Mining Journals to the Ground: An Exploratory Analysis of Newspaper Articles

Camila Leite da Silva, Lucas May Petry, Vinicius Freitas, Carina Friedrich Dorneles
Programa de Pós-Graduação em Ciências da Computação (PPGCC)
Universidade Federal de Santa Catarina (UFSC) — Florianópolis, Brazil
{camila.leite.ls, lucas.petry, vinicius.mctf}@posgrad.ufsc.br, carina.dorneles@ufsc.br

***Abstract*—In this work, we propose an exploratory approach to analyze the categorization, spread and replication of news content in journalistic web portals. With an increasing number of possible sources, these analyses may aid on identifying reliable content sources from which people may extract information. Our process involves web scraping news portals to gather data, representing them using several Natural Language Processing techniques, and evaluation through Clustering and Similarity analysis. We have represented news articles using *Bag of Words*, *TF-IDF*, *Paragraph Vector* and *Set of Named Entities*, and applied *Cosine* and *Jaccard* similarity measures to compare them. In categorization we have used *DBSCAN* and *Hierarchical Agglomerative* clustering algorithms. Our results show that many articles may be placed in more than one category in order to reach their desired audience. The most relevant news in the observed time period were identified in multiple portals; and we have also identified similar behavior among the analyzed representations when evaluating spreading and replication.**

***Index Terms*—Clustering; Data Journalism; String Similarity; Knowledge Discovery.**

I. Introduction

The multimedia revolution in the world has impacted multiple aspects of society. Especially in the journalistic process, content is produced and disseminated in a much faster and decentralized fashion than ever before. However, the quality of delivered information may be compromised in this process [1]. With the increasing amount of content available, readers are challenged to find and filter what is relevant and trustworthy.

Content quality and reliability are strongly related to their sources [2]. In addition, readers tend to access content from a few sources of information rather than exhaustively digging into several sources. In this scenario, understanding the reliability of content producers comes in handy for readers.

We link the quality of newspapers to three main aspects: (i) content categorization, (ii) content spreading, and (iii) content replication. Proper categorization of content makes it easier for readers to find what they are looking for [3], [4]. On the other hand, analyzing content spreading may reveal how effectively content is reaching community [5]. Lastly, high content replication within the same news outlet indicates multiple articles with no additional information on a subject; ultimately revealing increasing awareness on the subject [6].

This work was financed in part by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - CAPES, Finance code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

Data journalism is a novel approach for gathering, analyzing and reporting news; essentially, for making journalism [4]. This covers content analysis from newspaper web portals [7] to user-guided social media [5], as well as providing other structured and unstructured data sources to journalists [3]. Notably, the spread of *fake* news through social media is concerning as they shape and affect modern society [2], [8].

Data mining is enabling data journalism and knowledge discovery through approaches such as clustering, text mining and sentiment analysis [4], [6]. Nonetheless, most of the aforementioned works focus their research on the analysis of social media. We believe that extracting and mining content from newspaper web portals will lead to more reliable and complete sources of information.

In this scenario, we make two research hypotheses, which we want to corroborate in this work: (i) relevant news articles[1] will be published in multiple web portals; and (ii) high quality portals will always publish novel material. From these research hypotheses, we want to facilitate the answer to two complementary research questions: "How many sources must one follow to be up to date in the most relevant subjects?", and "Which sources are these?". In this paper, we make the following contributions:

- An exploratory approach to evaluate 3 important aspects of articles, *Categorization*, *Spreading*, and *Replication*;
- We evaluate four well-known Natural Language Processing (NLP) content representation techniques, measuring article similarity regarding the 3 aforementioned aspects;
- We present a study case on multiple local South-Brazilian news journals, where we extract insights on content quality based on real data collected from their web portals.

The remainder of this paper is organized as follows: in Section II we discuss about related works. In Section III we present an overview of our approach for data collection and analysis. In Section IV we introduce four content representation strategies and the data analysis tasks addressed in this work. In Section V we present the experimental design and metrics used. In Section VI we discuss about results and our findings. Lastly, in Section VII we present our final remarks and directions for future work.

[1]For simplicity, from now on we will refer to "news articles" simply as "articles".

II. Related Work

Recent telecommunication advances increased the information sources for local and global news, drastically enhancing the capabilities of keeping up to date with these news [1]. With wide access to reliable and unreliable information on the web, the need to discover and create quality content are main concerns of journalism in this new era. Data Mining is a key tool in achieving this objective [1], [3], [4].

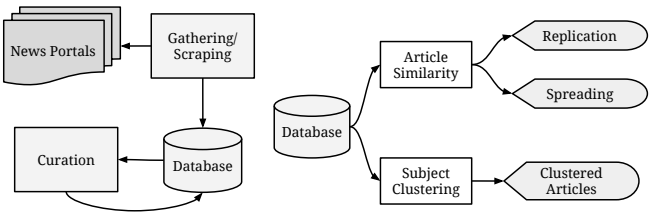
1) Content Categorization: The identification of the subjects that are discussed in a given article is mostly about how information is categorized, or tagged, and presented to the reader [7]. Bad categorization may lead to difficulty in finding useful content [3], and many works aim at enhancing the structure for content categorization. For example, the topics journalists use to present medical news are decisive for users deciding which articles to read; leading to a direct influence of categorization on entries that appear first in medical news [7].

2) Content Spreading: Content spreading refers on how information is disseminated through the web. It can be used to measure how successfully content reaches its target population, and the sources people trust when collecting data and sharing opinions. The wide use of social media may help on content spreading, as some users may influence their followers. Techniques can abstract these user connections as network graphs and clusters in order to identify who is spreading the most content [5]. While spreading high quality content may help disseminating information through the web, the phenomenon of *fake news* is showing the bad side of misinformed news spreading [8]. This is an example of how important it is to measure content circulation, as statistical analysis show their impact in scenarios such as the United States pre-elections, by the intense diffusion of false content to benefit some politicians or political parties [9], for example.

3) Content Replication: Similarly, the content *peddling*, or replication, plays an important role in the journalistic field, as it indicates if certain material is authentic or copied [6]. Starting from the identification of replicated or authentic content, it is possible to determine which sources contribute with the enrichment of original information on the web [1]. Likewise, we can identify which sources are important when deciding to change the heading of a subject presentation [7].

4) Text Mining: Data mining is a viable approach to perform the aforementioned analysis [4]. Tree [8] and Doc2Vec [10] representations, sentiment analysis [6], [11], clustering [12] and similarity measures [6] are widely used to extract meaning from journalistic data.

Moreover, the problematic scope for this work are topic association, news spreading and content replication, as these are pertinent on establishing quality information sources. Differently from other works, our focus is to compare the behavior of multiple representation techniques on each of the selected tasks, while considering the scope of articles published from well established Brazilian web portals.



(a) Consolidation (b) Analysis

Fig. 1. Overview of data gathering and analysis.

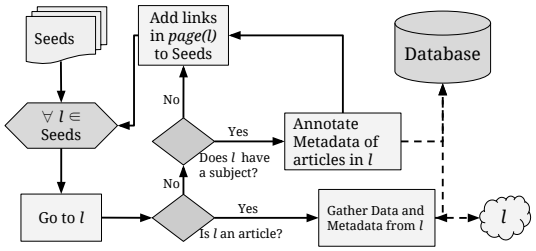


Fig. 2. Web Scraper basic workflow.

III. Work Outline

With the rise of information technology and availability in the media, there is an increasing demand for reliable sources of information, such as traditional newspapers, which take accountability for material they publish. However, even when accounting for these sources alone, there is still a much higher supply of news than any individual user is able to consume. We propose a two-part methodology to gather and analyze journalistic data, outlined in Fig. 1, in order to address the research hypotheses introduced in Section I.

Our work involves two main processes, data consolidation (Fig. 1a) and analysis (Fig. 1b). The consolidation process involves gathering the data and metadata from an article and, storing this information in a database and manipulating it so it best suits our needs. The analysis process, on the other hand, focuses on extracting meaning from gathered data. This is divided in two independent steps, the process of *Similarity Analysis* (e.g., understanding articles that are similar to one another), and *Subject Clustering* (e.g., finding which articles have the same subject). We use data mining and machine learning techniques in order to achieve both our objectives, which is better detailed in Section IV.

In order to answer our research questions, we chose a list of local South-Brazilian newspaper web portals. Since our list of target information sources was rather limited, this simplified our web scraper specifications, briefly described in Fig. 2. This spider software was developed using the *Scrapy* web crawling framework[2]. It will seek article web pages through our seed portals, while storing subject meta-data it finds along the way.

From the article pages, we extracted: (i) author names; (ii) subject; (iii) tags; (iv) title and subtitle; and (v) text body. We treated the first four data fields (i-iv) as meta-data, while the

[2]**Scrapy,** Scraping and Web Crawling framework: https://scrapy.org/

text body is the main piece of information we want to analyse. In development, we noticed some portals used *tags* as *subjects*, and did not separate their news in different subjects within the portal.

IV. Content Representation and Analysis

Texts are a very rich, yet unstructured type of data. Although it may be easy for humans to analyze text data as it is, most machine learning algorithms require processing text into a homogeneous and structured representation.

Common steps in text preprocessing include tokenization, lemmatization, stemming, and filtering stop words [13]. Afterwards, different techniques can be applied for representing text data in a structured manner for further analysis. With this in mind, in Section IV-A we present four different techniques for representing textual data, used in the data consolidation step (Fig. 1a). In Section IV-B, we describe the similarity metrics used and the data analysis tasks addressed in this work, used in the data analysis (Fig. 1b).

A. Content Representation

1) Bag of Words (BOW): BOW is one of the most common and naive approaches for representing textual data, where each document is represented by a vector of word occurrences, regardless of their order. As documents may contain thousands of different words, BOW vectors tend to be high-dimensional.

2) Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF [14] is a commonly used technique for term weighting in text mining. In a contrast to BOW, TF-IDF weighs the importance of words in each document in relation to the whole corpus of documents.

3) Paragraph Vector (Doc2vec): Doc2vec uses learning dense vector representations in documents with neural network models [10]. The models are based on the distributional hypothesis concept [15] in which words that occur in similar contexts tend to have similar meanings. Doc2vec represents documents as abstract feature vectors and, therefore, overcomes the high-dimensionality problem present in BOW and TF-IDF representations.

4) Set of Named Entities (SNE): Among the most meaningful fragments of texts are Named Entities (NE), e.g. people, organizations. NE Recognition techniques [16], [17] extract semantically relevant pieces of text that represent real-world entities.

B. Data Analysis and Knowledge Discovery

In order to extract valuable knowledge from articles, we must first define how to measure articles similarity depending on each of the four different representation techniques aforementioned.

We use the cosine similarity for measuring the similarity of articles represented by vector space models (BOW, TF-IDF, Doc2Vec), which is widely applied to textual data [18]. Differently from the euclidean distance, the cosine similarity does not take the magnitude of vectors into account. Hence, it is appropriate to our analysis, because we are interested in identifying articles on the same topic or subject that do not necessarily have similar length. For the SNE content representation technique, we apply the Jaccard Similarity Coefficient, as given by Equation 1. Considering the scope of news articles, it is reasonable to assume that if two articles mention the same NEs, then they probably talk about the same topic or subject.

$$\text{Similarity of two sets, } A \text{ and } B = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

We explore three contextually different tasks in this work: subject clustering, news replication, and news spreading.

1) News Subject Clustering: Clustering can be used to evaluate the power of different content representation strategies of articles. We may also use clustering techniques to find similar articles on the same topic from different portals. Thus, we designed a clustering experiment for evaluating the techniques described in Section IV-A.

2) News Replication: As we discussed earlier, we understand that high quality portals should always publish novel and relevant content. In this work, we measure how much each portal replicates the same article, or even slightly different articles with roughly the same content. The approach we use to measure content replication is given by Equations 2 and 3. Let $N_p \subset N$ be the set of all entries of N from a given portal p. Given two articles $m, n \in N_p$, their similarity with an arbitrary metric is denoted as $sim(m, n)$. Given a $match$ (i.e. sim exceeding a threshold ϵ, Equation 2), we count an entry as replicated within p.

$$match(m, n) = \begin{cases} 1, & sim(m, n) \geq \epsilon \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

$$replication(m, p) = \sum_{n \in N_p} match(m, n) \tag{3}$$

3) News Spreading: Most news readers expect to see relevant articles (e.g. an important national sports fact, the death of an influential person) and feel informed when accessing their favorite news portals. With statistics about news spreading, readers can take informed decisions to choose the news portals they should follow. Given an article, we measure its spreading by finding similar articles on different portals, as described in Equation 4.

$$spreading(m) = \sum_{p \in P} \begin{cases} 1, & \exists n \in N_p \mid match(m, n) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In summary, for an article m, we count how many portals $p \in P$ published an article n that matches m, and so the spreading of m ranges from 1 to $|P|$. As in news replication, $match(m, n)$ is given by Equation 2.

V. Experimental Evaluation

We have evaluated the text representation and analysis techniques discussed in Section IV using real data extracted from South-Brazilian news web portals. All source code, extracted and derived data are available in our GitHub repository[3]. Our

[3]https://github.com/lucaspetry/news-similarity

TABLE I
SUMMARY OF DATA AND SUBJECTS EXTRACTED FROM WEB PORTALS.

Subjects	DC	JSC	G1	ND	RIC
Economy	633	-	-	5	401
Health	-	2	-	5	-
Politics	573	-	43	208	1,619
Public Safety	3	-	1	92	-
Sports	-	418	-	2	587
Unclassified	-	-	412	284	823
Weather	-	-	-	16	23
World	671	1,703	-	17	272
Total	1,880	2,123	456	629	3,725

dataset is composed of 8,813 articles, collected from September 22^{nd} to October 22^{nd} 2018. Table I presents the article distribution by subject for each portal, after data curation[4]. Our target portals were, respectively: (i) *Diário Catarinense*, (ii) *Jornal de Santa Catarina*, (iii) *Globo G1 - Santa Catarina*, (iv) *Notícias do Dia Online*, and (v) *RIC Mais*.

There are a total of 1,519 unclassified articles in the dataset. We performed our clustering experiments considering only labelled articles, since unclassified articles could fall into any of the subjects and would affect accuracy results. For replication and spreading we analyzed the whole dataset, since subject information is not meaningful for these analysis.

A. Content Categorization Design

We evaluate the vectorization methods BOW, Doc2Vec, and TF-IDF, as well as SNE. For all vectorization methods, we preprocessed the news, removing stop words. For BOW and TF-IDF, we also performed word stemming. Doc2Vec was trained using the *gensim* [19] package. We trained the model for 250 epochs, with a window size of 3 and embedding size of 200. For the SNE method, we have used the *spacy* [20] package to recognize named-entities.

We evaluated content categorization with two different and commonly used clustering algorithms: Hierarchical Agglomerative Clustering (HAC) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), using the *scikit-learn* library [21]. HAC initially treats each element as a single cluster, then successively merges pairs of clusters that are most similar until all elements are within one single cluster. A hierarchy of clusters is built, which is then used for cluster extraction. DBSCAN, on the other hand, is a density-based algorithm that groups together elements that are most similar and belong to dense regions, identifying outliers during clustering.

In the HAC experiment we tested the complete and average linkage methods available in *scikit-learn*. While complete linkage merges clusters based on the maximum pairwise distance of their elements, average linkage considers the average of all distances. We performed several cuts on the generated dendrogram tree, extracting clusters with size ranging from 5 to 300. For DBSCAN, we ranged the maximum distance between samples of a neighborhood from 0.05 to 0.9. We also tried different number of samples in a neighborhood, ranging from 5 to 25.

[4]We standardized subjects throughout portals based on the categories and articles tags provided by them. More details can be found in our repository.

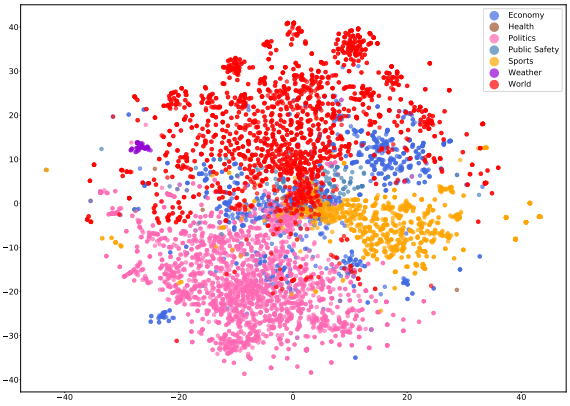


Fig. 3. Clustering visualization with t-Distributed Stochastic Neighbor Embedding (t-SNE).

B. Metrics

Spreading and replication were measured by their definitions presented in Sections IV-B2 and IV-B3. Similarity thresholds (ϵ) were varied in an interval $[0.6 - 0.9]$ iteratively incremented by 0.1.

We evaluate clustering results with *homogeneity*, *completeness*, and *V-measure*. Given a ground truth with labeled elements, the homogeneity measures if clusters contain only elements of the same class; whereas completeness measures if all the elements of each class are in the same cluster. V-measure is the harmonic mean of homogeneity and completeness, i.e., it measures the trade-off between intra and inter-cluster quality.

VI. RESULTS

In this section we discuss the results obtained for each evaluation task described in Section IV-B.

A. Subject Clustering

Table II shows the best clustering results obtained for each algorithm and each technique. The best results were selected based on their V-measure score. The best scoring techniques were BOW and Doc2Vec, with a V-measure of 0.58 on HAC.

The overall results show that the techniques performed systematically better on HAC compared to DBSCAN, as the latter enforces a maximum distance between a minimum number of elements in order to create a cluster. This shows that the articles representations used may be too sparse for DBSCAN. Even though the distances between articles can discriminate some of the different subjects, the articles are still too far from each other, which is shown by the poor results achieved by DBSCAN. The high number of outliers found by the algorithm also corroborates this claim.

As shown in Table II, the algorithms generated more clusters than the original number of subjects provided by the web portals, which suggests that they are able to detect specific

TABLE II
BEST CLUSTERING RESULTS FOR EACH TECHNIQUE, DISREGARDING UNCLASSIFIED ENTRIES. FOR HIERARCHICAL AGGLOMERATIVE, $P \in \{average, complete\}$. FOR DBSCAN, $P \in \{(0.05 - 0.9,\ 5 - 25)\}$.

Algorithm	P	Technique	Similarity	Clusters	Outliers (%)	Homogeneity	Completeness	V-Measure
HAC	average	BOW	Cosine	50	-	0.64	0.53	**0.58**
	average	Doc2Vec	Cosine	45	-	0.62	**0.54**	**0.58**
	average	SNE	Jaccard	85	-	0.58	0.49	0.53
	average	TF-IDF	Cosine	40	-	**0.68**	0.49	0.57
DBSCAN	(0.60, 20)	BOW	Cosine	27	63.4	0.29	0.29	0.29
	(0.35, 5)	Doc2Vec	Cosine	24	81.9	0.09	0.19	0.12
	(0.75, 15)	SNE	Jaccard	20	55.3	0.31	**0.33**	**0.32**
	(0.65, 5)	TF-IDF	Cosine	168	54.1	**0.42**	0.22	0.29

topics within broad subjects (e.g. articles about soccer originally labeled as sports). Moreover, articles may belong to distinct categories in different portals. For instance, an article entitled "Marketing director of Santander criticizes Neymar falls", published by RIC, was labeled with the Economy subject. However, a similar article on the same topic could have been labeled as Sports in a different portal. By analyzing the clustering results, we found that BOW, SNE, and TF-IDF actually assigned this article to clusters where the majority of the entries was about sports. Hence, clustering algorithms allied to these techniques are good alternatives to automatically label news articles.

Fig. 3 shows the distribution of articles according to the categories provided by the analyzed portals. News subjects are not well-separated sets, as there is a large overlap of categories. In fact, news articles are not restricted to a single subject. For instance, articles in the World category may refer simultaneously to Economy and Politics, hence characterizing a fuzzy clustering problem in which articles may belong to multiple clusters. However, due to the limitation to a single subject of the collected data, we used hard clustering techniques only.

B. News Spreading

In this analysis, articles from different news portals are compared pairwise, and we measure their similarity with different threshold values to assume content equality. Fig. 4 shows how many portals share similar articles, with results presented in logarithmic scale. Regardless of threshold, the majority of the content is exclusive to one single portal (class 1). Class 2 is also highly populated, which suggests that some content is spread among at least 2 portals.

By exploring spreading results, we learned that the portals DC and JSC share the highest amount of similar content. BOW, Doc2Vec, and TF-IDF identified about 33.5% of articles in both portals that have a corresponding article at least 90% similar in the other portal. Furthermore, we found that articles published in three or more portals are related to subjects that globally affect people, such as natural disasters and international political impasses. Examples of class 3, with *Doc2vec* and a similarity threshold of 0.8, are: "Kim Jong-un wants 'lasting peace', says South Korean president"; "Ecuador expels Venezuela ambassador after the chavist official criticizes the

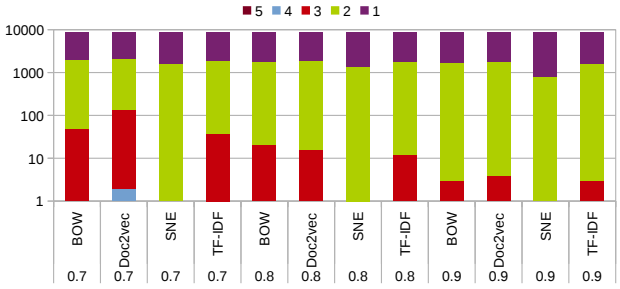


Fig. 4. Number of portals in which a given article was detected. Each column color represents a class, that is, if an article was detected in 2 portals, it is in class 2, if it was detected in 1 portal, in class 1, and so on. The y axis indicates the total number of articles, represented in a logarithmic scale.

President"; "Investigators try to understand the reasons for the massacre in Crimea"; "Florida awaits the 'extremely dangerous' hurricane Michael"[5].

Using a 0.8 threshold, nearly every pair of articles identified as similar by the analyzed metrics were actually discussing the exact same subject.Differently from vector space models, SNE is only able to detect articles that mutually exist in 2 web portals. However, our empirical analysis shows that the actual spreading of relevant news is much higher, suggesting that SNE is not a good technique for evaluating spreadness.

C. News Replication

Fig. 5 shows the relation of how much content was replicated within a news portal, using different threshold values to assume content equality. Percentages were calculated by dividing the amount of detected replicated pairs ($\mathcal{R}$, given by Eq. 3) by the number of possible pairs in the portal ($\mathcal{P}$): $\frac{\mathcal{R}}{\mathcal{P}}$. Some news portals replicate content more than others, e.g. *JSC*, *DC* and *RIC* are the portals with highest replication indices, regardless of threshold and representation technique.

JSC had the highest values of replication when compared with the other portals in all metrics and thresholds. *RIC*, on the other hand, has the highest volume of articles, but does not replicate as much content, indicating that a higher percentage of novel content is published. Even SNE, which has

[5]Original titles may be found in the dataset available in our git repository.

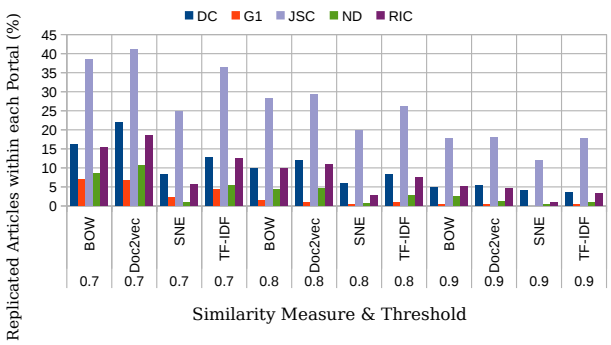


Fig. 5. Percentage of replicated articles within each portal, according to each technique and similarity threshold.

shown to be the most restrictive metric, was able to identify high levels of replicated content in *JSC*. This suggests that articles identified as replicated are indeed the same articles. We identified that this happens because some portals create new articles when they receive more details about a particular topic, instead of updating recently published articles on the same topic.

Some of the replicated news articles extracted from *JSC*, using SNE and similarity threshold of 0.8 are: "Rocket launched from the Gaza Strip falls in Israel"; "Turkish media say that Saudi journalist was 'beheaded'"; "13 dead in Mallorca floods".

D. Discussion

Our experiments show that categorization with DBSCAN leads to bad results with the analyzed metrics due to the sparsity of the dataset, while HAC had a better performance with all metrics.

When evaluating news spreading, SNE was unable to detect high similarity between articles due to its restrictiveness, which is unable to identify the context of cited named entities within text. On the other hand, BOW, TF-IDF, and Doc2vec portrayed similar behavior while analyzing content spreading among different portals.

Replication results were similar among the techniques analyzed, which suggests that we were able to achieve accurate insights. Additionally, the results show that volume of published content is not necessarily an indicator of content replication. Notably, JSC was the most replicating portal, while RIC had the highest volume of content published in the same period.

VII. Conclusion

We have presented a case study of text knowledge extraction in a real-world dataset, extracted from news web portals. We analyzed the behavior of the BOW, TF-IDF, Doc2vec, and SNE data representation techniques to perform article categorization (with clustering), and analyze news spreading and replication. Our analysis results have shown that the proposed approach is effective for measuring the quality of news portals.

Further analysis in the subject include using fuzzy clustering in content categorization, as our results show that content is often associated to more than one single subject [22]. Exploring the relation of portals and their social-media counter-parts may enable further understanding of news spreading.

References

[1] B. Van der Haak, M. Parks, and M. Castells, "The future of journalism: Networked journalism," *International Journal of Communication*, vol. 6, p. 16, 2012.

[2] K. Starbird, "Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter," in *International AAAI Conference on Web and Social Media*, 2017.

[3] A. Y. Halevy and S. McGregor, "Data management for journalism," *IEEE Data Eng. Bull.*, vol. 35, no. 3, pp. 7–15, 2012.

[4] K. Kirkpatrick, "Putting the data science into journalism," *Communications of the ACM*, vol. 58, no. 5, pp. 15–17, 2015.

[5] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks." *Icwsm*, vol. 10, pp. 90–97, 2010.

[6] I. Subašić and B. Berendt, "Peddling or creating? investigating the role of twitter in news reporting," in *European Conference on Information Retrieval*. Springer, 2011, pp. 207–213.

[7] L. M. Mahoney, T. Tang, K. Ji, and J. Ulrich-Schad, "The digital distribution of public health news surrounding the human papillomavirus vaccination: a longitudinal infodemiology study," *JMIR public health and surveillance*, vol. 1, no. 1, p. e2, 2015.

[8] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[9] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[10] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[11] H. Jin, J.-P. Hong, K.-H. Lee, and D.-W. Joo, "Diagnosis of corporate insolvency using massive news articles for credit management," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2019, pp. 4849–4854.

[12] S. Vairavasundaram, V. Varadharajan, I. Vairavasundaram, and L. Ravi, "Data mining-based tag recommendation system: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 3, pp. 87–112, 2015.

[13] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *preprint arXiv:1707.02919*, 2017.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[15] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[16] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proceedings of International Conference on Computational linguistics*, vol. 1. Association for Computational Linguistics, 2002, pp. 1–7.

[17] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum entropy approach," in *Proceedings of Conference on Natural Language Learning at HLT-NAACL*, vol. 4. Association for Computational Linguistics, 2003, pp. 160–163.

[18] A. Huang, "Similarity measures for text document clustering," in *Proceedings of New Zealand Computer Science Research Student Conference*, vol. 4, Christchurch, New Zealand, 2008, pp. 9–56.

[19] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *LREC Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010, pp. 45–50.

[20] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, 2017.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[22] S.-J. Lee and J.-Y. Jiang, "Multilabel text categorization based on fuzzy relevance clustering," *IEEE transactions on fuzzy systems*, vol. 22, no. 6, pp. 1457–1471, 2014.