cases, information reuse is even intended, e.g., by press agencies, which write articles that licensed publishers may copy-edit and publish. Copy-edited articles are often nearly identical and represent the only dependencies that established PD methods detect reliably [18, 31, 35]. The PD community calls this form of information reuse (1) copy and paste [21].

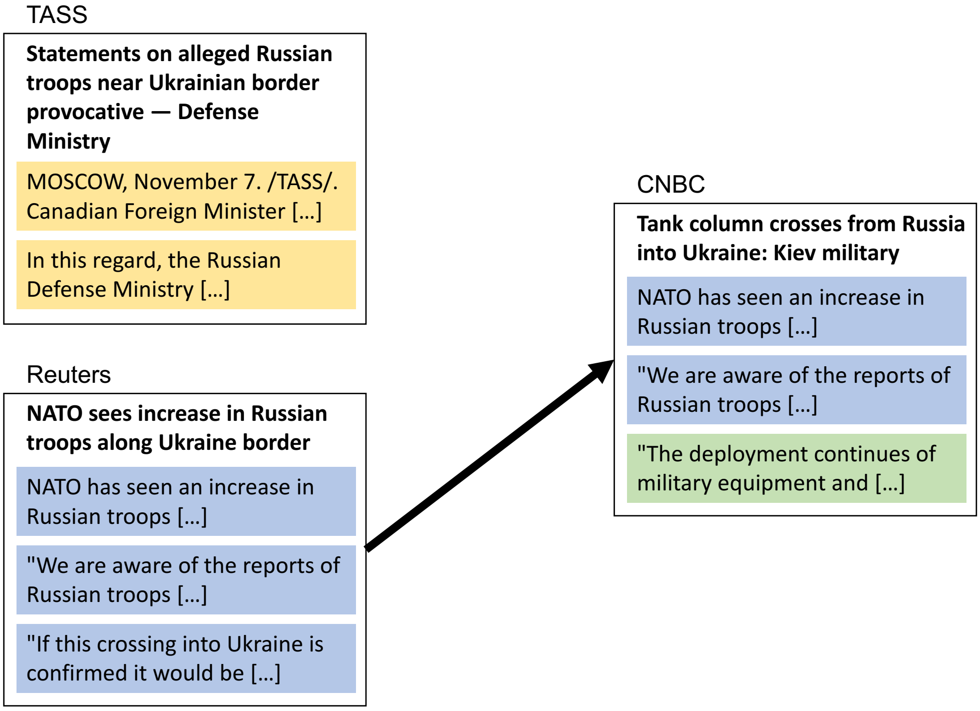

TASS
Statements on alleged Russian troops near Ukrainian border provocative — Defense Ministry
MOSCOW, November 7. /TASS/. Canadian Foreign Minister […]
In this regard, the Russian Defense Ministry […]
Reuters
NATO sees increase in Russian troops along Ukraine border
NATO has seen an increase in Russian troops […]
"We are aware of the reports of Russian troops […]
"If this crossing into Ukraine is confirmed it would be […]
CNBC
Tank column crosses from Russia into Ukraine: Kiev military
NATO has seen an increase in Russian troops […]
"We are aware of the reports of Russian troops […]
"The deployment continues of military equipment and […]

Figure 1: An example of content dependencies between news articles reporting on the same event. Each transparent box represents a single news article; each colored box within a transparent box represents one or more paragraphs of the article. Paragraph boxes from different articles having the same color contain (almost) identical information determined by manual inspection.

Figure 1 shows a real-world example of copy and paste information reuse in news. Three articles published on November 7, 2014 reported on an event that happened during the Ukraine crisis. The colored boxes in each rectangle represent individual paragraphs in the article. The figure shows that CNBC [43] took most of their text from an article previously published by Reuters (blue) [26]. In addition to the copy and paste text reuse, the CNBC article also contains novel information (green). The article published by TASS [42], a news agency owned by the Russian government, contains different information on the event (yellow), which, however, was not mentioned by any of the Western news outlets.

Forms of information reuse that are more complex than copy and paste also exist in news articles. For instance, journalists may (2) paraphrase other articles; (3) tightly copy and merge text segments with slight adjustments (shake and paste) [49], e.g., by substituting words with synonyms; and (4) translate articles written in other languages, e.g., from publishers in other countries [49]. In practice, information reuse in news coverage is a mixture of these four main forms.

1.2.2 Methods to detect information reuse

Established plagiarism detection methods can reliably find copy and paste, i.e., the most basic form of information reuse in news articles [18, 31, 36]. To detect disguised forms of plagiarism, including paraphrases, translations and structural plagiarism – in academic documents – researchers have proposed approaches that use syntactic analysis, semantic analysis, cross-language analysis, or a combination thereof often using machine learning [24, 25]. Syntax-based methods examine syntactic features to compare the structure of two documents [7, 47]. Because parsing and part-of-speech (POS) tagging is computationally expensive, syntax-based approaches are commonly not applied for plagiarism detection use cases, in which the potentially suspicious document is compared to a large collection.

Semantic detection approaches typically consider related terms [2, 4, 46, 48]. Often, semantic approaches employ pairwise comparisons of sentences and use a semantic network, such as WordNet [23], to retrieve terms that are semantically related to the terms in the sentences being compared. Using the set of exactly matching and related terms, the detection approaches derive similarity scores and flag documents as suspicious if the texts exceed a given similarity threshold. Some prior work has gone beyond comparing term-based semantic similarity by also considering similarity in the sentence structure [17, 29]. Such approaches apply semantic role labeling to identify the arguments of a sentence, e.g., the subject, predicate, and object, as well as how they relate, using a pre-defined set of roles from linguistic resources, such as PropBank [30], VerbNet [39], or FrameNet [3]. Existing detection approaches then typically combine the information on semantic arguments with the term-based semantic similarity. For instance, Osman et al. only consider exactly matching words and WordNet-derived synonyms for the similarity assessment if they belong to the same argument in both sentences [28]. Semantic plagiarism detection approaches have been shown to be more effective in identifying disguised plagiarism when compared to text matching approaches [28]. However, the computational effort of semantic approaches is significantly higher than that of character-based approaches. This makes them infeasible for large collections, and hence unsuitable for most practical use cases of plagiarism detection. For example, Bao et al. showed that considering WordNet synonyms, which exemplify a relatively straightforward semantic analysis, increased processing times on average by factor 27 compared to character-based approaches [4]. Cross-language plagiarism detection uses machine translation or other cross-lingual information retrieval methods [32]. Since machine translating text is computationally expensive, and thus infeasible for large document collections, cross-language plagiarism detection methods typically extract only keywords from the input text and query these keywords against an index of keywords extracted from documents in the reference collection. One or both sets of keywords are machine-translated – prior to being matched. Despite recent advances, e.g., the use of word embeddings [8, 24, 45], cross-language plagiarism detection is currently not reliable enough for practical use [32].

Aside from plagiarism detection, semantic textual similarity methods are useful to identify instances of information reuse. Many of the advances in semantic textual similarity research are due to the SemEval series, where the task is to measure the semantic equivalence of two sentences [1]. State-of-the-art semantic textual similarity methods use basic approaches, such as n-gram overlap, WordNet-based node-to-node distance, and syntax-based comparisons, e.g., comparing whether the predicate is identical in two sentences [37]. More advanced methods combine various techniques using deep learning networks and achieve a Pearson correlation to human coders of 0.78 [34] and F1 scores up to 91.6 on the STS datasets of the SemEval series [50]. Since these semantic textual similarity methods focus on sentence similarity, they are useful for the detection of complex, paraphrased instances (see Section **Error! Reference source not found.**). Yet, for news there are currently neither training nor evaluation datasets, making the use of language models difficult.

1.2.3 Approaches to detect information reuse in news

Few NIRD approaches have been proposed, but news dependencies are still non-transparent to users because none of the approaches visualizes the results. One approach employs methods from PD to find unauthorized instances of information reuse in articles written in Korean [35]. The approach computes the similarity of each article pair in a set of related articles. A high similarity between two articles suggests that the latter article contains information from the former article. To assess the similarity of two articles, the approach uses fingerprinting (see Section 2.2). A similar approach is qSign, which finds instances of information reuse in news blogs and articles using also fingerprinting [19]. However, since none of the approaches visualizes dependencies, the origin of information remains unclear in regular news consumption.

1.3 System overview

NewsDeps is a NIRD approach that aims to reveal information reuse on article level, i.e., identify and visualize the main sources of information for each of the articles. When applied to a set of related articles, NewsDeps shows which articles use information from previous articles, and which articles are unrelated. The workflow consists of three phases: (1) news import, (2) similarity measurement, (3) visualization.

1.3.1 Import of news articles

NewsDeps can import articles either from JSON files, the Common Crawl News Archive (also referred to as CC-NEWS in the literature) or by URL import. The system accepts any document that contains a title, main text, publisher, and publishing date and time. Additional fields, such as to define the background color of each article in the visualization, can also be processed.

The URL import allows users to import a list of URLs referring to online articles. The systems then retrieve the required fields using news-please, a web crawler and information extractor tailored for news [15]. News-please is also used to import news articles from the Common Crawl News Archive [27]. To do so, users define filter criteria, such as a date range for the publication date, search terms that headlines or the articles' main text must contain, or a list of outlets. These filter criteria are passed to news-please, which subsequently searches for relevant news articles in the Common Crawl News Archive.[1]

1.3.2 Measurement of information reuse in articles

Upon user request, NewsDeps computes a $k \times k$ document-to-document similarity (d2d) matrix, where k is the number of imported articles. Specifically, we compute only the upper right diagonal half of the d2d matrix, as information can only flow from previous articles to articles published afterward. Which article existed first and thus may have served as a source of information for articles published afterward is determined by the publishing date (see Section 1.3.1). NewsDeps currently supports two basic text similarity measures (1) TF-IDF & cosine and (2) Jaccard, and two plagiarism scores provided by the well-established approaches (3) Sherlock [16] and (4) JPlag [33], which commonly serve as base-line similarity measures in PD. The user can choose which similarity measure to apply. Sherlock and JPlag try to estimate how likely one of two documents was "plagiarized" from the other (first case in Section 4), while the text similarity measures TF-IDF & cosine and Jaccard are more suitable to group similar articles, e.g., by topic (second case). We linearly normalize all similarity values between 0 and 1, where 1 indicates maximum similarity between two articles.

1.3.3 Visualization of news dependencies

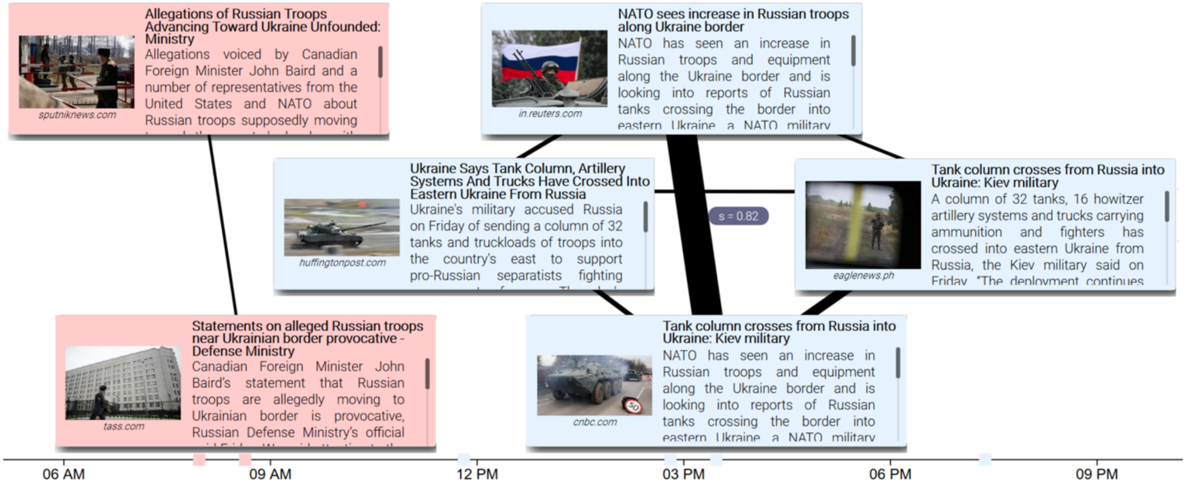
NewsDeps visualizes the dependencies of the imported articles in a directed graph (see Figure 1).

[1] news-please currently accesses the raw archive provided by the Common Crawl project. To speed up the search process, we are planning to preprocess the archive and import extracted articles in a database.

The articles are ordered temporally on one axis (by default the x-axis, but the user can swap the function of both axes). Each node represents a single article. We use the width of the edge between two articles as a visual variable to encode the pair's similarity, i.e., the more similar, the thicker the edge. Pairs below a user-defined similarity threshold are not connected visually.

The other axis, by default the y-axis, allows placing the article in a way that avoids occlusion. Technically, the graph is created by first placing articles at their temporal position. Afterwards, we apply force-direct node placement [10] in the other dimension, i.e., the articles are moved along the second axis until all articles are placed in a way that their distances to each other best represent their similarities. We call the resulting graph temporal-force-directed (TFD) graph. The primary purpose of the TFD graph is to reduce overlap of articles and edge crossings while adhering to chronological order. Note that most visualizations use force-directed graphs to show (dis-)similarity of elements, but the TFD graph's temporal placement of nodes may skew perceived distances between nodes. Therefore, we encode article similarity using the width of the edges.

NewsDeps follows the Information Seeking Mantra [41]: overview first, zoom and filter, then details on demand. Therefore, the approach offers four levels of detail (LOD) to display articles: none (the graph shows each article as a point only), source (the node displays an article's publisher), title (title only), detailed (title, publisher, main picture, and lead paragraph). An automated mode automatically chooses proper LOD, by balancing between showing more details and reducing overlap of visuals. NewsDeps provides various interaction techniques [41] to support the visual exploration by the user. For example, both axes and the graph can be dragged and zoomed to adjust the position and granularity of the time frame, details on demand are shown by hovering over important elements, such as articles and edges, and clicking on an article opens a popup showing the full article (see Figure 2). Each article has a linked axis-indication on each of the axes, which allows users to easily find the article that was published at a specific time (by hovering the article, the indications on both aces are highlighted), and vice versa. Users can set the opacity of nodes, edges, and axis-indications, which helps with occlusion that can occur in analyses of many articles. The edges are naturally directed as to the chronological order as articles can only reuse information from previous articles.



Allegations of Russian Troops Advancing Toward Ukraine Unfounded: Ministry
Allegations voiced by Canadian Foreign Minister John Baird and a number of representatives from the United States and NATO about Russian troops supposedly moving
sputniknews.com

NATO sees increase in Russian troops along Ukraine border
NATO has seen an increase in Russian troops and equipment along the Ukraine border and is looking into reports of Russian tanks crossing the border into eastern Ukraine, a NATO military
in.reuters.com

Ukraine Says Tank Column, Artillery Systems And Trucks Have Crossed Into Eastern Ukraine From Russia
Ukraine's military accused Russia on Friday of sending a column of 32 tanks and truckloads of troops into the country's east to support pro-Russian separatists fighting
huffingtonpost.com

Tank column crosses from Russia into Ukraine: Kiev military
A column of 32 tanks, 16 howitzer artillery systems and trucks carrying ammunition and fighters has crossed into eastern Ukraine from Russia, the Kiev military said on Friday. "The deployment continues
eaglenews.ph

Statements on alleged Russian troops near Ukrainian border provocative - Defense Ministry
Canadian Foreign Minister John Baird's statement that Russian troops are allegedly moving to Ukrainian border is provocative, Russian Defense Ministry's official
tass.com

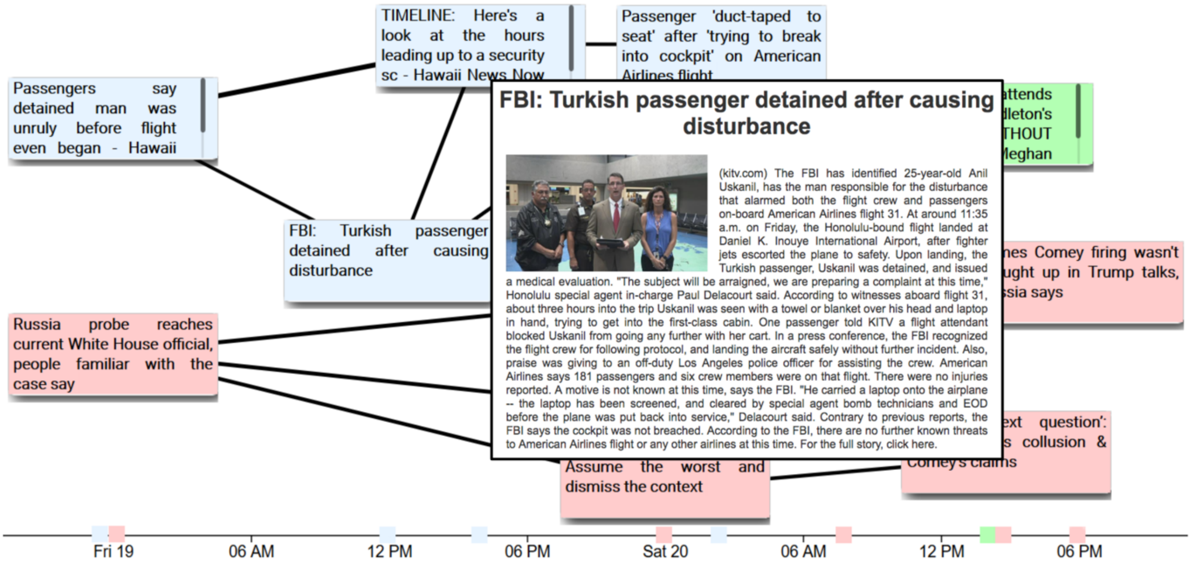
Tank column crosses from Russia into Ukraine: Kiev military
NATO has seen an increase in Russian troops and equipment along the Ukraine border and is looking into reports of Russian tanks crossing the border into eastern Ukraine, a NATO military
cnbc.com

Figure 2: NewsDeps shows articles as nodes in a temporal-force-directed graph

1.4 Case study

We analyze the strengths and weaknesses of our approach in a case study of two real-world news scenarios. In the first case, we investigate how NewsDeps supports exploration of news dependencies. Figure 2 shows the first case: six articles reporting on the same event during the Ukraine crisis on November 11, 2014. We use JPlag as we are interested in the likelihood that one article reused information from another. The edges in **Error! Reference source not found.** show that Western outlets (blue articles) reused information from other Western outlets but not or to a lesser extent from Russian outlets (red articles), and vice versa.[2] Hence, Western and Russian articles form separate, mostly not interconnected groups. The thick edge in the middle reveals that the CNBC article contains a large portion of information from the Reuters press release: the authors only changed the title and few other paragraphs towards the end of the article; the lead paragraph is a 1:1 copy, see the labels of both nodes. NewsDeps additionally reveals content differences through the detailed LOD: Western media reported that Russian troops invaded Ukraine, while the Russian state news agency TASS contrarily stated that even the claims that tanks were near the borders were false.

In the second case, we investigate how NewsDeps supports regular news consumption. We use TF-IDF & cosine since we are interested in grouping articles reporting on the same topic, and not to find instances of information reuse. Figure 3 shows three clusters of interconnected articles, each cluster reporting on a separate topic on May 19th, 2017. With the LOD set to title, the visualization helps to get an overview of the different topics quickly. The TFD graph helps to see when coverage on each of the topics began, and when new articles were published. The temporal placement also allows users to track when and what information is added to the coverage. For example, early articles of the blue topic state that a Turkish passenger was detained on a flight, while the last article adds more detailed information on how the passenger was fixated. Finally, the full article popup allows users to read the whole story.


TIMELINE: Here's a look at the hours leading up to a security sc - Hawaii News Now
Passenger 'duct-taped to seat' after 'trying to break into cockpit' on American Airlines flight
Passengers say detained man was unruly before flight even began - Hawaii
ttends
dleton's
THOUT
Meghan
FBI: Turkish passenger detained after causing disturbance
nes Comey firing wasn't ught up in Trump talks, ssia says
Russia probe reaches current White House official, people familiar with the case say
ext question': s collusion & Comey's claims
Assume the worst and dismiss the context
Fri 19
06 AM
12 PM
06 PM
Sat 20
06 AM
12 PM
06 PM

FBI: Turkish passenger detained after causing disturbance

(kitv.com) The FBI has identified 25-year-old Anil Uskanil, has the man responsible for the disturbance that alarmed both the flight crew and passengers on-board American Airlines flight 31. At around 11:35 a.m. on Friday, the Honolulu-bound flight landed at Daniel K. Inouye International Airport, after fighter jets escorted the plane to safety. Upon landing, the Turkish passenger, Uskanil was detained, and issued a medical evaluation. "The subject will be arraigned, we are preparing a complaint at this time," Honolulu special agent in-charge Paul Delacourt said. According to witnesses aboard flight 31, about three hours into the trip Uskanil was seen with a towel or blanket over his head and laptop in hand, trying to get into the first-class cabin. One passenger told KITV a flight attendant blocked Uskanil from going any further with her cart. In a press conference, the FBI recognized the flight crew for following protocol, and landing the aircraft safely without further incident. Also, praise was giving to an off-duty Los Angeles police officer for assisting the crew. American Airlines says 181 passengers and six crew members were on that flight. There were no injuries reported. A motive is not known at this time, says the FBI. "He carried a laptop onto the airplane -- the laptop has been screened, and cleared by special agent bomb technicians and EOD before the plane was put back into service," Delacourt said. Contrary to previous reports, the FBI says the cockpit was not breached. According to the FBI, there are no further known threats to American Airlines flight or any other airlines at this time. For the full story, click here.

Figure 3: Three groups of articles, each representing a different topic. The popup shows details of a selected article.

1.5 Discussion and future work

NewsDeps is a first step in the direction of semi-automated NIRD. The first case in our study demonstrated that NewsDeps helps readers to differentiate between articles that add new information or present the event from a different perspective and others that merely repeat previously published information. Hence, the visualization increases the transparency because users can follow the dissemination and determine the origin of information effectively in a group of related articles. NewsDeps can also be helpful to researchers concerned with relations and dependencies of articles and publishers, such as in the social sciences. The second case showed that NewsDeps also allows getting an overview of multiple topics, which is the primary purpose of popular news aggregators, such as Google News. However, to quickly get an overview of many articles, the visualization currently lacks cluster labels. By summarizing a topic with one label, readers would also be able to get an overview of large-scale news scenarios.

The first case of our case study showed that NewsDeps detects simple forms of information reuse. In the future, we plan to investigate how to identify also more complex forms, such as paraphrased

[2] This phenomenon is typically studied as *media bias by source selection* in the social sciences.

articles or when authors omitted single sentences or paragraphs. We plan to analyze the events described in articles: an event can be described by answers to the five journalistic W and one H-questions (5W1H), i.e., who did what, when, where, why, and how [14, 40]. Instead of analyzing all tokens in the text, we plan to conceive similarity measures that analyze and compare the 5W of described events in two articles. Another research direction will be to train a language model, such as BERT [6], and devise a neural architecture to identify information reuse in news. While datasets for training exist and have successfully been used on other domains [50], for the news domain a dataset must first be created.

NewsDeps currently allows users to analyze dependencies on article level. Following the overview first, details on demand mantra [41], we plan to add a visualization where users can select an article and compare it with its main sources. The visualization will then enable the users to view which information has been committed or omitted compared to the sources. PD visualizations commonly allow for detailed comparison of a selected document and suspicious documents [11], in our case the main sources.

1.6 Conclusion

In this paper, we described NewsDeps, the first information reuse detection system for news, which combines analysis and visualization of news dependencies on article-level. With the help of state-of-the-art similarity measures, our system determines the main sources of information for each of the analyzed articles. The visualization shows all articles in a temporal-force-directed (TFD) graph, which reveals if and how strongly articles relate to another. By showing how related articles reuse information, NewsDeps helps users to assess the authenticity. Additionally, NewsDeps increases the efficiency of news consumption as users can see if an article contains novel information or is a mere copy of other articles. In a case study, we demonstrated that NewsDeps reveals which articles repeat known information, and which articles contain new information or show a different perspective than previously published articles. Hence, the TFD graph helps to understand the dependencies in news coverage on article level.

1.7 References

[1] Agirre, E. et al. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (2016), 497–511.

[2] Alzahrani, S. and Salim, N. 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF 2010. CEUR Workshop Proceedings (2010), 1–8.

[3] Baker, C.F. et al. 1998. The Berkeley FrameNet Project. Proceedings of the 36th annual meeting on Association for Computational Linguistics (Stroudsburg, PA, USA, 1998), 86–90.

[4] Bao, J. et al. 2007. Comparing Different Text Similarity Methods.

[5] Christian, D. et al. 2014. The Associated Press stylebook and briefing on media law. The Associated Press.

[6] Devlin, J. et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv: 1810.04805. (2018).

[7] Elhadi, M. and Al-Tobi, A. 2009. Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. 2009 4th International Conference on Computer Sciences and Convergence Information Technology (Seoul, South Korea, 2009).

[8] Ferrero, J. et al. 2017. Using Word Embedding for Cross-Language Plagiarism Detection. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (2017), 415–421.

[9] Frank, R. 2003. “These crowded circumstances”: When pack journalists bash pack journalism. Journalism. 4, 4 (2003), 441–458. DOI:https://doi.org/10.1177/14648849030044003.

[10] Fruchterman, T.M.J. and Reingold, E.M. 1991. Graph drawing by force-directed placement. Software: Practice and Experience. 21, 11 (1991), 1129–1164. DOI:https://doi.org/10.1002/spe.4380211102.

[11] Gipp, B. 2014. Citation-based plagiarism detection. Springer Vieweg.

[12] Hamborg, F. et al. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries. 20, 4 (2019), 391–415. DOI:https://doi.org/10.1007/s00799-018-0261-y.

[13] Hamborg, F. et al. 2018. Bias-aware news analysis using matrix-based news aggregation. International Journal on Digital Libraries. (May 2018). DOI:https://doi.org/10.1007/s00799-018-0239-9.

[14] Hamborg, F. et al. 2019. Giveme5W1H: A Universal System for Extracting Main Events from News Articles. Proceedings of the 13th ACM Conference on Recommender Systems, 7th International Workshop on News Recommendation and Analytics (INRA 2019) (Copenhagen, Denmark, 2019).

[15] Hamborg, F. et al. 2017. news-please: A Generic News Crawler and Extractor. Proceedings of the 15th International Symposium of Information Science (2017), 218–223.

[16] Joy, M. and Luck, M. 1999. Plagiarism in programming assignments. IEEE Transactions on Education. 42, 2 (1999), 129–133. DOI:https://doi.org/10.1109/13.762946.

[17] Kent, C.K. and Salim, N. 2010. Web Based Cross Language Plagiarism Detection. 2010 Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM) (2010), 199–204.

[18] Kienreich, W. et al. 2006. Plagiarism Detection in Large Sets of Press Agency News Articles. Database and Expert Systems Applications, 2006. DEXA '06. 17th International Workshop on (2006).

[19] Kim, J.W. et al. 2009. Efficient overlap and content reuse detection in blogs and online news articles. Proceedings of the 18th international conference on World wide web (2009), 81–90.

[20] Marchi, R. 2012. With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity." Journal of Communication Inquiry. 36, 3 (2012), 246–262. DOI:https://doi.org/10.1177/0196859912458700.

[21] Maurer, H. et al. 2006. Plagiarism - A Survey. Journal of Universal Computer Science. 12, 8 (2006), 1050–1084. DOI:https://doi.org/10.3217jucs-012-08-1050.

[22] Meuschke, N. and Gipp, B. 2013. State-of-the-art in detecting academic plagiarism. International Journal for Educational Integrity. 9, 1 (2013), 50. DOI:https://doi.org/10.21913/IJEI.v9i1.847.

[23] Miller, G.A. et al. 1990. Introduction to wordnet: An on-line lexical database. International Journal of Lexicography. 3, 4 (1990), 235–244. DOI:https://doi.org/10.1093/ijl/3.4.235.

[24] Moreau, E. et al. 2015. Author verification: Basic stacked generalization applied to predictions from a set of heterogeneous learners. CEUR Workshop Proceedings (2015).

[25] Mozgovoy, M. et al. 2010. Automatic Student Plagiarism Detection: Future Perspectives. Journal of Educational Computing Research. (2010). DOI:https://doi.org/10.2190/ec.43.4.e.

[26] NATO sees increase in Russian troops along Ukraine border: 2014. http://www.reuters.com/article/us-ukraine-crisis-nato-idUSKBN0IR1KI20141107. Accessed: 2018-08-23.

[27] News Dataset Available: 2016. http://web.archive.org/save/http://commoncrawl.org/2016/10/news-dataset-available. Accessed: 2020-03-03.

[28] Osman, A.H. et al. 2012. An improved plagiarism detection scheme based on semantic role labeling. Applied Soft Computing. 12, 5 (2012), 1493–1502. DOI:https://doi.org/10.1016/j.asoc.2011.12.021.

[29] Osman, A.H. et al. 2012. Plagiarism detection scheme based on Semantic Role Labeling. International Conference on Information Retrieval & Knowledge Management (CAMP) (Kuala Lumpur, Malaysia, 2012), 30–33.

[30] Palmer, M. et al. 2005. The proposition bank: An annotated corpus of semantic roles. Computational Linguistics. 31, 1 (2005), 71–106. DOI:https://doi.org/10.1162/0891201053630264.

[31] Pera, M.S. and Ng, Y.K. 2011. SimPaD: A word-similarity sentence-based plagiarism detection tool on Web documents. Web Intelligence and Agent Systems. (2011). DOI:https://doi.org/10.3233/WIA-2011-0203.

[32] Potthast, M. et al. 2011. Cross-language plagiarism detection. Language Resources and Evaluation. 45, 1 (2011), 45–62. DOI:https://doi.org/10.1007/s10579-009-9114-z.

[33] Prechelt, L. et al. 2002. Finding Plagiarisms among a Set of Programs with JPlag. Journal Of Universal Computer Science. 8, 11 (2002), 1016–1038. DOI:https://doi.org/10.3217/jucs-008-11-1016.

[34] Rychalska, B. et al. 2016. Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (2016), 602–608.

[35] Ryu, C.-K. et al. 2009. A Detecting and Tracing Algorithm for Unauthorized Internet-News Plagiarism Using Spatio-Temporal Document Evolution Model. Proceedings of the 2009 ACM Symposium on Applied Computing (2009), 863–868.

[36] Sanderson, M. 1997. Duplicate detection in the Reuters collection. " Technical Report (TR-1997-5) of the Department of Computing Science at the University of Glasgow G12 8QQ, UK." (1997).

[37] Šarić, F. et al. 2012. Takelab: Systems for Measuring Semantic Text Similarity. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and

Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation.

[38] Scheufele, D.A. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. Mass Communication & Society. 3, 2–3 (2000), 297–316.

[39] Schuler, K.K. 2005. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. University of Pennsylvania.

[40] Sharma, S. et al. 2013. News Event Extraction Using 5W1H Approach & Its Analysis. International Journal of Scientific & Engineering Research. 4, 5 (2013), 2064–2068.

[41] Shneiderman, B. 1996. The eyes have it: a task by data type taxonomy for information visualizations. Proceedings 1996 IEEE Symposium on Visual Languages. (1996), 336–343. DOI:https://doi.org/10.1109/VL.1996.545307.

[42] Statements on alleged Russian troops near Ukrainian border provocative - Defense Ministry: 2014. http://tass.com/russia/758504. Accessed: 2018-08-23.

[43] Tank column crosses from Russia into Ukraine: Kiev military: 2014. http://www.cnbc.com/id/102155038. Accessed: 2018-08-22.

[44] The Media Insight Project 2014. The Personal News Cycle: How Americans Get Their News.

[45] Thompson, V. and Bowerman, C. 2017. Detecting Cross-Lingual Plagiarism Using Simulated Word Embeddings. CoRR. abs/1712.1, (2017).

[46] Tsatsaronis, G. et al. 2010. Identifying free text plagiarism based on semantic similarity. Proceedings of the 4th International Plagiarism Conference (Newcastle upon Tyne, UK, 2010).

[47] Uzuner, O. and Katz, B. 2005. Capturing expression using linguistic information. Proceedings of the 20th national conference on Artificial intelligence (Pittsburgh, Pennsylvania, USA, 2005), 1124–1129.

[48] Vu, H.H. et al. 2014. Sentence similarity by combining explicit semantic analysis and overlapping n-grams. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2014).

[49] Weber-Wulff, D. 2010. Test cases for plagiarism detection software. Proceedings of the 4th International Plagiarism Conference. (2010).

[50] Yang, Z. et al. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. (Jun. 2019).

Please use the following BibTeX code to cite this paper:

```
@incollection{Hamborg2021c,
      author = {Hamborg, Felix and Meschenmoser, Philipp and Schubotz,
      Moritz and Scharpf, Philipp and Gipp, Bela},
      booktitle = {Wahrheit und Fake im postfaktisch-digitalen Zeitalter},
      doi = {10.1007/978-3-658-32957-0},
      editor = {Klimczak, Peter and Zoglauer, Thomas},
      isbn = {978-3-658-32957-0},
      publisher = {Springer Vieweg},
      title = {{NewsDeps: Visualizing the Origin of Information in News
      Articles}},
      year = {2021},
      topic = {newsanalysis},
}
```