Computational Linguistics in the Netherlands Journal 10 (2020) 37–55 Submitted 10/2020; Published 12/2020

News Topic Classification as a First Step Towards Diverse News Recommendation

Orphée De Clercq ORPHEE.DECLERCQ@UGENT.BE
Luna de Bruyne LUNA.DEBRUYNE@UGENT.BE
Véronique Hoste VERONIQUE.HOSTE@UGENT.BE

LT³, Language and Translation Technology Team, Ghent University

Abstract

When developing an algorithm that uses news diversity as a key driver for personalized news recommendation it is crucial to focus on means to cluster news articles in a fine-grained manner, ideally by leveraging the content of the text. In this paper we investigate semantic classification of news articles in an unfiltered news stream. We first present an analysis of the EventDNA corpus: a collection of Dutch-language news articles annotated with event data according to a predefined typology. We found that the types assigned as features of events do not allow for such a semantic classification and investigate the IPTC News Media Topics standard as an alternative. By mapping event types with manually-assigned IPTC topics, we observe that a more diversified picture emerges, which leads us to conclude that the IPTC classification is a useful proxy. Based on a historical data sample of Dutch news articles covering the year 2018, we then perform a series of machine learning experiments in order to automatically predict the top two levels of the IPTC taxonomy. Various multi-label classification models are built with BERTje using a bottom-up and top-down approach. The results reveal that the top-down approach yields the best results, with an overall macro F-1 score of 86.4% and a Jaccard accuracy of 89.2% for the level-one topics and one of 83.7% and 87.5% for the level-two predictions.

1. Introduction

In today's digital world the news is omnipresent. Day after day, the media – newspapers, news sites, tv, radio and social media – report the most important events in the world, namely the events marking today's breaking news. More and more, this news stream is being personalized by automated decision-making and recommender systems following a commercial logic, i.e. based on calculated relevance to the user (Joris et al. 2020a).

In this respect, concerns have been raised about news personalization, selective exposure and 'filter bubbles' (Zuiderveen Borgesius et al. 2016, Haim et al. 2018). In the framework of the #NewsDNA project, a different logic is investigated to create a recommendation algorithm, one driven by diversity.[1] Since the concept of news diversity is hard to grasp, the project draws on widely used conceptualizations of news and media diversity (McQuail 1987, Napoli 1999, Voakes et al. 1996), see Joris et al. (2020b) for a more in-depth discussion.

In the research presented here, we take a first step towards content diversity which implies that the content being presented to the end user should reflect the diversity of topics and events occurring in an unfiltered news stream. Such an algorithm requires a deep textual understanding of the news and this is also the broader research context of this article. Extracting topics can give a first rough idea of the theme of a news article, whereas identifying the specific events the article mentions, allows for a semantic understanding of the text. In the past, the EventDNA corpus (Colruyt et al. 2020) has been created, which comprises 1,771 Dutch news articles annotated with events which were classified into topics following a predefined typology inspired by the rich ERE framework (Song et al. 2015). However, after the annotation effort, it was shown that a large majority of types

1. https://www.ugent.be/mict/en/research/newsdna

assigned as features to events in the EventDNA corpus were categorized as unknown event types, making semantic linking unfeasible.

Instead of designing a new typology also covering the unknown event types, we decided to investigate the IPTC (International Press Telecommunications Council) standard, which could also offer us such a semantic classification. IPTC Topics are a standardized taxonomy of news topics, comprising 17 top-level topics (e.g. `crime, law and justice`, `politics` or `education`) that are divided in increasingly granular subtopics (e.g. `law enforcement`, `election` or `higher education`). At the news article level these IPTC topics have been manually assigned in the EventDNA corpus. Based on a mapping between the events, assigned types and IPTC topics we will show that the latter offer a more diversified picture.

In the remainder of this paper, the creation of an IPTC Media Topics classifier is discussed, which is able to predict the top two levels of the IPTC taxonomy. Based on a historical data sample of Dutch news articles covering the year 2018, this classifier was created using state-of-the-art techniques for text classification. More specifically, all classification models were built with BERTje (de Vries et al. 2019), a Dutch version of the pre-trained transformer model BERT (Devlin et al. 2019). We experimented with two different approaches: a top-down and bottom-up approach. In the top-down approach, the labels of the top-level IPTC topics are predicted first, after which for each predicted label, a lower-layer classifier predicts the level-two topics. In the bottom-up approach, a single classifier is trained to first predict the level-two topics, the top-level topics of which are then derived using the IPTC taxonomy. For both approaches the same train, development and test splits were used and the best approach was then applied to the EventDNA corpus.

The results reveal that the top-down approach yields the best results, with an overall macro F-1 score of 86.4% and a Jaccard accuracy of 89.2% for the level-one topics and one of 83.7% and 87.5%, respectively, for the level-two predictions. When testing the best models on the EventDNA corpus, a drop in performance is observed. This can be partly explained by the manual annotations in the EventDNA corpus which are very extensive and often also include less relevant background topics. Nevertheless, the main level one topics have a good precision and are thus most often accurately predicted. The IPTC classifier is currently being used for labeling an unfiltered news stream and feeding a diversity-based recommendation algorithm with a variety of topics.

This paper is structured as follows, Section 2 offers some related work on event extraction from news and on transformer-based classification. Section 3 presents the EventDNA corpus and zooms in on the semantic categorization possibilities of the IPTC Media Topics. Section 4 describes the different experiments that were carried out to create an IPTC classifier and presents the results. Section 5 concludes this paper and offers suggestions for future work.

2. Related work

Approaches on how to automatically detect and classify events have been studied for a long time, mainly for English, in the MUC, ACE and ERE challenges.

Research on event extraction started with the Message Understanding Conferences (Grishman and Sundheim 1996). The ACE or Automatic Content Extraction program succeeded MUC in 1999, and event extraction was introduced in ACE in 2004 (Aguilar et al, 2014). Starting in 2014, the ACE scheme served as a design basis for the family of schemes used in DARPA's Deep Exploration and Filtering of Text (DEFT) program. A number of approaches to events were designed in the context of DEFT (Bies et al. 2016), the most important one being the Entities, Relations, and Events (ERE) scheme (Song et al. 2015, Aguilar et al. 2014), which builds on the concepts introduced by the ACE evaluation. Its starting point, Light ERE, is a consolidated version of ACE. Rich ERE, introduced in 2105, expands the number of semantic types that define the scope of events.

Typically, event mentions are first classified into semantic types such as *Conflict.Attack*, *Transaction.TransferOwnership*, etc. to facilitate mention detection. Both the ACE and ERE programs defined fixed semantic categories of events, such that event mentions are sorted into a taxonomy

of semantic categories (Aguilar et al. 2014). Events falling outside the typology were considered irrelevant. This closed-domain setting limits more advanced applications for event extraction (Araki and Mitamura 2018). It also forms a problem in transferring models to unrestricted data contexts, since existing training corpora used in these settings tend to be unnaturally skewed to contain more events from the typology (Grishman 2010). On the other hand, increasing the number of types also increases the complexity of annotation and could impact the quality and size of the corpus. To avoid this, it is possible to introduce generic event types, which, ideally, would be applied very sparingly.

Colruyt et al. (2020) were the first to apply such a typology to Dutch news text with the creation of the EventDNA corpus. However, in the final corpus they found that more generic types made up a large proportion of events, which led them to conclude that event type is not a meaningful event feature in this task. Rather, mapping event mentions and IPTC Topics may provide a more useful bridge to link events across articles. This is exactly what is being explored in this research.

In NLP, the rise of deep learning has altered the field and especially the huge general-purpose language models which can be fine-tuned to any downstream NLP task (Devlin et al. 2019) have become state of the art. The Bidirectional Encoder Representations from Transformers, also known as BERT, has achieved top results in many text classification tasks such as sentiment analysis and topic classification (Sun et al. 2019). Given the focus of our task, i.e. classification of news articles in predefined topics, we assume that such a methodology would also be a good fit for our classification task.

For languages other than English, Multilingual-BERT which is trained on Wikipedia in 104 different languages can be used (Devlin et al. 2019) or language-specific models using the same BERT architecture can be created. The latter is preferred as the language used on Wikipedia is rather domain-specific (de Vries et al. 2019). For Dutch, to the best of our knowledge two pre-trained BERT-like models have been created, BERTje (de Vries et al. 2019) and RobBERT (Delobelle et al. 2020). The specific BERTje model was trained on a 2.4B token corpus, consisting of Wikipedia, Twente News Corpus (Ordelman et al. 2007), SoNaR-500 (Oostdijk et al. 2013), a collection of contemporary and historical fiction novels and all articles of four Dutch news websites from January 1, 2015 to October 1, 2019. The RobBERT model was trained on the Dutch part of the 39GB OSCAR corpus, a part of the Common Crawl corpus (Suárez et al. 2019). As sub-word token input, BERTje uses WordPiece, whereas RobBERT uses byte-level Byte Pair Encoding (BPE). Both models are competitive, though RobBERT seems to perform slightly better on the DBRD dataset (van der Burgh and Verberne 2019) for sentiment analysis. Since we are dealing with edited Dutch news text, we decided to experiment with BERTje.

3. Semantic categorization of news events

The rationale behind this work is the creation of a news recommendation algorithm driven by diversity in the framework of the #NewsDNA project. This requires a way to measure diversity in a news articles dataset and rate the similarity between articles. It was decided to rely on news events as the unit of analysis and the EventDNA corpus was created. The corpus comprises 1,771 annotated news articles (Colruyt et al. 2020). All articles originate from a number of Flemish newspapers and were published in 2017 or 2018. They were provided to us by Mediahuis, a large media company which publishes major newspapers in both Belgium and the Netherlands. The articles were filtered such that only hard news – serious and urgent in nature (economics, politics, war and crime) – was retained, and preprocessed such that each document consists of a title and lead paragraph. Subsequently, all 1,771 documents were annotated with entities, events and IPTC topics following specific guidelines (Colruyt et al. 2019). As this paper's focus lies on events and IPTC topics, we will first zoom in on these two annotation layers in closer detail.

3.1 Events

Event mentions are anchored to a certain textual span. In EventDNA these spans can be entire clauses, implying that both verbal ("Coronavirus has seized control of New York") and nominal expressions ("COVID-19 outbreak in New York") can make up an event.

A distinction is made between main and background events: main events are the new information causing the reporter to write the article, whereas background events give context to the main event. Regarding the typology, many programs such as ACE and ERE have defined fixed semantic categories of events, such that event mentions are found and sorted into a layered taxonomy of semantic categories such as *Conflict-Attack* or *Transaction-TransferOwnership* (Aguilar et al. 2014). In such a closed-domain setting, events falling outside the typology are considered irrelevant.

EventDNA inherited the rich ERE typology (Song et al. 2015) which normally operates with a fixed typology of 9 types and 38 subtypes. Testing rounds brought to light a number of relevant events that could not be assigned a type at all, necessitating the creation of additional types. *Journalism.Publication* was created to cover events such as *"Turkish media report two incidents that occurred at the polls yesterday"*; *Journalism.Investigation* covers journalistic investigations, as in *"the Russian press dove into Toporovski's past"*. The type *Politics* with one subtype *Vote* covers events of all types of popular elections. Finally, the generic type *Unknown* (with subtype *Unknown*) covers any other event considered relevant for annotation that was not covered by another type. This resulted in a final typology with 12 types and 41 subtypes, see Colruyt (2020) for more details.

Especially, the addition of an *Unknown* type represents a break with the ACE/ERE tradition. In an unrestricted data context like the one #NewsDNA describes, relevancy is the key factor for deciding which events should be annotated. Throughout the annotation process, the guiding question the annotators used to determine whether an event should be annotated was "is this event *relevant* as a news item; in other words, can it occur in several articles? Is it interesting to use it as a basis to cluster articles?" This consideration weighed more heavily than the ability to categorize the event.

Table 1 presents an overview of all the main events that were annotated in the EventDNA corpus (4,244 in total) and their distribution among the used typology.

Type	Count	% of total
Manufacture	18	0.4
Journalism	43	1.0
Business	46	1.1
Politics	116	2.7
Personnel	195	4.6
Movement	199	4.7
Transaction	214	5.0
Life	251	5.9
Conflict	406	9.6
Justice	508	12.0
Contact	767	18.1
Unknown	1481	34.9

Table 1: Spread of main event types in the EventDNA corpus

What draws the attention is that a large majority, around 35%, of all main events were assigned to the category *Unknown*. With 18%, the second largest category are the *Contact* events. These two types are more generic in nature. *Unknown* events are often political and economic events which do not refer to concrete events or use fuzzy wording, as in *"breakthrough in Brexit negotiations"*. Interpreting these *Unknown* events is often dependent on recent developments, and they are abstracted to a degree that they belong more to a topic or sphere of interest (such as *"political activity"*) rather than to any discrete, bounded event type. Regarding the *Contact* events, we observed that the

lexicalization of the event is often generic in itself, such that it is not possible to recognize a concrete event type other than that some sort of communication is taking place (such as in e.g. *"May tries to calm down British citizens"*).

3.2 IPTC Media Topics

The International Press Telecommunications Council is an organization composed of more than 60 news organizations and companies from all over the world. Its activities have primarily focussed on developing and promoting industry standards for the exchange of news data of all common media types.[2] It has developed the IPTC Media Topics codes framework, this is a taxonomy of topics which can be used to classify news stories.[3] In total, it defines over 1,226 topics which can all be traced back to 17 top-level media topics. The tree branches up to five levels down; the deeper into the tree, the more granular and specific the topic definitions become. The "COVID-19 outbreak", for example, could be classified as *pandemic* at the fourth level, and traced back to `epidemic` (level 3), `communicable disease` (level 2) and `health` (level 1).[4]

In the EventDNA corpus, each news article has been manually enriched with any number of IPTC Media Topics. Annotators were provided with the taxonomy and asked to mark each document with all relevant IPTC topics, choosing the most specific ones applicable. The labels assigned could be traced back to the more general top-level topics. Annotators worked individually, but freely called on the help of an expert supervisor to resolve difficult cases. In total, 2,721 top-level topics were assigned to the 1,771 articles. Table 2 presents an overview of the number of top-level topics assigned to the articles, clearly illustrating that most articles deal with one or two topics.

Type	Count	% of total
1 topic	1011	57.1
2 topics	609	34.3
3 topics	118	6.7
4 topics	29	1.6
5 topics	2	0.1
6 topics	2	0.1

Table 2: Number of topics assigned per article

2. https://iptc.org/about-iptc/
3. https://iptc.org/standards/media-topics/
4. https://www.iptc.org/std/NewsCodes/mediatopic/treeview/mediatopic-en-GB.html

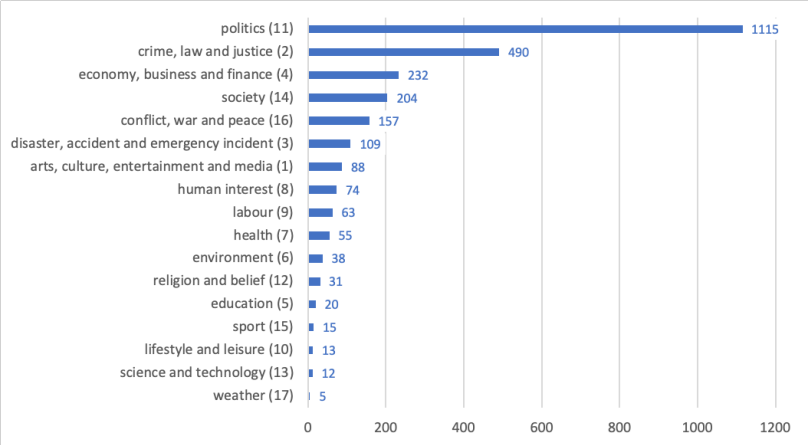

Figure 1: Distribution of top-level IPTC topics over articles

Figure 1 represents the distribution of those topics over the articles, underlining the focus on hard news as the top three topics are `politics`, `crime, law and justice` and `economy, business and finance`.[5]

3.3 Towards semantic categorization

Table 3 presents a frequency mapping between the event types and IPTC top-level topics. Looking at the two more generic event types, the rows *Contact* and *Unknown* highlighted in grey, we observe that a categorisation based on the IPTC Media topics allows for a much more diverse representation of the events represented in both these categories, with a majority of those events belonging to the IPTC media topic 11, i.e. `politics`. This semantic diversification is also nicely illustrated for the rather vague ERE *Transaction* event type, which is distributed over the three major IPTC topics, namely `politics` (110 events, as shown in example sentence [1]), followed by `economy, business and finance` (88 events, as exemplified in sentence [2]) and `crime, law and justice` (51 events, as shown in example [3]).

> *[1] Vlaamse regering maakt 9 miljoen vrij voor leerkrachten basisonderwijs (EN: Flemish government clears 9 million for primary teaching)*
>
> *[2] heeft warenhuisketen Colruyt beslist niet langer salami te verkopen van het Spaanse merk El Pozo (EN: department chain Colruyt decided to no longer sell salami of the Spanish brand El Pozo)*
>
> *[3] Dieven stelen diagnosetoestellen , bestelwagen en sigaretten bij garage (EN: Theft of appliances for diagnosis, delivery van and cigarettes at garage)*

Overall, we observe the same also holds for many of the other event types; the numbers printed in italics indicate the highest number of event types in a certain IPTC top-level category. Most of them are present in the category `politics`. The only two event types where most events cannot be attributed to the IPTC topic `politics` are the *Justice* and *Life* types. There the highest match is with IPTC topic 2 `crime, law and justice`. For the former event type this is an intuitive match, for the latter a closer inspection of the events revealed that many of them deal with murder and imprisonment (as exemplified in example sentence [4]).

5. The numbers in between brackets after each topic in the figure are just a manner to refer back to these topics in the remainder of the text.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Business	4	8	0	**14**	0	0	0	4	0	0	*27*	1	2	1	0	1	0
Conflict	9	150	9	36	6	2	19	10	22	0	*221*	5	0	73	3	**166**	0
Contact	54	133	19	82	10	7	17	18	28	3	*631*	28	0	86	1	49	0
Journalism	22	8	2	4	0	0	0	8	1	0	*24*	0	0	4	0	1	0
Justice	12	***447***	15	18	2	7	5	20	1	2	168	14	0	73	6	23	0
Life	5	*120*	73	0	3	4	41	11	0	6	70	5	0	103	4	55	0
Manufacture	0	4	0	9	0	0	0	0	0	0	*10*	0	0	0	0	0	0
Movement	9	49	28	15	0	13	1	7	3	6	*120*	4	2	23	12	19	1
Personnel	12	13	0	27	0	2	2	8	22	0	*161*	1	0	9	0	4	0
Politics	0	12	0	2	0	2	0	0	1	0	***113***	3	0	10	0	4	0
Transaction	4	51	2	88	0	1	12	8	11	8	*110*	1	1	11	0	12	0
Unknown	66	268	149	274	23	48	44	82	54	6	*902*	11	18	133	18	74	11

Table 3: Mapping between the event types and the IPTC top-level topics (see Figure 1 for an overview of which number belongs to which topic.

[4] De moord op een beroepsmilitair (EN: The murder on a soldier)

This brings us to the numbers indicated in bold in Table 3. These represent those event types which also have a matching IPTC top-level topic, i.e. *Business* with topic 4 `economy, business and finance`, *Conflict* with topic 16 `conflict, war and peace`, *Justice* with topic 2 `crime, law and justice` and *Politics* with topic 11 `politics`.

These findings made us confident that in order to allow for a semantic categorization of the events, we can rely on the IPTC media topics. However, one could argue that the IPTC topics are skewed towards `politics`. This is indeed the case for our corpus, as the focus of the EventDNA corpus was on hard news. On the other hand, every article could be labeled with as many topics as possible and as shown in Table 2 around 43% of all news articles was labeled with more than one topic. Figure 2 presents the co-occurrences in absolute numbers and in each row, a colour scale is applied so that the cell with the lowest value in that row has the lightest colour and the cell with the highest value the darkest colour. We observe that topic 11, `politics`, co-occurs with many of the other topics, most notable with topic 2 `crime, law and justice` and 4 `economy, business and finance`. An example of such a co-occurrence of topics 11 and 4 is given in example sentence [5]. Also topics 2 `crime, law and justice` and 14 `society` are topics which are often combined with others.

[5] De Amerikaanse president Donald Trump heeft zijn advocaat Michael Cohen de 130.000 dollar terugbetaald (EN: American president Donald Trump restituted the 130K dollars to hiw lawyer Michael Cohen)

4. IPTC classification

As the previous section illustrated IPTC Media Topics allow for a diversified semantic categorisation of news articles. To this purpose we created a multi-label IPTC classifier which is able to predict the top two levels of the IPTC taxonomy. We experimented with two different approaches: a top-down and bottom-up approach. In the top-down approach, the labels of the top-level IPTC topics are predicted first after which for each predicted label, a lower-layer classifier predicts the second-level topics. In the bottom-up approach, a single classifier is trained to predict the second-level topics, of which the top-level topics are then derived using the IPTC taxonomy. For both approaches, the same train, development and test splits were used and the best approach was then tested on the EventDNA corpus.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	88	25	1	9	0	2	2	7	5	0	43	1	0	14	1	5	1
2	25	490	10	26	4	4	21	16	8	0	202	10	0	72	4	53	0
3	1	10	109	9	0	2	8	4	0	2	17	0	0	20	2	4	3
4	9	26	9	232	2	5	2	3	24	1	102	0	1	3	0	9	0
5	0	4	0	2	20	1	0	0	0	0	7	3	1	4	0	3	0
6	2	4	2	5	1	38	0	3	0	1	20	0	0	4	0	1	0
7	2	21	8	2	0	0	55	4	2	1	18	1	1	17	2	7	0
8	7	16	4	3	0	3	4	74	1	0	19	0	0	15	0	5	0
9	5	8	0	24	0	0	2	1	63	0	23	0	1	4	0	4	0
10	0	0	2	1	0	1	1	0	0	13	6	0	0	2	1	0	0
11	43	202	17	102	7	20	18	19	23	6	1115	11	1	87	5	85	1
12	1	10	0	0	3	0	1	0	0	0	11	31	0	15	0	4	0
13	0	0	0	1	1	0	1	0	1	0	1	0	12	0	0	0	0
14	14	72	20	3	4	4	17	15	4	2	87	15	0	204	2	23	1
15	1	4	2	0	0	0	2	0	0	1	5	0	0	2	15	1	0
16	5	53	4	9	3	1	7	5	4	0	85	4	0	23	1	157	0
17	1	0	3	0	0	0	0	0	0	0	1	0	0	1	0	0	5

Figure 2: Co-occurrences in EventDNA corpus.

4.1 Dataset

In the framework of the #NewsDNA project we were given access to a historical datadump of the large Flemish mediaprovider DPG Media[6]. We will refer to this as the DPG corpus which comprises 240,764 Dutch news items from the year 2018. Each data instance consists of a news headline, a lead and the body of the news item separated into paragraphs. Moreover, all these news items were tagged with IPTC topics using a proprietary in-house IPTC classifier of the mediaprovider. A single article could have been assigned multiple topics, and to every assigned topic a confidence value was attached. Table 4 shows the number of predicted topics in the DPG corpus and the average confidence for the top two levels in the IPTC taxonomy.

IPTC Level	Number of topics	Total number of occurences	Average number of occurences per topic	Average confidence per occurence
Level one	17	428,187	25,187	0.704
Level two	109	338,566	3,106	0.665

Table 4: Number of labels in DPG corpus for the top two IPTC levels with its average confidence.

We decided to consider these automatically assigned topics as silver labels (in contrast to gold labels) and used them to train various multi-label classifiers. As the confidence labels also offer some sort of quality check, we experimented with two different confidence levels (cfr. Section 4.2).

After checking for duplicates, 235,726 news instances are left in the DPG corpus. The distribution of the top-level IPTC labels is shown in Table 5, for which we make a distinction between all labels and labels where the confidence score is at least 75%. For comparison's sake, the distribution of labels in the EventDNA corpus is shown as well. Most outspoken differences between the DPG and EventDNA corpus occur in the categories `arts, culture, entertainment and media`, `human interest`, `lifestyle and leisure` and `sport`, which are all underrepresented (this is also visualized by the colour gradation, the closer to red a cell is, the more represented that topic is in the corpus). By contrast, the EventDNA corpus has a relatively higher presence of news items labeled as `conflict, war and peace`. The labels `crime, law and justice`, `economy, business and finance` and `politics` are well-represented in both corpora.

6. To this purpose various Non-Disclosure Agreements have been negotiated between Ghent University and the mediaproviders prohibiting us to share the data with third parties

	IPTC topic	#items in DPG		#items in EventDNA
		conf $\geq$ 0.00	conf $\geq$ 0.75	
1	arts, culture, entertainment and media	48,906	26,711	88
2	crime, law and justice	40,638	24,736	490
3	disaster, accident and emergency incident	20,775	12,453	109
4	economy, business and finance	67,944	33,903	232
5	education	9,134	4,946	20
6	environment	8,658	2,961	38
7	health	10,670	4,997	55
8	human interest	45,294	16,857	74
9	labour	5,376	2,345	63
10	lifestyle and leisure	32,782	13,042	13
11	politics	46,796	26,353	1,115
12	religion and belief	4,043	1,150	31
13	science and technology	3,693	1,196	12
14	society	30,746	10,486	204
15	sport	45,762	35,385	15
16	conflict, war and peace	3,767	1,248	157
17	weather	3,203	1,381	5

Table 5: Distribution of labels in the DPG and EventDNA corpus.

4.2 Experimental setup

All classification models were built with BERTje (de Vries et al. 2019). As mentioned in the previous section we focus on the top two levels of the IPTC taxonomy: at level one the number of topics amounts to 17 and at level two to 109 (cfr. Table 4).

Two possible approaches were considered, a top-down and bottom-up approach. In top-down prediction, the level-one labels are predicted first. Then, for each predicted label, a lower-layer classifier makes the deeper level-two predictions. This approach requires a total of 18 classifiers. In bottom-up prediction, however, a single classifier will make predictions about the 109 possible labels of the second level at once. Because of the hierarchical structure of the IPTC taxonomy, the labels of the first level can be automatically derived as well. With only one classifier to be trained, the bottom-up approach is the simplest and least time-consuming. However, 109-label classification is more challenging than 17-label classification and errors made on the second level will be back-propagated to level one. Similarly, in top-down classification, errors made in the first level can be propagated to level two.

We will start with the most intuitive approach, namely top-down prediction. As we are working with a silver standard in the DPG corpus, the trade-off between quantity and quality of training labels is crucial. Therefore, we will both train a system using all labels and a system using only labels with a confidence score of 75% and more. The predictions of the best of these two models will be the starting point for the level two classification.

Next, we will test the bottom-up approach. Based on the findings of the first experiments, we will either continue with the 75%-and-more-confidence labels, or keep all labels including the low-confidence ones for the 109-label classification. From these predictions, the labels for the first level will be deduced automatically from the IPTC taxonomy.

Finally, the best approach (either top-down or bottom-up) will be selected and its trained model tested on the EventDNA corpus, which has gold-standard IPTC labels.

4.2.1 Preprocessing and data splits

For each data instance the news headline, lead and body were concatenated. Tokenization was done within the model, using the BERTje tokenizer and the maximum sequence length was 256 tokens. Instances in which the sequence length was lower were right-padded with the `[PAD]`-token.

For the level-one experiments, the labels were encoded to a 17-dimensional one-hot vector. For the level-two experiments, the data was separated into 17 parts based on the silver level-one label, so that each data part only consisted of items for which it had a positive level-one label. Please note that this does not exclude overlap between the parts, as the data is multi-labeled. After this operation, 18 versions of both corpora were available for experimenting.

The DPG versions were subsequently split into a training, development and test set according to a 8:1:1 ratio. To increase the quality of the development and test set, the labels with a confidence score lower than 75% were removed in all settings. For the training sets, two versions were kept, one in which all labels were preserved and a version with only the 75%-and-more-confidence labels. The removal of lower-confidence labels had as a consequence that some instances had no more positive labels, so these empty-label instances were removed from the respective dataset. Table 6 shows the number of instances for each of the final datasets.

		conf $\geq$ 0.00	conf $\geq$ 0.75		
	# labels	train	train	dev	test
level 1	*17*	188,579	150,880	18,824	18,864
level 2 (1)	*3*	30,562	30,374	1,941	1,909
level 2 (2)	*5*	28,868	28,737	1,741	1,790
level 2 (3)	*5*	15,558	15,484	1,057	1,124
level 2 (4)	*4*	44,806	44,578	2,526	2,536
level 2 (5)	*6*	3,748	3,716	144	149
level 2 (6)	*6*	3,715	3,694	134	90
level 2 (7)	*7*	5,927	5,889	284	257
level 2 (8)	*6*	14,783	14,652	619	724
level 2 (9)	*7*	2,138	2,124	61	63
level 2 (10)	*3*	17,593	17,507	733	712
level 2 (11)	*9*	12,681	12,619	490	483
level 2 (12)	*8*	2,716	2,704	77	81
level 2 (13)	*8*	1,809	1,798	39	28
level 2 (14)	*11*	13,204	13,122	353	390
level 2 (15)	*9*	34,305	34,103	3,336	3,352
level 2 (16)	*8*	1,093	1,088	4	7
level 2 (17)	*4*	1,053	1050	7	9

Table 6: Number of instances in datasets derived from the DPG corpus

4.2.2 Models and evaluation

The models were built with BERTje (de Vries et al. 2019)[7] and implemented with HuggingFace's Transformers library (Wolf et al. 2019). BERTje consists of an encoder with 12 transformer blocks, 12 attention heads and a hidden size of 768. We fine-tune the model on our target task with AdamW optimizer (Loshchilov and Hutter 2017) and the ReduceLROnPlateau learning rate scheduler. As we are dealing with multi-label classification, the loss to be optimized is binary cross-entropy loss. Hyperparameters were the same for all classifiers: maximum sequence length of 256 tokens and batch size of 16 instances, dropout of 0.2 and gelu as activation function in the implementation of Hendrycks

7. https://github.com/wietsedv/bertje

and Gimpel (2016). The [CLS]-token based on the concatenation of the last four layers of the model was used for classification, which first went through a pre-classifier (linear layer with 2,048 nodes) and then to the classification layer with Sigmoid activation function. The maximum number of epochs was set to 100 with a patience of 5 for early stopping.

The loss to be optimised was binary-cross entropy loss. The model with the lowest loss on the development set was saved and applied to the test set. In the top-down approach, the second-level classifier was applied to the positive level-one predictions of the test set. In the bottom-up approach, level-one predictions were deduced from level-two predictions. The best approach was then applied to the EventDNA corpus.

We evaluate by calculating macro F1-score, micro F1-score and Jaccard accuracy. The first two metrics evaluate each label individually, where in the macro setting all labels are considered equally important and in the micro setting the distribution of labels is taken into account. However, since this is a multi-label classification task, each news item can have one or more gold IPTC labels, and one or more predicted IPTC labels, which is why multi-label or Jaccard accuracy is also calculated and reported. Multi-label accuracy is defined as the size of the intersection of the predicted and gold label sets divided by the size of their union. Besides these three metrics, binary F1 (F1-score on the positive class) was included as additional metric for the level one predictions.

4.3 Results

In this section we report the results from the experiments as described in Section 4.2. We first share the results for the top-down approach, followed by those for the bottom-up approach. The best of these approaches is then tested on the EventDNA corpus.

4.3.1 Top-down approach

We started with the most intuitive approach, namely top-down classification, and used the predictions of level one to evaluate whether preference needs to be given to number of training instances or quality of training instances. Therefore, we trained a model based on only the training data where the label has a confidence score of 75% or more (conf $\geq$ 0.75), and a model where all training data, including the labels with lower confidence, is taken into account (conf $\geq$ 0.00). Please recall that testing is done on instances for which the confidence score of the label is at least 75%.

The results are presented in Table 7 and clearly indicate that the higher threshold (conf $\geq$ 0.75) leads to better results, in terms of macro and micro F1-score and Jaccard accuracy. With a Jaccard accuracy of 89.2%, performance is very satisfying.

When scrutinizing the results per label, represented as F1-score on the positive class, in Table 8, we observe that for the best setup (conf $\geq$ 0.75) labels 15 (`sport`), 3 (`disaster, accident and emergency incident`) and 2 (`crime, law and justice`) perform best, with an F1-score of respectively 98%, 96% and 95%. Also labels 1 (`arts, culture, entertainment and media`), 5 (`education`) and 11 (`politics`) score more than 90%. The lowest performing ones are labels 6 (`environment`), 14 (`society`) and 13 (`science and technology`).

Metric	conf $\geq$ 0.00	conf $\geq$ 0.75
Macro F1-score	0.779	0.864
Micro F1-score	0.828	0.903
Jaccard accuracy	0.804	0.892

Table 7: Results top-down approach level one

	F1 on positive class	
label	conf $\geq$ 0.00	conf $\geq$ 0.75
1	0.862	0.909
2	0.907	**0.949**
3	0.879	**0.961**
4	0.793	0.882
5	0.838	0.916
6	0.651	0.758
7	0.766	0.845
8	0.746	0.849
9	0.691	0.817
10	0.743	0.819
11	0.853	0.916
12	0.630	0.817
13	0.738	0.773
14	0.651	0.766
15	0.964	**0.979**
16	0.730	0.843
17	0.795	0.885

Table 8: Results per level one label (top-down approach)

The level two classification builds on the predictions of level one. As the 0.75 threshold data yielded the best results, the same threshold was used for these experiments. As explained in Section 4.2, seventeen separate classifiers, one for each previously predicted label, were trained to make the level-two predictions.

The results are presented in Table 9. With a macro F1-score of 83.7% and a Jaccard accuracy of 87.5%, performance of this level-two classification is only slightly lower. Looking at the level-two predictions of each topic in Table 10, expressed with macro-F1 as every level one topic can be subdivided into a number of level two topics, we observe that the classifiers never have a performance lower than 71%, and that for certain categories, regardless of the number of level-two labels or size of the training data, a macro F1 of more than 90% is achieved (labels 3, 5, 7, 12 and 16).

Metric	Score
Macro F1-score	0.837
Micro F1-score	0.887
Jaccard accuracy	0.875

Table 9: Results top-down approach level two

label	Macro-F1	label	Macro-F1
1	0.837	10	0.889
2	0.718	11	0.807
3	0.980	12	0.961
4	0.761	13	0.862
5	0.976	14	0.722
6	0.737	15	0.728
7	0.918	16	0.903
8	0.882	17	0.832
9	0.800		

Table 10: Results top-down approach level-two labels

4.3.2 Bottom-up approach

For the bottom-up approach, we start with a single classifier that predicts for each of the 109 level-two labels whether or not the topic is applicable to the news item. Although it seems not trivial to make predictions about such a high number of classes, the performance as depicted in Table 11 is

very similar to the outcome of the top-down approach. In fact, with a macro F1-score of 84.1% and a Jaccard accuracy of 89.3% it is even slightly higher.

Macro F1-scores within the level-two labels (Table 12) are best for the labels 3, 5, 7, 8, 12 and 16 (macro F1 > 90%) and worst for 2, 6, 14 and 15 (60% < F1 < 80%). This actually reveals a very similar pattern to the level-two predictions of the top-down approach. This suggests that the number of level-two labels in the respective level-one label and the nature of the labels seem to have more influence on the prediction performance than error propagation. The level-two labels of topic 2 (`crime, law and justice`), for example, are: `crime`, `judiciary`, `justice`, `law` and `law enforcement`, labels which are not so straightforward to distinguish from one another.

Metric	Score
Macro F1-score	0.841
Micro F1-score	0.905
Jaccard accuracy	0.893

Table 11: Results bottom-up approach level two

label	Macro-F1	label	Macro-F1
1	0.845	10	0.897
2	0.728	11	0.893
3	0.984	12	0.905
4	0.801	13	0.808
5	0.979	14	0.621
6	0.645	15	0.778
7	0.939	16	0.943
8	0.916	17	0.889
9	0.877		

Table 12: Results bottom-up approach level-two labels

As shown in Table 13, however, macro F1 drops to 73% when automatically deducing level one from the predictions of level two. When we compare Table 14 with Table 8, we observe that the bottom-up approach is more successful in classifying the level one labels 1, 2, 3, 4 and 15 and less successful in the others. In fact, this approach does not seem to suffer from error percolation when sufficient training data is available (cfr. red-coloured cells in Table 5), this also explains the high overall micro-F1 score and the fact that the Jaccard accuracy does not drop (89.9%). On the other hand, items from labels 16 and 17 (14.6% and 38.3% F1) but also from 6, 9, 11, 13 and 14 (F1 < 70%) are poorly classified. Although this cannot be directly linked to bad performance of these categories on level two, we do deal with some kind of error propagation here: as the second-level classifier often only predicts one out of 109 topics, a lot of level-one labels are missed as well.

Metric	Score
Macro F1-score	0.733
Micro F1-score	0.906
Jaccard accuracy	0.899

Table 13: Results bottom-up approach level one

label	F1 positive	label	F1 positive
1	0.931	10	0.833
2	0.940	11	0.681
3	0.968	12	0.821
4	0.924	13	0.600
5	0.772	14	0.649
6	0.596	15	0.986
7	0.828	16	0.146
8	0.855	17	0.383
9	0.545		

Table 14: Results per level-one label (bottom-up approach)

4.3.3 Performance on EventDNA corpus

Taking into account the performance of both the level-one and level-two predictions, we conclude that the top-down approach yields the best results. Moreover, the first experiments revealed that

using a smaller but more reliable dataset leads to better performance than a bigger dataset with less reliable labels. We will now test the top-down model trained on the 75%-and-more threshold on the EventDNA corpus.

When the level-one classifier of the top-down model is applied to the EventDNA corpus, a large drop in performance can be observed in macro F1, from 86.4% to 49% (Table 15). Jaccard accuracy also drops, from 89.2% to 62.2%, but this result is still acceptable. With a macro F1-score of only 33% and a Jaccard accuracy of only 27.7% the level two performance on the EventDNA corpus is even lower. This contrasts sharply with the results on the DPG corpus, where the level-two F1-score was on par with the level-one results. An error analysis is required to understand this large drop in performance, but we assume that the decisions made by the human annotators are quite different from the predictions made by the proprietary classifier our models were trained on.

Metric	level one	level two
Macro F1-score	0.490	0.330
Micro F1-score	0.663	0.341
Jaccard accuracy	0.622	0.277

Table 15: Results on EventDNA corpus (top-down approach)

Given the discrepancy between the results, we decided to perform one final experiment where we directly fine-tune BERTje on the EventDNA corpus. To this purpose we split the EventDNA corpus into a training, development and test set according to a 8:1:1 ratio. We compared performance on this test set by the best DPG model, the top-down model trained on the 75%-and-more threshold, and a newly created EventDNA top-down model. The results are presented in Table 16 below.

Metric	level one		level two	
	DPG	EventDNA	DPG	EventDNA
Macro F1-score	0.519	0.559	0.613	0.594
Micro F1-score	0.713	0.787	0.368	0.491
Jaccard accuracy	0.666	0.741	0.308	0.376

Table 16: Comparison when training and fine-tuning on either DPG or EventDNA

As expected, directly fine-tuning on the EventDNA corpus leads to better results. However, the difference is less outspoken than when testing on the entire EventDNA dataset.

4.4 Error Analysis

Precision and recall We start by looking at the level one predictions. Table 17 presents the number of gold labels per category, together with the number of predicted labels by the model and how many of those are indeed positive (i.e. true positives). The first thing to notice is that the number of gold labels (2,721) is much higher than the number of predicted labels (1,882). This is due to the different distribution of positive labels in the DPG corpus compared to EventDNA. In EventDNA, 2,721 labels are assigned on a total of 1,771 news items, which comes down to an average of 1.54 labels per instance. In the DPG training corpus, however, 176,156 labels are present for a total of 150,880 instances (1.17 labels per instance) and in the test set there are 22,076 labels on 18,864 instances (1.17 labels per instance as well). This difference in number of positive labels has a large impact on the classification performance, especially regarding recall. Only for a few categories, namely 4, 11 and 15, recall is reasonable. These are respectively the second, fourth and first largest categories in the training corpus.

Secondly, the portion of predicted labels that is actually correct (true positives), varies considerably between labels. In fact, precision is fairly good for most categories (macro-averaged precision of 66.4% and only 6 categories have a precision rate under 60%). Eight categories have a precision rate above 70% (2, 3, 5, 6, 7, 11, 12 and 15), of which categories 2, 3, 11 and 12 perform best and even have more than 80% precision. As categories 11 and 2 are also very frequent in the EventDNA corpus (see Figure 1) this is a satisfying result. Both categories were also highly present in the DPG training corpus (see Table 5), which could account for the high precision. The presence of category 3 in the training corpus, however, is not very pertinent, and even less so for category 15. However, the word use in news items with these topics could be what makes these items more salient and thus results in high precision. A manual inspection of these items should further clarify this.

Category	# Gold	# Predicted	# True positives	Precision	Recall
1	88	33	20	0,606	0,227
2	490	265	237	0,894	0,484
3	109	71	67	0,944	0,615
4	232	304	165	0,543	0,711
5	20	7	5	0,714	0,250
6	38	26	20	0,769	0,526
7	55	21	15	0,714	0,273
8	74	50	29	0,580	0,392
9	63	30	17	0,567	0,270
10	13	7	3	0,429	0,231
11	1,115	906	845	0,933	0,758
12	31	19	16	0,842	0,516
13	12	9	4	0,444	0,333
14	204	41	25	0,610	0,123
15	15	16	12	0,750	0,800
16	157	74	46	0,622	0,293
17	5	3	1	0,333	0,200
all	2,721	1,882	1,527	0.664	0.412

Table 17: Number of labels per topic in EventDNA corpus.

Manual inspection We selected 25 instances from the EventDNA corpus for a manual inspection of gold and predicted labels. We made sure that all level one labels appeared at least once as gold and predicted label.

The most notable observation is that the gold labels in the EventDNA corpus are very extensive. As part of the annotation effort also background topics were annotated and, as a consequence, not all labels are as relevant. The classifier on the other hand, mostly predicts the main level one topic. Here is an example:

> *[6] Twee Belgische toeristen overleden bij zwaar verkeersongeval in Portugal. Bij een zwaar verkeersongeval in het zuiden van Portugal zijn vrijdagnamiddag twee Belgen om het leven gekomen, twee andere mensen raakten lichtgewond. Dat melden Portugese media. (EN: Two Belgian tourists die in Portugal in tragic traffic accident.)*
>
> Gold labels (level 1): `disaster, accident and emergency incident`; `health`; `lifestyle and leisure`; `society`
> Predicted labels (level 1): `disaster, accident and emergency incident`

Although the classifier predicts only the main level-one topic, in a lot of cases, this one topic is in fact one of the gold labels, which explains why the Jaccard accuracy for level one is fairly

good (62.2%). For level two predictions, however, this is less the case. Apart from the problem that not enough labels are predicted, the predictions that are made are often different from the gold labels. However, a fine-grained classification of this kind is very complicated, and is often open to interpretation or preference. Example 7 illustrates this problem.

[7] 20.000 kankerpatiënten per jaar krijgen DNA-analyse terugbetaald: Belangrijke stap voor België. Minister Maggie De Block voorziet 4 miljoen euro per jaar. De regering trekt vier miljoen euro per jaar uit voor het terugbetalen van complexe DNA-analyses bij kankerpatiënten. Artsen weten zo tot in detail hoe de tumor van een patiënt in elkaar zit. Informatie die het aantal bijwerkingen kan beperken, maar vooral ook leidt tot betere behandelingen. (EN: Every year 20K cancer patients get their DNA-analysis refunded: an important step for Belgium.)

Gold labels (level 1): `health`; `politics`
Gold labels (level 2): `healthcare policy`; `health treatment`; `government`
Predicted labels (level 1): `health`
Predicted labels (level 2): `diseases and conditions`

Another problem that we encountered was the lack of world knowledge. In Example 8, for instance, world knowledge is necessary in order to know that Tomorrowland is a music festival and thus belongs to `arts, culture, entertainment and media`. Furthermore, the classifier made predictions based on words that it had seen in another context in the training data. This was the case with items containing words like 'car', 'bike' or 'woods', being classified as `lifestyle and leisure` or `human interest` (see Example 9). Finally, there was some content-related bias in the training data, which resulted for example in the assignment of the topic `election` to news about Donald Trump (Example 10), even when the news item had nothing to do with elections.

[8] Onderzoek naar ticketverkoop Tomorrowland. Bij de FOD Economie zijn de voorbije dagen verschillende klachten binnengekomen over de ticketverkoop van Tomorrowland. Dat heeft minister van Consumentenzaken Kris Peeters (CD&V) maandagochtend bekendgemaakt bij 'De Inspecteur' op Radio 2. (EN: Investigation of ticket sales for Tomorrowland)

Gold labels (level 1): `politics`; `arts, culture, entertainment and media`
Predicted labels (level 1): `economy, business and finance`

[9] Ook Natuurpunt kritisch voor voorstel Schauvliege. Er is vooral vraag naar meer natuur en bos dichtbij woonkernen. Meer natuur dichtbij, dat los je niet op met het veranderen van de toegangsregels van de bestaande bossen. Zo reageert Natuurpunt op het voorstel van minister van Natuur Joke Schauvliege om de regels in openbare bossen om te gooien. (EN: Also Natuurpunt is critical regarding Schauvliege's proposal.)

Gold labels (level 1): `environment`
Predicted labels (level 1): `environment`, `human interest`

[10] Trump duidelijk in handelsconflict: Er worden geen uitzonderingen gemaakt bij staalheffing. De Amerikaanse president Donald Trump heeft in gesprekken met wereldleiders gezegd dat er geen uitzonderingen worden gemaakt bij zijn geplande tariefverhogingen op staal en aluminium. Dat zei zijn minister van Handel Wilbur Ross zondag in een interview op de Amerikaanse televisie. (EN: Trump very clear in trade war: there will be no exceptions for steel levy.)

Gold labels (level 1): `economy, business and finance`; `politics`
Predicted labels (level 1): `politics`
Gold labels (level 2): `economic sector`; `government policy`; `government`; `international relations`;
Predicted labels (level 2): `election`

5. Conclusion

In this paper we investigated semantic classification of news articles in an unfiltered news stream, which, in a next step, can serve as input for a novel news recommendation algorithm which is driven by diversity. Based on an analysis of the EventDNA corpus (Colruyt et al. 2020), which is annotated with events according to a predefined typology, we found that the types assigned as features are not meaningful. A large part of the events were assigned to types which are rather generic, such as *Contact*, *Transaction* and *Unknown*. We investigated and found that the IPTC Media Topics standard offers an alternative semantic classification. By mapping events types with manually-assigned IPTC topics, a more diversified picture emerged and we thus concluded that the IPTC classification could be used as a proxy for the event typology.

In a next phase, we trained an IPTC classifier which is able to predict the top two levels of the IPTC taxonomy. Various classification models were built using BERTje (de Vries et al. 2019) and experiments following two different approaches were conducted. In a top-down approach, the labels of the top-level IPTC topics were predicted first, after which for each predicted label, a lower-layer classifier predicted the level-two topics. In the bottom-up approach, a single classifier was trained to first predict the level-two topics and subsequently derive the level-one topics using the IPTC taxonomy. For both approaches training data was derived from a large corpus of newspaper articles covering the year 2018. This DPG corpus had been automatically tagged with IPTC topics and every topic had also been assigned a confidence level. We used these topics as silver labels and for all experiments the same train, development and test folds were used. In the first approach, we also experimented with two different confidence levels to see which setup leads to better results. In a final experiment, the best approach was applied to the EventDNA corpus, which has gold-standard IPTC labels.

Our results reveal that the top-down approach overall leads to the best results when using a confidence level threshold of 75% and more. With a macro F1-score of 86.4% and a Jaccard accuracy of 89.2% for the level-one predictions and one of 83.7% and 87.5%, respectively, for the level-two predictions these results are very satisfying, especially given the large number of level-two labels to be predicted (109). When testing this approach on the EventDNA corpus we observed a large drop in performance, especially for the level-two topics. This can be partly explained by the manual annotations in the EventDNA corpus which were very extensive (1.54 labels per instance). The classifier which predicts on average 1.17 labels per instance thus failed to predict a large number of labels, which has an impact on recall. The precision, however, is fairly good to very good for the most frequently occurring topics in the EventDNA corpus. A manual inspection of 25 instances further underlined that the labeling in the EventDNA is very extensive. As part of that annotation effort also background topics were annotated and, as a consequence, not all labels are as relevant. The classifier on the other hand, mostly predicts the main level-one topic correctly.

In future work the IPTC classifier will be used to tag news articles which are then fed into a diversity-driven news recommendation algorithm. Regarding the current IPTC classifier, it could be interesting to make a combination of various level-two models of the top-down and bottom-up approach as we found that the bottom-up approach is better in classifying the less frequently occurring topics, whereas for the top-down approach the opposite is true. We also plan to continue work on event extraction using the EventDNA corpus and test the overall effectiveness of incorporating both IPTC topics and events into a news recommendation algorithm.

Acknowledgements

We thank the reviewers for their valuable comments. This work was supported by the Ghent University Multidisciplinary Research Partnership "#NewsDNA" [grant number BOFGOA2018000601].

References

Aguilar, Jacqueline, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis (2014), A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards, *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 45–53. https://www.aclweb.org/anthology/W14-2907.

Araki, Jun and Teruko Mitamura (2018), Open-domain event detection using distant supervision, *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 878–891. http://www.aclweb.org/anthology/C18-1075.

Bies, Ann, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, and Tim O 'gorman (2016), A comparison of event representations in DEFT, pp. 27–36.

Colruyt, Camiel, Orphée De Clercq, and Véronique Hoste (2019), EventDNA: Annotation guidelines for entities and events in Dutch news texts (v1.0), *Technical Report LT3 Technical Report - LT3 19.01*, Language and Translation Technology Team, Ghent University.

Colruyt, Camiel, Orphée De Clercq, and Véronique Hoste (2020), EventDNA: a dataset for Dutch news event extraction as a basis for news diversification, Manuscript under review.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), BERTje: A Dutch BERT Model, *arXiv preprint arXiv:1912.09582*.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), RobBERT: a Dutch RoBERTa-based language model, *ArXiv preprint arXiv:2001.06286*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://www.aclweb.org/anthology/N19-1423.

Grishman, Ralph (2010), The impact of task and corpus on event extraction systems, European Language Resources Association (ELRA), Valletta, Malta.

Grishman, Ralph and Beth Sundheim (1996), Message understanding conference - 6: A brief history, *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C96-1079.

Haim, Mario, Andreas Graefe, and Hans-Bernd Brosius (2018), Burst of the filter bubble?, *Digital Journalism* **6** (3), pp. 330–343, Routledge.

Hendrycks, Dan and Kevin Gimpel (2016), Gaussian error linear units (gelus), *arXiv preprint arXiv:1606.08415*.

Joris, Glen, Camiel Colruyt, Judith Vermeulen, Stefaan Vercoutere, Frederik De Grove, Kristin Van Damme, Orphée De Clercq, Cynthia Van Hee, Lieven De Marez, Veronique Hoste, Eva Lievens, Toon De Pessemier, and Luc Martens (2020a), News diversity and recommendation systems : setting the interdisciplinary scene, *in* Friedewald, Michael, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker, editors, *Privacy and Identity Management. Data for Better Living : AI and Privacy*, Vol. 576, Springer, pp. 90–105.

Joris, Glen, Frederik De Grove, Kristin Van Damme, and Lieven De Marez (2020b), News diversity reconsidered: A systematic literature review unraveling the diversity in conceptualizations, *Journalism Studies* **21** (13), pp. 1893–1912, Routledge.

Loshchilov, Ilya and Frank Hutter (2017), Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101*.

McQuail, Denis (1987), *Mass communication theory: An introduction (2nd ed.)*, Studies in Computational Intelligence, Thousand Oaks, CA, US.

Napoli, Philip M. (1999), Deconstructing the diversity principle, *Journal of Communication* **49** (4), pp. 7–34.

Oostdijk, Nelleke, Martin Reynaert, Veronique Hoste, and Ineke Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, Springer, pp. 219–247.

Ordelman, Roeland J.F., Franciska M.G. de Jong, Adrianus J. van Hessen, and G.H.W. Hondorp (2007), TwNC: a multifaceted Dutch News Corpus, *ELRA Newsletter*.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma (2015), From light to rich ERE: Annotation of entities, relations, and events, *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015*, ACL, pp. 89–98.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary (2019), Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures, *Proceedings of the 7t Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang (2019), How to fine-tune BERT for text classification?, *arXiv preprint arXiv:1905.05583*.

van der Burgh, Benjamin and Suzan Verberne (2019), The merits of Universal Language Model Fine-tuning for Small Datasets – a case with Dutch book reviews, *arXiv prepring arXiv: 1910.00896*.

Voakes, Paul S., Jack Kapfer, David Kurpius, and David S. Chern (1996), Diversity in the news: A conceptual and methodological framework, *Journalism & Mass Communication Quarterly* **73** (3), pp. 582–593.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew (2019), Huggingface's Transformers: State-of-the-art natural language processing, *ArXiv preprint arXiv:1910.03771*.

Zuiderveen Borgesius, Frederik J., Damian Trilling, Judith Möller, Balázs Bodó, Claes H. de Vreese, and Natali Helberger (2016), Should we worry about filter bubbles?, *Internet Policy Review*.