The Epistemology of Fine-Grained News Classification

Enrico Motta[a,b], Enrico Daga[a], Aldo Gangemi[c,d], Maia Lunde Gjelsvik[b], Francesco Osborne[a,e] and Angelo Salatino[a]

[a]*Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, United Kingdom*
[b]*MediaFutures Centre, University of Bergen, Lars Hilles gate 30, Bergen, Norway*
[c]*Department of Philosophy and Communication Studies, University of Bologna, Via Azzo Gardino, Bologna, Italy*
[d]*STLab, Institute for Cognitive Sciences and Technologies, National Research Council, 40126 Bologna, Italy*
[e]*Department of Business and Law, University of Milano-Bicocca, Milan, Italy*

Abstract. The process of news digitalization over the past decades has released massive amounts of news content, revolutionizing consumer access to news and disrupting traditional business models. These radical changes have also introduced new opportunities for media content analysis, potentially opening up new scenarios for ambitious large-scale media analytics initiatives, which can go well beyond the relatively small-scale studies currently carried out by media scholars and practitioners. However, take-up of computational methods to support media content analysis activities has been rather modest, reflecting a degree of disconnect between the needs of scholars and practitioners for task-specific and usable software solutions and the state of the art in computational techniques for news media analysis. In this paper we perform an initial step towards bridging this gap, by looking in detail at the task of *fine-grained news classification*. In particular, we propose a typology of *news topics,* which is formally specified and realised into a family of reusable ontologies. The proposed model has been validated empirically, through an analysis of a multilingual news corpus, as well as formally, in terms of the functional and logical properties of the ontologies. Our analysis brings together the media and computer science literature, connecting the formal definitions provided in this paper to the concepts used by media scholars.

Keywords: news classification, ontologies, knowledge engineering, formal specifications, semantic technologies

1. Introduction

The process of news digitalization over the past decades has released massive amounts of news content, revolutionizing consumer access to news and disrupting traditional business models [1]. For example, in the old age of print media, accessing multiple newspaper sources was relatively costly for consumers, hence most readers were essentially locked in the particular bundle of articles associated with a single newspaper. In the internet age, consumers can easily hop from one source to another, selecting the articles they wish to read from a variety of sources, which include both the digital versions of traditional newspapers, as well as offerings from digital-only news outlets. At the same time control over distribution and access to users has shifted from news producers to platform owners, weakening the link between readers and news sources and making it easier for rogue players to spread disinformation [46].

This explosion in news content availability has also introduced new opportunities for large-scale scholarly investigations [42]. In particular, both commercial and open access services are available, such as, Quantexa News Intelligence (www.aylien.com) and GDELT (www.gdeltproject.org). These services index and provide access to tens of thousands of news sources, thus enabling new solutions for large scale news monitoring and intelligence gathering.

Nonetheless, current solutions for automated news content analysis have enjoyed limited take-up by domain experts, such as political and media scientists [14]. As pointed out by Sjøvaag and Kvalhiem [73], current approaches to news classification suffer from either low granularity or high noise, thus reflecting a degree of disconnect between the needs of scholars for task-specific and usable software and the state of the art in advanced computational techniques for news media analysis [37]. Naturally, we agree with the aforementioned authors that there are issues related to both the usability and the performance of current computational solutions for news classification and, more in general, news analytics. However, we would also argue that a more fundamental problem concerns the lack of a strong epistemological foundation to the problem of classifying news items. In other words, we believe that, as a precondition to the development of effective computational solutions for *fine-grained news classification*[1], it is first necessary to characterise this task more precisely. To this purpose, in this paper we analyse the notion of *news topic*. In contrast with the topic modelling literature [39], where this concept defines a technical term denoting a vectorial representation of a set of documents, here we characterise it informally as "a matter dealt with in a text, discourse, or conversation; a subject"[2]. In particular, this paper provides the following contributions: we propose a conceptual framework covering the different types of news topics; we provide a specification of the framework in formal logic, which is then realised as a reusable ontology; and we validate the proposed approach both empirically, through an analysis of a multilingual news corpus, as well as formally, by verifying the functional and logical properties of the ontology.

The rest of the paper is organised as follows: in the next section we articulate the motivation for this work, highlighting the limitations of current computational approaches to news classification and the existing gap between the needs of media scholars and practitioners and the solutions available to them. In section 3 we illustrate the components of the proposed framework, which is then formalised in section 4. In section 5 we present a family of ontologies that instantiate the formal specification provided in section 4. In section 6 we illustrate an initial validation of the framework on a news corpus comprising both Norwegian and British news. Finally, in section 7 we discuss the relevant literature, while in the concluding section we reiterate the main contributions of this paper and discuss the next steps of this research.

2. Approaches to news classification

An established solution to classifying very large document collections, including news corpora, is provided by topic modelling approaches [39]. Their popularity is due to two main features of these algorithms: i) they can effectively scale up to very large document collections and ii) they do not require training data. However, there is also a broad consensus in the research community that such approaches suffer from an "interpretability problem" [18]: while their stated objective is to group documents into coherent topics (expressed as probability distributions over terms used in the document corpus), in reality it is not at all clear which notion of "topic" these approaches actually capture. Indeed, researchers have shown both that i) the outputs generated by topic modelling algorithms are not necessarily meaningful to humans [18] and also that ii) the evaluation metrics used in the topic modelling literature are neither robust across application scenarios [39], nor they correlate with human judgement [26].

An approach to addressing the issue concerning the semantic opaqueness of topic modelling algorithms focuses on generating succinct labels that are meaningful for humans [11] [3]. For example, the paper by Alokaili et al. [3] presents a state-of-the-art *topic labelling* algorithm that is able to generate labels, such as, "biofuel" from vectors, such as, <oil energy gas water power fuel global price plant natural>. Analogously, the paper by Bhatia et al. [11] includes a number of examples, such as, generating the label "criminal investigation" from the vector <investigation fbi official department federal agent investigator charge attorney evidence>. However, the issue here is that while these labels may be correct, they tend to have relatively little discriminatory power. For instance, in the latter example, the generic label "criminal investigation" identifies a broad topic rather than the specific one being discussed in the news, which ought be characterised more concretely in terms of a specific investigative event, carried out by a specific investigative agency, etc.

An alternative to the traditional topic modelling approaches is provided by modern neural network solutions, such as transformer-based language models, which have been shown to improve the state of the art

[1] We characterise the task of fine-grained news classification as the identification of the specific news topics (e.g., entities, events, situations and other types of concepts), which provide the main focus of a news item. This characterization will be further elaborated in sections 3 and 4.

[2] Google's online English dictionary.

in a variety of natural language processing tasks [23]. However, in order to produce high-quality results, these approaches have to be trained or fine-tuned on annotated news corpora, and unfortunately current datasets are not granular enough to support fine-grained news classification. For example, the widely used AG's news corpus[3] only classifies news items with respect to very high-level categories, such as Sports, Business and Sci/Tech.

The same granularity issue also applies to current vocabularies for news classification, which, while providing broad coverage across a wide variety of topics, are rather coarse-grained — i.e., only focus on rather generic categories. For instance, the *IPTC Media Topic NewsCodes*[4] taxonomy provides about 1300 concepts, starting from top-level nodes, such as "politics", "science and technology" and "crime, law and justice", and expanding down through six levels to leaf nodes, such as "capital punishment" and "suspended sentence". However, the primary purpose of these vocabularies is to provide a certain level of coarse-grained interoperability among news providers, and they do not address our goal of supporting the identification of the fine-grained topics (e.g., specific entities, events, situations and other types of topics) that a news item may focus on.

Other approaches have instead focused on identifying specific types of concepts inside the body of news items, such as *named entities* [81] [71] and *events* [41]. Indeed, because news coverage tends to be *event-centric* [73], event extraction is a key capability that is required to support any approach to fine-grained news classification. However, while these techniques are essential to support effective solutions for fine-grained news classification, they do not provide a complete solution, as not all news items necessarily centre on a particular event or entity. For instance, a news item may discuss an issue, such as the refugee crisis, without necessarily focusing on a particular event or entity. Indeed, even when a news item focuses on a particular entity — e.g., Italy, the discussion would usually centre on a particular aspect of the entity in question — e.g., its financial or political situation, or its natural attractions.

Given this state of affairs, it is not surprising that, as already pointed out, current solutions for automated news content analysis have enjoyed limited take-up by domain experts [14] [37]: in addition to the issues of performance and usability mentioned earlier, they also lack granularity and completeness.

Hence, we believe that there is a need to investigate the task of fine-grained news classification more comprehensively than it has been done so far in the literature. As already pointed out, the Computer Science literature by and large focuses on specific computational methods for individual elements of the overall puzzle — e.g., through extensive research on event extraction [43]. However, it has failed to provide a comprehensive analysis of the typology of concepts that can play the role of topic in news items, a key precondition for developing more robust and complete solutions for media analytics. Taxonomies, such as IPTC, provide a valuable resource for content interoperability, but they are mere vocabularies and, therefore, do not shed much light on the types of concepts that may need to be identified by computational engines attempting to perform automatic news classification. Hence, the work presented in this paper aims to provide both a fundamental (i.e., epistemological) and pragmatic (i.e., computational) value. In particular, a formal characterization of the types of concepts that can play the topic role in a news item provides a principled way to identify the current gaps in computational support for news classification.

3. A Framework for News Classification

In this section we discuss the types of concepts that provide the subject matter for news items. To this purpose, we define the task of fine-grained news classification as *the identification of the salient elements in a news story — i.e., the relevant news topics*. This definition has commonalities with the notion of *agenda setting* in the media literature [50], which is also concerned with the salience of issues in the media. However, agenda setting focuses primarily on the impact of the media agenda on the public and of course characterizations of this notion do not include the kind of epistemological analysis that is the focus of this paper.

In particular, our framework for fine-grained news classification distinguishes five generic types of news topics: *Entities*, *Events, Situations*, *Categorical Topics* and the *Commentary*. These are discussed in the following sections.

[3] https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset.

[4] https://iptc.org/standards/media-topics/.

3.1. Entities

Many news items focus on a particular entity. In principle this can be anything, including a person, an animal, a plant[5], a mineral[6], an organization, a country, a fictitious entity[7], a geographical place and several others. *Named Entity Recognition* [81] and *Entity Linking* [71] are well understood tasks, for which highly performant off-the-shelf methods are available, which are in routine use in commercial data services. For example, the aforementioned *Quantexa* news service automatically links entities in a news item to the relevant *Wikidata* entries [80].

However, to say that a news item is about an entity does not necessarily provide the most granular classification, as news may focus on a particular *aspect* of an entity. For example, while it would be correct to classify the topic of the news item at http://tinyurl.com/mu6h6nuv as "Donald Trump", a more accurate classification would indicate that the actual topic is "Donald Trump's wealth" or "Donald Trump's financial status". Analogously, as already pointed out, a news article about Italy is unlikely to focus on all aspects of this country, but most likely will focus on its economic or political situation, or its natural resources, or its artistic heritage, etc. The notion of entity aspect discussed here is related to the second level of *agenda setting theory*, which deals with the salience of the attributes of the entities that are the focus of attention in the media [50].

Entities can also be related to other entities and such relations can themselves be the focus of a news item. A typical case happens when a newspaper investigates the relationship between a politician and a businessman in the context of a corruption enquiry. A more unusual example concerns the story about the "Honduran Maradona"[8], whose core subject is actually the relationship between the writer and a Honduran football player.

In sum, while entities can be the focus of a news item, an *entity aspect* or a *relation between entities* can also play this role. Hence, while methods for *Named Entity Recognition* and *Entity Linking* play an important role in supporting fine-grained news classification, they do not necessarily provide the complete solution, even if we only focus on the subtask of entity-centric classification. Here, additional computational techniques are needed — e.g., methods for relation extraction [47][81][31].

3.2. Events

As already mentioned, news coverage focuses to a large extent on *events* [73], hence these play a key role in any news classification framework. To identify occurrences of events in textual content, the research community has over the years developed domain-independent event extraction solutions [41][44], which can achieve good performance, especially in the sub-task of *event detection*[9]. In addition, extensive event taxonomies [83] are also available, which can be used to support general-purpose event extraction solutions. While such solutions are essential to enable computational approaches to news classification, here we abstract from specific computational methods and domains to focus instead on the generic types of events that provide the focus of a news story. These are described in what follows.

3.2.1. Individual events

These provide the basic building block for a discussion about events. An individual event is simply something that is believed to have happened, such as a car crash, a bank robbery, a football match, an election, etc. Here, we use the formulation "believed to have happened" to include events that may not have actually taken place but are discussed as if they were real — e.g., because they are associated with specific entities and have spatio-temporal coordinates. A well-known historical example of false reporting in a newspaper is the 1948 headline of the Chicago Daily Tribune, "Dewey defeats Truman". This headline mistakenly declared Republican candidate Thomas Dewey as the winner of the 1948 U.S. presidential election when, in fact, incumbent President Harry Truman had won.[10] Other common examples come from novels, movies, and the theatre, where fictitious events are presented as if they were real. Hence, while detecting fake news and fake events are very important tasks covered by a vast literature — see, e.g., [40], in the context of this discussion the notion of 'individual event' comprises both real and

[5] https://www.theguardian.com/environment/2022/sep/23/gran-abuelo-chile-world-oldest-living-tree-alerce.

[6] https://eu.usatoday.com/story/news/world/2022/10/07/pink-diamond-auctioned-per-carat-world-record/8209350001/.

[7] https://eu.cincinnati.com/story/news/2019/11/18/mickey-mouse-birthday-disneys-iconic-character-turns-91/4226969002/.

[8] https://tinyurl.com/2p8rcza3.

[9] *Event detection* focuses on identifying references to events in the text, while *event extraction* requires identifying both an event and its *arguments*, i.e., the entities involved in the event.

[10] https://en.wikipedia.org/wiki/Dewey_Defeats_Truman.

imaginary events, as its purpose is to characterise any event that is discussed in the news, independently of its grounding in the real world.[11]

Individual events are normally *composite events* — i.e., they have sub-events. For instance, a bank robbery would include a variety of more specific events, such as the robbers entering and exiting the bank, threatening the staff in the bank, grabbing the money, etc. However, a journalist may decide that these more granular events are not *salient* enough to warrant too much attention and the story itself should instead centre on the bank robbery as a whole. That is, while a specific individual event may ontologically comprise a variety of sub-events, a journalist may decide that the more granular sub-events are not interesting enough and therefore ignore them – in this example treating the bank robbery event effectively as an *atomic event*.

3.2.2. Collections of events

Often, news stories focus on collections of events, rather than individual ones. For instance, let's consider a story that discusses sightings of unidentified drones near oil and gas fields in Norway[12]. From a journalistic point of view, it makes sense to group all these events together because i) they are obviously of the same type and ii) doing so increases the importance of the story. This aspect is related to the *Impact news angle*[13], sometimes also called *Prominence* [72], which emphasises that the value of a news item depends on the size or impact of an event. For instance, one person getting food poisoning at a wedding party is unlikely to make the news, while 100 people getting food poisoning is much more newsworthy. As Shoemaker and Reese point out, "the importance of a story is measured in its impact: how many lives it affects" [72]. Hence, analogously to the earlier discussion about individual vs composite events, also in the case of a collection of events the key classification criterion is journalistic rather than ontological. That is, while a variety of different events may exhibit significant commonalities, we talk about a collection of events only when a journalist has grouped a number of events together, either to enhance the impact of a news story or because the events naturally form a collection for reporting purposes — e.g., when we talk about all the football matches played in a particular round of the football league.

When talking about a collection of events, a key element is what Carriero et al. [17] refer to as the *unifying factor*, the criterion that determines membership of the collection. For instance, when grouping together multiple drone sightings, the unifying factor may abstract from the specific sighting modality — e.g., detection through a radar screen vs direct sightings by humans.

Here, it is important to emphasise that an event involving multiple agents does not necessarily define a collection of events. For example, an individual terrorist attack may injure or kill many people, however this can be treated as an individual event, if we are talking about an individual attack in a specific spatiotemporal location, regardless of the size of the casualties.

3.2.3. Negative events

When talking about events, intuitively we are inclined to think of actual events, which involve a certain number of agents, and take place in a specific location at a specific time. Hence, the notion of *negative event* — i.e., something that has not happened — is somewhat counter-intuitive and has been much debated in philosophy.

In particular, as discussed in [61], "if one's doing of something is an event, then surely one's not doing it is the absence of an event". However, as also pointed out by Payton [61], the problem with characterising negative events as absences is that in reality "we can manifest our agency just as much by not doing things as by doing them". For instance, a situation in which mutinying soldiers omit to fire at the enemy cannot be characterised simply as the absence of an event, but itself defines an important event that may warrant journalistic coverage. Therefore, our typology also includes this class of events. Like ordinary

[11] Another interesting case relevant to this discussion is one in which different news sources provide different accounts of the same real-world event. Again, in the first instance, we would be simply concerned with identifying these various event descriptions in the news, even though of course it would make sense also to add a second order reasoning module able to identify all event descriptions in the news that talk about a particular real-world event and to reason about possible discrepancies in the different conceptualizations. An approach to aligning such heterogenous event data is described in [32].

[12] https://www.pbs.org/newshour/world/unidentified-drones-over-norways-offshore-platforms-fuel-fears-of-russian-threat.

[13] A news angle is a journalistic framing device that is used "both i) to assess whether something is newsworthy and also ii) to shape the structure of the resulting news item" [53].

(i.e., positive) events, these can be characterised in terms of the relevant agents, although it is often tricky to locate negative events spatio-temporally. For instance, if a person decides not to cast their vote in the local elections, it is clear enough that they are expressing agency by not voting. However, it is more tricky to decide where and when this event of non-voting is situated – see [61] for a detailed discussion of these issues.

Naturally, just like positive events, not all negative events are necessarily newsworthy – indeed, the vast majority are not. However, in contrast with positive events, it is also the case that not all negative events necessarily make sense. For instance, if one of the authors of this paper is currently writing this page, they are also not doing an infinite number of other plausible things, such as playing the saxophone or riding a bicycle. However, they are also trivially not doing a variety of other far less plausible actions, such as standing upside down in Puerto Rico or talking to the Prime Minister of Papua New Guinea in Tok Pisin. Hence, while there are relatively few things that happen, there are practically an infinite number of things that don't happen. For this reason, as pointed out in [61], when talking about negative actions, we are primarily interested in *omissions*, either deliberate decisions of not doing something or simply failing to do something that is *expected to happen*. This characterization limits negative events to a meaningful subset of all conceivable negative actions, from which a journalist would then choose the ones that are newsworthy.

3.2.4. Dependent events

These are events that journalistically only make sense in the context of some other event. For instance, an event associated with a jury producing a verdict in a criminal trial only makes sense in the context of the broader set of events that together constitute a trial. A fundamental issue here is that, at least in the physical world, no event is independent in an ontological sense – see also [13] for a discussion about dependent entities in ontology engineering. For instance, a criminal trial trivially depends on a defendant being born. However, as in the case discussed earlier of collective events, the viewpoint here has less to do with ontology than with journalistic practice. Hence, while the accused in a trial can only be a person who was involved in a birth event, this particular event is usually not relevant to the discussion. The notion of dependency here is related to the notion of *background* or *context* in journalistic guidelines[14]: if the main event of a story is predicated on other events, which are essential to understand the event in question, then we say that this is a *dependent event*. Dependencies can be taxonomic, as in the case of the verdict event depending on a super-event (the trial), but can also be based on other principles, such as causality or preconditions — e.g., a trial took place only because a referral to trial was issued.

3.2.5. Predictions

A special type of meta-event that occurs regularly in the news is *Prediction*, as in the headline "Pupil numbers in England set to shrink by almost 1 million in 10 years"[15]. Here, it is useful to distinguish this type of event, because usually events in the news refer to things that are claimed to have happened (even when such events turn out to be fabricated) or to negative events where some agent or group of agents have expressed agency by not doing something. Hence, this type of event is distinguished from the other types discussed earlier both because of its metalevel nature, and also because it is the only type that does not focus on events that have or have not happened in the past.

3.3. Situations

A *situation* can be characterised as *a state of affairs*, typically resulting from one or more events. As such, events and situations are closely linked and indeed the definitions of these notions in the philosophical literature emphasise this close coupling. For instance, in [66] situations are defined as boundaries between events, with the latter characterised as "motions and actions" that engender a transition from one situation to another [66]. As an example, if railway workers are striking (an event), there may be no public transport options between certain cities affected by the strike (a situation).

Situations play an important role in news classification, given that news items often focus on the consequences of major events. For instance, as a result of an earthquake, a city may be without power and, depending on whether the focus of a news item is on the earthquake or its consequences, we could characterise it as either focusing on an event or a situation. An interesting case here concerns scenarios in which situations are expressed in negative terms — e.g., as the

[14] See, e.g., https://www.americanpressinstitute.org/journalism-essentials/makes-good-story/good-stories-provide-context/.

[15] https://www.theguardian.com/education/2022/jul/14/pupil-numbers-in-england-set-to-shrink-by-near-1m-in-10-years.

inability to do something. For instance, let's consider the news item entitled *"Third of young women and girls in UK can't access free period products"*[16]. One could consider whether this story should be characterised in terms of a negative event (young women in UK are not accessing free period products) or a situation characterised by the impossibility for young women in UK to access free period products. Our view is that a negative event implies an element of agency [61]: either a deliberate choice of not carrying out an action or a failure to do so — e.g., because the agent in question has simply forgotten about the action. However, in this case we are not talking about young women expressing agency, but simply about a situation where it is not possible for them to access free period products. Hence, we prefer to use the notion of situation, rather than event, to characterise this type of scenarios.

3.4. Categorical topics

As discussed in the previous section, situations define state of affairs, which normally can be directly linked to one or more events, which have led to the state of affairs in question. For instance, the UK's exit from the European Union (an event) has led to a 4% drop in the UK's overall GDP (a situation). However, while such 4% drop in UK's GDP can be characterised as a specific situation, the country's GDP or, more in general, its financial profile can be seen as a *categorical topic*, that is a topic for discussion and journalistic analysis that tends to be relevant and newsworthy regardless of any contingent situation. Such topics include social, economic and political issues, such as crime, poverty, taxation, finances, economic, foreign and defence policies, immigration, party politics, and many others. Indeed, many of these topics are captured by existing taxonomies, such as the aforementioned *IPTC Media Topic NewsCodes*, which cover generic categories in science, politics, arts and entertainment, education, health and other fields. While the role of these "categorical topics" is to provide coarse-grained aggregations of news items that cover the same domain, in contrast with the focus in this paper on fine-grained news classification, we nonetheless include them in our framework, as indeed it is useful to be able to connect fine-grained and coarse-grained news classification mechanisms. For example, we may want to associate the IPTC news code "politics" to a news item which focuses on a political figure or to specify that the topic of a news item that talks about a particular individual's approach to maintaining a healthy work-life balance is an example of a more generic "work-life balance" topic. The latter can be seen as a sub-topic of more generic topics, such as "lifestyle" or "wellness", which are covered in the IPTC taxonomy.

3.5. The commentary element: viewpoints.

A key aspect of the news universe is the *commentary* — i.e., the set of *viewpoints* expressed on a particular issue or topic, which are covered by the media. Indeed, the acid test for a democratic media landscape is related to *viewpoint diversity* [8], namely the extent by which media sources provide citizens with a robust range of alternative interpretations on a given issue. Viewpoint diversity in turn is closely related to *actor diversity*, in the sense that "the representation of a plurality of active actors in a news article seems to go hand in hand with a more diverse range of viewpoints" [48]. Accordingly, Masini et al. [49] show that the debate on immigration rarely includes the voices of the immigrants themselves and therefore this key element of the debate is heavily under-represented in news coverage. Hence, the ability to identify viewpoints in the news is essential in order to develop robust computational approaches to assessing whether individual media sources or a media landscape as a whole — e.g., the set of UK's mainstream media sources, fulfil their democratic role to inform readers about alternative views on a particular issue. In addition, in the context of a fine-grained news classification framework, it is essential to consider viewpoints for two main reasons: i) a viewpoint on a topic can itself be the main focus of a news article and ii) it is appropriate to extend the classification framework to include not just topics but also the topic-viewpoint dynamics, as a necessary precondition to enable automatic approaches to analysing viewpoint diversity in the media.

While fine-grained frameworks for analysing argumentation networks have been available in the scientific community for a long time [15], these are not necessarily appropriate to the context of analysing the dynamics of topics and viewpoints. In this scenario, the goal is less to try and capture the fine-grained distinctions between the different positions than to abstract from these to capture the main viewpoints associated with a topic and assess to what extent these are covered by media sources. As pointed out in [8], when analysing the news discourse, we are interested

[16] https://tinyurl.com/298nns2r.

in identifying viewpoints that "open up different perspectives" and "construct different meaning". Analogously, Masini et al. [49] cluster the variety of fine-grained positions on the topic of immigration around four main distinct viewpoints. For instance, they abstract a "victimisation" viewpoint out of a number of positions reported in the media, which characterise immigrants as victims. These individual positions may be articulated differently — e.g., immigrants may be victims because of racism, traffickers, or unjust government policies, but they all share an emphasis on the difficulties experienced by immigrants in different EU countries, which ought to elicit a sympathetic rather than hostile approach to the issue.

Consistently with these proposals, in this paper we characterise the notion of viewpoint as an abstraction of a number of fine-grained positions about a topic. Here, we adopt the same approach that we used to characterise the notion of "collection of events" and define a viewpoint as the result of clustering together a number of positions expressed in the media about an issue, which satisfy the same *unifying factor*. Such a unifying factor defines a viewpoint-specific criterion – see section 4.5 for details on how viewpoints are formally characterised in our framework.

Finally, we should also point out that viewpoints are expressed not only by people whose opinion is presented in the news — e.g., when journalists report on a politician who expresses their view on an issue, but also by journalists who write opinion pieces and even by the news sources themselves — e.g., in the traditional editorials published by many newspapers, which are attributed to the news outlet itself, rather than to a specific journalist.[17]

4. A formal model of news classification

In this section we provide a formal specification of the concepts introduced in the previous section, to provide a more robust characterization of our framework for fine-grained news classification. In particular, we model definitions as First Order Logic (FOL) statements, using a notation which mirrors standard representations for knowledge graphs, such as RDF[18]. Hence, we limit ourselves to binary relations and we also make use of the standard taxonomic relations, *type (?instance ?class)* and *subclassOf (?class1 ?class2)*. Question marks are used to represent logic variables. Unless otherwise stated, free variables are universally quantified. Atomic statements are reified by assigning an identifier to them and then using the relations *subject*, *object* and *predicate* to retrieve the components of the statement, following standard RDF notation. In terms of naming convention, classes are capitalised, individuals are expressed in lowercase and we usually prefix relations with a verb — e.g., *hasTopic.* Rare exceptions to this rule are made for i) relations that are borrowed from the RDF vocabulary — e.g., *subject*, *predicate*, *object* and *type*, ii) a few relations where the verb and the directionality of the relation are obvious and the verb can be omitted without creating an ambiguity — e.g., *TopicInNewsItem,* and iii) relation *topicRole,* which corresponds to the *T* operator, which will be introduced in Section 4.1.1. Because the *T* operator defines a bijective function, we prefer to use a function-like naming style for this relation.

Formally a news classification function takes as input a news item and generates the set of topics associated with it. Hence, given a set *N* of news items and a space *T* of topics, a news classification function *NC* can be specified as follows:

NC: N $\rightarrow$ *{t_1 .... t_n}, where each $t_i \in T$*

In what follows, we will formally define the space of news topics, following up from the informal characterization provided in section 3. A synoptic view of the main relations and classes in our formal model is shown in Figure 1. The solid lines, labelled with relation *topicRole*, connect the class *Topic* to the other classes in our model that can play the role of topic in a news item. To minimise cluttering, some classes have been duplicated in the figure. These are shown in italics.

4.1. Characterising entities, entity aspects and entity relations as news topics

4.1.1. Entities as news topics

We assume a knowledge base, *kb*, which contains a vast range of entities of different types, including people, organizations, countries, etc. For instance, we may have an entity, *jf_kennedy,* belonging to *kb,* of type *Person.* Let's also assume then that we have a news item, *ni12345*, which talks about this entity. Hence, we want to say that *jf_kennedy* is one of the (possibly multiple) news topics associated with *ni12345*. In order to do this, we introduce a relation, *hasTopic (NewsItem Topic),* and its inverse, *topicInNewsItem (Topic NewsItem)*, and we state:

s1: hasTopic (ni12345 T(jf_kennedy))

[17] See, e.g., http://tinyurl.com/yfs2974x.

[18] https://www.w3.org/RDF/.

The notation “*id: statement*” indicates both that *statement* is asserted in our knowledge base — i.e., the statement *hasTopic (ni12345 T(jf_kennedy)* in the example, and also that *id*, *s1* in the example, is the identifier reifying *statement*. Hence, the notation in the above example provides us with a concise way to add to our knowledge base both a domain-level statement instantiating relation *hasTopic* and also the following (meta-)statements[19]:

type (s1 Statement)

subject (s1 ni12345)

object (s1 T(jf_kennedy))

predicate (s1 hasTopic)

The operator *T* used in the definition associates an entity to its corresponding topic. *T* is needed to ensure that, for example, we correctly distinguish in our model the person *J.F.Kennedy* from the topic *T(J.F. Kennedy)*. This is essential to allow us to distinguish, for instance, the time span of the person, 1917-1963, from the time span of the topic, which is still very much in the news.

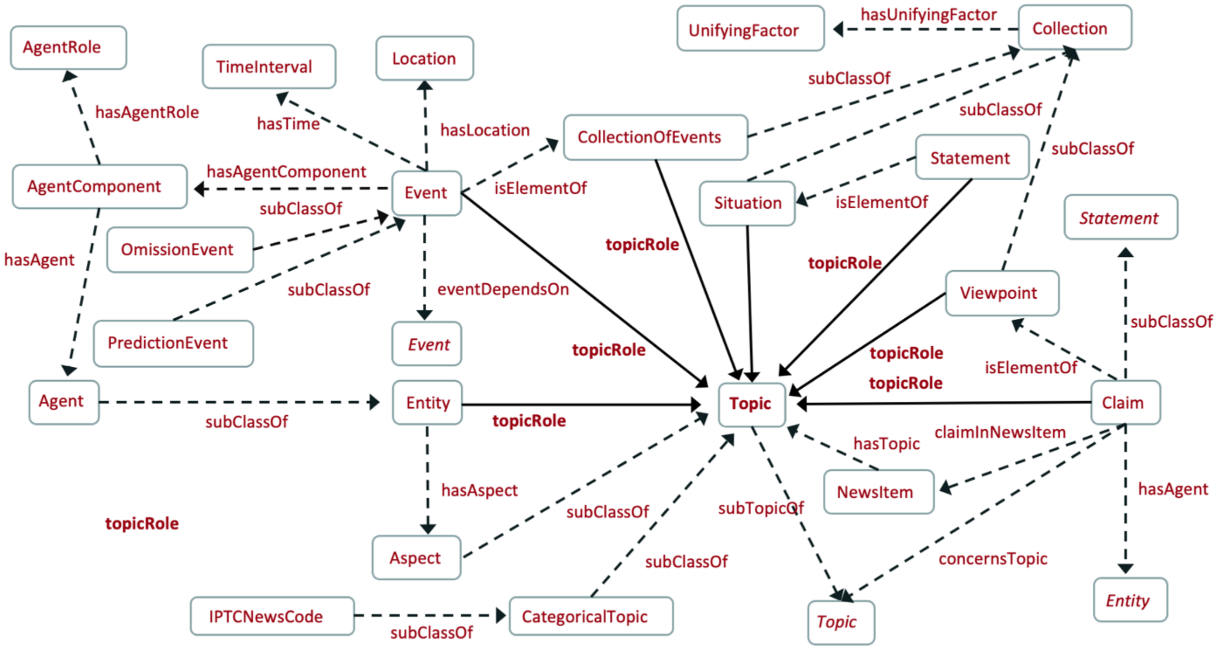


Fig. 1. Synoptic view of the main classes and relations in our formal model of news classification.

More in general, the *T* operator is needed every time an individual in our knowledge base, which is not ontologically a topic (e.g., a person, an event, a statement, etc.) plays the role of a topic in a logical expression. To this purpose, we include the following axioms:

T(?e) = ?t → type (?t Topic)

T(?e) = ?t ↔ topicRole (?e ?t)

sameAs (?e1 ?e2) ↔ sameAs (T(?e1) T(?e2))

The relation *topicRole* connects an entity in the knowledge base to the individual that represents the entity as a topic. In addition, we also specify that the mapping from an entity to the associated topic is 1-1.

4.1.2. Characterising entity aspects and relations as news topics

Let’s consider first the case in which a relation between entities (e.g., business links between a politician and a business person) is the focus of a news item. This situation can be handled as shown below, where we use a domain relation to express the business connection between the two people and then we take advantage of our reification mechanism.

s2: hasBusinessConnection
(politician1 businessperson1)

s3: hasTopic (ni55342 T(s2))

[19] To minimise verbosity, in what follows we will only reify statements that are themselves arguments of other (meta-)statements.

That is, the statement itself declaring the business connection between *politician1* and *businessperson1* becomes the focus of the news item. Again, the operator *T* is used in *s3* to ensure that there is no confusion between the individual *T(s2)*, which is an instance of class *Topic*, and the individual *s2*, which is an instance of class *Statement*.

Let's now consider the news item discussed in section 3.1, which concerns Donald Trump's financial status. This can be represented as follows:

hasAspect (donald_trump dt_financial_status)

hasTopic (ni37239 T(dt_financial_status))

subTopicOf
(T(dt_financial_status) T(financial_status))

The relation *hasAspect (Entity Aspect)* associates an entity to a relevant aspect. In addition, we also state that the specific topic, *T(dt_financial_status)*, discussed in *ni37239*, is a sub-topic of a more general *T(financial_status)* topic.[20] [21]

Finally, we add the following axioms to our model, to indicate that if an entity's aspect or a relation between entities is a topic in a news item, then both the entities in question and the predicate are also topics of the same news item. In particular, including the predicate as a topic makes it possible to easily retrieve all relation topics typed through the same predicate.

hasAspect (?e ?a) ∧ hasTopic (?ni T(?a)) → hasTopic (?ni T(?e))

hasTopic (?ni T(?st)) ∧ subject (?st ?s) ∧ predicate (?st ?p) ∧ object (?st ?o)
→ hasTopic (?ni T(?s)) ∧ hasTopic (?ni T(?p))
∧ hasTopic (?ni T(?o))

The implication of the above axioms is that aspects need to be entity-specific. That is, if we were to associate the entity *donald_trump* to a generic aspect *financial_status*, which may apply also to other entities, then we would infer incorrectly that the topic *T(donald_trump)* is the subject of a news item that may be discussing somebody's else financial status.

4.2. Events as news topics

Given the broad scope of our model, which is not restricted to a particular class of events, we want our characterization to be generic enough to cover all types of events. For this reason, we base our formalization on the one used by the *News Angle Ontology* [53], which is in turn based on the *Simple Event Model* [36]. Essentially, an event in this model is characterised in terms of time, location and the *agents* involved in the event. The relation between events and agents is mediated by *agent components*, which specify the roles played by the agents involved in the event. Multiple views are possible for the same event — e.g., an invading army can be seen as liberators or oppressors depending on whose viewpoint is being represented [36].

As discussed in Section 3.2.4, from a journalistic point of view, certain events only make sense in the context of other events. In our model we represent such dependencies by introducing the relation *dependsOn (?e_or_s1 ?e_or_s2),* which specifies that an event (or situation) depends on another event (or situation). In the context of events, we consider two types of dependencies, those brought in by *subEventOf* relations and those that are associated with non-hierarchical relations between events — e.g., causal dependencies between events. Accordingly, we add the following axioms to our model:

subEventOf (?ev1 ?ev2)
→ eventDependsOn (?ev1 ?ev2)

eventDependsOn (?e ?e_or_s)
→ dependsOn (?e ?e_or_s)

dependsOn (?e_or_s1 ?e_or_s2)
↔ preconditionFor (?e_or_s2 ?e_or_s1)

preconditionForEvent (?e_or_s ?e)
↔ eventDependsOn (?e ?e_or_s)

Needless to say, this rather minimalist characterization of event dependencies is only meant to highlight the important role played by event dependencies in journalistic scenarios, as discussed in Section 3.2.4, and to provide an initial set of relations. More

[20] The relation *subTopicOf* in our model corresponds to *skos:broaderTransitive* in the SKOS model (https://www.w3.org/2004/02/skos/).

[21] This is of course a bit of a shortcut, because the hierarchy of "financial status" topics would likely be much more elaborated, separating for instance the branch dealing with the financial status of individuals, *T(financial_status_of_individual)*, from that of organizations, *T(financial_status_of_organization)*, and countries, *T(financial_status_of_geopolitical_entity)*. Here, we are simply making the point that highly specific topics, such as *T(dt_financial_status)* would be classified as sub-topics of broader ones. Whenever possible, the latter should be retrieved from general-purpose topic taxonomies, such as *IPTC*.

comprehensive characterizations of event relations can be found in the literature, including the work by Mirza et al. [52], which focuses on causal relations and that by Rebboud et al. [64], which provides a rich set of event relations, including both causal relations and other types.

4.2.1. Collection of events

As discussed in section 3.2.2, a collection of events is a set of distinct events that are brought together by some *unifying factor.* To this purpose, we introduce a class *Collection* in our model, which corresponds to the class *dul:Collection* in the *DOLCE Ultralite*[22] ontology *(DUL)*. This characterises collections as "any container of entities that share one or more properties". Accordingly, we formally specify that membership of a *Collection* is predicated on meeting the conditions associated with the associated unifying factor [17]. The following axioms capture these notions:

type (?x Collection)
→ ∃ ?uf hasUnifyingFactor (?x ?uf)

type (?c Collection) ∧ hasUnifyingFactor (?c ?uf)
∧ isElementOf (?e ?c)
→ satisfiesUF (?e ?uf)

type (?c Collection) ∧ hasUnifyingFactor (?c ?uf)
∧ satisfiesUF (?e ?uf)
→ isElementOf (?e ?c)

The above definitions specify that all and only the elements that satisfy the relevant unifying factor are members of a collection.

We can now define the class *CollectionofEvents* simply as a collection whose members are events:

subclassOf (CollectionofEvents Collection)

type (?x CollectionofEvents) ∧
isElementOf (?ev ?x)
→ type (?ev Event)

As an example, we can now represent the collection of events associated with the multiple drone sightings in Norway in September and October 2022. In particular, we introduce an axiom specifying the criteria for including an event, *?e*, in the collection of events *drone_sightingsNorwaySeptOct2023*. The axiom specifies that *?e* has to be a *DroneSightingEvent*, that the location of the event has to be in Norway and that the event must have taken place in September or October 2022. Here we make use of the appropriate temporal relations from the *OWL Time Ontology*[23] and we also use the general-purpose *dul:hasLocation* predicate[24], which is provided by the *DUL* ontology.

type (drone_sightingsNorwaySeptOct2023 CollectionOfEvents)

hasUnifyingFactor
(drone_sightingsNorwaySeptOct2023 UFDroneSightingsNorwaySeptOct2023)

satisfiesUF (?x UFDroneSightingsNorwaySeptOct2023) ↔
type (?x DroneSightingEvent) ∧ dul:hasLocation (?x ?y) ∧ dul:hasLocation (?y Norway) ∧
time:hasTime (?x ?t) ∧ [time:intervalAfter (?t Aug2022) ∨ time:intervalBefore (?t Nov2022)]

4.2.2. Predictions

Predictions are events that are associated to another event or situation, which is predicted to happen in the future. Hence, we can simply model this class by defining the appropriate subclass of class *Event* and adding to it a slot pointing to the predicted event or situation.

subclassOf (PredictionEvent Event)

type (?x PredictionEvent)
→ ∃ ?ev_or_sit hasPrediction (?x ?ev_or_sit)

type (?ev1 PredictionEvent)
→ ∃ ?ev_or_sit
hasPrediction (?ev1 ?ev_or_sit)
∧ time:hasTime (?ev1 ?t1)
∧ time:hasTime (?ev_or_sit ?t2)
∧ time:intervalBefore (?t1 ?t2)

The above axiom states that each *PredictionEvent* is associated with a predicted event or situation and that the time associated with the *PredictionEvent* must be earlier than the time associated with the predicted event or situation.

4.2.3. Negative Events

Following [61], here we focus on *omissions* — i.e., negative events that are expressions of agency. Hence, we introduce a class, *OmissionEvent* and we state that each *OmissionEvent* is associated with another event (the omitted event) through the relation *hasOmittedEvent.*

subclassOf (OmissionEvent Event)

[22] The DOLCE Ultralite ontology (http://www.ontologydesignpatterns.org/ont/dul/DUL.owl – prefix = dul) is part of the family of DOLCE foundational ontologies [13].

[23] https://www.w3.org/TR/owl-time/.

[24] Of course, far more granular models of spatial representation can be found in the literature –see for instance the work by Chiatti et al. [19] and Cohn and Renz [20]. Nonetheless, the general-purpose *dul:hasLocation* predicate used here suffices for this simple example.

type (?ev1 OmissionEvent) → ∃ ?ev2 hasOmittedEvent (?ev1 ?ev2)

Payton also argues that if an agent omits to carry out an action that was supposed to be executed during a time interval, *t*, then such omission is also situated in the same time interval. We can therefore formalise this constraint by means of the following axiom:

hasOmittedEvent (?ev1 ?ev2) ∧ time:hasTime (?ev2 ?t)
→ time:hasTime(?ev1 ?t)

4.3. Situations as news topics

Situations can be represented in terms of a set of logical statements. For instance, let's assume a situation where a number of key executives are leaving a company, say *company1*. This is represented by the following three statements:

s4: quitsJob (executive1 company1)

s5: quitsJob (executive2 company1)

s6: quitsJob (executive3 company1)

We can then create an instance of class *Situation*, which aggregates these three statements:

type (situation1 Situation)

isElementOf (s4 situation1)

isElementOf (s5 situation1)

isElementOf (s6 situation1)

That is, *situation1* is defined as the collection of the three statements representing the departure of the company executives. More in general, we can define the class *Situation* as a *Collection*, whose elements are all instances of class *Statement*:

subclassOf (Situation Collection)

type (?x Situation) ∧ isElementOf (?s ?x)
→ type (?s Statement)

As pointed out by Gangemi and Mika [29], situations are collection of statements that are subject to an interpretation process. For instance, in this case an external observer may infer that *company1* is in a bleak situation, because several key people have left. Hence, analogously to the model proposed by Gangemi and Mika, we introduce the notion of *Description* to support such interpretation process and we introduce a relation, *characterises*[25], which connects a description to the relevant situation. In particular, we can represent our example as follows:

s7: hasFutureProspects (company1 bleak)

type (d1 Description)

hasElement (d1 s7)

characterises (d1 situation1)

Finally, analogously to our characterization of class *Situation*, we also represent class *Description* as a collection of statements:

subclassOf (Description Collection)

type (?x Description) ∧ isElementOf (?s ?x)
→ type (?s Statement)

4.4. Categorical topics

Let's consider a news item, *ni34265*, which discusses poverty in Italy. Consistently with the discussion in section 3.1, we can model this scenario as a case where a particular *aspect* (poverty) of a particular *entity* (Italy) is the topic of the news item in question. In addition, as discussed in section 3.4, we consider "poverty" to be a categorical topic and indeed this category is included in the IPTC taxonomy. Hence, analogously to the way we modelled entity aspects in section 4.1.2, we can represent the fact that a news item discusses poverty in Italy as follows:

hasTopic (ni34265 T(poverty_in_italy))

hasAspect (italy poverty_in_italy)

subTopicOf (T(poverty_in_italy) IPTC:poverty)

Highly generic categories in the *IPTC* taxonomy, such as "politics" play a similar role to the one played by highly generic fields of research in scholarly taxonomies. For instance, the topic "Computer Science" provides the most generic category in the CSO Ontology [68] and essentially defines its scope — i.e., the CSO Ontology maps the space of research topics in Computer Science and does not cover other fields of study. Hence, as research papers are unlikely to focus on a highly generic topic such as "Computer Science" (they would normally focus on a far more granular sub-topic), analogously it is unlikely that news items would focus on a generic category such as "politics" and they would normally focus instead on a specific political event or issue. Here, we want to connect the specific topics that are the focus of news items to the more generic 'topic containers' in our model. In the

[25] Gangemi and Mika introduce a relation named *satisfies*, to associate situations to descriptions. Here we use a different relation, *characterises*, to focus instead on the *interpretations* introduced by descriptions.

above example we showed how to do this when modelling entity aspects but of course we would like to do this for all relevant concepts. As an example, here we show how we can achieve this for classes of entities, such as *Politician,* or event types, such as *PoliticalEvent,* by connecting these to the relevant categorical topic, in this case *IPTC:politics.*

hasTopic (?ni T(?ev)) ∧ type (?ev PoliticalEvent)
→ hasTopic (?ni IPTC:politics)

hasTopic (?ni T(?e)) ∧ type (?e Politician)
→ hasTopic (?ni IPTC:politics)

For the sake of scalability these types of connections need to be derived through automated mechanisms and indeed, in [22], the authors provide an initial approach to automatically map IPTC codes to event data, thus starting to bridge the gap between coarse-grained news classification and event extraction engines. However, more work is needed to provide more comprehensive solutions that integrate generic taxonomies, such as, IPTC, to the variety of relevant specific topics.

4.5. Modelling Viewpoints

4.5.1. Viewpoints as coherent collections of claims

As pointed out in section 3.5, when modelling viewpoints in the news we can distinguish between a micro level characterised by individual claims and positions about a topic and a macro level where we aggregate individual claims into meaningful alternative perspectives on an issue, consistently with standard practice in media analytics [8][49].

Following the approach by Buckingham Shum et al. [15], a claim can be characterised as a *statement* expressed by an *agent* with some *justification.* We also associate a claim with the *news item* where the claim has been stated. In practice, a justification is an optional component, as claims in the media are not necessarily articulated with supporting evidence, in contrast with common practice in the scientific literature. Hence, we can characterise claims as follows:

subclassOf (Claim Statement)

type (?c Claim) → ∃ ?c hasAgent (?c ?a)

type (?c Claim) → ∃ ?c concernsTopic (?c ?t)

type (?c Claim) → ∃ ?n claimInNewsItem (?c ?n)

type (?c Claim) ∧ hasJustification (?C ?j) → type (?j Justification)

That is, a claim is a statement made by an agent in a news items, which concerns a topic. A claim can also have an optional justification. As an example, let's consider the case in which a scientist states that the Pfizer vaccine is effective against the delta variant of the covid-19 virus.[26] We can then state:

s5: isEffectiveAgainst (pfizer_vaccine covid19_delta_variant)

hasAgent (s5 scientist1)

concernsTopic (s5 T(covid19_vaccination))

hasJustification (s5 pfizer_trials_db)

claimInNewsItem (s5 ni_reuters_pfizer_vaccine_240621)

Let's assume then that we have several statements in the news supporting the efficacy of the Pfizer vaccine against the covid-19 virus. We can then aggregate these statements into a particular viewpoint as follows.

We first define the class *Viewpoint*:

subclassOf (Viewpoint Collection)

type (?v Viewpoint) ∧ isElementOf (?s ?v)
→ type (?s Claim)

We can then define a viewpoint that is in favour of the Pfizer vaccine by aggregating all statements that claim its effectiveness against covid, abstracting from the specific agent, covid variant, and possible justification[27].

type (pro_Pfizer_vaccine_viewpoint Viewpoint)

hasUnifyingFactor
(pro_Pfizer_vaccine_viewpoint pro_Pfizer_vaccine_viewpoint_uf)

satisfiesUF (?x pro_Pfizer_vaccine_viewpoint_uf)
↔
type (?x Claim) ∧
predicate (?x isEffectiveAgainst) ∧
subject (?x pfizer_vaccine) ∧ object (?x ?y) ∧

[26] http://tinyurl.com/bdzf99sm.

[27] The representation of this example can be seen as a bit simplistic, as it assumes that all statements supporting the effectiveness of the Pfizer vaccine against covid are expressed using the same vocabulary. In reality, it is OK to make this assumption: while individual claims would be expressed in news items using different terminologies and argumentations, they may nevertheless be mapped by an ontology-based relation extraction engine to a standardised representation, e.g., using the *isEffectiveAgainst* relation.

type (?y Covid19_Variant) ∧
concernsTopic (?x T(covid19_vaccination))

4.5.2. Viewpoints as news topics

The relations *hasClaim* and *hasViewpoint* connect a news item to a claim or viewpoint that is mentioned in the news item in question. The following axiom states that if a news item includes a claim, then it also includes any viewpoint the claim belongs to:

hasClaim (?ni ?c) ∧ isElementOf (?c ?v) ∧ type (?v Viewpoint) → hasViewpoint (?ni ?v)

Analogously, if a claim is a topic of a news item, then any associated viewpoint will also be a topic of the news item, as stated by the following axiom:

hasTopic (?ni T(?c)) ∧ type (?c Claim) ∧ isElementOf (?c ?v)
∧ type (?v Viewpoint) → hasTopic (?ni T(?v))

Hence, if we state that the claim, *s5*, about the effectiveness of the Pfizer vaccine, is the topic of the news item *ni_reuters_pfizer_vaccine_240621,* then we can also derive that the pro-Pfizer-vaccine viewpoint is also a topic of the same news item – see the two statements below.

hasTopic (ni_reuters_pfizer_vaccine_240621 T(s5))

hasTopic (ni_reuters_pfizer_vaccine_240621 T(pro_Pfizer_vaccine_viewpoint))

5. The News Classification Ontology

The model discussed in the previous section has been realised as an *OWL vocabulary* [21]. The resulting *News Classification Ontology (NCO)* follows *Linked Data principles* [12] and uses the namespace *http://data.open.ac.uk/ontology/newsclassification#*, which also provides the Web address of the ontology document. Because the structure of NCO follows closely the model presented in Section 4, there is no need here to provide a complete description of the ontology and we will instead focus the discussion on the key technical design elements relevant to the realization of the formal model in an OWL ontology.

In addition to the *nco* namespace, in the course of the discussion we will also refer to the namespace *http://data.open.ac.uk/ontology/ncoexamples#,* with prefix *nco_ex*, which provides a suite of test cases to validate the *NCO* ontology.

5.1. Relation to other ontologies

NCO imports both the *OWL Time Ontology* and the *SKOS model*[28]. The former is needed to allow us to characterise time-indexed entities, such as events, while the latter provides the foundation for representing topics. In particular, we characterise i) the class *nco:Topic* as a specialization of *skos:Concept* and ii) the properties *nco:hasSubTopic* and *nco:subTopicOf* as specializations of *skos:narrowerTransitive* and *skos:broaderTransitive.* We also reuse the representation of events which is provided by the *News Angle Ontology,* as discussed in Section 4.2.

In addition, we also provide two extended versions of the *NCO* ontology, which align with other relevant ontologies. The first one is *NCO-IPTC*, accessible at *http://data.open.ac.uk/ontology/nco-iptc#*. This version integrates the full *IPTC* taxonomy into *NCO*. It is provided as a distinct ontology from *NCO*, as we recognise that not all users of *NCO* may wish to import the rather large *IPTC* taxonomy. The second one, *News2D0*, provides a full alignment with the *DOLCE-Zero* ontology[29] and can be accessed at *http://www.ontologydesignpatterns.org/ont/news/news2d0.owl*.

NCO-IPTC classifies each IPTC news code as an instance of class *nco:CategoricalTopic* and models the hierarchy of news codes by means of the appropriate *nco:subTopicOf* property assertions. No other extensions to *NCO* are realised in this ontology, hence it is not necessary to describe it further in this paper. *News2D0* is instead described in section 5.5.

In what follows we discuss the design of *NCO*.

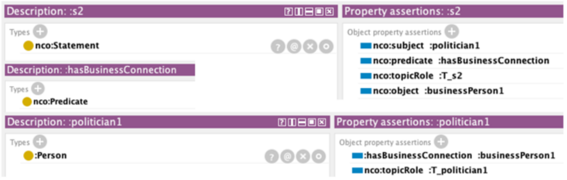
5.2. OWL representation of the main concepts in NCO

A key requirement from the specification presented earlier is that the ontology needs to support statement reification, both to enable the correct modelling of claims, situations and descriptions, and also to make it possible to characterise relations between individuals as topics, in accordance with the example provided in Section 4.1.2. To this purpose *NCO* includes a class *nco:Statement*, whose instances are reifications of statements included in the ontology. The object properties *nco:object*, *nco:predicate* and *nco:subject* are provided to connect an instance of class *nco:Statement* to the elements of the relevant triple in the knowledge base. Here we take advantage of the punning capability provided by OWL 2 [21] and, in particular, we define an individual of type *nco:Predicate* for each

[28] http://www.w3.org/2004/02/skos/core#.

[29] http://www.ontologydesignpatterns.org/ont/d0.owl (prefix = d0).

predicate included in a triple that has been reified. As an example, in Figure 2 we represent the case discussed in Section 4.1.2, where the business connection between a politician and a business person is itself a news topic. That is, the triple <*nco_ex:p1 nco_ex:hasBConn nco_ex:b1*> is added to our knowledge base and we use the punning feature to define an individual, *nco_ex:hasBConn,* of type *nco:Predicate*. Having done this, we can then reify the triple by defining an instance of class *nco:Statement* and adding the relevant property assertions, as shown in Figure 2. The figure also shows examples of the use of property *nco:topicRole*. As discussed in Section 4.1.1, this property is needed to characterise correctly the role an entity plays when it becomes a news topic, in particular by distinguishing an entity from its associated topic. The property *nco:topicRole* is defined as both *functional* and *inverse functional* in the *NCO* ontology.


Description: :s2
Types
nco:Statement
Description: :hasBusinessConnection
Types
nco:Predicate
Description: :politician1
Types
:Person
Property assertions: :s2
Object property assertions
Property assertions: :politician1
Object property assertions

nco:subject :politician1
nco:predicate :hasBusinessConnection
nco:topicRole :T_s2
nco:object :businessPerson1

:hasBusinessConnection :businessPerson1
nco:topicRole :T_politician1

Fig. 2. Coverage with respect to the different categories in the framework.

The class *nco:Statement* is also needed to model correctly claims, situations and descriptions. In particular, claims are represented in the ontology as instances of class *nco:Claim*, which in turn is a subclass of *nco:Statement*, while descriptions and situations are represented as collections of statements. To this purpose, *NCO* introduces class *nco:Collection,* whose subclasses include *nco:CollectionOfEvents*, *nco:Situation*, *nco:Description* and *nco:Viewpoint.*

Finally, aspects are represented as instances of class *nco:Aspect*, while the object property *nco:hasAspect*, with domain *nco:Entity* and range *nco:Aspect*, connects entities to their associated aspects.

5.3. Property chains in NCO

In *NCO* we make extensive use of *OWL property chains* to represent several axioms that are included in the model. These are presented in what follows.

5.3.1. Characterising membership of a collection

As discussed in Section 4.2.1, an entity can be a member of a collection if and only if it satisfies the relevant unifying factor. This constraint is represented in *NCO* through the following property chains:

ObjectProperty: nco:hasElement
  SubPropertyChain: nco:hasUnifyingFactor ***o*** *nco:ufSatisfiedBy*

ObjectProperty: nco: satisfiesUF
  SubPropertyChain: nco:isElementOf ***o*** *nco:hasUnifyingFactor*

In the above definitions, the property *nco:isElementOf* is the inverse of *nco:hasElement*, while *nco:ufSatisfiedBy* is the inverse of *nco:satisfiesUF*.

5.3.2. Using property chains to ensure correct propagation of nco:hasTopic rules

As discussed in Section 4.1.2, if an entity aspect or a relation between entities are the topics of a news item, then also the entity associated with the aspect and the constituent entities of the relation are topics of the news item in question. These requirements are captured through the following property chains:

ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic ***o*** *nco:topicRoleOf* ***o*** *nco:isAspectOf* ***o*** *nco:topicRole*

ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic ***o*** *nco:topicRoleOf* ***o*** *nco:subject* ***o*** *nco:topicRole*

ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic ***o*** *nco:topicRoleOf* ***o*** *nco:predicate* ***o*** *nco:topicRole*

ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic ***o*** *nco:topicRoleOf* ***o*** *nco:object* ***o*** *nco:topicRole*

In the above definitions we use the predicate *nco:hasTopic* to express the relation between an instance of class *nco:NewsItem* and a topic relevant to the news item. We also use the properties *nco:topicRoleOf* and *nco:isAspectOf*, which are respectively the inverse of *nco:topicRole* and *nco:hasAspect*.

In addition, if a news item, say *ni*, has been associated to a topic, say *T1*, and *T1* is a *nco:subTopicOf T2*, then we also want to associate *ni* to *T2*. To this purpose, the property chain below is also included in our model:

ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic ***o*** *nco:subTopicOf*

Finally, we also want to enforce the axiom discussed in Section 4.5.2, which specifies that if a claim, say *c*, is the topic of a news item and *c* belongs to the cluster of claims defining a viewpoint, say *v*, then *v* is also a topic for the news item in question. This axiom is represented as follows:

```
ObjectProperty: nco:hasTopic
  SubPropertyChain: nco:hasTopic o nco:topicRoleOf o nco:isClaimOf o nco:topicRole
```

Here we make use of the property *nco:isClaimOf*, which connects a claim to a viewpoint that includes it.

5.3.3. Other property chains included in the NCO Ontology

Another property chain models the situation where a claim, say *c*, appears in a news item (without necessarily being a topic of the news item) and therefore we also want to associate the news item in question to any viewpoint to which *c* belongs. This is shown below:

```
ObjectProperty: nco:hasViewpoint
  SubPropertyChain: nco:hasClaim o nco:isClaimOf
```

Finally, the *NCO* ontology also includes two property chains enforcing the time constraints associated with classes *nco:PredictionEvent* (see Section 4.2.2) and *nco:OmissionEvent* (see Section 4.2.3). In particular, we require the time of a *nco:PredictionEvent* to be *time:intervalBefore* that of the associated predicted event or situation and we also state that the time associated with an omitted event is the same as that of the relevant *nco:OmissionEvent*. These two constraints[30] are defined as follows:

```
ObjectProperty: time:intervalBefore
  SubPropertyChain: nco:isTimeOf o nco:hasPrediction o nco:hasTime
ObjectProperty: nco:hasTime
  SubPropertyChain: nco:isOmittedEventOf o nco:hasTime
```

The above definitions use the property *nco:hasTime* (whose inverse is *nco:isTimeOf*), which specialises *time:hasTime* by connecting instances of classes *nco:Event* or *nco:Situation* to a *time:TemporalEntity*. Finally, *nco:isOmittedEventOf* is the inverse property of *nco:hasOmittedEvent*.

5.4. Ontology Evaluation

The logical and structural consistency of the *NCO* ontology was checked by means of the HermiT 1.4.3.456 reasoner running in Protégé 5.6.3. The reasoner classifies all classes and object properties without reporting any errors. In total the *NCO* ontology includes 50 classes and 1622 axioms. In addition, we also tested the ontology against the set of formal requirements expressed in the formal specification provided in Section 4. In particular, these include (but are not limited to):

- Correct propagation of topic classification assignment from aspects to associated entities.
- Correct propagation of topic classification assignment from reified statements to the components of the relevant triple, in accordance with the requirement specified in Section 4.1.2.
- Correct propagation of topic classification assignment through topic hierarchy.
- Correct propagation of topic classification assignment from claims to relevant viewpoints.
- Correct modelling of the axioms characterising collections, as for specification in Section 4.2.1.
- Correct enforcement of constraints about the time indexing of prediction and omission events.
- Correct realization of meta-modelling machinery supporting the representation of reified statements, claims, situations, descriptions and viewpoints.
- Correct importing of IPTC news codes.

To this purpose we defined a suite of test cases, comprising the *ncoex* OWL knowledge base, which allowed us to check that the aforementioned requirements were correctly realised. In addition, we also carried out additional checks to ensure that, for instance, the ontology correctly supports generic queries about the metamodelling framework. Hence, we defined SPARQL queries able both to retrieve all triples in the *ncoex* knowledge base that had been reified and also the ones which hadn't. For instance, through these queries we are able to check that the knowledge base provides a correct and complete representation, at both object and meta level, of all triples expressing relations between entities, which are themselves news topics, regardless of whether they have been asserted or inferred through an OWL reasoner. More broadly, through this set of test queries, we also checked that the *NCO* ontology correctly supports queries that encompass both domain triples and their reified representation. As an example, we provide below the SPARQL representation of a query that retrieves all

[30] Arguably, property chains work better as inferential mechanisms rather than constraint-checking ones. However, by expressing all axioms in OWL (rather than through other formalisms, such as rules or constraint languages) we maximise the reusability and portability of the ontology.

triples in *ncoex* that are both statements in the knowledge base and have also been reified.

```
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX owl: http://www.w3.org/2002/07/owl#
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
PREFIX nco: http://data.open.ac.uk/ontology/newsclassification#
PREFIX ncoex: http://data.open.ac.uk/ontology/ncoexamples#
SELECT ?st ?sub ?p ?obj
WHERE {
 ?p a owl:ObjectProperty .
 ?sub ?p ?obj .
 ?st a nco:Statement .
 ?st nco:subject ?sub .
 ?st nco:predicate ?p .
 ?st nco:object ?obj
 }
```

Finally, we evaluated the ontology's compliance with the FAIR principles for scientific data management [82], using the FAIR-Checker validator [33]. This identified a few minor shortcomings that were addressed in a revised version of the ontology.

5.5 Aligning NCO with DOLCE

As mentioned earlier, we have also produced a separate version of the *NCO* ontology that imports and is fully aligned with the *DOLCE-Zero* foundational ontology (*D0*). *D0* is built on top of the *DOLCE Ultralite* ontology (*DUL*) and is designed to deal effectively with the systematic polysemy of many lexical items, whose multiple senses may create problems when used as OWL classes. To this purpose, it provides a more relaxed semantics for a number of key definitions — e.g., by allowing the modelling of lexical items that can carry a sense of physical or abstract location, event or event type, etc.

In what follows we provide a brief outline of the way the top classes and properties of *NCO* have been aligned to the relevant entities in *D0*[31].

First of all, a number of classes in *NCO* are equivalent[32] to or direct subclasses of homonymous classes in *D0*. These include *nco:Entity*, *nco:Situation*, *nco:Agent*, *nco:AgentRole* (subclass of *dul:Role*), *nco:Location*, *nco:Collection*, *nco:Description*. For instance, *nco:Collection* is defined as equivalent to *dul:Collection* and therefore its subclasses, including *nco:Viewpoint*, *nco: Description* and *nco:Situation* all become subclasses of *dul:Collection*.

Other classes are instead interpreted according to the specific semantics they bear in *NCO*. For instance, *nco:NewsItem* is a subclass of *dul:InformationEntity*, while *nco:Statement* and its subclass *nco:Claim* are subclasses of *dul:Situation.* Here we consider a statement as denoting a situation, rather than considering it as an information item. The class *nco:AgentComponent* is also a subclass of *dul:Situation*, since it reifies the n-ary relation between an agent, its role in a context, time, etc. The class *nco:UnifyingFactor* is instead a subclass of *dul:Description*, following the collection semantics of *DUL*.

A special case is *nco:Aspect*, which can be any entity, and is aligned as a subclass of *dul:Entity*.

The alignment of properties follows from the class alignment. For instance, *nco:characterises* is a sub property of *dul:isSatisfiedBy*; *nco:concernsTopic* is a sub property of *d0:hasFocus*; *nco:dependsOn* is interpretable as a sub property of *dul:isPreconditionOf*; *nco:hasClaim* is semiotically interpretable as a sub property of *dul:expresses*; *nco:hasElement* is equivalent[33] to *dul:hasMember*; *nco:hasJustification* is a sub property of *dul:hasInScope,* as it associates a claim situation to its justifying situation.

The result of the alignment can be seen as providing a different semantics for *NCO*, in terms of the foundational entities defined by *D0*. In addition, the alignment has been validated by showing that the resulting ontology, *News2D0*, remains coherent and its reasoning capabilities can be safely applied to news annotation.

6. Empirical Validation of the Framework

An initial validation of the framework was carried out by manually classifying a corpus of 224 news articles. These were retrieved from two news outlets, Aftenposten, a Norwegian newspaper, and The Guardian, a British newspaper. The articles were collected

[31] Needless to say, the term "entities in *D0*" refers both to entities native to the *D0* ontology (i.e., with prefix *d0*) and also to entities in *DUL*, which are of course also included in *D0*.

[32] Two classes, say *A* and *B*, are equivalent in OWL if they have the same extension — i.e., *A* is a subclass of *B* and *B* is a subclass of *A*.

[33] Two properties, say *p* and *q*, are equivalent in OWL if they have the same extension — i.e., *p* is a sub property of *q* and *q* is a sub property of *p*.

by visiting news outlets' websites on different days and collecting links for all the stories published on that particular day. This was done to maximise diversity in the corpus, under the assumption that the news on a particular day tend to be dominated by events that have occurred in the previous 24 hours. Because of an imbalance in the number of articles published each day in the two news outlets, the collection of Aftenposten news articles required more days than The Guardian, as the aim was to produce a reasonably balanced corpus. In total, 100 news items were collected from the Aftenposten and 124 from The Guardian.

Table 1

Coverage with respect to the different categories in the framework.

	AftenPosten	**The Guardian**
#Newsitems	**100**	**124**
Entity	5	13
Entity Aspect	2	9
Relation between Entities	1	4
Individual Event	97	119
Collection of Events	0	2
Negative Event	4	4
Dependent Event	36	48
Prediction	9	9
Situation	7	12
Viewpoint #newsitems covering viewpoints	37	50
Viewpoint #viewpoints expressed in corpus	39	67

As shown in Table 1, all categories in the framework were represented in the news corpus, hence providing an initial confirmation that the framework appears to cover all the topics that news items focus on. The only category that was sparsely represented was "Collection of Events", which did not appear in any AftenPosten story and only appeared in a couple of Guardian stories, even though we have additional evidence of the value of this category from other contexts — e.g., see sample story mentioned in section 2.2.2.

The annotated corpus is publicly available at https://bit.ly/newscorpus2023.

7. Related Work

To our knowledge, there has not been any attempt in the literature at mapping out a comprehensive model of what types of concepts provide the main subject matter for news items. Indeed, the vast majority of relevant computational research has focused on developing formal representations and information extraction methods for specific classes of relevant concepts, such as named entities and events, without necessarily addressing the broader picture.

Research in media science has instead focused on notions such as *news values* [38], *news angles* [53], and *news frames* [24], which help to characterise the types of stories that tend to be newsworthy (e.g., stories about celebrities) and the way journalists frame them. Hence, these analyses are somehow orthogonal to the work presented in this paper, which focuses not on the style of communication in the news domain, but on characterising the generic types of concepts that provide the subject matter for news items. As already mentioned, the exception here is the notion of *agenda setting*, which is also concerned with the salience of issues in the media. However, this notion primarily focuses on the impact of this topic selection process on the public, rather than on the epistemology of news topics.

7.1. Entities, Events and Situations

As discussed in Section 3.1, highly performant methods for *Named Entity Recognition* and *Entity Linking* are already in routine commercial use, even though this is still a very active area of research — e.g., see recent zero-shot approaches based on neural architectures [84]. In addition, methods for relation extraction [47][81][31] are also available. Here, a very promising approach entails the adoption of large-scale language models, based on the transformer architecture, such as GPT-4 [58], LaMDA [75], and LLaMA 2 [76], among others. These models have demonstrated efficacy in extracting entities and relationships to generate knowledge graphs from textual corpora [62]. This is typically achieved either through task-specific fine-tuning or by employing a few-shot learning approach [51]. However more research in this area is needed to reach the level of performance and usability required for effective fine-grained news classification, in particular with respect to relation extraction.

Researchers in open domain event extraction have in recent years taken advantage of large-scale semantic resources, such as FrameNet [9] and Wikidata events [67], which provide generically applicable schemas that can support the event extraction process. In particular, Huang et al. [41] have developed a state-of-the-art technique, showing that it is possible to take advantage of the semantic structure of known events to learn the extraction of new event schemas, using a zero-shot transfer learning approach. More recently, Fincke et al. [28] have shown that by reframing event extraction as a question-answering task and by "priming" a language model depending on the question being asked, they were able to improve the performance of an event extraction module in a zero-shot cross-lingual setting.

These improvements in event extraction have gone hand in hand with the development of formal models for event representation. The *Simple Event Model* provides a foundational ontology for events, which is independent of any particular domain and is "designed with a minimum of semantic commitment" [36]. Thanks to its simplicity and flexibility, this model has been very successful, providing the basis for a variety of large-scale event extraction initiatives in the news domain, such as (among others) the NewsReader project [65] and EventKG [34], a large-scale knowledge base that includes about 700K events and over 2.3 million temporal relations. The EventKG model extends SEM by supporting the specification of temporal relations between entities and between entities and events, and also by providing mechanisms to state the provenance of event information — e.g., by linking an event to the source from which the event has been extracted. EventKG also provides the foundation for a more recent large-scale event knowledge base, the Open Event Knowledge Graph (OEKG), which augments EventKG with a variety of other datasets [35]. Another initiative developing a comprehensive large-scale ontology for events is the Rich Event Ontology [83], which builds on DOLCE [13], integrates a variety of semantic resources, including FrameNet [9] and VerbNet [70], and provides thousands of event classes.

However, in the context of the framework proposed in this paper, it is important to emphasise that current event ontologies tend to focus on events as "things that have happened", while very little attention has been given to negative events, intended as occurrences that have not happened as a result of an agent's deliberate decision of not performing an action. Hence, more work is needed to improve our ability to identify this type of events in the news domain and other contexts. Analogously, while there is much work in the literature on formal representations of situations — e.g., see the work by Gangemi and Mika cited earlier [29], the information extraction field has not traditionally considered situations as a separate epistemological entity from events and therefore research in this area is lacking. The only exception is the work on situational awareness in domains such as smart cities [55] and cybersecurity [56], where situations however tend to be characterised in a domain-dependent way — e.g., as a set of relevant data points in a smart city system.

7.2. Categorical topics

Approaches to classifying news in terms of generic categories, such as the ones provided by the IPTC news codes, have been available for several years [7] and indeed commercial services, such as Quantexa News Intelligence, already classify news items automatically in terms of the relevant IPTC categories. However, these taxonomies are manually generated and therefore evolve rather slowly. Hence, there is a need for accurate computational solutions, which can speed up the evolution process and ensure that these taxonomies are able to keep up with the variety of generic topics that regularly emerge in the media. This type of algorithms are now available to support the automatic evolution of taxonomies of research areas [59] and in principle could provide the basis for analogous solutions for automatically generating comprehensive taxonomies of media topics.

Another issue we have already mentioned concerns the need to integrate coarse-grained and fine-grained classification mechanisms, taking as starting point the work by De Clercq et al. [22], which associates IPTC codes to event data.

7.3. Viewpoints

As discussed in section 3.5 and 4.5, we consider viewpoints as positions expressed in the media which open up different perspectives on an issue. This leads to the formal definition expressed in section 4.5, where a viewpoint is characterised as a collection of claims that subscribe to the same position — i.e., a set of claims that do not "construct different meanings", according to the theoretical framework proposed by Baden and Springer [8].

Argumentation frameworks for characterizing networks of claims have been available for a long time [77][34] and have formed the basis for a number of formal representations for modelling arguments [15][63]. Compared to the extensive set of relations defined in the framework of Rhetorical Structure Theory [45], both these formal models and argument-mining tools [16] tend to focus on a small set of key relations, such

[34] The first edition of the influential book by Stephen Toulmin cited here was published in 1958.

as those that link a *claim* to its *premise* and the *attacks/supports* relations between claims. However, while in principle *supports* relations between claims can be used to identify congruent claims that belong to the same viewpoint, to our knowledge our recent experiment on capturing the viewpoint dynamics in the news [54] provides the only example in the literature that, consistently with the framework presented in this paper, connects the notion of viewpoint expressed in the media literature to a concrete computational approach, able to identify the viewpoints relevant to a specific topic and characterise them with respect to a set of congruent claims.

In addition, research on argument mining tend to focus on claim and relation identification in a rather context-independent way, while the news domain is characterised by a degree of redundancy, where multiple news sources often discuss the same topic at the same time, expressing converging or diverging viewpoints. This feature of the news domain is exploited in the work by Park et al. [60], who observe that initial news items about an event or issue tend to be similar, while later articles from different sources are more likely to introduce diverse viewpoints. They also take advantage of the structure of a news item, giving more weight to the *head* of the article in question. However, despite introducing these interesting heuristics and realising the approach into a concrete news browser, *NewsCube*, their approach is keyword-based and therefore prone to noise. Vilares and He [79] go beyond the solution proposed by Park et al., by adopting an unsupervised LDA-based approach that attempts to jointly identify topics and viewpoints. They also generate readable summaries of the main viewpoints, by identifying sentences associated with the most discriminative words in the relevant topic-viewpoint model. However, as with the approach by Park et al., they also use a rather syntactic (i.e., keyword-based) approach to modelling and moreover the quality of the generated summaries tends to vary significantly, often highlighting sentences that do not necessarily express a viewpoint, in the sense of providing a *contrastive opinion.* The approach by Trabelsi and Zaïane [78] exhibits a performance improvement, by taking advantage of the dialectic structure of posts in a forum. They also use effective heuristics for summarizing viewpoints, such as focusing on *verbal expressions* and choosing expressive summaries out of a clustering process of candidate phrases. However, their approach capitalises on interactions between different post creators on social media and therefore is not directly applicable to our news scenario. Indeed, as pointed out by Doan and Gulla [25] in the context of identifying political viewpoints, “the current state of the art falls somewhat short of our goal with automatic political viewpoint identification” [25]. We believe that the same remark can be made about the state of the art concerning viewpoint identification in the news domain. Our aforementioned work on capturing the viewpoint dynamics in the news [54] provides an initial step towards tackling the challenge of developing effective solutions for this task.

Another research area that is relevant here is *stance detection* [2], which focuses on identifying the attitude (stance) expressed by an *agent* towards a *target*. While stance detection has originally focused on rather restricted scenarios (e.g., identifying positive and negative reviews for a product), more recent work is tackling scenarios that are closer to the one described in this paper – in particular, by considering claims expressed in a news item or social media posts as targets for a stance detection method and including both a topic classifier and a *Topic-Guided Stance Detection* module in the architecture [4]. However, more work is needed to customise and extend these techniques to support effective viewpoint identification in the news domain.

7.4. Related work in ontology engineering

A variety of ontologies in the literature cover the notions of events and situations, including both upper-level ontologies [13] and also more specific proposals that focus on these concepts [6][36][69][83]. As already pointed out, our characterization of events is based on our earlier work on the *News Angle Ontology* [53], which describes events in terms of agents, location, and time. As far as situations are concerned, here we subscribe to the design proposed by Gangemi and Mika [29], which distinguishes between descriptions and situations and is compatible with a representation of these concepts as collections of statements. The paper by Gangemi and Presutti [30] is relevant to our formalization of viewpoints, as it models a *perspective* as a cognitive device that makes it possible to impose multiple *lenses* on an event or situation by taking a particular *cut* over the event or situation in question. In contrast with our characterization of viewpoints as collections of semantically congruent claims, this definition focuses on characterising the narrative-centric process of constructing a perspective. The notion of viewpoint is also implicitly tackled in argumentation ontologies [15][74], which model *positions* or *claims* concerning an issue and then provide relations to state which positions/claims are in agreement or disagreement. In contrast with these approaches, we identify the group of claims that constitute a viewpoint through a *unifying factor* associated with a viewpoint, rather than by stating agree/disagree relations between claims. Our approach, which follows the model proposed by Carriero et al. [17], has the advantage of making explicit the criterion associated with a viewpoint. In addition, it can also be easily integrated with representations where agree/disagree relations are asserted between claims.

As far as the news domain is concerned, ontologies for annotating news content have been developed by major media organizations, such as the BBC. In particular, the *BBC Storyline Ontology*[35] centers on the notion of *storyline*, which groups together the various elements of a journalistic narrative. Thus, a storyline may include a number of news items and also cover different but related events. This ontology also covers the notion of *topic*, however it limits it only to *entities*, such as people and organizations, and *themes*. Our earlier work on the *News Angle Ontology* [53] has instead focused on characterising *news angles*, which can be seen as design templates that can be used to shape the narrative around an event or set of events. The ontology engineering literature also comprises broader multimedia ontologies [5], however these focus primarily on the process of annotating digital content — e.g., a JPEG image, rather than topic classification.

8. Conclusions

In this paper, we have discussed the limitations of current solutions for news classification and highlighted the gap between the current state of the art in computational solutions for news content analysis and the needs of media scholars and practitioners. Crucially, we have also argued that in order to address this gap it is necessary to develop a better understanding of the task of fine-grained news classification, in particular by identifying the various categories of entities that can be the focus of news items. To this purpose, in this paper we have presented a formal framework that characterises news topics in terms of a typology comprising entities, events, situations, categorical topics and viewpoints. The framework has been realised into a family of open source ontologies and empirically validated by manually annotating a corpus of news items randomly drawn from Norwegian and British newspapers.

Having developed the framework, the next step of this research will focus on applying it to support effective computational methods for fine-grained news classification. In particular, while a variety of information extraction methods already exist for certain types of news topics, in particular entities and events, much novel work is needed to develop effective techniques to recognise other elements of our framework, such as negative events, situations and viewpoints. To tackle this challenge we plan to capitalise on recent advances in large language models, which have paved the way for new opportunities in information extraction. Crucially, these techniques need to be guided by robust domain representations, in order to yield verifiable and high-quality outcomes [85][27], while avoiding hallucinations [10]. Hence, a key research hypothesis underpinning this work is that the formal framework presented in this paper may play an effective role in enabling novel model-driven information extraction solutions, tailored to the task of fine-grained news classification.

Another challenge in this context concerns the development of an effective solution able to identify correctly the news topic – that is, able not only to extract, for example, the correct representation of an event reported in a news item, but also to conclude accurately that the event in question is indeed the correct focus of the news item. Here, we expect to be able to take advantage of the writing style used by journalists, which typically uses the title, byline and lead paragraph of a news item to emphasise the focus of a story.

While the key goal of our research is to support better news analytics, it would also be interesting to examine whether the combination of this framework and appropriate computational linguistics methods can be used in *news synthesis* — e.g., to generate news summaries that take advantage of the structured representation proposed in our framework. This is indeed an important challenge in the media industry, where effective methods for reformulating and summarizing content in different contexts (e.g., social media, news feeds) define essential capabilities in the modern, highly heterogeneous media landscape. In particular, our hypothesis (yet to be tested) is that, by explicitly capturing the variety of key constituent elements in a news item, such as, topics, actors, events, and viewpoints, it may be possible to generate more effective alternative formulations or summaries of news content compared to current methods [57].

In conclusion, the proposed conceptual framework for news classification defines the first step of our

[35] https://www.bbc.co.uk/ontologies/storyline-ontology/.

research agenda, whose ultimate goal is to develop better solutions to enable a variety of user audiences, including media scholars and practitioners, commercial media companies, and policy makers and regulators, to effectively make sense of the dynamics of topics and viewpoints in the media. We are very much looking forward to the next phases of this work.

References

[1] Abramson, J. (2019). Merchants of truth: Inside the news revolution. Random House.

[2] Allaway, E. and McKeown, K. (2023). Zero-shot stance detection: Paradigms and challenges. Frontiers in Artificial Intelligence, 5.

[3] Alokaili, A., Aletras, N. and Stevenson, M. (2020). Automatic generation of topic labels. 43rd ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1965-1968.

[4] Arakelyan, E., Arora, A. and Augenstein, I. (2023). Topic-Guided Sampling For Data-Efficient Multi-Domain Stance Detection. arXiv:2306.00765.

[5] Arndt, R., Troncy, R., Staab, S., Hardman, L. and Vacura, M. (2007). COMM: designing a well-founded multimedia ontology for the web. In Aberer, K. et al. (Eds.), 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, 11-15 November 2007, Busan, Korea. LNCS 4825, Springer.

[6] Attard, J., Scerri, S., Rivera, I. and Handschuh, S. (2013). Ontology-based situation recognition for context-aware systems. Proceedings of the 9th International Conference on Semantic Systems, pp. 113-120.

[7] Bacan, H., Pandzic, I. S. and Gulija, D. (2005). Automated news item categorization. 19th Annual Conference of The Japanese Society for Artificial Intelligence, pp. 251-256.

[8] Baden, C. and Springer, N. (2017). Conceptualizing viewpoint diversity in news discourse. Journalism, 18(2), pp. 176-194.

[9] Baker, C. F., Fillmore, C. J. and Cronin, B. (2003). The structure of the FrameNet database. International Journal of Lexicography, 16(3), pp. 281-296.

[10] Beutel, G., Geerits, E. and Kielstein, J.T. (2023). Artificial hallucination: GPT on LSD?. Critical Care, 27(1), p.148.

[11] Bhatia, S., Lau, J. H. and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. International Conference on Computational Linguistics. Association for Computational Linguistics, ACL Anthology.

[12] Bizer C., Vidal M. E. and Skaf-Molli H. (2018). Linked Open Data. In: Liu L., Özsu M.T. (eds), Encyclopedia of Database Systems. Springer, NY.

[13] Borgo, S., Ferrario, R., Gangemi, A., Guarino, N., Masolo, C., Porello, D., Sanfilippo, E. M. and Vieu, L. (2022). DOLCE: A descriptive ontology for linguistic and cognitive engineering. Applied Ontology, 17(1), pp.45-69.

[14] Boumans, J. W. and Trilling, D. (2016). Taking Stock of the Toolkit. Digital Journalism, 4(1), pp. 8-23.

[15] Buckingham Shum, S., Motta, E., and Domingue, J. (2000). ScholOnto: an ontology-based digital library server for research documents and discourse. International Journal on Digital Libraries, 3, pp. 237-248.

[16] Cabrio, E. and Villata, S. (2018). Five years of argument mining: A data-driven analysis. Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 5427-5433.

[17] Carriero, V. A., Gangemi, A., Nuzzolese, A. G. and Presutti, V. (2021). An Ontology Design Pattern for Representing Recurrent Situations. In Advances in Pattern-Based Ontology Engineering, pp. 166-182. IOS Press.

[18] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

[19] Chiatti, A., Bardaro, G., Motta, E. and Daga, E. (2022). A Spatial Reasoning Framework for Commonsense Reasoning in Visually Intelligent Agents. AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition, 15-17 Jun 2022, Örebro, Sweden, CEUR.

[20] Cohn, A. G. and Renz, J. (2008). Qualitative spatial representation and reasoning. Foundations of Artificial Intelligence, 3, pp. 551-596.

[21] Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P. and Sattler, U. (2008). OWL 2: The next step for OWL. Journal of Web Semantics, 6(4), pp. 309-322.

[22] De Clercq, O., De Bruyne, L. and Hoste, V. (2020). News topic classification as a first step towards diverse news recommendation. Computational Linguistics in the Netherlands, 10, pp. 37-55.

[23] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019, pp. 4171-4186.

[24] De Vreese, C. H. (2005). News framing: Theory and typology. Information Design Journal & Document Design, 13(1).

[25] Doan, T. M. and Gulla, J. A. (2022). A survey on political viewpoints identification. Online Social Networks and Media, 30.

[26] Doogan, C. and Buntine, W. (2021). Topic model or topic twaddle? re-evaluating semantic interpretability measures. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3824-3848.

[27] Emelin, D., Bonadiman, D., Alqahtani, S., Zhang, Y. and Mansour, S. (2022). Injecting domain knowledge in language models for task-oriented dialogue systems. arXiv preprint arXiv:2212.08120.

[28] Fincke, S., Agarwal, S., Miller, S. and Boschee, E. (2022). Language model priming for cross-lingual event extraction. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), pp. 10627-10635.

[29] Gangemi, A. and Mika, P. (2003). Understanding the semantic web through descriptions and situations. On the Move to Meaningful Internet Systems, pp. 689-706. Springer Berlin Heidelberg.

[30] Gangemi, A. and Presutti, V. (2022). Formal Representation and Extraction of Perspectives. In Vossen, P. and Fokkens, A. (eds), Creating a More Transparent Internet: The Perspective Web, pp. 208-228.

[31] Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A.G., Draicchio, F. and Mongiovì, M. (2017). Semantic web machine reading with FRED. Semantic Web, 8(6), pp.873-893.

[32] Gao, Y., Song, S., Zhu, X., Wang, J., Lian, X. and Zou, L. (2018). Matching heterogeneous event data. IEEE Transactions on Knowledge and Data Engineering, 30(11), pp. 2157-2170.

[33] Gaignard, A., Rosnet, T., de Lamotte, F., Lefort, V. and Devignes, M. (2023). FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. Journal of Biomedical Semantics, 14. https://doi.org/10.1186/s13326-023-00289-5.

[34] Gottschalk, S. and Demidova, E. (2018). EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, pp. 272-287. Springer.

[35] Gottschalk, S., Kacupaj, E., Abdollahi, S., Alves, D., Amaral, G., Koutsiana, E., Kuculo, T., Major, D., Mello, C., Cheema, G.S. and Sittar, A. (2023). OEKG: The Open Event Knowledge Graph. arXiv preprint arXiv:2302.14688.

[36] van Hage, W. R., Malaisé, V., Segers, R., Hollink, L. and Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). Journal of Web Semantics 9(2), pp. 128-136.

[37] Hamborg, F., Donnay, K. and Gipp, B. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries, 20(4), pp. 391-415.

[38] Harcup, T. and O'Neill, D. (2017). What is news? News values revisited (again). Journalism Studies, 18(12), pp. 1470-1488.

[39] Harrando, I., Lisena, P. and Troncy, R. (2021). Apples to apples: A systematic evaluation of topic models. International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 483-493.

[40] Helmstetter, S. and Paulheim, H. (2018). Weakly Supervised Learning for Fake News Detection on Twitter. Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 274–277.

[41] Huang, L., Ji, H., Cho, K., Dagan, I., Riedel, S. and Voss, C. (2018). Zero-Shot Transfer Learning for Event Extraction. 56th Annual ACL Meeting (Volume 1: Long Papers), pp. 2160-2170.

[42] Jacobi, C., Van Atteveldt, W. and Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), pp. 89-106.

[43] Lai, V. D. (2022). Event extraction: A survey. arXiv preprint arXiv:2210.03419.

[44] Liu, J., Chen, Y., Liu, K., Bi, W., and Liu, X. (2020). Event extraction as machine reading comprehension. Empirical Methods in Natural Language Processing (EMNLP) 2020, pp. 1641-1651.

[45] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 8(3), pp. 243-281.

[46] Martens, B., Aguiar, L., Gomez-Herrera, E. and Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. Joint Research Centre Digital Economy Working Paper, 2018-02.

[47] Martinez-Rodriguez, J. L., López-Arévalo, I. and Rios-Alvarado, A. B. (2018). OpenIE-based approach for knowledge graph construction from text. Expert Systems with Applications, 113, pp. 339-355.

[48] Masini, A. and van Aelst, P. (2017). Actor diversity and viewpoint diversity: Two of a kind? Communications, 42(2), pp. 107-126.

[49] Masini, A., van Aelst, P., Zerback, T., Reinemann, C., Mancini, P., Mazzoni, M., Damiani, M. and Coen, S. (2018). Measuring and explaining the diversity of voices and viewpoints in the news: A comparative study on the determinants of content diversity of immigration news. Journalism Studies, 19(15), pp. 2324-2343.

[50] McCombs, M. E., Shaw, D. L. and Weaver, D. H. (2018). New directions in agenda-setting theory and research. In Advances in Foundational Mass Communication Theories, pp. 131-152. Routledge.

[51] Mihindukulasooriya, N., Tiwari, S., Enguix, C.F. and Lata, K. (2023). Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. arXiv preprint arXiv:2308.02357.

[52] Mirza, P., Sprugnoli, R., Tonelli, S. and Speranza, M. (2014). Annotating causality in the TempEval-3 corpus. Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pp. 10-19.

[53] Motta, E., Daga, E., Opdahl, A.L. and Tessem, B. (2020). Analysis and design of computational news angles. IEEE Access, 8, pp. 120613-120626.

[54] Motta, E., Osborne, F., Pulici, M. M. L., Salatino, A. A. and Naja, I. (2024). Capturing the Viewpoint Dynamics in the News Domain. 24th International Conference on Knowledge Engineering and Knowledge Management, EKAW-24. Amsterdam, November 2024.

[55] Neshenko, N., Nader, C., Bou-Harb, E. and Furht, B. (2020). A survey of methods supporting cyber situational awareness in the context of smart cities. Journal of Big Data, 7(1), pp. 1-41.

[56] Nikoloudakis, Y., Kefaloukos, I., Klados, S., Panagiotakis, S., Pallis, E., Skianis, C. and Markakis, E. K. (2021). Towards a machine learning based situational awareness framework for cybersecurity: an SDN implementation. Sensors, 21(14) – https://doi.org/10.3390/s21144939.

[57] Oliveira, H. and Lins, R. D. (2024). Assessing Abstractive and Extractive Methods for Automatic News Summarization. Proceedings of the ACM Symposium on Document Engineering 2024, pp. 1-10.

[58] OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.

[59] Osborne, F. and Motta, E. (2015). Klink-2: integrating multiple web sources to generate semantic topic networks. 14th International Semantic Web Conference, Bethlehem, PA, USA.

[60] Park, S., Lee, S. and Song, J. (2010). Aspect-level news browsing: Understanding news events from multiple viewpoints. Proceedings of the 15th international conference on Intelligent User Interfaces, pp. 41-50.

[61] Payton, J.D. (2018). How to identify negative actions with positive events. Australasian Journal of Philosophy, 96(1), pp. 87-101.

[62] Peng, C., Xia, F., Naseriparsa, M. and Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review 3, pp.1-32.

[63] Peroni, S., Shotton, D. and Vitali, F. (2012). Faceted documents: describing document characteristics using semantic lenses. Proceedings of the 2012 ACM Symposium on Document Engineering, pp. 191-194.

[64] Rebboud, Y., Lisena, P. and Troncy, R. (2022). Beyond causality: Representing event relations in knowledge graphs. International Conference on Knowledge Engineering and Knowledge Management, EKAW 2022, pp. 121-135. Cham: Springer International Publishing.

[65] Rospocher, M., Van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T. and Bogaard, T. (2016). Building event-centric knowledge graphs from news. Journal of Web Semantics, 37, pp.132-151.

[66] Riker, W. H. (1957). Events and situations. The Journal of Philosophy, 54(3), pp. 57-70.

[67] Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R. and Tannier, X. (2019). Searching news articles using an event knowledge graph leveraged by Wikidata. In Companion Proceedings of the 2019 World Wide Web Conference, pp. 1232-1239.

[68] Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F. and Motta, E. (2020). The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence, 2(3), pp. 379-416.

[69] Scherp, A., Franz, T., Saathoff, C. and Staab, S. (2012). A core ontology of events for representing occurrences in the real world. Multimedia Tools and Applications, 58, pp. 293-331.

[70] Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania.

[71] Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A. and Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. Semantic Web, 13(3), pp. 527-570.

[72] Shoemaker, P.J. and Reese, S.D. (1991). Mediating the message: Theories of influences on mass media content. Guilford Publications.

[73] Sjøvaag, H. and Kvalheim, N. (2019). Eventless news: Blindspots in journalism and the 'long tail' of news content. Journal of Applied Journalism and Media Studies, 8(3), pp. 291-310.

[74] Tempich, C., Pinto, H. S., Sure, Y. and Staab, S. (2005). An argumentation ontology for distributed, loosely-controlled and evolving engineering processes of ontologies (DILIGENT). 2nd European Semantic Web Conference, Heraklion, Crete, Greece, May 29–June 1, 2005. Springer.

[75] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y. and Li, Y. (2022). LaMDA: Language models for dialog applications. arXiv preprint arXiv:2201.08239.

[76] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[77] Toulmin, S. E. (2003). The Uses of Argument. Cambridge University Press.

[78] Trabelsi, A. and Zaïane, O. R. (2019). Phaitv: A phrase author interaction topic viewpoint model for the summarization of reasons expressed by polarized stances. Proceedings of the International AAAI Conference on Web and Social Media, 13, pp. 482-492.

[79] Vilares, D. and He, Y. (2017). Detecting perspectives in political debates. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1573-1582.

[80] Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledge base. Communications of the ACM, 57(10), pp. 78-85.

[81] Wadden, D., Wennberg, U., Luan, Y. and Hajishirzi, H. (2019). Entity, Relation, and Event Extraction with Contextualized Span Representations. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5784-5789.

[82] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), pp. 1-9.

[83] Windisch Brown, S., Bonial, C., Obrst, L. and Palmer, M. (2017). The Rich Event Ontology. Proceedings of the Events and Stories in the News Workshop, pp. 87-97.

[84] Wu, L., Petroni, F., Josifoski, M., Riedel, S. and Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6397-6407.

[85] Yang, J., Xiao, G., Shen, Y., Jiang, W., Hu, X., Zhang, Y. and Peng, J. (2021). A survey of knowledge enhanced pre-trained models. arXiv preprint arXiv:2110.00269.