Implementing Accuracy for Responsible AI in Newsrooms

Roxana L. Quintanilla Portugal[1,2], Juliana Delle Ville[3], Leandro Antonelli[3,4]

[1]Departamento Académico de Informática, UNSAAC, Cusco, Perú
[2] Department of Media and Communication, LMU, Munich, Alemania
[3]Lifia, Fac. de Informática, UNLP, La Plata, Bs As, Argentina
[4]UAI CAETI – Facultad de Tecnología Informática, UAI, Bs As, Argentina

roxana.quintanilla@unsaac.edu.pe
{juliana.delleville, leandro.antonelli}
@lifia.info.unlp.edu.ar

Abstract. This paper explores the intersection between software development and journalism, highlighting the fundamental importance of implementing non-functional requirements to achieve a balance between immediacy and accuracy. In software development, requirements encompass user needs and demand precise and continuous updates to meet time-to-market demands. In journalism, audience engagement drives the need for precise news coverage, especially with the growth of artificial intelligence (AI). Non-functional requirements (NFRs) have gained relevance, emphasizing the need for effective balance. This work investigates the integration of technologies such as Named Entity Recognition (NER) and topic modeling into news updating processes as means to enhance both efficiency and precision. Additionally, this strategy, beneficial in requirements management and applicable across domains, is explored. The article is structured to delve into the imperatives of news writing, the proposed strategy, potential applications, and directions for future research.

Keywords: Non-functional Requirements, Artificial Intelligence (AI), Natural Language Processing (NLP), Journalism, Requirements Engineering.

1. Introduction

In the realm of software development, requirements are paramount for ensuring the quality and success of a product. These requirements manifest in various forms, encompassing the needs and expectations of users, the necessity for reengineering or refactoring existing software applications, and compliance with laws and regulations.

In journalism, audience engagement with published news performance holds significant weight in deciding upon updates, achieved through the acquisition of information from reliable sources. Consequently, newsrooms strive for heightened *accuracy* in their coverage, particularly with the advent of artificial intelligence (AI), which has escalated the demand for producing *transparent*, *precise*, and *reliable* news for the audience [1]. This evolution in software requirements has underscored the

Proceedings of the 27th Workshop on Requirements Engineering (WER24), August 07-09, 2024, Buenos Aires, Argentina.
DOI: 10.29327/1407529.27-30

significance of Non-Functional Requirements (NFRs) and the need to balance them effectively [2].

The research challenge arises when exploring how technologies such as Named Entity Recognition (NER) and topic modeling can be effectively integrated into the news updating process within media outlets, aiming to balance both efficiency and precision in identifying relevant information and adapting news content based on emerging sources and discoveries. It's worth noting that this strategy is also beneficial in requirements management by incorporating information obtained from big data, thus improving both knowledge and the requirements themselves [3]. Additionally, we find that the contribution of this strategy is useful in other areas such as the legal field, where it seeks jurisprudence on specific facts [4].

The remainder of the document is organized as follows: Section 2 delves into the imperative for *precision* and *immediacy* within newsrooms. Section 3 outlines the strategy devised to fulfill this imperative. Section 4 examines potential applications of this strategy in diverse domains. Lastly, we conclude by suggesting potential avenues for future research.

2. Background

The problem we are addressing falls within the realm of AI-driven news curation [5], which involves using AI technology to assist in tasks such as gathering, organizing, filtering, and presenting news. In this context, there are Non-Functional Requirements (NFRs) that influence or constrain these tasks, with three key metrics regarding audience interest: *recency*, *relevance*, and *diversity* of stories [6]. In this work, we aim to balance the NFRs demanded by journalistic rigor, such as *recency* (also referred as immediacy or efficiency), with precision (also referred as accuracy).

The first author's experience with the journalism industry[1] confirms this demand by eliciting requirements for responsible fact-checking with AI, using an ad hoc process [7]. The Softgoals Interdependence Graph (SIG) [8] shown in Fig. 1 indicates that *immediacy* negatively impacts *accuracy*. To address this discrepancy, the journalists involved in the project have worked on operationalizations that can help balance *accuracy* with *immediacy*. Fig. 2 shows four operationalizations that help to improve *accuracy*: i) the four-eyes principle, ii) preventing factual errors, iii) error-free text, and iv) not forgetting news updates. The latter is divided into two operationalizations: a) correction updates and b) coverage updates. This work implements the coverage of updates.

In addition to addressing the *recency/immediacy* of news, our proposal seeks operationalizations to improve the *transparency* of these automations, facilitating their

[1] www.algorithmic.news

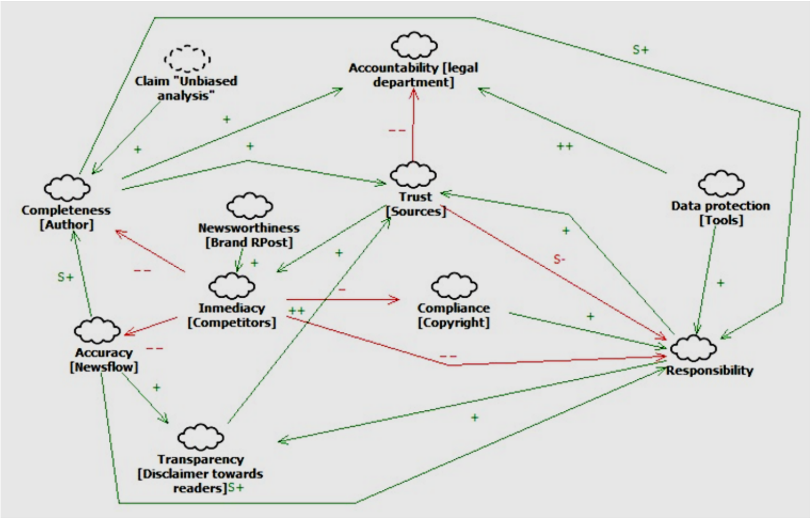


Figure 1. SIG for Responsible Fact Checking

use, understanding, and intervention by journalists when necessary. In this sense, we aim to overcome the limitations of cutting-edge algorithmic techniques that, by operating opaquely, do not contribute to the goals shown in Fig.1. To the best of our knowledge, this problem is being addressed through Machine Learning methods [9] mainly focused on speeding up news production. However, the issue of *responsibility* [1], as often demanded by journalists, is a less explored path, especially in terms of a *transparent* implementation [10].

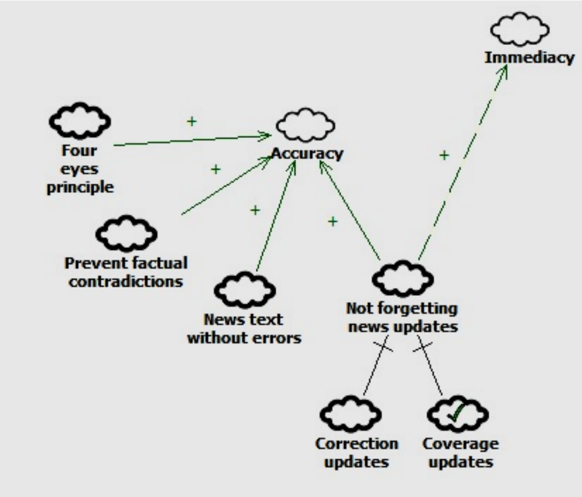


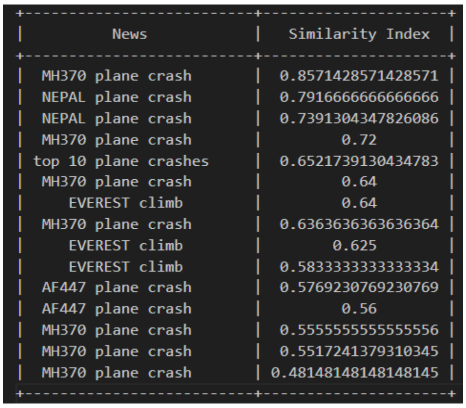
Figure 2. Operationalization Contributing to Accuracy in Responsible Fact-Checking

3. Strategy implementation

In text analysis, it is essential to think about reducing the complexity of automatically summarizing documents, which would allow efficient comparison of news to obtain greater coverage. In this sense, key elements, such as verbs and nouns, help identify concepts related to people, companies, organizations, locations, dates, and more. These elements are commonly known as entities, which Artificial Intelligence frameworks identify using modules known as Named Entity Recognition (NER). NER comprises trained models that can identify specific concepts such as people and organizations, using predefined rules to recognize dates.

Another valuable technique for extracting high-level concepts from text is topic modeling, which encapsulates the main themes discussed in a given text. Various algorithms, including both supervised and unsupervised machine learning models, are employed to conduct topic modeling. For instance, Sajid et al. [11] utilize topic modeling to summarize short texts, Kim et al. [12] leverage it to uncover new topics for text mining, and Li et al. [13] employ it for tagging purposes. One widely used model for topic extraction is Latent Dirichlet Allocation (LDA), an unsupervised model that identifies topics based on the text corpus.

To achieve a balanced implementation that satisfices[2] *accuracy*, *transparency*, and *efficiency* in identifying news articles that enhance coverage of a pivot news item, we evaluate strategies to find the similarity of texts. The greater the similarity of a document to the pivot news, the less additional value it contributes to the overall coverage. First, we combine named entity recognition (NER) techniques with latent Dirichlet allocation (LDA) to improve *accuracy* and *efficiency*, although these techniques cannot guarantee *transparency*. Our methodology streamlines the document comparison process through a structured pipeline. Initially, the texts undergo preprocessing, which involves converting the text to lowercase, removing stop-words to eliminate irrelevant words, and stemming to unify conjugated forms. Following preprocessing, NER is applied to construct a set of entities for each document. Subsequently, LDA is utilized on these entities to generate a set of topics for each document. Finally, Jaccard's similarity measure is employed to compare the sets of topics across documents. To evaluate the proposed strategy, 15 news articles were tested, including reports on incidents such as MH370, AF447, and Nepal, all related to flights and climbing.



```
+-------------------------+---------------------+
|          News           |  Similarity Index   |
+-------------------------+---------------------+
|  MH370 plane crash      |  0.8571428571428571 |
|  NEPAL plane crash      |  0.7916666666666666 |
|  NEPAL plane crash      |  0.7391304347826086 |
|  MH370 plane crash      |         0.72        |
| top 10 plane crashes    |  0.6521739130434783 |
|  MH370 plane crash      |         0.64        |
|     EVEREST climb       |         0.64        |
|  MH370 plane crash      |  0.6363636363636364 |
|     EVEREST climb       |        0.625        |
|     EVEREST climb       |  0.5833333333333334 |
|  AF447 plane crash      |  0.5769230769230769 |
|  AF447 plane crash      |         0.56        |
|  MH370 plane crash      |  0.5555555555555556 |
|  MH370 plane crash      |  0.5517241379310345 |
|  MH370 plane crash      | 0.48148148148148145 |
+-------------------------+---------------------+
```

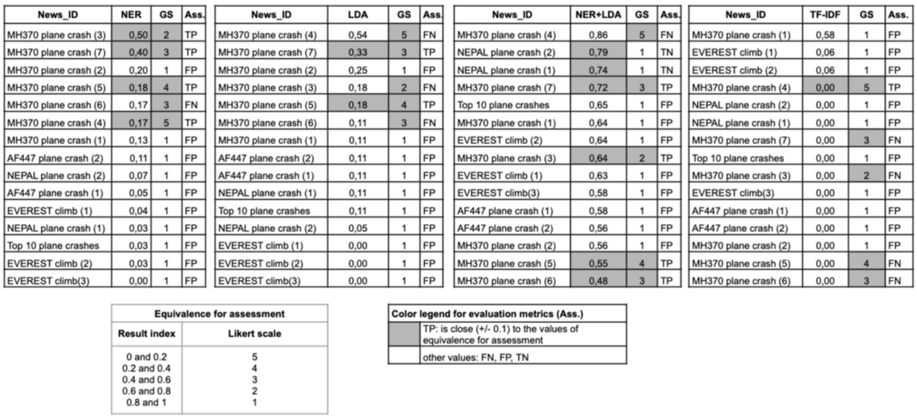
Figure 3. Similarity index of NER + LDA + Jaccard strategy

The results in Fig. 3 indicate that at least the second half of the texts could give more coverage to the pivotal news, however this is not *accurate* because in that second half there are texts from AF447 and the Everest Climb that are not related to the pivot news. Although the syntactic strategy used suggests an index of dissimilarity in the contents that could indicate possibilities of having greater coverage of a news story, nuanced distinctions arise when examining the specificity of the flights boarded. To address these distinctions, we evaluate two individual strategies: NER and LDA. We want to identify if the texts mention the same or more entities or topics than those

[2] This term was coined by Herbert Simon, a combination of two words: “satisfy” and “suffice”, thus when making decisions a person chooses the best enough option satisficing an NFR. Therefore, when dealing with NFRs, we can satisfice them only to some degree, unlike functional requirements (FRs) that can be fully achieved.

existing in the pivot news. It should be noted that Jaccard is applied in all of them as a final step. Additionally, to improve *transparency*, we process our texts using Term Frequency-Inverse Document Frequency (TF-IDF), this strategy allows one to verify the values assigned to each document term in a corpus.

Fig. 4 presents the results of all tested strategies. To evaluate them, a Gold Standard (GS) was carried out to identify which texts give greater coverage to the pivot news. The larger the Likert scale, the greater the coverage of the text. Then we created an equivalence table to align the similarity index resulting from the strategies with the Likert scale measurement. A news item rated with a higher scale should be related to a low similarity index with the pivotal news item. With this table we classify the results using the values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Finally, we evaluate them using the Recall, Precision and F1- score metrics used in data mining and/or artificial intelligence strategies.



News_ID	NER	GS	Ass.
MH370 plane crash (3)	0,50	2	TP
MH370 plane crash (7)	0,40	3	TP
MH370 plane crash (2)	0,20	1	FP
MH370 plane crash (5)	0,18	4	TP
MH370 plane crash (6)	0,17	3	FN
MH370 plane crash (4)	0,17	5	TP
MH370 plane crash (1)	0,13	1	FP
AF447 plane crash (2)	0,11	1	FP
NEPAL plane crash (2)	0,07	1	FP
AF447 plane crash (1)	0,05	1	FP
EVEREST climb (1)	0,04	1	FP
NEPAL plane crash (1)	0,03	1	FP
Top 10 plane crashes	0,03	1	FP
EVEREST climb (2)	0,03	1	FP
EVEREST climb(3)	0,00	1	FP

News_ID	LDA	GS	Ass.
MH370 plane crash (4)	0,54	5	FN
MH370 plane crash (7)	0,33	3	TP
MH370 plane crash (2)	0,25	1	FP
MH370 plane crash (3)	0,18	2	FN
MH370 plane crash (5)	0,18	4	TP
MH370 plane crash (6)	0,11	3	FN
MH370 plane crash (1)	0,11	1	FP
AF447 plane crash (2)	0,11	1	FP
AF447 plane crash (1)	0,11	1	FP
NEPAL plane crash (1)	0,11	1	FP
Top 10 plane crashes	0,11	1	FP
NEPAL plane crash (2)	0,05	1	FP
EVEREST climb (1)	0,00	1	FP
EVEREST climb (2)	0,00	1	FP
EVEREST climb(3)	0,00	1	FP

News_ID	NER+LDA	GS	Ass.
MH370 plane crash (4)	0,86	5	FN
NEPAL plane crash (2)	0,79	1	TN
NEPAL plane crash (1)	0,74	1	TN
MH370 plane crash (7)	0,72	3	TP
Top 10 plane crashes	0,65	1	FP
MH370 plane crash (1)	0,64	1	FP
EVEREST climb (2)	0,64	1	FP
MH370 plane crash (3)	0,64	2	TP
EVEREST climb (1)	0,63	1	FP
EVEREST climb(3)	0,58	1	FP
AF447 plane crash (1)	0,58	1	FP
AF447 plane crash (2)	0,56	1	FP
MH370 plane crash (2)	0,56	1	FP
MH370 plane crash (5)	0,55	4	TP
MH370 plane crash (6)	0,48	3	TP

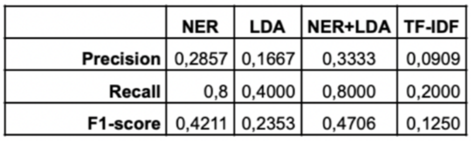
News_ID	TF-IDF	GS	Ass.
MH370 plane crash (1)	0,58	1	FP
EVEREST climb (1)	0,06	1	FP
EVEREST climb (2)	0,06	1	FP
MH370 plane crash (4)	0,00	5	TP
NEPAL plane crash (2)	0,00	1	FP
NEPAL plane crash (1)	0,00	1	FP
MH370 plane crash (7)	0,00	3	FN
Top 10 plane crashes	0,00	1	FP
MH370 plane crash (3)	0,00	2	FN
EVEREST climb(3)	0,00	1	FP
AF447 plane crash (1)	0,00	1	FP
AF447 plane crash (2)	0,00	1	FP
MH370 plane crash (2)	0,00	1	FP
MH370 plane crash (5)	0,00	4	FN
MH370 plane crash (6)	0,00	3	FN

Equivalence for assessment	
Result index	**Likert scale**
0 and 0.2	5
0.2 and 0.4	4
0.4 and 0.6	3
0.6 and 0.8	2
0.8 and 1	1

Color legend for evaluation metrics (Ass.)	
▨	TP: is close (+/- 0.1) to the values of equivalence for assessment
	other values: FN, FP, TN

Figure 4. Comparison of results of all tested strategies

To the best of our knowledge, we are addressing a common text similarity problem with various approaches by considering non-functional requirements (NFRs) as first-class requirements. Therefore, with the results in Fig. 4, we have that the NER + LDA strategy produces the best *precision*[3], particularly when evaluating the harmonic mean F1 score (Fig. 5). But, the strategies hardly satisfy *precision*.



	NER	LDA	NER+LDA	TF-IDF
Precision	0,2857	0,1667	0,3333	0,0909
Recall	0,8	0,4000	0,8000	0,2000
F1-score	0,4211	0,2353	0,4706	0,1250

Figure 5. Confusion matrix of the results

On the other hand, *efficiency* is achieved in all strategies given the current corpus is limited to 15 texts. However, neither strategy can satisfy the *transparency* criterion, as both NER and LDA

[3] In this work, equivalent to *accuracy*. It is important to note that while both *accuracy* and *precision* have specific definitions in the field of data analysis, *accuracy* is the term used by journalists in this context. However, the measurement used to evaluate the performance of an algorithm provides results in terms of *precision*.

algorithms are often perceived as black boxes due to the complexity inherent in the *interpretation* of their results.

A granular analysis of each strategy reveals that NER and NER +LDA achieve the highest Recall with 80%, positioning them as strategies for further improvement, especially when considering the importance of retrieval tasks in natural language processing (NLP) for the engineering. requirements. As cited by D. Berry [14], "a tool that falls short of close to 100% recall, applied to the development of a high-dependability system, may even be useless, because to find the missing information, a human has to do the entire task manually anyway".

The proposed strategy is implemented through a Python script utilizing NLP libraries such as Spacy [15] and NLTK [16], along with the Keras [17] library for creating and training an unsupervised LDA model. This algorithm code and the text corpus are available on GitHub through Zenodo [18].

4. Applications beyond journalism

The described text analysis workflow, which comprises techniques such as entity extraction, topic modeling, and result ranking, extends its relevance beyond journalism into diverse domains, including the legal sector. In legal contexts, this approach streamlines document review processes, aiding in the analysis of legal documents, contracts, and court rulings. By extracting entities such as legal terms, entities, and dates, legal professionals can efficiently identify relevant information, track case precedents, and conduct comprehensive legal research. Moreover, it facilitates *compliance* monitoring by automatically analyzing regulatory documents and identifying compliance risks [19]. The ability to uncover patterns, trends, and relationships within textual data empowers legal professionals to make informed decisions, streamline workflows, and ensure regulatory adherence. Thus, the application of this text analysis workflow in the legal domain can enhances *efficiency*, *accuracy*, and *compliance*, offering significant benefits to legal practitioners and organizations alike.

5. Conclusion

Our proposal is distinguished from the state of the art by not depending on trained models that could overlook relevant information as seen in Fig. 3 where texts with less similarity that would indicate greater coverage are not necessarily so, since they may be sharing the same lexicon (entities or topics), but they are not necessarily referred to the pivot news. The limitation of the strategies used shows that such algorithms can lead to a biased or incomplete presentation of information, thus affecting the quality and objectivity of journalistic reports. In Fig. 4 we evaluate the strategies that, being more syntactic and independent of the domain, can provide greater *efficiency*, however they still do not give results with good *accuracy* and much less *transparency*.

The work presented contributes to showing that by seeking a balance between qualities, there is greater *awareness* of the results produced by an algorithm before searching for the next algorithm with the highest performance that exists. By evaluating each strategy, we are helping to design solutions that best meet the demands of journalists who seek to intervene and verify each step of the process to have greater *accountability* for their decisions in the face of AI tools.

As threats to the validity of the results of this work, we have that the GS was carried out by only one of the authors. Additionally, given the diversity of algorithms for NER or LDA, results may vary.

Finally, it is worth noting that this strategy can be useful not only in the journalistic field but also for requirements management. Identifying similar requirements, merging them, or enriching them can help create products that depend on time to market [20] for their success. The ability to continually adapt and improve requirements throughout the product lifecycle is crucial to ensuring competitiveness and relevance in an ever-changing business environment.

As for future work, our goal is to implement a minimum viable product that can be used in newsrooms for validation. In addition, we seek to further explore the implementation of NFRs identified in the journalistic environment, as well as within the framework of the aforementioned project. Along this path, we look for other approaches to identify more RNFs early, such as that of Cysneiros and Leite [21] that uses the identification of the lexicon and its relationships in a graph as a mechanism to find not only NFRs but also operationalizations (FR). With current algorithms like NER, we can overcome the manual work they faced in that work.

Acknowledgment

We extend our sincere gratitude to the Volkswagen Foundation for sharing data, and Portugal appreciates the support of the research project "Towards Responsible AI in Local Journalism" with grant number 9B761 and 9B390.

References

1. Diakopoulos, N.: Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62 (2016). https://doi.org/10.1145/2844110
2. Portugal, R. L. Q.: Speeding-Up Non-Functional Requirements Elicitation. Doctoral Thesis (2020) PUC-Rio University. https://renati.sunedu.gob.pe/bitstream/sunedu/2137208/1/QuintanillaPortugalRL.pdf
3. Henriksson, A., Zdravkovic, J.: Holistic data-driven requirements elicitation in the big data era. Softw Syst Model **21**, 1389–1410 (2022). https://doi.org/10.1007/s10270-021-00926-6.
4. Mentzingen, H., António, N., & Bacao, F. (2023).: Automation of legal precedents retrieval: findings from a literature review. International Journal of Intelligent Systems (2023). https://doi.org/10.1155/2023/6660983
5. Manisha, R., & Acharya, K.: The Impact of Artificial Intelligence on News Curation and Distribution: A Review Literature. Journal of Communication and Management, 2(01), 23-26 (2023). https://doi.org/10.58966/JCM2023214
6. Chakraborty, A., Luqman, M., Satapathy, S., & Ganguly, N.: Editorial algorithms: Optimizing recency, relevance and diversity for automated news curation. In Companion Proceedings of the The Web Conference 2018 (pp. 77-78) (2018). https://doi.org/10.1145/3184558.3186937

7. Portugal, R.L.Q., Wilczek, B., Eder, M., Thurman, N., & Haim, M.: Design Thinking for Journalism in the AI age: Towards an Innovation Process for Responsible AI Applications. In The Joint Computation+ Journalism European Data & Computational Journalism *Conference* (pp. 22-24) (2023). https://openaccess.city.ac.uk/id/eprint/30698/
8. Chung, L., Nixon, B. A., Yu, E., & Mylopoulos, J.: Non-functional requirements in software engineering (Vol. 5). Springer Science & Business Media (2012). https://doi.org/10.1007/978-1-4615-5269-7
9. Wang, H. C., Chen, C. C., Li, T. W.: Automatic content curation of news events. Multimedia Tools and Applications, 81(8), 10445-10467 (2022). https://doi.org/10.1007/s11042-022-12224-4
10. Portugal, R. L. Q., Engiel, P., Roque, H., do Prado Leite, J. C. S.: Is there a demand of software transparency?. In *Proceedings of the XXXI Brazilian Symposium on Software Engineering* (pp. 204-213) (2017). https://doi.org/10.1145/3131151.3131155
11. Sajid, A., Jan, S., & Shah, I. A.: Automatic topic modeling for single document short texts. *In 2017 International Conference on Frontiers of Information Technology (FIT) (*pp. 70-75). IEEE (2017). https://doi.org/10.1109/FIT.2017.00020
12. Kim, H. D., Castellanos, M., Hsu, M., Zhai, C., Rietz, T., & Diermeier, D.: Mining causal topics in text data: iterative topic modeling with time series feedback. In Proceedings of the 22nd ACM international conference on information & knowledge management (pp. 885-890) (2013). https://doi.org/10.1145/2505515.2505612
13. Li, F., Shen, H., & He, T.: Tag-topic model for semantic knowledge acquisition from blogs. In 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (pp. 221-226) IEEE (2011). https://doi.org/10.1109/NLPKE.2011.6138198
14. Berry, D. M.: Evaluation of tools for hairy requirements engineering and software engineering tasks. School of Computer Science, University of Waterloo, Tech. Rep. (2017). https://cs.uwaterloo.ca/~dberry/FTP_SITE/tech.reports/EvalPaper.pdf
15. Spacy, Industrial-Strength Natural Language Processing, https://spacy.io/, last accessed 2024/03/30
16. NLTK, Natural Language Toolkit, https://www.nltk.org/, last accessed 2024/03/30
17. Keras, https://keras.io/, last accessed 2024/03/30
18. Luminicen. Luminicen/ImplementingAccuracyQualityforResponsibleAIinNewsrooms: ImplementingAccuracyQualityforResponsibleAIinNewsrooms (1.1) (2024). Zenodo. https://doi.org/10.5281/zenodo.12607376
19. Engiel, P., Leite, J. C. S. D. P., Mylopoulos, J.: A tool-supported compliance process for software systems. In 11th International Conference on Research Challenges in Information Science (RCIS*)* (pp. 66-76) IEEE (2017). https://doi.org/10.1109/RCIS.2017.7956519
20. Portugal, R. L. Q., do Prado Leite, J. C. S., Almentero, E.: Time-constrained requirements elicitation: reusing GitHub content. In IEEE Workshop on Just-In-Time Requirements Engineering (JITRE) (pp. 5-8) IEEE (2015). https://doi.org/10.1109/JITRE.2015.7330171
21. Cysneiros, L. M., do Prado Leite, J. C. S.: Using the Language Extended Lexicon to Support Non-Functional Requirements Elicitation. In *WER* (pp. 139-153) (2011). https://www.inf.puc-rio.br/wer01/NFu-Req-2.pdf