Bus Inf Syst Eng
https://doi.org/10.1007/s12599-023-00843-6

RESEARCH PAPER

SJORS: A Semantic Recommender System for Journalists

Ángel Luis Garrido · Maria Soledad Pera · Carlos Bobed

Received: 11 December 2022 / Accepted: 28 October 2023

Abstract Recommender Systems support a broad range of domains, each with peculiarities that recommendation algorithms must consider to produce appropriate suggestions. In the paper, we bring attention to a little-studied scenario related to the news domain: recommendations catering to media journalists. Based on the particular needs inherent to a newsroom, the authors introduce SJORS, a wire news Recommender System that takes into account the activities of each journalist as well as other critical factors that arise in this particular domain, such as wire news recency. Given the nature of the items recommended, SJORS deals with the inherent ambiguity of natural language by exploiting different semantic techniques and technologies. The authors have conducted several experiments in a media company, which validated the performance and applicability of the system. Outcomes emerging from this work could be extended to other domains of interest, such as online stores, streaming platforms, or digital libraries, to name a few.

Keywords Recommender systems · Semantics · Machine learning · NLP · Journalists

Accepted after 3 revisions by Natalia Kliewer.

Á. L. Garrido (✉) · C. Bobed
University of Zaragoza, Zaragoza, Spain
e-mail: garrido@unizar.es

Á. L. Garrido
Universitat Politecnica de Valencia, Valencia, Spain

M. S. Pera
TU Delft, Delft, The Netherlands

1 Introduction

In recent years, the world of journalism has undergone a process of transformation motivated by the growing use of digital devices for the consumption of news, the consequent change in the advertising model, and the 2008 financial crisis. This process has led to a scenario where the workforce at newspapers has been reduced, with journalists being expected to undertake more complex work due to the tough competition among different media, and the immediacy expected by customers (Siles and Boczkowski 2012).

Editors use external news agencies, which send thousands of articles daily. A sample of these pieces of news, called *wire news* (Whitney and Becker 1982), can be seen in Fig. 1. The amount of data received translates into complex search tasks for journalists, who must go through the available wire news to locate the most suitable ones to prepare their articles. Unfortunately, information overload is not the only problem journalists must face. They also constantly need high quality, and above all, recent materials, making temporal considerations fundamental. The wire news gain importance the more recent they are, which is why the latest ones are generally deemed more useful. In addition, the topics covered by the journalists in the mid/long term also influence what they perceive as a helpful piece of wire news.

Recommender Systems (RS), which aim at streamlining the decision-making process (Ricci et al. 2011), can help journalists by enabling the discovery–among the many available–of wire news of interest to them bypassing the need to search for them. Simultaneously addressing the aforementioned requirements in a RS, however, is not a trivial task (Feng et al. 2021). This is compounded by the fact that the *wire news* to be recommended are described using extended, domain-dependent metadata (e.g., dates,

Published online: 21 December 2023

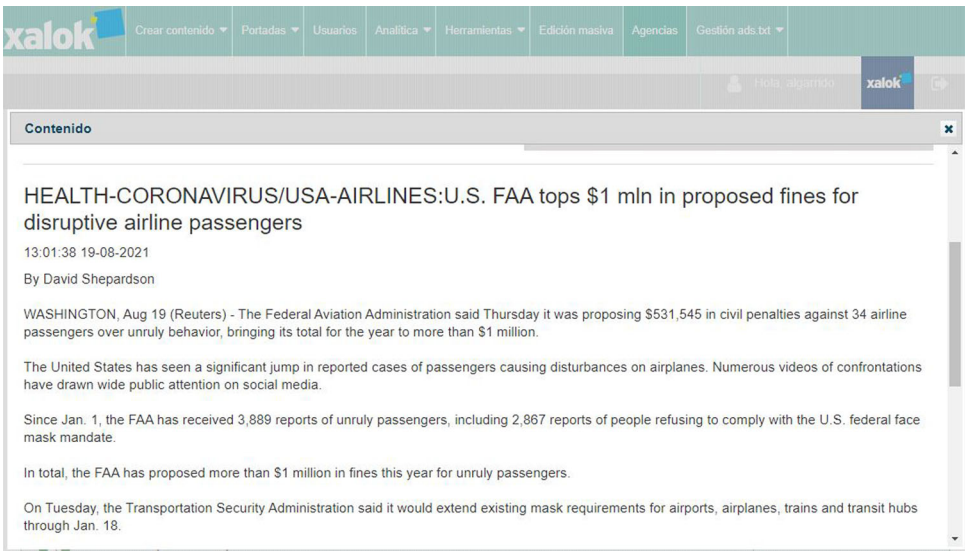

Fig. 1 Sample of a piece of wire news, a text-based piece of news, in a typical presentation format within a production software of a newsroom

authors, tags, locations) as well as natural language (e.g., the content of the piece of wire news or the description of their attached photographs). The latter introduces further difficulties that the recommender algorithm must deal with–interpretation problems due to the inherent ambiguity of the natural language (Navigli 2009). A good example is the word *Washington*, which can refer to a city, a state, a person, or a sports team[1]. Disambiguation is crucial, as it would enable the system to avoid recommending a sports journalist a piece of wire news focused on the geographical place.

In this paper, we introduce *SJORS*, a Semantic JOurnalists Recommender System. SJORS identifies the top-N most relevant wire news for a journalist at any given time, which he/she can use in composing news articles to publish. To do so, SJORS adopts a novel recommendation strategy that depends upon: (1) the activity of the journalists, including their searches on the agencies' databases, their access to the different previous wire news, and their edited and published articles, (2) the recency and completeness of the wire news, and (3) the actual piece of wire news content, i.e., the natural text which can be permeated by ambiguity.

Nowadays, newsrooms demand their journalists to regularly change the topics of the news they cover, to maximize coverage (Westlund and Ekström 2019). SJORS leverages the historical interaction of the journalists with the searching and editing system to identify and prioritize wire news that best suit the needs and domain of expertise of a given journalist. Although SJORS offers wire news deemed suitable for journalists' current tasks, it also pays attention to the issues previously worked on by the journalists. Moreover, SJORS prioritizes the most recent wire news because they have more up-to-date and complete information, using the dates and times of entry of the wire news to the information systems of the newsroom. To offer personalized suggestions, SJORS analyzes the content of wire news using semantic techniques based on the representation of the text using distributional semantics (i.e., *word embeddings*, representation of words using real-valued vectors that encode the meaning of each word (Turian et al. 2010)) and on the detection and disambiguation of the entities in the content. This latter task is known as *Named Entity Linking* (Sekine and Ranchhod 2009), which includes detecting *named entities* (persons, places, organizations, products, etc.) in the text and grounding them to a knowledge base (in our case, *Linked Data* (Bizer et al. 2011)).

The objectives of this work are twofold: on the one hand, investigating the most suitable strategy to represent best the data received in a newsroom from external agencies. On the other hand, designing a tailor-made recommendation system capable of accomplishing the specific task of assisting journalists with their news selection work. To the best of our knowledge, there are no works similar to the one we present in this work, including the exploration of representation strategies for a new domain such as this one.

The main contributions of the work presented in this paper are:

- We propose SJORS, a novel RS for the news domain which, instead of targeting news consumers, targets news creators, i.e., the journalists.
- We introduce an algorithm that integrates both temporal and semantically enriched textual information

[1] Ambiguity applies to the descriptive text and metadata if the latter contains pure textual information, i.e., it is not grounded to an entity like a *knowledge base or a knowledge graph*.

seamlessly in the recommendation schema. The algorithm also leverages interest models for both the wire news and the journalists, which results in more adequate suggestions.
- We carried out an in-depth study of the characteristics of the application domain of the RS: the newsroom of a real newspaper and its daily use of wire news. This study was conducted within a real production newsroom in close collaboration with the journalists from Henneo[2], one of the most relevant media companies in Spain.
- Given the lack of datasets and benchmarks in this domain, we collected data to enable analysis and evaluation by capturing activity in a real newspaper newsroom and by collecting the news published in that period, as well as the wire news received. This resulted in a dataset which we will share with the research community.[3]
- We discuss the results of the exhaustive experimentation we conducted to showcase the correctness and validity of SJORS in close collaboration with real journalists, using data obtained from a real newsroom. We designed an evaluation environment that allowed us to perform: (1) offline examinations to tune the models employed as part of the recommendation process, and (2) online assessments, to showcase the feasibility of SJORS.

The rest of this paper is structured as follows. In Sect. 2, we present background and related literature. In Sect. 3, we offer insights on the application domain of our proposed recommender, the newsroom. In Sect. 4, we detail the design methodology of SJORS. In Sect. 5, we describe our evaluation framework, as well as the results of the empirical study conducted to evaluate the performance of SJORS. Lastly, in Sect. 6, we offer concluding remarks and directions for future work.

2 Related Work

In this section, we review existing literature focused on news RS. Given the textual nature of the items SJORS suggests, we also offer background on the language modeling techniques that SJORS exploits.

2.1 News RS

The literature on RS in the news domain is extensive (Karimi et al. 2018; Özgöbek et al. 2019; Raza and Ding 2022; Wu et al. 2023). News recommendation differs from many traditional recommendation domains such as e-commerce or entertainment due to traits inherent to the domain: (1) the relevance of news items can change very rapidly (i.e., decline or increase due to recent real-life events), in contrast to other domains like movies or books, (2) user's interest can dynamically change, depending on different contextual factors like the time of the day, the features of the user's device (e.g., mobile phone vs. desktop), or the user's current location, (3) the potential need to surprise readers, and (4) the capacity of yielding more diverse reading behavior (Bodó 2019).

Regarding news recommendation strategies themselves, the most prominent ones are those based on click analysis (Zheng et al. 2018), collaborative filtering (Boutet et al. 2013), semantic analysis (Cantador and Castells 2009; Wu et al. 2019), association rules (Golian and Kuchar 2017), context analysis (Gabriel De Souza et al. 2019), deep learning (De Souza Pereira Moreira 2018), and exploration of social network activity (Kazai et al. 2016). Recent works also consider side issues, for example, the multiple stakeholders which drive the development of news recommenders (users, journalists, editors, product owners, etc.) (Smets et al. 2022), or the influence that this type of technology can have on society (Heitz et al. 2022). For an outline of open problems in this area, see Abdollahpouri et al. (2021).

What these works have in common is that the news consumer is always the major stakeholder. This means that issues related to news recommendations and the methodologies applied to generate such suggestions are centred on the user for whom the news recommendations are intended.

2.2 RS for Journalists

In recent years, we have gradually seen an increase in the proposal of tools and systems for newsrooms in the media based on artificial intelligence (Zhang and Pérez Tornero 2021). For example, Berven et al. (2020) developed an architecture and prototype called News Hunter that uses knowledge graphs, natural language processing, and machine learning together to support journalists. Niarchos et al. (2022) introduced a system for operating a set of drones integrated with a data retrieval and semantics processing system, aiming to mediate real-time breaking news coverage.

In the case of SJORS, the main stakeholder is the journalist, who relies on the RS to distill, from the constant influx of resources emerging from multiple news agencies and data sources, the suitable pieces of text that can drive the timely writing of new articles. Even though the literature from this perspective is limited, we discuss research outcomes closely aligned with SJORS below.

[2] http://www.henneo.com.

[3] The dataset will be made available by directly reaching out to authors, given the use of proprietary data.

Montes-García et al. (2013) introduced a context-aware RS for journalists that identifies similar topics across different sources. Cucchiarelli et al. (2017, 2019) proposed a three-fold knowledge representation in which an explicit, semantic-rich domain knowledge vector space representation is incorporated between the user and item spaces. The strategy introduced in Montes-García et al. (2013) favors proximity (i.e., prioritizes news items that occur in locations of geographical proximity to that of the user) and uses videos as its use case. In contrast, the work described in Cucchiarelli et al. (2017, 2019) suggests aspects related to an event that remain still uncovered (based on information reported on Twitter or Wikipedia), as opposed to data sources that journalists should consider in writing their articles. Voskarides et al. (2021) studied the task of news article retrieval in the context of event-centric narrative creation of journalists. They proposed an automatic dataset construction procedure combining lexical and semantic rankers taking into account the chronological order. It is also worth mentioning Finsense (Liou et al. 2021), a topic-specific system which assists investors and journalists in sorting out the information in financial news articles.

All the aforementioned strategies overlook some of the intrinsic characteristics of a journalist: the typology of their publications, their historical and recent activity, and the sudden changes in the topics they have covered. The latter can be due to the polyvalence imposed on them by the editors due to limited newsrooms' workforces, or simply because of the need to cover absences or vacations.

In changing the focus to the journalist, rather than news consumers and the news articles themselves, there is an evident limitation on the use of collaborative filtering strategies as journalists are known to lack the time and motivation to perform rating tasks within the frenetic pace of work that is imposed in newsrooms (Bordogna et al. 2006; Montes-García et al. 2013). Moreover, journalists write about different matters (i.e., the topics they cover change over time and other journalists' topics might not ever overlap), so it might not be possible to extract patterns of *similar users*, which is critical for collaborative filtering strategies.

In recent years, another critical issue has disrupted the world of journalism: "fake news". These are artificially fabricated news that mimics media content to serve a particular non-informational purpose (Lazer et al. 2018). Research related to fake news has primarily drawn attention in a political context, but also on topics like vaccination, nutrition, and stock values (Vo and Lee 2018; Nan et al. 2021). Note that in our scenario, we deal with curated and trustworthy sources. Otherwise, we would incorporate some algorithmic functionalities to detect and filter out fake news (Sharma et al. 2019).

Overall, to benefit a journalist, an RS should: (1) evaluate the evolving profile of the user, (2) consider the previous publications of the journalists, (3) interpret the information contained in the wire news themselves, and (4) place the importance of all these factors in time. As far as we can ascertain, there is an absence of coverage of these issues in existing recommendation strategies, which evidences the need for works that study this interesting and challenging use case in greater depth, such as the one we present in this work.

2.3 Language Models

In these last years, the use of different language modeling techniques based on continuous representations (i.e., word embeddings) has attracted the attention of researchers focused on enhancing the recommendation process (Vasile et al. 2016; Greenstein-Messica et al. 2017; Lv et al. 2017; Caselles-Dupréet al. 2018; Khattar et al. 2018; Ren et al. 2019). These unsupervised techniques provide a method to represent each word in a text corpora by a vector in a continuous space. Vectors can be built using many different methods. Still, they all share the capability of capturing the semantics of the words as they follow the *distributional semantics* hypothesis proposed by Firth (1957), *"a word is characterized by the company it keeps"*. Thus, seminal works such as Word2Vec (Mikolov et al. 2013) or Glove (Pennington et al. 2014) have boosted the advances in many different tasks in the field of Natural Language Processing (NLP). These embedding techniques provide *static* representations of the words (i.e., a word has a representation regardless of the context where it appears).

Lately, these embedding models have evolved to provide *contextual word representations*, where each word/sentence vector depends on the context where it appears. Since 2018, these pre-trained language models (PLM), which utilize self-supervised learning over raw large-scale texts, have received special attention. Some samples of this evolution are ElMO (Peters et al. 2018), a contextualized word representation that models complex characteristics of word use (e.g., syntax and semantics), and how these uses vary across linguistic contexts. Followed by BERT (Devlin et al. 2018), whose main technical innovation was the application of bidirectional training through the use of Transformers (Vaswani et al. 2017). More recently, we find the different versions of Chat-GPT (Wu et al. 2023), a chatbot that uses natural language processing to create humanlike conversational dialogue. Chat-GPT is an integration of multiple technologies such as deep learning, unsupervised learning, instruction fine-tuning, multi-task learning, in-context learning and reinforcement learning. GPT's implementations first learn a general language model on unlabeled raw texts, and then are fine-tuned

according to specific tasks. All these models must always be adjusted for performing particular tasks, and their behavior in work scenarios where the modeling of the data and the actors must be combined with a strict temporal contextualization is still under study.

In SJORS, we adopt static models as they are easier to obtain for our particular domain, and have been successfully applied for domain-independent keyword disambiguation (Buey et al. 2021). These models can be easily extended to sentences, e.g., SIF (Arora et al. 2019) or document representations, e.g., Doc2Vec (Le and Mikolov 2014), which we use in this work. Once the use case has been analyzed and understood, one of the next steps will be to explore the advantages of using LLM to model both journalists and wire news.

3 Understanding the Domain

To better contextualize the reach and applicability of SJORS, we present an overview of the domain of this work. In particular, we describe the usual work environment and conditions journalists face in a newsroom.

Modern digital newsrooms are smaller, more nimble, and less hierarchical than their analogue predecessors (Wu and Lee 2013). Journalists are more dynamic and versatile than before, and the topics they write about can change over time (Linden 2017). The fact that a journalist does not have a fixed beat at the newsroom is already an important limitation for recommending wire news. The issue goes beyond preference change over time as it depends on the multiple topics that are currently either of interest or the focus for a journalist, and how those can change back and forward. For instance, a journalist working in the sports area may be writing about different topics and then, due to the characteristics of the texts, evolve, but always within sports. When leaving the area and working on an article on international politics, the characteristics of the texts and the entities changed so much that there is no possible transfer. Therefore, the only reliable information about a journalist relates to the activities they carry out–the more recent, the better.

Due to staff reduction, editors know their reporting limitations and do not try to cover everything on their own (Carson and Muller 2017; Kramp and Loosen 2018). Journalists use software tools to access *wire news*, i.e., pieces of news from agencies and other external sources, such as correspondents, special reporters, and/or photographers. These tools are usually either the agencies' own websites, or proprietary tools that manage the content, index it, and provide search tools. In a typical newsroom, it is easy to receive more than 15,000 external wire news per day.

It is difficult for journalists to find the most helpful information relevant to their needs. Even more so if the wire news is not purged from the editorial system, impacting the number of them available for consideration. Thus, in this work environment, an RS independent of the agencies' search system is a perfect fit, customizing the content to each journalist.

Figure 2 shows an overview of the typical information flows in any media newsroom (Gürsel 2012; Cohen 2019; Johnson and Radosh 2023). The presence of an RS such as SJORS could help journalists by exploiting the information stored in the wire news database and production news database.

Figure 2 shows the following elements:

1. On the left side, we find the external information sources of the newsroom: agencies and collaborators. They send text-based wire news, which are stored in the *Wire news Database* through an input manager system of a specific *Wire news' Input Software*. As most media newsrooms work with several agencies, the presence of an integrated tool that enables access to all their sources becomes a key requirement to improve efficiency (*wire news Search Engine* in the figure).
2. Journalists search and select wire news to read. Wire news appearing in the searches and wire news accessed are usually stored into a *Wire news Database*, kept by the system to avoid losing any source.
3. Journalists prepare their articles working on the *News Editing System*. The intermediate work they do before publishing the definitive piece of news is difficult to record since it can also be done outside the system, using other office tools (word processors, notepads, etc.).
4. Journalists produce the final *Media News* by using the *News Editing System*. These pieces of news are stored in the *Production News Database*.

4 SJORS: Our Proposed News Recommender for Newsrooms

We discuss SJORS' design; its complete architecture is depicted in Fig. 3. SJORS consists of four modules: three for modeling wire news, published news, and journalists, and another one for recommendation generation. SJORS also includes a subsystem that produces the different vector types that are used for modeling purposes.

Following, we first describe the research objectives and the data considered for the design of SJORS. We then explain how to model the data elements, presenting how the vectors and different techniques are obtained. In addition, we explain the approach used to model the journalists

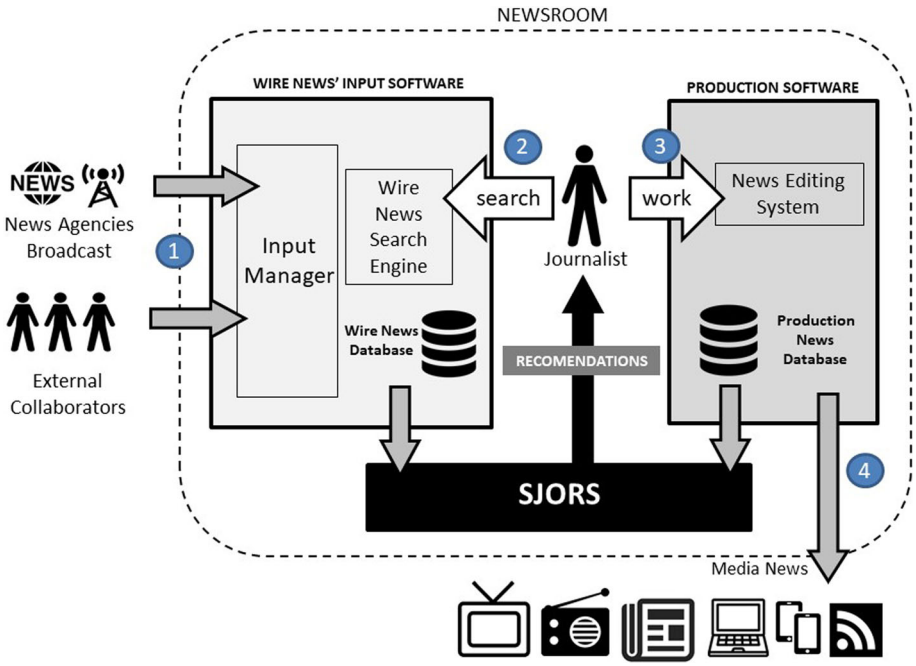

Fig. 2 SJORS' application context. The grey arrows represent the flow of the wire news and the pieces of news; the white ones represent the journalist's actions. Also noted are the different steps in a common newsroom's workflow: (1) wire news entry, (2) searches and inquiries, (3) journalists' work, and (4) news publication

themselves. Lastly, we outline the recommendation process.

4.1 Research Objective and Considered Data

The research objective of this work is to develop and evaluate a system that can help journalists by identifying the most relevant wire news received by external agencies. This is done for a given journalist at a given time, and within the context of a particular newsroom. To accomplish this task, we propose a system (SJORS) able to:

1. Analyze articles previously authored by the journalist using the newsroom production software. The information that can be extracted from these pieces of news provides SJORS with a view of the long-term profile of this journalist, modeling the diverse topics that he/she has recently covered.
2. Obtain all the wire news recovered by the journalist's searches from the agency news database. This information provides SJORS with a broad view of the topics that the journalist might be currently interested in.
3. Consider the journalist's interactions with the wire news retrieved in response to prior searches (i.e., accessed to read). This provides SJORS with a fine-grained view of the journalists' particular and recent interests.

As shown in Fig. 2, we propose that SJORS can access both databases (i.e., the piece of wire news and production databases) for generating personalized recommendations to journalists.

SJORS models a journalist profile based on information that increasingly varies in its immediacy: from more static, i.e., the domains that the journalist usually covers, to more dynamic and volatile, i.e., the wire news that has captured his/her attention in the stream of news that is continuously arriving at the newsroom. SJORS then considers the long-term profile (publications) and a short-term one (searches and accesses), respectively. In practice, they are intertwined in time. For this reason, SJORS does not consider them separately.

With the information extracted from the different pieces of news and wire news considered (all of them text-based elements), SJORS can capture the specifics and possible modifications/evolution of the topics the journalist has focused on during a specific time frame. This time frame may vary depending on the characteristics of each newsroom.

As stated in Sect. 3, journalists mainly use wire news as their source of documentation. Thus, the recommended wire news, as well as the journalists' interests, are represented in our system through the content in text format that accompanies these wire news themselves. We also acknowledge the importance of multimedia elements (photos, videos, graphics, etc.), but, as they require different technical solutions to be managed, we consider them out of the scope of this work, leaving them as future work.

4.2 Representing Wire News and Pieces of News

To handle and recommend the wire news, SJORS simultaneously leverages different content-representation

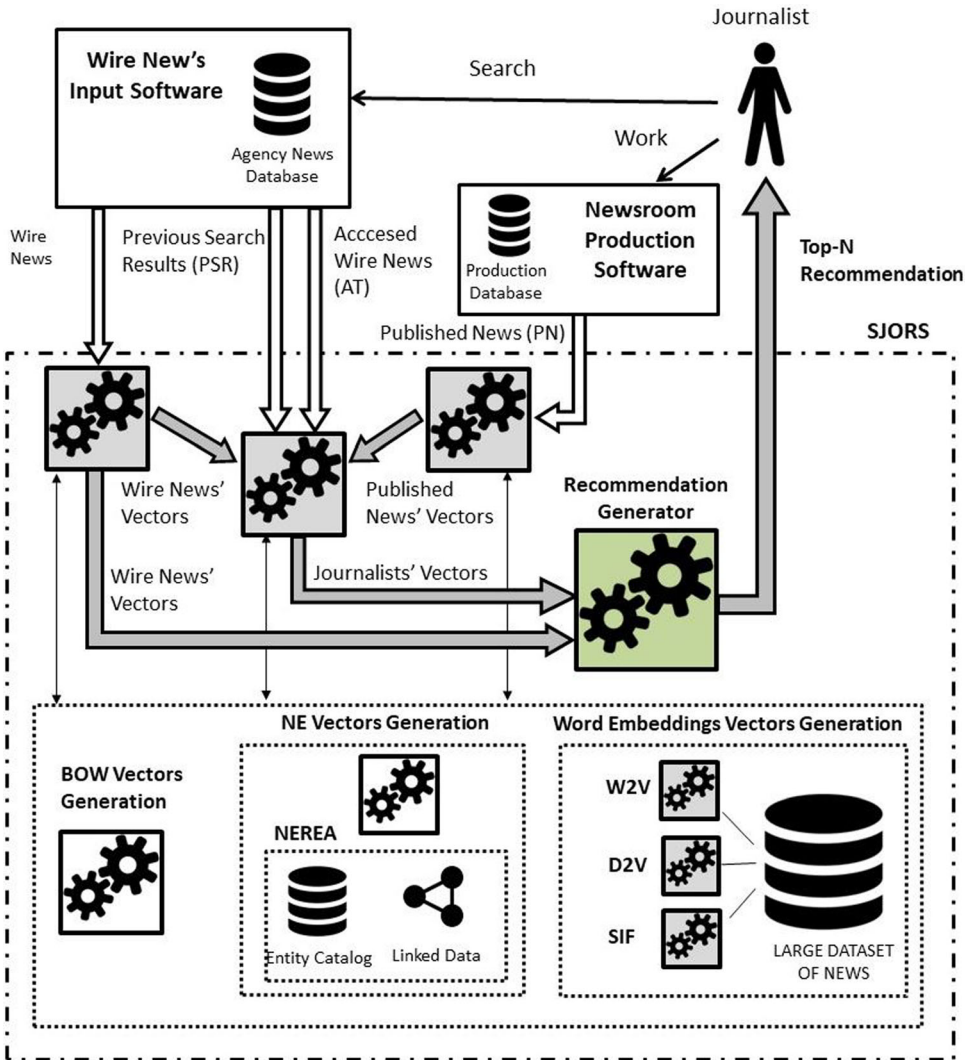

Fig. 3 Overview of the RS SJORS modeling the data through several approaches (BOWs, Named Entities, and Word Embeddings)

strategies that provide a vector representation of the text of the wire news with diverse levels of semantics (Camacho-Collados and Pilehvar 2018). We describe such modeling approaches in detail as they are also the basis for the journalist's representation.

4.2.1 Modeling Using BOWs

We first describe the simple, yet effective, Bag of Words (BOW) technique Zhang et al. (2010). For each piece of wire news, SJORS uses together the title and the body text of the news item to build a BOW representation with a configurable amount of words paired with their weights. Given that this vectorization is the most straightforward one, it is treated as our baseline in our experiments.

To achieve this representation, SJORS includes a dedicated module in its architecture called *BOW Vectors Generation* (see Fig. 3). The first step is to *lemmatize*[4] all the words in the text. This process simplifies the task of obtaining keywords and reduces the number of words the system has to examine. Moreover, SJORS eradicates *stop words*: prepositions, conjunctions, articles, and other words with low semantic information. After lemmatization and stop word removal, the system has a list of significant keywords for each piece of wire news. To capture the importance of keyword k within a text T, SJORS uses the well-known TF-IDF weighting strategy (Salton and Buckley 1988):

$$TF{-}IDF(k,T) = TF * log_{10}\left(\frac{|C|}{C_k}\right) \quad (1)$$

where TF is the frequency of k in T (raw count of a term in a document divided by the raw count of all terms), $|C|$ is the total number of documents in a given collection C, and C_k is the number of texts in C that contains k. Each piece of

[4] A lemma can be defined as the canonical form representing each word, for example "going", "gone", or "went" has the same lemma, "go".

wire news is represented by a vector containing lemmatized words with their respective TF-IDF weights, which we call *vector* v_{tBOW}. The use of these vectors implies that, when performing operations over the wire news, these operations are executed over vectors of wire news with dimensions that are the union of the vocabularies/words that composed them. However, for the sake of efficiency and without loss of generality, SJORS considers the 100 most frequent components. The system stores these top 100 terms after their calculation to obtain the final recommendation. This amount of terms is sufficient to model wire news, as empirically verified in different works related to news (Garrido et al. 2016, 2017).

4.2.2 *Modeling Using Named Entities and Linked Data*

The second vectorization technique used in SJORS examines Named Entities (NE) (Sekine and Ranchhod 2009). In the text of the wire news, it is easy to find names of people, places, or companies. For example, if the piece of wire news contains the phrase "San Francisco" then together the words could be considered as a single named entity to capture its actual meaning. This is a better option than considering the words "San" and "Francisco" independently.

We explicitly tackle this problem in SJORS by exploiting Linked Data repositories, which make available structured data on the Web (Bizer et al. 2009). We do so by leveraging NEREA, an automatic Named Entity Recognizer and Disambiguation system (Garrido et al. 2016, 2017)[5], which we briefly discuss below for the sake of completeness, ensuring that the work presented is self-contained.

4.2.2.1 4.2.2.1 NEREA Overview NEREA aims to recognize relevant entities from a text in a local document database and disambiguate them. The system uses three types of knowledge bases: (1) local classification resources, (2) Linked Data, and (3) its catalog denominated *Entity Catalog*. Local information can have different formats: a close list of terms, a thesaurus, or an ontology. The result of the process is an *Entity Catalog* that contains a set of disambiguated relevant entities within the local repository, which are used to represent people, organizations, or locations.

For entity recognition, NEREA:

1. Receives as the input a named entity, and its context, i.e., the whole text where that named entity is located. The context will help to select the entity referenced by the named entity.
2. For each named entity detected in a text, NEREA checks if it already exists in the named entity catalog. In case it does not, NEREA searches for information related to the named entity in the local classification resources, or through Linked Data sources. In each of these sources, NEREA locates the possible candidates that match the named entity, resulting in a list of candidates.
3. With the most relevant words of the text where the named entity appears, NEREA calculates its BOW and builds a weighted context vector using TF-IDF for each word.
4. Compares each vector to the context vector generated by the same procedure for each candidate, using the cosine similarity, a common method used to measure the similarity using the BOW model.

To illustrate the NEREA application, suppose we found the named entity "Barcelona" in a news item. It is a term that can have several meanings, out of which we focus on three very different ones: 1) the Spanish city, 2) the soccer team of that city, and 3) the botanist Julie F. Barcelona. In the content of each news item, a series of words will appear that, due to their relevance, will form the context vector. In the case of the city, these words can be: "ramblas", "beach", "town hall", etc. In the case of the soccer team, we find: "match", "result", "coach", etc. Finally, in the case of the person, we find, for example, "botany", "scientist", "New Zealand". Both the city and the soccer team will almost certainly appear referenced in the local classification resources of any Spanish media, and with the context information generated from the published news, it will be possible to detect whether the news is about the city or the soccer team. But if the news is related to the biologist, we can use DBpedia[6] (Auer et al. 2007) (one of the main hubs of Linked Data), where the entry[7] provides enough context (using the *abstract* field) to discover that the news item actually refers to her.

In summary, the purpose of this process is to obtain a local catalog of unique and unambiguous entities, linked with local metadata or external Linked Data, which can provide knowledge to the entity. This enables NEREA to perform the disambiguation tasks automatically, with that catalog as the main resource of the local environment.

4.2.2.2 Application of NEREA to SJORS To disambiguate the entities in the incoming wire news, SJORS builds an *Entity Catalog* using the aforementioned methodology. The Entity Catalog stores a list of records,

[5] NEREA was chosen for NE extraction due to its applicability to the news domain. Prior experiments verifying its effectiveness (70% at F-Measure, as reported in Garrido et al. (2016)).

[6] https://www.dbpedia.org/, although we use the Spanish version.

[7] https://es.dbpedia.org/page/Julie_F._Barcelona.

each one formed by the following elements: (1) the entity itself (unique), (2) a list of related named entities (they can appear in other records), (3) an URL to link to an external source in Linked Data Format (e.g., DBpedia) that contains textual information about the entity, and (4) a set of keywords obtained from the textual information about the Linked Data. The creation of the Entity Catalog is an offline process that has to be performed before deployment. However, despite the high number of different entities that can be found in the wire news received, its update is efficient and can be done continuously and incrementally as the information goes into the newsroom's news agency database.

SJORS uses the named entities contained in the wire news for composing the vectors, called *item vector* v_{tNE}. To obtain this representation, SJORS includes a dedicated module in its architecture called *NE Vectors Generation* (see Fig. 3). After the lemmatization process, with the help of a morphological analyzer, the system selects those words detected as named entities[8]. Unlike BOW, we do not remove stop words, as they can be part of NEs Using again TF-IDF as defined in Eq. (1), the system obtains the frequencies of appearance of the named entities to generate the weighted NE vectors. It is important to note that the TF-IDF approach is usually associated with power-law distributions, and its use is appropriate in the context of the news and the named entities (Zhang and Skiena 2014), having been validated empirically in the past by Garrido et al. (2016).

4.2.3 Modeling Using Word Embeddings

For modeling wire news, SJORS also considers techniques based on *static* word embedding techniques (WE) (Turian et al. 2010) that produce representations at different levels: word, sentence, and document embedding. All of them require a neural network model to learn word associations from a large corpus of text, and once trained, the model is used to identify words whose meaning is similar regardless of syntax. Below is a brief description of each of these approaches:

1. *Word2vec (W2V):* Word-level embeddings are based on the embedding approach described in Mikolov et al. (2013). In this case, two-layer neural networks are trained to reconstruct linguistic contexts of words and produce a vector space, typically of several hundred dimensions, where each unique word in the input corpus is assigned to a corresponding vector in the space. Words that share a common context in the corpus are located close to one another in the vector space. The vector representation of each piece of wire news is obtained using the W2V representation of its words, which are averaged using their centroid.
2. *SIF:* Building on W2V, SIF (Arora et al. 2019) represents a sentence by a weighted average vector of its word vectors, from which the most frequent component of the document corpus is subtracted. The most frequent component is obtained using PCA/SVD[9] over the whole vocabulary vector representation, considering that this vocabulary representation already captures the information seen in the corpus. This component may encompass words that frequently occur in a corpus but lack semantic content, acting similarly as IDF term in TF-IDF scoring scheme.
3. *Doc2vec (D2V):* This embedding method has a different learning strategy (Le and Mikolov 2014). It exploits the idea that the prediction of neighboring words for a given word strongly relies on the document as well. It is an extension to the W2V models, adding another feature vector, which is document-unique, and its goal is to directly embed each document in the vector space, regardless of its length. In this case, the vector of each piece of wire news is directly calculated by the trained model given the text of a piece of wire news.

Regardless of the embedding technique used, SJORS can generate WE vectors denominated *item vector* v_{tWE}. For this purpose, SJORS uses a dedicated module in its architecture called *Word Embeddings Vectors Generation* (see Fig. 3). This module is also in charge of building the required models offline. SJORS needs access to the Agency News Database to build an internal specific large dataset of news. Note that the vocabulary, the named entities, and the writing style (in this case, journalistic style) from the text of this large dataset have to be aligned, given the specificity of each wording and domain it covers, with the text data from the Agency News Database and the Production Database.

4.3 Modeling Journalists

To model journalists, SJORS explicitly accounts for the relationship that the journalist has with the accessed wire news and his/her published articles. To capture the journalist's interests, SJORS relies on three different vectors:

[8] For NE recognition, SJORS uses Freeling (http://nlp.lsi.upc.edu/freeling), a natural language tool capable of processing the Spanish Language.

[9] Principal component analysis (PCA) and singular value decomposition (SVD) are commonly used linear dimensionality reduction approaches in exploratory data analysis and Machine Learning. They attempt to find linear combinations of features in the original high dimensional data matrix to construct a meaningful representation of a dataset.

- Previous Search Results (PSR). SJORS keeps track of the previous searches performed by the journalist, obtained from the Wire news's Input Software (see Fig. 3). From each search, SJORS uses the top-k wire news (i.e., the most relevant) for creating a set of items called *PSR* set[10]. This set reflects the context of the different searches that the journalist has conducted on the Wire news's Input Software. Using the wire news in *PSR*, SJORS builds v_{PSR}, a vector obtained by averaging the respective wire news' vectors.
- Accessed Wire news (AT). Apart from *PSR*, SJORS keeps track of the items that the journalist has accessed for a complete reading, also obtained from the Wire news's Input Software (see Fig. 3), called the *AT* set. This set captures the actual journalists' current focus of research/interest. *AT* is then used by SJORS to build v_{AT}, a vector obtained by averaging the vectors of the accessed wire news in *AT*.
- Published News (PN). Lastly, SJORS turns to a set that contains the texts of the pieces of news published by the journalist. These news articles capture the general domain of the recent activity of the individual. The set, called *PN*, is obtained from the Newsroom Production Software (see Fig. 3) and it is used to build the vector v_{PN}.

SJORS models a journalist as a combination of these three vectors (i.e., v_{PSR}, v_{AT}, v_{PN}), calculating a linear combination of the components into one vector called *journalist vector* v_J:

$$v_J = A \times v_{PSR} + B \times v_{AT} + C \times v_{PN} \quad (2)$$

with A, B, C between 0 and 1 and $A + B + C = 1$.

Due to the nature of the items they contain, the sets of wire news that are used for modeling have different lifespans, which have to be adapted to the particular work pace of the newsroom. On one hand, the contents of PSR and AT have to be updated frequently depending on the number of wire news entering from outer agencies, as well as the inner maintenance policy (i.e., the lifetime of a piece of wire news in the system). On the other hand, the temporal window (or the number of items considered) for PN also has to be adjusted to reflect the actual subject flexibility requirements of the journalists. For example, when deploying SJORS in a real-world newsroom, we observed that a journalist writes an average of one or two news articles each day, which means that 20 news articles represent between two to four working weeks (considering days off). This amount of time was enough to define the topics recently addressed by that journalist as shown in our experiments (see Sect. 5). In our setup, going further back in time does not seem to have much interest, since the topics to be published in a media change pretty frequently.

[10] In our setup, we experimentally set K = 10.

4.4 Recommendation Generation

The *Recommendation Generator* is the last step in Fig. 3. Using the vectors representing the wire news and the vector capturing the interests of the journalists (v_t and v_J, respectively), SJORS determines the relevance of a given piece of wire news t for a given journalist J as follows:

$$rel(J,t) = \alpha \times \left(1 - \frac{(dt_0 - dt_t)^e}{T^e}\right) + \beta \times cos(v_J, v_t) \quad (3)$$

where dt_0 the current time, dt_t the time when the piece of wire news was received in the system, T the maximum time that the item is kept in the system, e a parameter that defines the novelty and importance of the item for the ranking, *cos* is the popular cosine similarity (Li and Han 2013; Gunawan et al. 2018), and v_J and v_t the vectors associated to the journalist and the piece of wire news, respectively.

Equation (3) has three coefficients (α, β, e) that modulate the importance of each of the two dimensions that form it, *time and similarity*:

- α represents the weight of the novelty of a piece of wire news since in our approach we assume that the newest wire news is going to be more interesting to a journalist.
- β represents the weight of the proximity of the topic of the wire news on the topics the journalist is interested in (the cosine similarity).
- e is a coefficient to adjust the weight of the importance of the recentness of the piece of wire news for the journalists. Values between 0 and 1 result in a smooth importance decay, while values above 1 translate into a sudden drop in the importance of a piece of wire news as soon as it takes a certain time in the system, flattening the values of the recommendation. We have observed in the experiments that value close to 1 are best suited to the newsroom domain.

These coefficients are determined empirically, adapting them to the particularities of each newsroom. Once they are set up, SJORS generates the list of Top-N wire news for each of the journalists in the newsroom, ordering them by the *rel*(J, *it*) value.

It is possible to find duplicate elements among wire news considered for recommendation–it is very common for a piece of wire news number to be sent several times sent by one or several agencies. Thus, SJORS eradicates duplicates by removing from the Top-N list items that have an overlap greater than 90% as determined using *ROUGE-N* (Lin 2004), a well-known metric used for comparing

texts. In those cases, the piece of wire news of smaller size is eliminated from the top-N list.

5 Experiments and Analysis

In this section, we first describe the dataset we have collected for assessment purposes. Then, we describe the metrics employed to quantify the performance of our proposed recommender as well as the experimental framework. Finally, we discuss the results of the empirical assessment that we have performed to analyze the design and overall performance of our approach.

5.1 Standard Dataset and Established Baseline

To our knowledge, there is no standard dataset in the domain under study. Thus, to gather data necessary for evaluation, while preserving the environment as real as possible, we established a collaboration with Henneo, the seventh Spanish communication group by turnover volume, with several media devoted to local and national scope.

We built a dataset, called *HenneoNR21*, that captures a real newsroom work environment. We registered the activity of a group of 20 randomly-selected journalists for a one-month period (30 days) in the newsroom of the newspaper *Heraldo de Aragón*. Both the news and the wire news contained in the dataset are in Spanish. The activities tracked from journalists in *HenneoNR21* include information from the external agencies database and the production database:

1. *Searches*: The external news search engine only allows to query the database using date range and free text filters, obtaining a list of headlines of the wire news included within the scope of the search. Whenever a journalist performed a search, a trace with the list of the wire news retrieved was stored along with his/her ID.
2. *Accesses*: When browsing search results, the wire news that the journalist selected to further read (i.e., clicked on the headlines) were also traced, indicating that they have been accessed for further analysis.
3. *Publications*: The production software kept track of the authorship of the different pieces of news written, which allowed us to access the last pieces of news published by each journalist on the media.

We summarize in Table 1 the main topics often covered by the journalists who participated in our data collection process. Recall that journalists in newsrooms can be versatile and work on several topics at the same time.

As wire news from news agencies are purged after four days, we created a copy of the input coming from the agencies. We archived over the month period 164,628 wire news in a test database outside the production system. We integrated this test database along with the journalists' activity information to make it possible to completely replicate journalists' behavior on real data. This allowed us to analyze and probe the applicability of SJORS in a controlled environment.

5.1.1 Capturing Data

To create a gold standard that could be used for evaluating SJORS, we randomly selected a total of 100 publications from these 20 journalists, published during 30 days along which we built the dataset. More than one piece of wire news can be appealing to a journalist J in terms of helping him/her compose a publication P. These wire news are denominated "relevant".

5.1.2 Relevance Assessment

Requiring journalists to label as "relevant" or "not relevant" wire news presented to them is not feasible due to daily time concerns. We argue that semi-automating this process can be achieved by looking for wire news that present a high amount of overlapping content with respect to a corresponding publication because the wire news have been searched and accessed *before* the publication of the pieces of news. To quantify the degree of overlap between a piece of wire news (T) and the news published by the journalist (PN) we use ROUGE-N (Eq. 4).

$$ROUGE{-}N = \frac{\sum_{i=1}^{gram_n \in I} Count_{match}(gram_n)}{\sum_{i=1}^{gram_n \in I} Count(gram_n)} \tag{4}$$

where n is the length of the n-gram, $Count_{match}(gram_n)$ is the maximum number of different n-grams co-occurring between T and PN, and $Count(gram_n)$ is the number of different n-grams in PN.

In the newsroom domain, a recommended item (a piece of wire news) T is treated as *relevant* to a given journalist J, if T's topic and content are closely related to the news that the journalist will later publish. It is not practical for journalists to manually verify the relevance of each of the aforementioned 100 publications. Thus, we identify as the "ideal" (i.e., relevant) wire news that should be suggested to the corresponding journalist the top-10 with the highest ROUGE value (from those collected four days[11] before each of the publications). We do so given that we knew

[11] Four days is the average time in which an article can be worked on before its publication, according to the information obtained from real journalists.

Table 1 Detailed information about the journalists from the newspaper *Heraldo de Aragón* whose activity has been collected to create the dataset used for evaluation purposes

Main topics	Journalists	Frequency	Size (words)
Local	4	Daily	300–900
Sports	4	Daily	300–900
National-international	2	2–5 a week	> 900
Culture	2	Weekly	300–900
Events chronicle	2	2–5 a week	< 300
Economy and business	2	2–5 a week	300–900
Editorial	2	Weekly	< 300
Entertainment	2	Weekly	> 900

what the journalist finally published. ROUGE only serves to verify a-posteriori, or as in this case, to help us to produce our *Gold Standard*.

To evaluate the approaches based on the use of word embeddings, a database with a large volume of news was necessary to be able to generate the models. For this, wire news from different agencies were stored for several months until reaching *two millions*.

Given that the resources in *HenneoNR21* refer to private data of Heraldo de Aragón and the news agencies, it is not possible to make this dataset publicly accessible. However, researchers interested in accessing *HenneoNR21* for experimental purposes are encouraged to email the corresponding author to provide access through a private agreement.

5.2 Metrics

To quantify the performance of our proposed recommender, we use two common performance measures: Normalized Discounted Cumulative Gain (*nDCG*) (Järvelin and Kekäläinen 2002) (Eq. 5) and Mean Reciprocal Rank (*MRR*) (Croft et al. 2010) (Eq. 7). The former consider the correctness of the recommendations and penalizes *relevant* recommendations positioned lower in the ranking. The penalization is based on a reduction, which is logarithmically applied to the position of each *relevant* item in a ranked list. The latter captures the average number of suggested items a user has to scan through to identify a relevant one.

The *nDCG* for a recommendation list L generated for a user J is calculated as:

$$\mathrm{nDCG}_{L,J} = \frac{\mathrm{DCG}_{L,J}}{\mathrm{IDCG}_J} \tag{5}$$

with $\mathrm{DCG}_{L,J}$ defined as:

$$\mathrm{DCG}_{L,J} = \mu_J(T_1) + \sum_{i=2}^{|L|} \frac{\mu_J(T_i)}{\log_2(i+1)} \tag{6}$$

where T_i is the i-th item in L, $\mu_J(T_i)$ is J's utility for T_i (i.e., 1 if relevant, 0 otherwise); and IDCG_J is computed as $\mathrm{DCG}_{L,J}$, with respect to a list consisting only of the items *relevant* to J in non-increasing order of utility according ROUGE value.

The MRR for L given J is computed as:

$$MRR_{L,J} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{7}$$

where $rank_i$ is the ranking position of the first relevant recommended piece of wire news for a sample of queries Q.

5.3 Experimental Setup

We detail the experiments conducted to assess SJORS.

5.3.1 Ablation Study

We have performed an ablation study using five different vector representations in five experiments for obtaining the ideal top-N recommendations. The objective is to see which is the best methodology, both at a general level and in particular according to the topics covered by the journalists. In the first experiment of the study, the wire news and the journalists were modeled using the BOW vectors. In the second one, the modeling was performed using NE vectors disambiguated through Linked Data. Experiments 3, 4, and 5 cover the use of three different techniques based on word embeddings: Word2vec, Doc2vec, and SIF, respectively. Finally, we evaluate a linear combination of the five techniques to try to find combinations that offer better results.

5.3.2 Time Considerations and Calculations

In all these experiments, once the wire news and journalists were modeled, for each publication P made by a journalist J in *HenneoNR21*, we identified the corresponding best top-N wire news recommendation following the Eq. (2) presented in Sect. 4.3. To simulate the behavior of a newsroom as much as possible in our experiments, we considered as candidate wire news to be recommended those available at 20 P.M. of the day before the publication P was included in the newspaper. We selected this cut-off since our observations revealed that it is the peak hour for

writing the articles to publish them in the newspaper of the next day[12]. SJORS gets the 10 more relevant wire news and calculates *nDCG* and *MRR* to evaluate each recommendation list using Equations (5) and (7), respectively. When a journalist J performs an action, the journalist vector v_j is recalculated using Eq. (2), and then SJORS generates the top-N recommendations by examining each available piece of wire news and determining its relevance to J using Eq. (3). Thus, the top-N recommendations for a working journalist are continuously recalculated for taking into account his/her activity and the continuous entry flow of wire news.

5.3.3 Dataset Split

For experiment purposes, we divided *HenneoNR21* into two parts: one for development and one for testing. For the *development dataset*, we select the data from the first 24 days of the month, with their corresponding 80 publications from the selected journalists. The *testing dataset* consists of the information belonging to the remaining 6 days, including 20 publications edited by the journalists. We did not employ cross-validation in our evaluations as it we would break the adequacy of the recommendation to the temporal thread.

5.3.4 Parameter Exploration

Using the *development dataset*, we followed a traditional parameter sweeping strategy (with increments of 0.1 among coefficients) and set SJORS' coefficients. Using different coefficient values for A, B, C, α, and β, SJORS generated recommendations for (journalist, publication) pairs in the *development dataset*, which were evaluated based on the gold-standard (i.e., top-10 ideal wire news) using nDCG and MRR. As argued in Sect. 4.4, we established parameter e from Eq. (3) to 1, to have a linear decrease in the importance of the news according to their recentness[13]. The computed nDCG and MRR were averaged over all (journalist, publication) pairs in *HenneoNR21*. Lastly, we settled for coefficient values for each vector modeling, which led to overall better suggestions. These coefficients, shown in Table 2, are used as the default parameters for SJORS for subsequent evaluation purposes.

[12] Note that every time recommendations are generated, the corresponding published pieces of news are compared with an average number of 20,000 wire news.

[13] We tested higher values of parameter e, but the decrease was too steep and flattened the recommendations favoring just the most recent items.

5.4 Results, Analysis, and Discussion

Following the experimental framework discussed in Sect. 5.3, we evaluate the effectiveness of our proposed recommendation. For each model setup, and performing a sweep, we looked for the set of coefficients that lead to the best performance, as shown in Table 2. The best result is provided by the study carried out with word embeddings generated by SIF, and we can observe the importance of the accessed wire news (parameter B), while the recentness of the piece of wire news (parameter *alfa*) and the similarity of the topics between the piece of wire news and the journalist (beta parameter) are compensated.

Using the best coefficients for each model, and considering separately the topic groups of Table 1, we evaluated our approach on the testing dataset. A summary of the results is presented in Table 3.

Finally, to best take advantage of the different representation techniques to model journalists and wire news, we consider another approach for a recommendation based on a weighted linear combination of the different methods. We run a parametric sweep of the different methods, each of them with their best parameters appearing in Table 2, and then point to which one is the best option. Best global results are given in Table 4, where the importance of methods based on word embeddings can be appreciated.

Results summarized in Table 2 reveal that, on average, SJORS works better when vectors are modeled with semantic approximations based on NE and SIF. Moreover, it emerges from the results reported in Table 4 that SJORS' best performance is the result of combining all the modeling techniques, as all contribute to representing items, users, and context from complementary perspectives. It should be also noted that the values of semantic approximations based on WE are always relevant in the best configurations.

5.4.1 Study of the Coefficients

When observing the importance of the coefficients of Eq. (3) in Table 2, another significant finding is that, in general, when SJORS builds the journalists' vector v_j using the Eq. (2), the most important is *his/her access to the detail of the wire news* (v_{AT}, coefficient B), while the searches (v_{PSR}, coefficient A), and the publications he/she made in the past, (v_{PN}, coefficient C) are less relevant. Nevertheless, we cannot directly get rid of them as the profile of the journalist. The low relevance of coefficient C might be due to the dynamism of the newsroom. If journalists work for several sections, or there is high turnover in the newsroom staff, the value of the C coefficient is low. In any case, the value for these parameters

Table 2 Summary of the micro-averaged results obtained by SJORS on the testing dataset using the best coefficients determined via parameter sweep using the development dataset. Parameters A, B, and C are the weight of, respectively, Previous Search Results, Accessed Items, and Published Items. Parameter *alpha* represents the time, and *beta* the similarity between journalists' interests and wire news

Model	nDCG	MRR	A	B	C	α	β
BOW	0.501	0.633	0.2	0.8	0	0.3	0.7
NE	0.53	0.675	0	1	0	0.2	0.8
W2V	0.48	0.62	0	0.7	0.3	0	1
D2V	0.481	0.703	0.1	0.8	0.1	0.4	0.6
SIF	**0.544**	**0.783**	**0**	**0.8**	**0.2**	**0.4**	**0.6**

The values that provide the best results are shown in bold

should be tuned for the different deployment scenarios to make sure whether the historic profile is relevant or not.

5.4.2 Temporal Analysis

If we look in Table 2 at the parameter *alpha*, which captures the importance of how recent a piece of wire news is, compared to the beta parameter, which measures the degree of similarity between the topics of the piece of wire news and those covered by the journalist, we see that although in general the parameter *beta* has a little more weight, the temporal aspect cannot be neglected. That indicates that the novelty factor of a piece of wire news is something to take into consideration.

Table 4 Best results of SJORS on the testing dataset by using a linear combination of the item modeling techniques

BOW	NE	W2V	D2V	SIF	nDCG	MRR
0.2	**0.1**	**0.2**	**0.4**	**0.1**	**0.606**	**0.816**
0.1	0.2	0	0.5	0.2	0.600	0.816
0.2	0	0.1	0.2	0.5	0.560	0.833
0.2	0	0	0.2	0.6	0.555	0.833

The values that provide the best results are shown in bold

5.4.3 Topic Oriented Analysis

Based on the conducted preliminary exploration, given the relatively low number of published news, we observe that in micro and macro-averaged results SIF beats the rest of the approaches (see Table 3). But it is also noteworthy the low performance of WE-based approaches for economy and business, culture, and sports topics, which can be explained by the advantage that keyword and NE-based approaches can have when proper names are the most relevant elements.

From results reported in Table 3, we also see that are two topics for which SJORS apparently has not given the expected results: editorial and entertainment. On the one hand, editorial articles provide a specific perspective on an issue and often use a vocabulary and writing style that greatly differs from the straightforward, fact-based reporting often seen in wire news. On the other hand, the entertainment section typically includes interviews, information about health, beauty, religion, hobbies, books, and authors. It is important to note, however, that this section only appears in the newspaper once a week, i.e., on

Table 3 Micro and macro-averaged exploratory results obtained by SJORS on the testing dataset grouped by topics using the best coefficients determined via parameter sweep using the development dataset. The subtotal represents the number of pieces of news of each topic published by the journalists

Method →	BOW		NE		W2V		D2V		SIF	
Main topic (subtotal) ↓	nDCG	MRR	nDCG	MRR	nDCG	MRR	nDCG	MRR	nDCG	MRR
Local (5)	0.819	1	0.832	1	0.853	1	0.925	1	**0.935**	**1**
Economy and business (2)	0.597	1	**0.683**	**1**	0.585	1	0.585	1	0.437	1
Events chronicle (2)	**1**	**1**	0.765	1	0	0	0.322	0.2	1	1
Culture (3)	**0.47**	**1**	0.437	1	0.25	0.2	0.3	0.33	0.45	0.5
Sports (3)	0.376	0.5	0.571	0.75	**0.717**	**1**	0.584	1	0.52	1
National-international (2)	0.554	0.33	0.61	0.25	**0.83**	**1**	0.379	1	0.508	1
Editorial (2)	0	0	0	0	0	0	**0.213**	**0.5**	0.137	0.33
Entertainment (1)	0	0	0	0	0	0	0	0	0	0
Micro-average	0.501	0.633	0.53	0.675	0.48	0.62	0.481	0.703	**0.544**	**0.783**
Macro-average	0.477	0.603	0.487	0.625	0.404	0.525	0.413	0.628	**0.498**	**0.728**

The values that provide the best results are shown in bold

weekends. Wire news is seldom useful for this type of content, as the topics covered in the entertainment section are heterogeneous and can greatly vary from week to week, making previously published news in this section unrelated to current stories.

5.4.4 Applicability to Journalists

A positive aspect regarding SJORS' behavior is its ability to quickly suggest interesting content to the journalist. The influence of the journalist's past publications in obtaining good suggestions among the recommender items is very limited in most cases (note the low weight of parameter C, see Table 2). This suggests that SJORS does a great job in the short term, which is why despite having little data when a new journalist joins the newsroom, soon wire news suitable for their activity will be recommended. Another insight that seems to emerge from the presented results is that SJORS yielded better results for journalists who more frequently work in the newsroom (as we can see in Table 3, for example, Local vs Entertainment), which is logical because SJORS have more information about them.

5.4.5 Comparison

Comparisons with existing works focused on RS for which journalists are the main stakeholders could not be addressed. This is because the information considered in other works, discussed in Sect.2, is quite different from SJORS, and, consequently, the datasets are incompatible. As a result, those systems are hardly comparable in an offline setting, and it has not been possible to make coherent comparisons. On the one hand, other works present a different task than SJORS. Moreover, neither strategy bears in mind what in SJORS have been considered as modeling characteristics of a journalist: the typology of their publications, the historical and recent activity over the wire news received at the newsroom, and the changes in the topics they have covered. Another important fact is the situation in the timeline of the news, an aspect that is not usually considered in other works. The datasets used in these works do not include all this information, so a rigorous comparison cannot be made.

6 Conclusions and Future Work

In this work, we have introduced SJORS, a novel Recommender System in the news domain. Unlike the vast majority of research work targeting this area, SJORS is customized to journalists, instead of news consumers. In fact, SJORS aims at identifying wire news that journalists can use to ease their news article writing process.

The main contributions of this work are:

- We have provided an in-depth study of a particular domain and use case: the recommendation of wire news to journalists in a newsroom. The objectives of the users, the implications of choosing one or another news, and the factors involved in the recommendation are quite different from media news' readers.
- We have proposed an approach that considers two of the main nuances affecting RS in this domain: news recency and language ambiguity. The former is addressed through a simple, yet effective strategy that prioritizes the newer wire news. The latter one is faced, on the one hand, by performing an in-depth analysis of the past activity of the journalist, and on the other hand, by carrying out a semantic analysis of the wire news.
- Due to the lack of benchmark datasets that can be used to evaluate RS such as SJORS, we have conducted a study in a real newsroom context, generating *HenneoNR21*, a dataset in Spanish with information about the work in a newsroom of 20 journalists during one month interacting with more than 150,000 wire news. This new dataset, available under request, will allow researchers to conduct real-time studies and A/B testing to continue to understand the challenges/limitations inherent to this use case and develop new methods to better serve journalists.
- Based on this newly-created dataset, we performed several experiments in a real newsroom, which validated the performance of SJORS with 0.606 at nDCG and 0.816 at MRR.

As opposed to a simple search or dealing with duplicated and subtly varied resources from agencies, a recommender such as SJORS is meant to reduce the overload of information for journalists, and to minimize the time journalists spend looking for resources used for writing news. The advantages of SJORS' recommendation strategy are twofold: it allows the journalist to discover wire news of interest without carrying out an explicit search; further, it improves the journalist's experience by overcoming possible limitations of the news editing system. For example, SJORS can be integrated as follows: (1) The recommender's suggestions appear at the bottom of the searches in a different widget, (2) instead of personalizing the search results, wire news related to what the editor is currently writing (along with their long- and short term interests) are suggested, (3) users can set alerts that periodically send them wire news of interest by mail. Besides, another advantage of SJORS is that it is capable of adapting to a specific environment without the need for prior labeling, both at the level of the typology of the texts and the specific needs regarding the temporal requirements of the users.

One of the advantages of the architecture of SJORS is that, given its modularity, we could add and test new modeling techniques by adding new modules which provide different vectorization techniques. This has allowed us to be able to consider and compare easily all the previous modeling techniques seen in Sect. 5.

The applicability of this work is not limited to the journalistic field. In fact, we posit that the proposed methodology could be applied in other works targeting environments where there is a need for a recommender that works on multiple text-based elements (descriptions or content) for users whose only available information is their searches and/or their publications, and where the time factor is relevant. For example, SJORS' applicability could be studied in other areas such as online stores, and streaming platforms; it could be used by archivists, or even to recommend books or articles to users of a digital library.

However, this work also has some limitations that we hope to alleviate in the following approaches. On the one hand, since the system's functioning is linked to the journalist's activity, the lack of recording of that activity would hinder performance. Too many subjects covered by each journalist would also have an adverse effect on the results. On the other hand, the fact that the analysis is based on data constructed from a particular media organization could imply certain limitations to the study results: It could have a particular bias, which will be analyzed through its future applications in other media.

In future work, we plan to explore the use of pretrained Language Models (PLMs) such as BERT (Devlin et al. 2018) and its derivations, and the GPT-like (Wu et al. 2023) ones (this is, both masked and generative models) through the use of prompting (Liu et al. 2023) to reduce the data requirements, facilitate matching, and alleviate the ambiguity problem. However, it needs to be clarified how to include temporal constraints in the prompts, and how to include temporal knowledge within the PLM by updating the captured knowledge (Onoe et al. 2023). In this context, some dense retrieval techniques could be explored (Zhao et al. 2022), although they suffer from the update problem as well.

Finally, we want to extend our work to multimedia. We live in a multimedia world, one where information is no longer limited to text. Incorporating new types of elements (as photographs, videos, infographics, etc.) into SJORS' recommendation process through textual representation (based, for example, on advances in image semantics recognition) would surely lead to higher-quality recommendations.

Acknowledgements This work has been supported by Spanish national Project PID2020-113903RB-I00 (AEI / FEDER, UE) and DGA / FEDER. We want to thank HENNEO for their collaboration in different stages of the project. We also thank Maria G. Buey and Jorge Bernad for their valuable help.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Abdollahpouri H, Malthouse EC, Konstan JA, Mobasher B, Gilbert J (2021) Toward the next generation of news recommender systems. In: Proceedings of the web conference (www), pp 402–406

Arora S, Liang Y, Ma T (2019) A simple but tough-to-beat baseline for sentence embeddings. In: Proceedings of the international conference on learning representations (ICLR)

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: a nucleus for a web of open data. The semantic web. Springer, Heidelberg, pp 722–735

Berven A, Christensen OA, Moldeklev S, Opdahl AL, Villanger KJ (2020) A knowledge-graph platform for newsrooms. Comput Indust 123(103):321

Bizer C, Heath T, Berners-Lee T (2009) Linked data: the story so far. Semantic services, interoperability and web applications: emerging concepts. Inf Sci Ref pp 205–227

Bizer C, Heath T, Berners-Lee T (2011) Linked data: the story so far. In: Semantic services, interoperability and web applications: emerging concepts, IGI global, pp 205–227

Bodó B (2019) Selling news to audiences-a qualitative inquiry into the emerging logics of algorithmic news personalization in european quality news media. Digit J 7(8):1054–1075

Bordogna G, Pagani M, Pasi G, Villa R (2006) A flexible news filtering model exploiting a hierarchical fuzzy categorization. Flex Query Answ Syst pp 170–184

Boutet A, Frey D, Guerraoui R, Jegou A, Kermarrec AM (2013) Whatsup: a decentralized instant news recommender. In: Proceedings of the IEEE international symposium on parallel & distributed processing (ISPDC), IEEE, pp 741–752

Buey MG, Bobed C, Gracia J, Mena E (2021) A domain independent semantic measure for keyword sense disambiguation. In: Proceedings of the international. symposium on applied computing (SAC), ACM, pp 1883–1886

Camacho-Collados J, Pilehvar MT (2018) From word to sense embeddings: a survey on vector representations of meaning. J Artif Intell Res 63:743–788

Cantador I, Castells P (2009) Semantic contextualisation in a news recommender system. In: Proceedings of the workshop on context-aware recommender systems (CARS), ACM, vol 1068, pp 19–25

Carson A, Muller D (2017) The future newsroom. Centre for Advancing Journalism. University of Melbourne. Tillgänglig.

https://arts.unimelb.edu.au/__data/assets/pdf_file/0003/2517726/20913_FNReport_Sept2017_Web-Final.pdf, accessed 07 Nov 2023

Caselles-Dupré H, Lesaint F, Royo-Letelier J (2018) Word2vec applied to recommendation: hyperparameters matter. arXiv preprint arXiv:1804.04212

Cohen NS (2019) At work in the digital newsroom. Digit J 7(5):571–591

Croft WB, Metzler D, Strohmann T (2010) Search engines. Pearson Education, London

Cucchiarelli A, Morbidoni C, Stilo G, Velardi P (2017) What to write? a topic recommender for journalists. In: Proceedings of the EMNLP workshop: Natural language processing meets journalism (NLPJ), ACL, pp 19–24

Cucchiarelli A, Morbidoni C, Stilo G, Velardi P (2019) A topic recommender for journalists. Inf Retr J 22(1):4–31

De Souza Pereira Moreira G (2018) CHAMELEON: a deep learning meta-architecture for news recommender systems. In: Proceedings of the international conference on recommender systems (RecSys), pp 578–583

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Feng S, Meng J, Zhang J (2021) News recommendation systems in the era of information overload. J Web Eng 16:459–70

Firth J (1957) A synopsis of linguistic theory 1930-1955. Stud Linguist Anal pp 1–32

Gabriel De Souza PM, Jannach D, Da Cunha AM (2019) Contextual hybrid session-based news recommendation with recurrent neural networks. IEEE Access 7:169185–203

Garrido AL, Ilarri S, Sangiao S, Gañán A, Bean A, Cardiel Ó (2016) NEREA: named entity recognition and disambiguation exploiting local document repositories. In: Proceedings of the international conference on tools with artificial intelligence (ICTAI), IEEE, pp 1035–1042

Garrido AL, Sangiao S, Cardiel O (2017) Improving the generation of infoboxes from data silos through machine learning and the use of semantic repositories. Int J Artif Intell Tools 26(05):1760022

Golian C, Kuchar J (2017) News recommender system based on association rules @ CLEF NewsREEL 2017. In: Proceedings of international conferenece of the CLEF initiative, CEUR

Greenstein-Messica A, Rokach L, Friedman M (2017) Session-based recommendations using item embedding. In: Proceedings of the international conference on intelligent user interfaces (IUI), ACM, pp 629–633

Gunawan D, Sembiring C, Budiman MA (2018) The implementation of cosine similarity to calculate text relevance between two documents. In: Journal of physics conference series, IOP Publishing, vol 978, p 012120

Gürsel ZD (2012) The politics of wire service photography: infrastructures of representation in a digital newsroom. Am Ethnol 39(1):71–89

Heitz L, Lischka JA, Birrer A, Paudel B, Tolmeijer S, Laugwitz L, Bernstein A (2022) Benefits of diverse news recommendations for democracy: a user study. Digit J 10(10):1710–1730

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446

Johnson K, Radosh J (2023) The broadcast news toolkit: inside the digital newsroom. Taylor & Francis, Milton Park

Karimi M, Jannach D, Jugovac M (2018) News recommender systems-survey and roads ahead. Inf Process Manag 54(6):1203–1227

Kazai G, Yusof I, Clarke D (2016) Personalised news and blog recommendations based on user location, facebook and twitter user profiling. In: Proceedings of the international conference on research and development in information retrieval (SIGIR), ACM, pp 1129–1132

Khattar D, Kumar V, Varma V, Gupta M (2018) Weave &rec: a word embedding based 3-d convolutional network for news recommendation. In: Proceedings of the international conference on information and knowledge management (CIKM), pp 1855–1858

Kramp L, Loosen W (2018) The transformation of journalism: from changing newsroom cultures to a new communicative orientation? Communicative figurations. Palgrave Macmillan, Cham, pp 205–239

Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D et al (2018) The science of fake news. Science 359(6380):1094–1096

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the international conference on machine learning (ICML), PMLR, pp 1188–1196

Li B, Han L (2013) Distance weighted cosine similarity measure for text classification. In: Proceedings of the international conference on intelligent data engineering and automated learning (IDEAL), Springer, Heidelberg, pp 611–618

Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Proceedings of the ACL workshop: Text summarization branches out, ACL, vol 8

Linden CG (2017) Decades of automation in the newsroom: why are there still so many jobs in journalism? Digit J 5(2):123–140

Liou YT, Chen CC, Tang TH, Huang HH, Chen HH (2021) Finsense: an assistant system for financial journalists and investors. In: Proceedings of the international conference on web search and data mining (WSDM), ACM, pp 882–885

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9):1–35

Lv P, Meng X, Zhang Y (2017) Fere: exploiting influence of multi-dimensional features resided in news domain for recommendation. Inf Process Manag 53(5):1215–1241

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Montes-García A, Álvarez-Rodríguez JM, Labra-Gayo JE, Martínez-Merino M (2013) Towards a journalist-based news recommendation system: the Wesomender approach. Exp Syst Appl 40(17):6735–6741

Nan Q, Cao J, Zhu Y, Wang Y, Li J (2021) Mdfend: multi-domain fake news detection. In: Proceedings of the international conference on information & knowledge management (IKM), pp 3343–3347

Navigli R (2009) Word sense disambiguation: a survey. ACM Comput Surv CSUR 41(2):10

Niarchos M, Stamatiadou ME, Dimoulas C, Veglis A, Symeonidis A (2022) A semantic preprocessing framework for breaking news detection to support future drone journalism services. Future Intern 14(1):26

Onoe Y, Zhang MJ, Padmanabhan S, Durrett G, Choi E (2023) Can lms learn new entities from descriptions? Challenges in propagating injected knowledge. arXiv preprint arXiv:2305.01651

Özgöbek Ö, Kille B, Gulla JA, Lommatzsch A (2019) The 7th international workshop on news recommendation and analytics (INRA 2019). In: Proceedings of the international conference on recommender systems (RecSys), pp 558–559

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the international conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of the international conference of the north American chapter of the association for computational linguistics (ACL), Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237, 10.18653/v1/N18-1202, https://aclanthology.org/N18-1202

Raza S, Ding C (2022) News recommender system: a review of recent progress, challenges, and opportunities. Artific Intell Rev pp 1–52

Ren J, Long J, Xu Z (2019) Financial news recommendation based on graph embeddings. Decis Support Syst 125(113):115

Ricci F, Rokach L, Shapira B (2011) Introduction to recommender systems handbook. Recommender systems handbook. Springer, Heidelberg, pp 1–35

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523

Sekine S, Ranchhod E (2009) Named entities: recognition, classification and use, vol 19. John Benjamins, Amsterdam

Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol TIST 10(3):1–42

Siles I, Boczkowski PJ (2012) Making sense of the newspaper crisis: a critical assessment of existing research and an agenda for future work. New Med Soc 14(8):1375–1394

Smets A, Hendrickx J, Ballon P (2022) We're in this together: a multi-stakeholder approach for news recommenders. Digit J 10(10):1813–1831

Turian J, Ratinov L, Bengio Y (2010) Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the annual meeting of the association for computational linguistics (ACL), pp 384–394

Vasile F, Smirnova E, Conneau A (2016) Meta-prod2vec: product embeddings using side-information for recommendation. In: Proceedings of the international conference on recommender systems (RecSys), ACM, pp 225–232

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762

Vo N, Lee K (2018) The rise of guardians: fact-checking url recommendation to combat fake news. In: Proceedings of the international conference on research & development in information retrieval (SIGIR), pp 275–284

Voskarides N, Meij E, Sauer S, de Rijke M (2021) News article retrieval in context for event-centric narrative creation. In: Proceedings of the international conference on theory of information retrieval (ICTIR), pp 103–112

Westlund O, Ekström M (2019) News organizations. In: The handbook of journalism studies, 2nd edition, Routledge, London

Whitney DC, Becker LB (1982) keeping the gates for gatekeepers the effects of wire news. J Q 59(1):60–65

Wu C, Wu F, An M, Huang J, Huang Y, Xie X (2019) NPA: neural news recommendation with personalized attention. In: Proceedings of the international conference on knowledge discovery & data mining (SIGKDD), ACM, pp 2576–2584

Wu C, Wu F, Huang Y, Xie X (2023) Personalized news recommendation: methods and challenges. ACM Trans Inf Syst 41(1):1–50

Wu J, Lee J (2013) The transformation of media reporters in the perspective of new media: from traditional to versatile. In: Proceedings of the international conference workshop on computer science in sports (IWCSS), Atlantis Press, Dordrecht, pp 214–216

Wu T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y (2023) A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin 10(5):1122–1136

Zhang W, Pérez Tornero JM (2021) Introduction to ai journalism: framework and ontology of the trans-domain field for integrating ai into journalism. J Appl Journal Media Stud

Zhang W, Skiena S (2014) News-based group modeling and forecasting. arXiv preprint arXiv:1405.2622

Zhang Y, Jin R, Zhou ZH (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1(1–4):43–52

Zhao WX, Liu J, Ren R, Wen JR (2022) Dense text retrieval based on pretrained language models: a survey. arXiv preprint arXiv:2211.14876

Zheng G, Zhang F, Zheng Z, Xiang Y, Yuan NJ, Xie X, Li Z (2018) DRN: a deep reinforcement learning framework for news recommendation. In: Proceedings of the world wide web conference (WWW), ACM, pp 167–176