Exploring Implicit Sentiment Evoked by Fine-grained News Events

Cynthia Van Hee, Orphée De Clercq and Véronique Hoste
LT^3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication, Ghent University
Ghent, Belgium
firstname.lastname@ugent.be

Abstract

We investigate the feasibility of defining sentiment evoked by fine-grained news events. Our research question is based on the premise that methods for detecting implicit sentiment in news can be a key driver of content diversity, which is one way to mitigate the detrimental effects of filter bubbles that recommenders based on collaborative filtering may produce. Our experiments are based on 1,735 news articles from major Flemish newspapers that were manually annotated, with high agreement, for implicit sentiment. While lexical resources prove insufficient for sentiment analysis in this data genre, our results demonstrate that machine learning models based on SVM and BERT are able to automatically infer the implicit sentiment evoked by news events.

1 Introduction

Why do we read the news that we read and how are news articles received by their audiences? Both research questions are important in the domains of news personalization, framing theory and sentiment and emotion analysis, among others. Digitization and globalization have profoundly changed the media ecology (Mitchelstein and Boczkowski, 2009; Deuze, 2003). There is an increasing trend to consume news via the internet (54%, as opposed to 22% consuming print media[1]), and more specifically via newspaper websites, smartphone apps, social media, etc.

This (partial) shift to online news consumption assigns much more responsibility to citizens, who select from a wide variety of news sources, distributors and topics. Recommendation algorithms do part of the work by filtering, out of the extensive offer of information, news that sparks citizens' interest. Most commonly, such algorithms apply collaborative filtering, which is based on users' past reading behaviour and similar interests in their network. A detrimental side effect of this interplay between algorithms and user behaviour, especially on social media platforms, is that it may lead to a less diverse news consumption, a phenomenon often referred to as the 'filter bubble' (Parser, 2013).

A game changer in this respect are algorithms that use content diversity as the key driver for personalized news recommendation. To date, however, content-based filtering is largely based on topic clustering and keyword matching (Adnan et al., 2014; Liu et al., 2010) without considering semantic information including sentiment and controversy. The present study is part of the #NewsDNA project which aims to investigate and develop a news recommendation algorithm that is driven by content-based diversity[2]. However, before implementing this type of diversity into a recommender, we need to be able to automatically derive sentiment from newswire text. To this end, we explore whether news events evoke implicit sentiment in the reader and, if so, whether this implicit sentiment can be derived automatically using lexicon-based and machine learning techniques. We focus on text spans that describe hard news events (i.e. covering important topics in the public debate, such as politics, finance and economics, war and crime, as well as international news (Shoemaker and Cohen, 2005; Patterson, 2000; Tuchman, 1973)).

This paper is the first initiative to model the semantics of written editorial content (devoid of topic restrictions), where fine-grained news events' implicit sentiment is examined manually, and where attempts are made to model this sentiment automatically. Besides presenting a novel dataset for implicit sentiment detection in news texts, we aim

[1] Figures from the yearly imec.digimeter report (Vandendriessche and De Marez, 2019), publishing recent information about media and technology use in Flanders.

[2] https://www.ugent.be/mict/en/research/newsdna.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 138–148
April 19, 2021. ©2021 Association for Computational Linguistics

to answer the following research question:

- *Can we automatically detect the implicit sentiment evoked by fine-grained news events?*

2 Related Research

While sentiment and emotion analysis have a long history in review analysis and recommendation applications using user-generated content, one of the first studies on subjectivity analysis focused on newswire text (Bruce and Wiebe, 1999). This work, among others, has inspired researchers to apply similar techniques to other data genres, and with the rise of Web 2.0, user-generated content (UGC) quickly became widely investigated, and the main genre under consideration for sentiment research. Compared to the high number of sentiment prediction pipelines that have been established for UGC analysis (Li and Hovy, 2017), not a great deal of research has been done into sentiment analysis at a fine-grained (i.e. below the sentence) level, sentiment analysis in factual data, multi-modal data or in figurative language like irony and humour, etc. (Mohammad, 2017). With this paper, we aim to tackle two of the above-mentioned challenges simultaneously by predicting implicit sentiment evoked by fine-grained (factual) news events.

Sentiment analysis has a broader application range than detecting explicit sentiment clues in subjective texts. Objective utterances can express sentiment as well, be it indirectly by either specific language use (i.e. words that activate emotional values), or by the sentiment certain events evoke through cultural or personal emotional connection. This distinction brings up the terminological confusion around *sentiment* and *opinion*. As pointed out by Liu (2015), the difference between the two is quite subtle, but dictionary definitions of both terms indicate that opinions represent a person's concrete view, whereas sentiments are more of a person's feeling. Although both are not completely independent of one another, it is worthwhile to mention this distinction so as to have a good understanding of the related research.

Implicit sentiment can thus be analyzed from the author's perspective (i.e. implicit opinions), as well as from the reader's (i.e. implicit sentiment). Research on implied opinions is prevalent in research areas such as electoral politics (e.g. Bansal and Srivastava, 2018; Chiu and Hsu, 2018), political viewpoints and argumentation mining (e.g. Chen et al., 2010) and stock market predictions (e.g. Khedr et al., 2017), but it is also gaining research interest in typical UGC analysis, for instance to detect irony and sarcasm (e.g. Van Hee et al., 2018), and for analyzing newswire text.

Looking at the mere impact of news events on their audiences without having readers' reactions at hand, the focus of this research lies on detecting implicit sentiment rather than implied opinions. Irrespective of potential framing, when consuming news, readers may infer a positive or negative impression of an event or topic based on world knowledge, cultural background, historical context or even personal experiences. Such text spans are known as "statements or phrases that describe positive or negative factual information about something without conveying a private state" (Wilson, 2008, p. 2741). Later, Toprak et al. (2010) coined the term 'polar facts' to refer to such statements. In what follows, we discuss some seminal studies on sentiment analysis in factual text from both the author's and readers' perspectives.

2.1 Implicit sentiment analysis from the author's perspective

Balahur et al. (2010) performed sentiment analysis on quotations in English newswire text. They defined the sentiment of named entities in quotations by applying sentiment lexicons to varying context windows inside the quotes. Jiang et al. (2017) combined a clustering algorithm with lexicon-based sentiment analysis using SentiWordNet (Baccianella et al., 2010) at the sentence level to distinguish between positive and negative attitudes from UK news sources towards climate change-related topics. A similar methodology was applied by Burscher et al. (2016) to analyze the framing of the topic of nuclear power in English news articles. They found that within the frame of nuclear accidents or waste, articles were much more negative compared to articles that focused on the effects of nuclear power on climate change, or its economic aspects.

Nozza et al. (2017) presented a multi-view corpus enriched with different variations of sentiment annotations; including objective versus subjective labels, implicit versus explicit sentiment, emotion categories, irony annotations, and so on. While the study presents clear definitions of the categories, the accompanying corpus examples are rather confusing (e.g. with "Tonight @CinemaX #SuicideSquad!! Come to see #HarleyQuinn :)" as an example of an objective text and "I went out the cinema after 15 minutes #suicidesquad" as an example of an implied opinion). Low inter-rater agreement scores also confirm the difficulty to distinguish between implicit and explicit opinions. Chen and Chen (2016) explored implicit aspect-based sentiment analysis in Chinese hotel reviews following the premise that implicit opinion expressions are located nearby explicit opinions. Fang et al. (2020) proposed an aspect-based approach to implicit opinion analysis of Chinese car reviews. They applied similarity metrics and clustering algorithms to extract and categorize feature expressions and aggregated their implicit sentiment based on pointwise mutual information (PMI).

2.2 Implicit sentiment analysis from the readers' perspective

Henley et al. (2002) investigated framing effects on violence perception in news reporting of homophobic attacks. Apart from investigating author's perceptions, they performed a manual content analysis to investigate readers' viewpoints regarding the framing of events. It was shown that, for instance, more homophobic newspapers reported on violence against gay people more vaguely compared to violence against straight people, as a result of which the former incidents were perceived less harmful. Conversely, more neutral newspapers were found to report on all types of violence in the same manner. In 2007, a shared task was set up by Strapparava and Mihalcea (2007) focusing on valence and emotion classification of English newspaper headlines. The SemEval-2015 task on implicit sentiment detection of events (Russo et al., 2015) focused on predicting whether structured events (i.e. newspaper sentences containing the pattern "I—we + [verbal/nominal keyword]") are considered pleasant or unpleasant.

While most work has been done on English data, similar approaches to detect sentiment and emotions in news from the readers' perspective have been applied to Czech (Burget et al., 2011), Chinese (Lin et al., 2008) and Dutch (Atteveldt et al., 2008). Related research has also focused on sentiment analysis of named entities in news (Godbole et al., 2007) and sentiment analysis for fake news detection (Kula et al., 2020; Bhutani et al., 2019).

Most similar to the present research is the work by Atteveldt et al. (2008), who classify implicit sentiment evoked by Dutch news on national political elections with a classification accuracy of 0.56 F_1-score. To the best of our knowledge, they are the first to perform such sentiment classification at a more fine-grained level (i.e. considering entity relations, evaluations and performances), as opposed to the document or sentence level. However, their approach is limited in that only specific event structures are considered, and that the data are collected within one well-defined domain, i.e. political elections. Given that the ultimate goal of the present research is to detect sentiment in any kind of hard news, our corpus is not restricted to political events, but encompasses a wide variety of news topics. In addition, we aim to not only detect positively or negatively evoked sentiment, but also consider 'neutral' and 'conflict' as sentiment labels (see Table 4).

3 Corpus Construction

Striving for diversification driven by content analysis, our research focus is on fine-grained news events (see Section 1), and more specifically the implicit sentiment they evoke in the reader. In the following paragraphs, we zoom in on the data collection and annotation process and present the results of inter-rater-agreement experiments.

3.1 Data Collection and Preparation

We collected a large set of Dutch news articles from major Flemish newspapers published in 2017 and 2018[3]. As mentioned before, our focus was on collecting hard news. Moreover, all articles were reduced to the title and lead, which include the most relevant information as defined by the inverted-pyramid structure applied in journalism.

A first step in the annotation process involved the identification of text spans that present news events. This was done as part of an important effort to create a new Dutch dataset in this research area by Colruyt et al. (2020). Once identified, all news events were subsequently annotated for implicit sentiment (see Section 3.3.1). Since identifying the sentiment that is evoked by an isolated chunk of text is quite an arduous task, all events were presented to the annotators in their original context, being the news articles' titles and leads. In total, 1,735 articles were annotated with fine-grained news events and their implicit sentiment, as well as the sentiment triggers.

[3] The data were provided as JSON files by Mediahuis, a media company that publishes national and regional newspapers in Belgium, the Netherlands and Ireland.

3.2 Data Annotation

All annotations were executed in the web-based annotation tool WebAnno (Eckart de Castilho et al., 2016) and by making use of a novel annotation scheme for implicit sentiment evoked by news events (Van Hee et al., 2021). To sum up, news events are pieces of text that describe an event, situation or description that is newsworthy, i.e. that caused the reporter to write the article. In the first step of the annotation process, the annotators indicated the implicit sentiment evoked by each event (e.g. *in 2040 stevige opwarming aarde [EN: in 2040 robust increase in global warming]*). All events were assigned a sentiment label out of the following: 'positive', 'negative', 'neutral' and 'conflict'. Where 'positive' and 'negative' were used to mark events evoking a positive and negative sentiment in the reader, the 'neutral' label was used when no specific sentiment was elicited.

As the reception and evaluation of news events may largely depend on personal factors (e.g. sociocultural and historical background), we provided the annotators with an extra guideline stating that annotations should be made from a European/Western viewpoint. The annotators were instructed to use 'conflict' labels sparingly, and only in cases where an event's implicit sentiment was ambiguous or depended too heavily on the annotator's personal interests, background, ideology, etc.

Once an event was assigned a non-neutral sentiment, the annotators marked all words or word groups that are indicative of this sentiment. In the annotation scheme, such text spans are referred to as 'sentiment triggers', which have either a 'positive', 'negative', or 'conflict' sentiment and can be flagged as ironic if the annotator judges irony is involved. The challenge in annotating sentiment triggers resides in the fact that these are, given the data genre, no explicit subjectivity markers, but rather *polar facts* (see Section 2). Figure 1 shows an annotation example where events are linked to their sentiment triggers. Importantly, sentiment triggers can be, but are not necessarily, part of the event span and they can be non-consecutive spans.

3.3 Inter-annotator Agreement

An inter-annotator agreement study was conducted on a subset of the corpus to verify the consistency of sentiment annotations across the annotators and hence to substantiate the feasibility of annotating implicit sentiment evoked by newswire text. Forty randomly selected documents were reserved for this experiment, which were annotated by three annotators, independently from one another. The annotations were carried out after briefing and training the annotators for the task and before the remainder of the corpus was labeled so as to allow the guidelines to be revised or clarified where deemed necessary.

3.3.1 Implicit Sentiment of News Events

	Pos	**Neg**	**Neu**	**Conf**	**Total**
Rater 1	22	93	51	5	171
Rater 2	33	106	23	9	171
Rater 3	28	93	49	1	171
Average	27.67	97.33	41	5	171

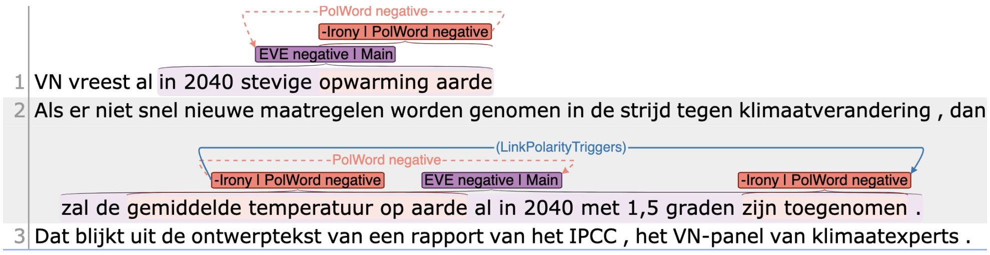
Table 1: Event distribution by rater and sentiment.

Tables 1 and 2 present the data distribution statistics and inter-rater agreement scores, respectively. It is clear from Table 1 that most events in the IAA set evoked a negative implicit sentiment. More specifically, on average 97 out of 171 of the events or 57% were attributed a negative sentiment and 28 or 16% a positive one. On average, 5 events, or 3% of the events, were attributed the 'conflict' label, meaning that the event's evoked sentiment depended too heavily on its broader context or on the annotator's personal viewpoints. The above reveals that more than 3 in 4 news events in the corpus evoke a sentiment in the reader and can hence be considered *polar facts*. By contrast, on average 41 or 24% of the events were annotated as neutral.

Inter-rater agreement scores were calculated using the cloud-based version of AgreeStat360[4], a software package for advanced statistical analysis of agreement among multiple raters (Gwet, 2014). The software allows to calculate the Krippendorff's Alpha (Krippendorff, 2011) with all of its weights and coefficients, including Fleiss' Kappa (Fleiss, 1971), used for multiple raters' agreement calculation and AC1 (Gwet, 2014), which is a variation of Kappa that corrects an expected agreement in skewed data distributions that is artificially high.

Table 2 presents agreement scores between the three raters in terms of defining individual news events' evoked sentiment. All metrics considered and following the interpreting guidelines by Landis and Koch (1977), we can conclude that the annotations show a high level of agreement.

[4] http://agreestat360.com.



1 VN vreest al in 2040 stevige opwarming aarde
2 Als er niet snel nieuwe maatregelen worden genomen in de strijd tegen klimaatverandering , dan zal de gemiddelde temperatuur op aarde al in 2040 met 1,5 graden zijn toegenomen .
3 Dat blijkt uit de ontwerptekst van een rapport van het IPCC , het VN-panel van klimaatexperts .

Figure 1: Pre-annotated events *"in 2040 robust increase in global warming"* and *"by 2040 the average temperature on earth will have risen by 1.5 degree"* are linked to their sentiment triggers *"robust increase in global warming"* and *"average temperature on earth"* + *"will have risen"*.

Method	Coeff	StdErr	95% C.I.	P-Value
AC 1	0.77	0.03	(0.71,0.83)	0.00e+00
Fleiss' Kappa	0.69	0.04	(0.62,0.76)	0.00e+00
Krippendorff's Alpha	0.69	0.04	(0.62,0.76)	0.00e+00
Percent Agreement	0.82	0.02	(0.77,0.86)	0.00e+00

Table 2: Inter-rater agreement coefficients for implicit sentiment annotation of news events.

3.3.2 Sentiment Triggers

Apart from annotating implicit sentiment evoked by news events, the annotators also marked in the same sentence all sentiment triggers that influenced their decision. Given the data genre, identifying such words was expected to be difficult because this depends on (i) the amount of context available and (ii) the extent to which an event has an intrinsic sentiment that humans are aware of by context or world knowledge.

No specific guidelines were defined for the annotation of sentiment triggers; they could be single words or phrases of any type of syntactic structure. Annotators were, however, asked to select the minimal word span. Calculating inter-rater agreement for sentiment triggers requires a strategy to align the text spans between the different annotators. Matching text spans between two outputs (whether they are human annotations or system predictions) is a familiar challenge in sentiment annotation and detection tasks, and especially known in the field of aspect-based sentiment analysis. Depending on the importance of exact span overlap, text spans can be evaluated by searching for an exact or a relaxed match at the start and ending token boundaries. In Lee and Sun (2019), an exact match imposes a 100% overlap between two text spans, whereas a relaxed match imposes that 1) either side of the boundaries is matched with at least one token or 2) at least one token overlaps between the spans.

As no detailed guidelines were provided for the syntactic composition of sentiment triggers, an exact span match evaluation would affect the inter-rater agreement too negatively. In fact, we looked at the 100% overlap ratio for sentiment triggers when annotated for a specific event and found that (one or more) sentiment trigger(s) were annotated for 148 events. For 38 events (26%), all annotators indicated the exact same sentiment triggers. For 12 events (8%), half of the sentiment triggers were identical amongst the three annotators. For 92 events (62%), there were no exact matches. For the remaining 6 events (4%), 1 out of 3 or 4 sentiment triggers were annotated by all three raters.

Partial matches were not taken into account for the above statistics. In a second and more detailed examination, we considered two annotations to match if at least one character index between the two text spans overlapped, regardless of matching boundaries. Table 3 shows these agreement results as F_1-scores per annotator pair, where the first rater mentioned served as the gold standard for the evaluation. As can be deduced from this table, with an average F_1-score of 0.72 over all events, the inter-rater agreement for sentiment triggers is quite high. It means that out of 10 sentiment triggers annotated by the gold standard 7 are also found by a second, independent rater.

	Raters 2 and 1	Raters 2 and 3	Raters 3 and 1
F_1-score	0.72	0.74	0.70
Average		0.72	

Table 3: Inter-rater scores for sentiment trigger annotation amongst three annotator pairs.

4 Corpus analysis and experiments

In the following paragraphs, we thoroughly investigate our research question by analyzing the full corpus and conducting experiments to examine whether detecting implicit sentiment evoked by news events is a feasible task.

4.1 Annotating Implicit Sentiment in News

Table 4 reveals the sentiment distribution in the full corpus, which contains 7,652 news events in total. We observe that, comparable to the results of the inter-rater experiment (see Section 3.3), most events have a negative sentiment or are considered neutral. Only 1 in 10 news events evokes a positive sentiment, and 4% were considered ambiguous.

	Event sentiment			
	Pos	Neg	Neu	Conf
# events	849	3,699	2,789	315
Percentage	11%	48%	36%	4%

Table 4: Event sentiment distribution in the full corpus.

A qualitative, manual analysis of the annotations was performed to gain more insights into the differences between neutral and non-neutral news events. This analysis revealed that words that occur more frequently in neutral events compared to positive and negative events are topical words that occur frequently in the news bulletin, like 'government', 'minister', 'European' and 'American'. Neutral events also more often contain time indicators such as 'Monday', 'last week', 'today' and verbs expressing locutionary acts (e.g. 'said', 'asks', 'communicated'), compared to non-neutral events. Negative events more often contain nouns and adjectives like 'murder', 'attack', 'shooting', 'war', 'famine', and verbs including 'judging', 'arrested' and 'wounded' than positive and neutral events. The noun 'increase' also occurs most frequently in negative events, mostly associated with terms like 'tension' or terms related to addiction and disease. Frequently occurring terms in positive events are more difficult to pinpoint at the word level, but it is observed that words like 'solution', 'approved' and 'new' occur more frequently in positive events compared to negative and neutral ones. An analysis of the conflict events revealed that often, these mention highly topical nouns and named entities like 'Brexit' (56 out of 315 events), 'Trump' (24/315), 'Catalonia' (23/315), 'referendum' (16/315), 'Jerusalem' (12/315) and 'nuclear exit' (10/315). These are all examples of concepts that evoke ambivalent feelings depending on the reader and on the broader context, hence the events they occur in were labeled as 'conflict'.

An analysis of the sentiment triggers (underlined in the examples) showed that they are mostly (>99% of the cases) included inside the event span, as shown in example 1. Interestingly, sentiment triggers outside of the event span (example 2) are often part of a subjective statement by the author or a quotation.

(1) [Brother of the presumed Giant of the Brabant Killers provides investigators with new tips]$_{\text{event}}$.

(2) [The billion-dollar takeover of 21st Century Fox]$_{\text{event}}$ creates a new major power.

4.2 Automatically Predicting Implicit Sentiment in News

Having a news article corpus in place, in which annotators differentiated between neutral events and events that evoke a particular sentiment, we were able to investigate the feasibility of implicit sentiment detection. Filtering out the doubles lead to an experimental corpus of 7,425 events, which was split in a training partition of 6,683 events and a test set of 742 events. The label distributions in both sets remained the same as in Table 4.

4.2.1 Lexicon-based Approach to Event Sentiment Detection

We first explored the effectiveness of two lexicon-based approaches to automatically determine implicit sentiment in news events. For the first approach, we relied on four sentiment lexicons for Dutch, including the Pattern lexicon (De Smedt and Daelemans, 2012) composed of 3,223 qualitative adjectives, an in-house sentiment lexicon with size n= 434 composed of manual review annotations, the Duoman lexicon (Jijkoun and Hofmann, 2009) composed of 8,757 wordforms and the NRC Hashtag Lexicon (Mohammad and Turney, 2013) including 13,683 entries[5]. All lexicons were manually checked to filter irrelevant entries. The order in which these lexicons were consulted was determined by preliminary experiments (i.e. when a word had no match in the Pattern lexicon, the next step was to consult the in-house lexicon, next Duoman and finally NRC).

For the second approach, we used SenticNet (Cambria and Hussain, 2015), an automatically constructed semantic knowledge resource based on common sense knowledge from the Open Mind Common Sense initiative (Singh et al., 2002) and

[5] The original lexicon of 14,182 unigrams, which had been automatically translated to Dutch, was manually filtered by a Translation student.

GECKA (Cambria et al., 2015), combined with affective knowledge from WordNet-Affect (Strapparava and Valitutti, 2004). SenticNet entries provide sentiment information for concepts of varying n-gram length, such as "accomplish goal", "celebrate special occasion", "be on cloud nine", etc. We considered it a potentially valuable resource for our task as it is not restricted to explicit sentiment terms, which are probably hard to find in newswire text. For our experiments, we made use of the SenticNet 5 API (Cambria et al., 2018), which returns sentiment values for the concepts it receives.

Table 5 presents the results of the lexicon-based sentiment analysis approaches. Overall, the scores are low, with a top F_1 score of 0.47 obtained with the four combined lexicons that outperformed SenticNet with 16%. Looking at the performance per class, we can conclude that the results are clearly better for the negative and neutral instances. Intuitively, we expected SenticNet to be better suited for the task, given the data genre and SenticNet's inclusion of implicit polar concepts. However, there are several hypotheses as to why it was outperformed by the other lexicons. Firstly, a qualitative analysis revealed that the coverage largely differs, with on average 3 or more matches per event for the regular lexicons, and only 1 for SenticNet. Secondly, all entries in the combined lexicons were manually verified, either by the authors of the lexica or by the authors of this paper, unlike SenticNet's entries, which are automatically collected from a small annotated seed set. Thirdly, as SenticNet contains concepts rather than words, all text needed to be pre-processed using a concept parser (Rajagopal et al., 2013)[6]. As such a parser is currently unavailable for Dutch, we decided to translate all events to English using Google Translate[7]. Automatic translation, however, means that some of the semantics may be lost, which may have affected the results of this approach.

4.2.2 Machine Learning Approach to Event Polarity Detection

Using machine learning, we investigated a feature-based and end-to-end architecture. For the feature-based approach, we applied Support Vector Machines using the LibSVM library (Chang and Lin, 2011). For the latter approach, we applied two state-of-the-art transformer-based architectures for Dutch, i.e. BERTje (Vries et al., 2019) and RobBERT (Delobelle et al., 2020). While both models are based on the BERT architecture originally released for English (Devlin et al., 2019), they were each pre-trained on different corpora. BERTje is pre-trained on a 12 GB Dutch corpus composed of different genres, including books, social media data, Wikipedia and -especially relevant for our task- newswire text. By contrast, RobBERT is based on the Dutch section of the OSCAR corpus (Ortiz Suárez et al., 2019), a 39 GB large subcorpus of the Common Crawl corpus[8], the largest web crawl corpus available. Although the latter is pre-trained on much more data, we expect BERTje to be better suited for the current task.

SVM parameter settings for the classifier and feature extraction were simultaneously optimized using a grid search in a nested cross-validation setup. For the classification algorithm, we varied the kernel type, cost and $gamma$ parameters and tested equal versus balanced class weighting. Regarding feature engineering, we varied the n-gram length and type (i.e. words versus characters and uni-/bi-/trigrams) and tested with a maximum feature threshold (i.e. None; 5,000; 10,000; 20,000). In both transformer setups, 3 epochs were defined with preliminary experiments. However, actual training did not even require that many epochs, as from epoch 1 (BERTje) and 2 (RobBERT) onwards, validation loss surpassed training loss, which may suggest overfitting. For all classifiers, the parameter settings and feature combinations that yielded the best results in the cross-validation experiments were used to train the final model that was subsequently applied to the held-out test set[9].

The results (Table 6) reveal that all three classifiers perform in a similar way, especially when considering the weight-averaged F_1 scores. Ranging between F_1= 0.69 and 0.72, the scores clearly outperform the combined sentiment lexicons approach and the majority baseline (predicting the negative class only). The SVM classifier seems to handle the underrepresented classes 'positive' and 'conflict' better than RobBERT and BERTje.

We also conducted a qualitative analysis, the results of which are presented in Table 7. Predictions for the first event suggest that the SVM predic-

[6] https://github.com/SenticNet/concept-parser.

[7] Translations done on 22/09/2020.

[8] https://commoncrawl.org/

[9] Best SVM parameters and features: linear kernel with cost C=10; balanced class weighting; 45,973 uni- and bigram word n-grams without threshold. Best settings for BERT: dropout: 0; sequence length: 128; learning rate (Adam): 5e-05; batch size 64; number of epochs: 3.

	Performance held-out test set			Performance per class			
Classifier	Accuracy	Macro-avg F_1	Weight.-avg F_1	F_1Pos	F_1Neg	F_1Neu	F_1Conf
Pattern+in-house +Duoman+NRC	**0.44**	**0.33**	**0.47**	**0.28**	**0.58**	**0.40**	0.04
SenticNet	0.26	0.21	0.31	0.15	0.45	0.20	0.04

Table 5: Scores obtained by the lexicon-based approaches on the held-out test set (n= 742).

	Performance held-out test set			Performance per class			
Classifier	Accuracy	Macro-avg F_1	Weight.-avg F_1	F_1Pos	F_1Neg	F_1Neu	F_1Conf
SVM	0.69	**0.61**	0.69	**0.49**	0.76	0.65	**0.53**
RobBERT	0.72	0.60	**0.72**	0.48	0.80	0.70	0.43
BERTje	**0.74**	0.54	**0.72**	0.48	**0.81**	**0.71**	0.15
Majority baseline	0.50	0.17	0.33	0.00	0.66	0.00	0.00

Table 6: Scores obtained by the machine learning approach on the held-out test set (n= 742).

Gold	Event	SVM	BERTje	RobBERT
negative	*In the night of July 14th, the day of national celebration, and after the country's World Cup win, 845 vehicles went up in flames.*	positive	negative	negative
conflict	*a speech Trump gave to the NRA gun lobby*	negative	neutral	neutral
negative	*in Antwerp they currently remain cautious*	positive	positive	positive
negative	*an inferno*	negative	neutral	negative
positive	*a deal*	negative	neutral	positive

Table 7: Qualitative analysis examples: gold label, event text and predictions by SVM, BERTje and RobBERT.

tion might be triggered by the positive words in the event, whereas they do not influence the predictions of the BERT models. The second example shows the difficulty of nested events, i.e. "RNA gun lobby" which is annotated as "conflict", but is nested inside the neutral event "a speech Trump gave (...)". Here as well, the SVM seems rather triggered by purely lexical items. The third example demonstrates the importance of context for accurate sentiment prediction at a more fine-grained level. The event's context is a proposition to make two Belgian ports work more closely together, which is welcomed by one party, but not by the port of Antwerp. The last two events ("een inferno" and "een deal" in Dutch) are examples of correct predictions by RobBERT while the predictions by BERTje are incorrect. An explanation could be that the web crawl data RobBERT is trained on is more likely to contain English terms, unlike the cleaner corpus at the basis of BERTje. Lastly, while some events are extensive in terms of context (example 1), others are more constrained, which complicates their prediction.

5 Conclusion and Future Work

With this paper, we investigated the detection of implicit sentiment evoked by Dutch newswire text. While related research approaches the task mainly at the document or sentence level using lexicon-based methods, we focused on fine-grained events below the sentence level and experimented with lexicon-based approaches and machine learning. For the latter, we compared the performance of SVMs, which have proven successful in sentiment analysis tasks, with two transformer-based models. Our results demonstrate that the machine learning approach performs accurately with a top F_1 score of 0.72 and shows a considerable improvement over the majority baseline. The experiments also demonstrate that machine learning clearly outperforms the lexicon-based approach, even when extensive (implicit) sentiment lexicons are used. Furthermore, we created and manually annotated a Dutch corpus of news events and were able to show high inter-rater agreement for event sentiment and sentiment span annotations. In future research, it will be interesting to explore whether additional context, including named entities and co-referring events, inside and across sentence boundaries, can improve implicit sentiment detection further.

Acknowledgments

We thank Siel Debouver and Annaïs Airapetian for their invaluable contribution to the corpus annotation and Bram Vanroy for his code contribution. This research was funded by the Special Research Fund of Ghent University and supported by #News-DNA.

References

Md Adnan, Mohammed Chowdury, Iftifar Taz, Tauqir Ahmed, and Mohammad Rahman. 2014. Content based news recommendation system based on fuzzy logic. In *International Conference on Informatics, Electronics and Vision (ICIEV 2014)*, pages 1–6.

Wouter Atteveldt, Jan Kleinnijenhuis, Nel Ruigrok, and Stefan Schlobach. 2008. Good news or bad news? conducting sentiment analysis on dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics*, 5:73–94.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

Alexandra Balahur, Raf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

Barkha Bansal and Sangeet Srivastava. 2018. On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science*, 135:346 – 353. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5.

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5:187–205.

Radim Burget, Jan Karasek, and Zdeněk Smékal. 2011. Recognition of emotions in czech newspaper headlines. *Radioengineering*, 20:39–47.

Bjorn Burscher, R. Vliegenthart, and C. D. Vreese. 2016. Frames beyond words. *Social Science Computer Review*, 34:530–545.

Erik Cambria and Amir Hussain. 2015. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*, 1st edition. Springer Publishing Company, Incorporated.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Erik Cambria, Dheeraj Rajagopal, Kenneth Kwok, and Jose Sepulveda. 2015. Gecka: Game engine for commonsense knowledge acquisition. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology Journal*, 2(3).

Bi Chen, Leilei Zhu, Daniel Kifer, and Dongwon Lee. 2010. What is an opinion about? exploring political standpoints using opinion scoring model. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1007–1012. AAAI Press.

Huan-Yuan Chen and Hsin-Hsi Chen. 2016. Implicit polarity and implicit aspect recognition in opinion mining. pages 20–25.

Shu-I Chiu and Kuo-Wei Hsu. 2018. Predicting political tendency of posts on facebook. In *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, ICSCA 2018, page 110–114, New York, NY, USA. Association for Computing Machinery.

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

Tom De Smedt and Walter Daelemans. 2012. "vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3568–3572, Istanbul, Turkey. ELRA.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model.

Mark Deuze. 2003. The web and its journalisms: Considering the consequences of different types of newsmedia online. *New Media & Society*, 5(2):203–230.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Zhao Fang, Qiang Zhang, Xiaoan Tang, Anning Wang, and Claude Baron. 2020. An implicit opinion analysis model based on feature-based implicit opinion patterns. *Artificial Intelligence Review*, 53:4547–4574.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *ICWSM 2007 - International Conference on Weblogs and Social Media*.

Kilem L. Gwet. 2014. *Handbook of Inter-Rater Reliability (Fourth Edition), The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, Gaithersburg, USA.

Nancy M. Henley, Michelle D. Miller, Jo Anne Beazley, Diane N. Nguyen, Dana Kaminsky, and Robert Sanders. 2002. Frequency and specificity of referents to violence in news reports of anti-gay attacks. *Discourse & Society*, 13(1):75–104.

Ye Jiang, Xingyi Song, Jackie Harrison, Shaun Quegan, and Diana Maynard. 2017. Comparing attitudes to climate change in the media using sentiment analysis based on Latent Dirichlet Allocation. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 25–30, Copenhagen, Denmark. ACL.

Valentin Jijkoun and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, page 398–405, USA. ACL.

Ayman Khedr, S.E. Salama, and Nagwa Yaseen. 2017. Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, 9:22–30.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Online: https://repository.upenn.edu/asc_papers/43/.

Sebastian Kula, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, and Michał Woźniak. 2020. Sentiment analysis for fake news detection by means of neural networks. In *Computational Science (ICCS 2020)*, pages 653–666, Cham, Switzerland. Springer International Publishing.

Richard J. Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Grace E. Lee and Aixin Sun. 2019. A study on agreement in pico span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1149–1152, New York, NY, USA. ACM.

Jiwei Li and Eduard Hovy. 2017. Reflections on sentiment/opinion analysis. In Erik Cambria, Das Dipankar, Sivaji Bandyopadhyay, and Antonio Feraco, editors, *A Practical Guide to Sentiment Analysis*, chapter 3, pages 41–61. Springer International Publishing AG, Cham, Switzerland.

K.H.-Y Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226.

Bing Liu. 2015. *Sentiment analysis: mining opinions, sentiments, and emotions*, 1st edition. New York: Cambridge University Press.

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 31–40, New York, NY, USA. ACM.

Eugenia Mitchelstein and Pablo Boczkowski. 2009. Between tradition and change: A review of recent research on online news production. *Journalism*, 10(5):562–586.

Saif Mohammad. 2017. Challenges in sentiment analysis. In Erik Cambria, Das Dipankar, Sivaji Bandyopadhyay, and Antonio Feraco, editors, *A Practical Guide to Sentiment Analysis*, chapter 4, pages 61–85. Springer International Publishing AG, Cham, Switzerland.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Eli Parser. 2013. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited, London, England.

Thomas E. Patterson. 2000. *Doing Well and Doing Good: How Soft News Are Shrinking the News Audience and Weakening Democracy*. Harvard University Press, Cambridge, MA.

Dheeraj Rajagopal, Erik Cambria, Daniel Olsher, and Kenneth Kwok. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pages 565–570.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. SemEval-2015 task 9: CLIPEval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450, Denver, Colorado. Association for Computational Linguistics.

Pamela Shoemaker and Akiba Cohen. 2005. News around the world: Content, practitioners, and the public. *News Around the World: Content, Practitioners, and the Public*, pages 1–409.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. ACL.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. ELRA.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. ACL.

Gaye Tuchman. 1973. Making news by doing work: Routinizing the unexpected. *American journal of Sociology*, 79(1):110–131.

Cynthia Van Hee, Orphée De Clercq, and Véronique Hoste. 2021. Guidelines for annotating implicit sentiment evoked by fine-grained news events (version 1.0). Technical report, LT3, Faculty of Arts, Humanities and Law, Ghent University (Ghent, Belgium).

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. We usually don't like going to the dentist : using common sense to detect irony on twitter. *Computational Linguistics*, 44(4):793–832.

Karel Vandendriessche and Lieven De Marez. 2019. imec.digimeter 2019: Measuring digital media trends in flanders. Online: https://www.imec.be/nl/expertises/imec-digimeter/digimeter-2019.

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

Theresa Wilson. 2008. Annotating subjective content in meetings. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2738–2745, Marrakech, Morocco. ELRA.