What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools

Boukes, M.; van de Velde, B.; Araujo, T.; Vliegenthart, R.

DOI
10.1080/19312458.2019.1671966

Publication date
2020

Document Version
Final published version

Published in
Communication Methods and Measures

COMMUNICATION METHODS AND MEASURES
2020, VOL. 14, NO. 2, 83–104
https://doi.org/10.1080/19312458.2019.1671966

Routledge
Taylor & Francis Group

What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools

Mark Boukes, Bob van de Velde, Theo Araujo, and Rens Vliegenthart

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, the Netherlands

ABSTRACT

This article scrutinizes the method of automated content analysis to measure the tone of news coverage. We compare a range of off-the-shelf sentiment analysis tools to manually coded economic news as well as examine the agreement between these dictionary approaches themselves. We assess the performance of five off-the-shelf sentiment analysis tools and two tailor-made dictionary-based approaches. The analyses result in five conclusions. First, there is little overlap between the off-the-shelf tools; causing wide divergence in terms of tone measurement. Second, there is no stronger overlap with manual coding for short texts (i.e., headlines) than for long texts (i.e., full articles). Third, an approach that combines individual dictionaries achieves a comparably good performance. Fourth, precision *may* increase to acceptable levels at higher levels of granularity. Fifth, performance of dictionary approaches depends more on the number of *relevant* keywords in the dictionary than on the number of valenced words as such; a small tailor-made lexicon was not inferior to large established dictionaries. Altogether, we conclude that off-the-shelf sentiment analysis tools are mostly unreliable and unsuitable for research purposes – at least in the context of Dutch economic news – and manual validation for the specific language, domain, and genre of the research project at hand is *always* warranted.

Automated content analysis of all kinds of texts has become usual practice in a variety of professional (Puschmann & Powell, 2018) and academic fields (González-Bailón & Paltoglou, 2015; Jongeling, Sarkar, Datta, & Serebrenik, 2017). It provides a cost-effective solution (Boumans & Trilling, 2016) and is particularly useful when analyzing large numbers of texts (Scharkow, 2017), which would demand major budgets if it had been done *manually*. The tone – also referred to as sentiment, valence, affect or textual polarity (Soroka, Young, & Balmas, 2015) – of a text is often considered a key characteristic for investigation. The various operationalizations of *automatically* measuring the tone of news coverage, however, differ widely. The current study provides an overview of off-the-shelf dictionary-based sentiment analysis tools[1] through which the tone of texts can automatically be assessed and compares them to two sets of manually coded news items: headlines and full texts of economic news. Thereby, we explore the agreement between different sentiment analysis techniques and how strongly (or weakly) they agree with human-coded news and among each other.

Surprisingly, the existing literature lacks such an explicit in-depth comparison, especially in non-English contexts. Although there are many sentiment analysis tools and datasets available for English language sentiment analysis (see, e.g., Ribeiro, Araújo, Gonçalves, Gonçalves, & Benevenuto, 2016), there are far fewer dictionary-based options for smaller languages (e.g., German, see Rauh, 2018; or Dutch, see Trilling & Boumans, 2018). Automatically coding the tone of texts in a small language

CONTACT Mark Boukes markboukes@gmail.com Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, 1018WV Amsterdam, the Netherlands

Supplemental data for this article can be accessed here.

comes with additional challenges. The different sentiment analysis tools are less extensively (or even not) validated compared to their English counterparts (Trilling & Boumans, 2018). Moreover, many languages are inherently more complex (e.g., in terms of grammar or sentence structure) than the relatively straightforward English language (Rauh, 2018; Rudkowsky et al., 2018).

Lacking alternatives, the few available tools which support a non-English language are frequently adopted for reasons of feasibility rather than validity. Whereas tools for the English language have been validated and verified at least to a certain extent (e.g., González-Bailón & Paltoglou, 2015; Ribeiro et al., 2016; Young & Soroka, 2012), this is far from true for most other languages. We will do so in the context of the Dutch language. Although perhaps relatively small by the absolute number of speakers, the Netherlands hosts one of the largest communication science communities worldwide: The Netherlands is the third largest country in terms of international publications (behind the U.S. and U.K., see Günther & Domahidi, 2016), and fourth largest country within the International Communication Association both in membership and conference attendees. The conclusions and recommendations of the current study, nevertheless, reach beyond the Dutch context and apply to any language for which multiple sentiment analysis tools are available. Economic news from the Netherlands is merely used as an example, but our approach can be replicated in any other language and/or domain.

We specifically focus on the measurement of tone in economic news as this is a frequent subject of automated analyses (Van Atteveldt, Kleinnijenhuis, Ruigrok, & Schlobach, 2008, p. 74, and see, e.g.; De Boef & Kellstedt, 2004; Hollanders & Vliegenthart, 2011; Soroka, 2012). The economic topic suits such automated approaches particularly well, because economic developments are, usually, uniformly understood as positive or negative. For example, rising unemployment levels will be understood by anyone as something negative, whereas increasing income levels would be understood positively. Compared to other topics, the economy is a straightforward issue and automated content analysis, hence, should perform better than on more ambiguous topics (e.g., politics).

Approaches to Measure the Sentiment in Economic News

Existing research that operationalizes the tone of economic news can roughly be divided into four approaches; all of these are top-down approaches in the sense that they do not include supervised machine learning techniques (i.e., specifically training an algorithm based on manually coded data) – and neither will our investigation. First, a considerable number of published works have manually coded the tone of economic news by trained coders (Boomgaarden, Van Spanje, Vliegenthart, & De Vreese, 2011; Boukes & Vliegenthart, 2017; Fogarty, 2005; Goidel & Langley, 1995; Goidel, Procopio, Terrell, & Wu, 2010; Hester & Gibson, 2003; Kalogeropoulos, Svensson, Van Dalen, de Vreese, & Albæk, 2015; Soroka, 2006). The coders read (or watch) a news item, and judge whether its tone is negative, neutral or positive. Most often, this is done on a 3-point scale, in other instances wider scales are used.

Second, some studies simply infer the tone of an article by the presence of one particular keyword. Blood and Phillips (1995), for example, simply counted the number of headlines that include "recession" in a certain period. Similarly, Wu, Stevenson, Chen, and Güner (2002) counted the number of articles that contained a reference to "recession" in either the headline or the lead of an article. This, arguably, oversimplifies the actual nuance that exists in news coverage (Fogarty, 2005): An article that writes the "recession is over" would still be counted as an indication of negative news. Moreover, many negative articles regarding the economy may *not* explicitly mention "recession" and instead use alternative terms to describe similar economic circumstances (e.g., unemployment, inflation, or crisis). This measurement, arguably, lacks both *precision* (are the detected articles really negative?) and *recall* (does it detect all the negative articles?).

A third approach is to infer the tone of economic news by assessing the presence of *multiple* keywords. One study, for example, counted the number of articles that referred to *at least one* of the multiple negative economic developments (i.e., recession, economic crisis, shrinking economy,

economic down turn or fall) (Hollanders & Vliegenthart, 2011). Expanding the list of words that Kleinnijenhuis, Schultz, Oegema, and Van Atteveldt (2013) used to analyze business news about financial institutions, Damstra and Boukes (2018) used a tailor-made list of 65 words particularly focusing on *the sentiment* within economic news. Thereby, their operationalization distinguished between hope-related words (e.g., hope, confidence, enthusiasm, inspiration, relief, rescue, and recovery) and fear-related words (e.g., fear, shock, panic, danger, worry, stress, tension, and anxiety).[2] Including words as these (e.g., emotions, mental states) rather than specific economic terms in their dictionary, Damstra and Boukes (2018) captured the general sentiment in economic news rather than the presence of negative economic terms that may be negated in the text (e.g., "the recession is finally over").

Fourth, several studies have applied dictionaries to automatically measure the tone of economic news (e.g., De Boef & Kellstedt, 2004; Shapiro, Sudhof, & Wilson, 2019; Soroka, 2012; Tetlock, 2007; Van Dalen, de Vreese, & Albæk, 2017). Their procedures were straightforward and relied on so-called "bag-of-words" approaches (see Scharkow, 2017): Counting the number of words in a text that are categorized as positive in a pre-established dictionary; counting the number of negative words in the same text; and eventually subtracting these from each other (see Young & Soroka, 2012, for detailed explanation) – mostly without taking the syntactic structure of sentences into account.

Automatic Measurements of Sentiment

Sentiment analysis tools are often specialized: Their performance depends on the *domain* and *genre* for which they were created (see, e.g., Ribeiro et al., 2016; and the discussion in; Van Atteveldt & Peng, 2018). Words such as "good" and "bad" may seem stable regardless of context; yet, the valence of "lie" or "cool" depends on their context (Rauh, 2018; Van Atteveldt et al., 2008; Young & Soroka, 2012). For sentiment analysis tools to function well, they must recognize the appropriate valence of words and word-combinations in the particular domain that they are employed and avoid importing valence judgments that *do not* transfer between domains (Muddiman, McGregor, & Stroud, 2019).

Quite often, the *domain* in which a sentiment analysis tool is developed is within the context of easy-to-obtain datasets, such as movie, product or restaurant reviews (for an overview, see Medhat, Hassan, & Korashy, 2014). Although reviews may be a valuable resource to validate automatic assessments of tone – reviews are often straightforwardly negative or positive and focused on one topic – the question is how well this translates to less specific domains (e.g., the financial one, see Loughran & McDonald, 2011). For instance, news articles are usually longer and less focused on one single subject compared to reviews, which accordingly complicates automatically coding its sentiment (González-Bailón & Paltoglou, 2015). The obvious downside of using freely available datasets for the validation of automated sentiment analysis tools, hence, is the lack of insight whether their dictionaries will function well within alternative domains.

The *genre* of messages that are used in the creation and evaluation of sentiment analysis tools might constrain the applicability to other genres. Sentiment analysis tools are generally constructed based on datasets of clearly informal, user-generated and short texts (e.g., movie reviews, tweets, forum comments), and may therefore yield less accurate results when applied to texts written in different styles (e.g., more formal and longer news articles with a less narrow scope). The *genre* (i.e., text properties, such as function and style, that are unrelated to the domain) has been found to have an even stronger impact on algorithm accuracy than a text's topic (e.g., sports, politics, and science) (Van der Wees, Bisazza, Weerkamp, & Monz, 2015). As many sentiment analysis tools, such as *SentiStrength* and *Pattern*, are built specifically for texts written by a lay audience (i.e., genre), one should question these tools' reliability and validity when automatically coding the tone of journalistically produced news articles.

In this paper, we examine commonly applied sentiment analysis tools tailored toward the Dutch language: (1) the Dutch translation for the *LIWC* (Pennebaker, Boyd, Jordan, & Blackburn, 2015; Dutch translation, see Zijlstra, Van Meerveld, Van Middendorp, Pennebaker, & Geenen, 2004) that

has been developed to detect emotional, cognitive, and linguistic dimensions of texts; two more complex "contextual" rule-based dictionaries, i.e. (2) *SentiStrength*[3] and (3) the *Pattern* library[4]; (4) the relatively new induced dictionary *Polyglot*[5]; and (5) the translated version of the *ANEW* word list (abbreviated as DANEW, see Moors et al., 2013).

LIWC

Linguistic Inquiry and Word Count (*LIWC*) aims at analyzing texts and detecting emotional, social, cognitive words as well as standard linguistic dimensions (e.g., usage of pronouns, numbers, etc.) of texts. The original versions of *LIWC* were developed as part of a project studying language and disclosure (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), and these versions were later expanded and translated (Pennebaker et al., 2015). The Dutch version, a translation from the English original (Zijlstra et al., 2004), includes a dictionary of 6,568 words distributed across 66 categories. Among the categories provided by *LIWC*, positive (e.g., happy, grateful) and negative (e.g., despair, sadness) emotion scores (reported as the percentage of total words in the text) were used to operationalize the sentiment analysis for this paper. According to the authors of the translated version, 89% of the word categories show high to very high correlations between English and Dutch. *LIWC* is, in contrast to the other sentiment analysis tools tested in this manuscript, not freely available.

SentiStrength

SentiStrength was developed as an extension of (a) the *LIWC* lexicon, and (b) the General Inquirer list of sentiment words, with (c) ad-hoc additions that developers came across when testing the tool (Thelwall, 2013). In *SentiStrength*, words are stemmed ("amazing" becomes "amaz*") and validated on a corpus of 2,600 social media comments originating from *MySpace*. The task of *SentiStrength* is short-text classification, meaning that words have individual sentiment values, which are modified by context features, such as adjectives ("very"), negations ("not good"), and punctuation ("!"). Scores are assigned on the level of an entire (short) text, with separate scores ranging between −5 and −1 for negativity and +1 to +5 for positivity based on the strongest present sentiment combination for that orientation. The results are often summed to get the overall text score. *SentiStrength* is not based on automatically learned word combinations, instead operating on fixed rules formulated by experts rather than statistics (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). The Dutch version is a translation of the English *SentiStrength* algorithm and has not been specifically validated for Dutch texts.

Pattern

The *Pattern* package was built as a "full pipeline" tool and includes functionality for web-retrieval, text-analysis, and prediction (De Smedt & Daelemans, 2012). Eventually, the sentiment analysis functionality of *Pattern* overshadowed its data collection aspects. *Pattern*'s lexicon was built on a labeled corpus of 14,000 online Dutch book reviews retrieved from the popular website *Bol.com*. Additional words were inductively added by comparing word-distributions in the positive and negative texts using *K*-nearest neighbors. *Pattern* supports negations, adjectives, and punctuation as modifiers of sentiment value.

Polyglot

The Dutch sentiment functionality of *Polyglot* was created using an "induced" lexicon approach (Chen & Skiena, 2014). The starting point for the non-English lexicons lies in a knowledge-graph based dictionary expansion. Concretely, *Polyglot* draws from a hand-built and structured dataset (the

knowledge) that, similar to *Wikipedia*, contains word information in the form of links between words (hence the *graph*). Here, words are connected as synonyms, antonyms, translations, and transliterations. Starting with known English sentiment words, sentiment is assigned through the links that words have to their Dutch counterparts and their related words. This approach does *not* take word context into account. This means negations or modifiers are neglected when generating scores. *Polyglot* includes the 100,000 most frequent words in a language (Rudkowsky et al., 2018).

DANEW

Affective Norms for English Words (*ANEW*) is a psychometric lexicon that indexes the valence, arousal, and dominance of words (Bradley & Lang, 1999). The three-dimensional representation of emotional responses to words is based on the factor decomposition of judgments on verbal queues (Osgood, Suci, & Tannenbaum, 1957) showing that these three dimensions account for variance in responses. *DANEW* is the Dutch version of *ANEW*, compiled by administering a translated questionnaire among 224 students in Flanders and the Netherlands and covers 4,300 Dutch words. The value of words in this lexicon is continuous, based on the questionnaire responses on 7-point Likert scales (Moors et al., 2013): Respondents ranked words from "very negative/unpleasant" to "very positive/pleasant" (valence/pleasantness factor), from "very passive/calm" to "very active/aroused" (activity/arousal factor) as well as from "very weak/submissive" to "very strong/dominant" (power/dominance factor). It is important to note that these words were presented out-of-context to respondents and each word was coded by 16 different assessors (eight female, eight male).

Comparing Automatic Measurements of Sentiment

The sentiment analysis packages differ in approach from multi-annotator metrics (*DANEW*) to iteratively build lexicons without (*LIWC*) or with (*SentiStrength*) combination rules, machine-extended existing dictionaries with (*Polyglot*) and without (*Pattern*) machine learning approaches. Genres of messages used to create the tools also vary, from general word impressions of citizens (*DANEW*) to reviews (*Pattern*), social media posts (*SentiStrength*), and knowledge bases (*Polyglot*). Hence, the quality of tone measurements may diverge widely. This could be either due to the domain and genre for which they were constructed and eventually applied to (González-Bailón & Paltoglou, 2015; Jongeling et al., 2017) or the dimensions of textual polarity that they focus on (Soroka et al., 2015). Concretely, the reason could be that "dictionaries (…) show stunningly little overlap, and where they do overlap codes are often discrepant" (Young & Soroka, 2012, p. 211). Most importantly, none of these sentiment packages explicitly specifies a domain to which it would apply; thus, making generalizability and validity in other contexts uncertain. We assess their performance in the context of economic news.

On the one hand, economic news should be relatively straightforward. This is a well-edited domain, less prone to typos and slang than reviews or social media messages. Moreover, sentiment seems relatively unambiguous because a "good" or "bad" economy is a fairly uncontested value compared to other topics in the news. On the other hand, economic news may be jargon-laden; thereby, potentially signaling valence with economy-specific jargon not included in off-the-shelf dictionaries (e.g., potentially words as "recession", "downturn" or "stagflation" may not be part of the lexicon). Moreover, certain words generally understood to be negative, may not be negative in the economic context and, thus, introduce measurement noise (Loughran & McDonald, 2011). In addition, the sentiment surrounding specific companies, stocks or trades might be easily conflated with performance of the economy in general.

Young and Soroka (2012) found that different sentiment analysis tools for the English language not only yield different levels of performance in terms of correlations with human coding, they also correlate rather weakly amongst each other (at average: $r = .33$). For the (largely) unvalidated Dutch tools, this could arguably even be worse. Interestingly, tools with more words in their respective dictionary do not necessarily perform better than those with smaller dictionaries (González-Bailón & Paltoglou, 2015; Rauh, 2018; Young

& Soroka, 2012). The reason is that not only recall matters, but that precision is vitally important too: Dictionaries that are too large will easily detect valence in words that are irrelevant to the topic of interest.

We assess the performance of automated tone dictionaries by analyzing their agreement with a human-coded tone. Moreover, we assess their recall (no false negatives, i.e., ability to identify *all* items that belong to a category), precision (no false positives, i.e., how often are the predictions for a specific category *correct*) and F_1-scores (i.e., balancing precision and recall in a harmonic mean, see Chinchor, 1992). To investigate how well different automated measurement instruments for sentiment analysis perform, but also and how strongly they actually agree with each other (or not, see Shapiro et al., 2019), we will answer the following research questions:

RQ_1: *Which automatic measurements of tone achieve the best performance?*

RQ_2: *How strongly do automatic measurements of tone agree amongst each other?*

Arguably, a more accurate automated measurement instrument for sentiment analysis can be created by combining and standardizing already existing sentiment analysis tools for a certain language (Rauh, 2018; Young & Soroka, 2012). In machine learning, *ensemble approaches* that combine multiple classifiers have been shown to outperform individual classifiers (Burnap & Williams, 2014; Dietterich, 2000; Tuarob, Tucker, Salathe, & Ram, 2014). To test this within the realm of off-the-shelf dictionaries, we analyze the added value of combining the scores of the individual automatic measurements that performed relatively well on themselves. Consistent with general measurement theory, more reliable measurements should be yielded when multiple measures *of the same construct* are combined (Gonçalves, Araújo, Benevenuto, & Cha, 2013). However, this will only occur *if* they indeed measure the same underlying component (Scharkow, 2017), which is likely not always the case for off-the-shelf sentiment analysis tools (Soroka et al., 2015). As a combination of tone measurements potentially corrects for the weaknesses and unreliability of individual tools, we expect the following:

H_1: *A combined approach will more strongly agree with human coding than the individual automatic measurements.*

In news stories, the headline can easily be distinguished from the body of an article. In many newsrooms, these headlines are even written by different journalists – reporters are responsible for the full text, whereas copy editors write the headline. Three reasons can be given to expect that automatic sentiment analysis tools may perform better on short texts as the headlines than on long texts as the full body of a news article. First, most of the sentiment dictionaries were *developed* based on short texts (e.g., social media posts) or just individual words (*DANEW*). Second and related to this, most of these tools were *engineered* and validated on messages that had a clear focus or topic (e.g., reviews). Headlines of articles about economic topics will mostly refer to how these are performing (i.e., focus) – positively or negatively – and not refer to other topics. Full texts, on the other hand, will be more nuanced (i.e., balancing both sides of a story) and may contain a majority of words that only relate to this topic indirectly. Accordingly, the tone of coverage regarding a specific subject (i.e., the economy) does not necessarily equal that of the overall textual polarity of the full article. And third, headlines are used to grasp the attention of the audience. Accordingly, a *condensed value* can be expected in these short texts – headlines normally have a clear focus and will explicitly mention positive or negative aspects. Therefore, we expect:

H_2: *Automatic measurements of tone in headlines agree more strongly with hand-coded tone assessments than automatic measurements of tone in the full body of a news article.*

Automated sentiment analyses are often part of longitudinal studies. Regarding economic news, for example, studies were interested in how the tone of coverage developed in periods of multiple years (e.g., Shapiro et al., 2019). In such studies, the granularity of analysis is generally coarser than the individual

article. Researchers care more about daily, weekly or even monthly averages of the tone in (economic) news than about the sentiment of an individual article. With such aggregation of data, errors in individual articles may cancel each other out (Van Atteveldt et al., 2008) and "inbuilt neutrality bias" (i.e., texts contain many more unique words than present in a dictionary) can be avoided (Rauh, 2018, p. 8). Whereas the coding of individual items may be susceptible to the unreliability of automated approaches, taking the average coding of larger sets of items should be more precise. This could make automated sentiment analysis more precise on higher levels of aggregation (Van Atteveldt et al., 2008) and leads to the following expectation:

H_3: *The correlations between automatic measurements of tone and manually coded tone assessments will be stronger at higher levels of data aggregation.*

Method

This study relies on a combination of content analysis methods. First, a manual content analysis has been conducted on two large sets of economic news texts (i.e., newspaper articles and news website items) of which the headlines and full bodies of text have been coded separately. After that, automatic sentiment analyses are run on the same sets of data.

Manual Sentiment Analysis

The content analysis – part of a broader research project (see, e.g., Boukes & Vliegenthart, 2017; Boukes, Damstra, Vliegenthart, 2019) – was performed on the economic news coverage of a variety of print and online news outlets. A team of 22 student assistants (i.e., coders) has analyzed news articles published on Mondays through Saturdays (i.e., Sundays are excluded in the current analysis due to the absence of newspapers) between February 1 and July 7, 2015. This period was characterized by early signs of economic recovery after the severe crisis commencing in 2008. As economic recovery was still uncertain but became obvious throughout the year, the data captured a period with varying economic states.

Using a search string that focuses on economic news,[6] newspapers articles were downloaded via *LexisNexis* and stored in the *Amsterdam Content Analysis Toolkit* (AmCAT, see Van Atteveldt, Ruigrok, Takens, & Jacobi, 2014). Ten newspaper outlets were coded, including four quality outlets (*Volkskrant, NRC, Trouw, Financieel Dagblad*), three popular newspapers (*Telegraaf, AD, Metro*) and three regional ones (*Noordhollands Dagblad, Dagblad van het Noorden, Gelderlander*). Additionally, economic news from five websites (popular: *NU.nl, Telegraaf.nl*; quality: *NOS.nl, Volkskrant.nl*, and *NRC.nl*) was collected and stored within the *Infrastructure for Content Analysis*-tool (INCA, see Trilling et al., 2018) and also coded for its tone. Whereas all newspaper articles (and Nu.nl's) were manually analyzed, a subset of 25% randomly selected website items were manually analyzed due to budgetary and time constraints. Accordingly, a larger number of news items was analyzed for the printed news media (n = 4,845) than for the news websites (n = 961).

Coders were explicitly instructed to assess the tone of an article's headline *before* they read the article's text; so, the full text's coding could not influence the judgment of the headline. The coding instruction read as follows (translated from Dutch): "What is the tone of the headline with regards to the economy?" It, thus, asked for a general evaluation of an economic topic and *not* of the text as a whole. This is in accordance with the operationalizations applied in previous research, which used the text of news articles as an indication for how well or badly the economy was performing. The following options (with already recoded values) were given to measure the tone of a headline: (−1) negative; (0) neutral, no tone, or mixed positive and negative, uncertainty, or ambiguity; and (+1) positive. Headlines that did not refer to the economy were set to missing (no tone present). This resulted in a set of negative, neutral, and positive headlines (see Table 1), in which a slight majority was neutral toward the economy. The balance between positive and negative news is similar in newspaper articles compared to news website items.

Table 1. Overview of tone measurements of headlines per medium.

Tone	Newspaper		News websites		Total	
−1	1,280	(33%)	244	(31%)	1,524	(33%)
0	1,640	(43%)	368	(47%)	2,008	(43%)
+1	935	(24%)	173	(22%)	1,108	(24%)
Total	3,855		785		4,640	

One central element within the overall coding process was the tone with which the national economy was described. The codebook, therefore, included the following question: "Is the current economic situation of the Netherlands (or one of its regions) assessed in a normative sense? If so, how?" Coders answered the question on the following scale: (0) No, the economy is not evaluated in a normative (good, bad) sense, (1) negative (e.g., the economic situations worsens/are bad, unemployment is [too] high), (2) mixed but mainly negative, (3) neutral (not good, not bad), (4) mixed (negative as well as positive developments), (5) mixed but mainly positive, (6) positive (e.g., the economy recovers, unemployment is decreasing, economic growth is expected). This scale was recoded into a −2 to +2 scale, where the articles that did not evaluate the economy (score 0) were set to missing. Including these in the analysis would imply that the automated sentiment analysis tools were searching for a tone that did not exist, which would make the comparison between manually and automatically coded news invalid. Table 2 gives an overview of the full-text data, which again found a similar pattern for print and online news.

Intercoder-reliability tests were performed using Nogrod 1.1 (Wettstein, 2018) on a subset of 148 articles (both print and online) that were analyzed by *at least* 3 of our coders. This resulted in a dataset of 802 articles (the average article was coded by 5.63 coders). Intercoder reliability statistics showed that agreement among human coders was acceptable. Both the tone of the headline (Krippendorff's $\alpha = .80$) and the tone of the full texts (Krippendorff's $\alpha = .69$) were assessed reliably in the manual content analysis. Thus, human coders mostly agreed with each other about the tone in the texts.

Automatic Sentiment Analysis

Automated sentiment analysis was done in a Python 3 environment with the exception of *LIWC*, which was run as a standalone program. *Pattern* and *Polyglot* are both available as Python packages, and both do their own tokenization. A custom script was implemented to calculate scores based on the tailor-made approach of Damstra and Boukes (2018; presence of 30 positive and 35 negative keywords). *DANEW* was implemented by binarizing word-sentiment scores based on above and below mean values and the score was calculated as the number of positive minus the number of negative scores (similar to Damstra & Boukes' specification). In both cases, words were lowered and tokenized by splitting on all non-alphanumeric characters. *SentiStrength* was used through a custom wrapper around the JAVA distribution, again relying on the tokenizer embedded in *SentiStrength* itself. *LIWC*'s positive emotion and negative emotion scores were summed (as more negative is expressed in more negative numbers). Eventually, the combined approach was created by computing the mean score of four of the five off-the-shelf tools (i.e., thus not the recession and Damstra &

Table 2. Overview of tone measurements in full texts per medium.

Tone	Newspaper		News websites		Total	
−2	265	(23%)	61	(24%)	326	(23%)
−1	225	(19%)	43	(17%)	268	(19%)
0	120	(10%)	10	(4%)	130	(9%)
+1	274	(23%)	64	(25%)	338	(24%)
+2	287	(25%)	77	(30%)	364	(26%)
Total	1171		255		1426	

Boukes' dictionaries); *Pattern* was excluded, because it generally performed very poorly (see Results, and Shapiro et al., 2019, for similar approach).

All tools (except for the binary measure of "recession" presence) provided continuous or ordinal estimates for the sentiment or valence of each text in the form of a numeric output. To make a valid comparison between approaches, all sentiment variables were standardized in a manner appropriate for the respective analysis: *Z*-score scaling (mean = 0; standard deviation = 1) was applied, so all tone scores were measured on a comparable scale. Subsequently, the *Z*-score scaling values of each tool were trinarized (−1, 0, +1) using 0.5 as a threshold (values below −0.5 were considered −1, values above 0.5 were considered +1, and values between −0.5 and 0.5 were considered as being 0) to allow F_1-comparisons and the calculation of Krippendorff's α.[7] Supplementary Files A and B provide an overview of the data using jittered scatterplots. The complete dataverse has been made accessible via https://github.com/uvacw/EconomicNews_sentiment_CMM

Results

Performance of the Different Off-The-Shelf Sentiment Analysis Tools

Within the genre of news and domain of the economy, we find that none of the automated tone measurements achieved acceptable reliability levels to detect the tone in economic news when following the thresholds that Krippendorff (2004, p. 241) recommended: i.e., α ≥ .800 or at least α ≥ .667 for tentative conclusions. The automated sentiment analysis tools with the relative best performance (but still unsatisfactorily low) for the headlines also achieve the most agreement with the manual coding of the full texts (see Table 3). Regarding headlines, we find the highest Krippendorff's α-values for *Polyglot* (α = .25) and *LIWC* (α = .23), with *DANEW* following closely (α = .22). The performance of the frequently used measure *SentiStrength* (α = .18) and the less popular *Pattern* (α = .17) is even worse. However, the worst findings are yielded for the small word list approaches: Recession (one word; α = − .01) and Damstra & Boukes (65 words; α = .07).

Whereas the simple approach of only searching for "recession" performed badly overall (i.e., negative α-values), it is interesting to see that the tailor-made dictionary of Damstra and Boukes (2018) performs slightly better for the full text (α = .25). Albeit 0.10-points less than *LIWC*, it follows *Polyglot* very closely (α = .26) and has a considerably higher α-score than the other three off-the-shelf tools. This finding shows that word-context is important and *adequate* lexicon size vital to achieve relatively more valid measurements. *SentiStrength* (α = .16) and *DANEW* (α = .15) perform poorly, but *Pattern* in particular yields the least reliable results for the sentiment in full texts (α = .09).

These findings for headlines and full texts are replicated in analyses that use *correlations* instead of alphas; these results can be found in the analysis of different levels of granularity (Table 6). The agreement between manual coding and automated sentiment analysis tools in detecting the tone of economic news, thus, is unacceptably low following the standards normally imposed on manual content analysis. Answering Research Question 1: The best results are found for *LIWC* and *Polyglot*

Table 3. Agreement between human and automated sentiment scores for headlines vs. full texts.

	Headline	Full Text	*Significant difference*
Recession	−0.010 [−0.017 − 0.005]	−0.024 [−0.046, −0.003]	No
Damstra and Boukes (2018)	0.073 [0.055, 0.092]	0.247 [0.205, 0.287]	Yes
LIWC	0.226 [0.200, 0.249]	0.349 [0.301, 0.394]	Yes
SentiStrength	0.179 [0.152, 0.209]	0.155 [0.103, 0.202]	No
Pattern	0.169 [0.140, 0.195]	0.087 [0.036, 0.136]	Yes
Polyglot	0.250 [0.222, 0.276]	0.258 [0.211, 0.304]	No
DANEW	0.218 [0.191, 0.243]	0.154 [0.104, 0.205]	Yes
Combined approach	0.298 [0.270, 0.324]	0.315 [0.268, 0.360]	No

Agreement-based predictions for each tool for negative, neutral or positive categories. Krippendorff's α informed for ordinal variable. The 95% confidence intervals were calculated using *R*-script as proposed by Zapf, Castell, Morawietz, and Karch (2016).

(and Damstra & Boukes for the full texts), but also their overlap with human coders signals a too low level of reliability (α < .35, where ≥ .67 would be required for tentative conclusions).

This brings us to Research Question 2: How strongly do automatic measurements of tone agree amongst each other? Table 4 displays the agreement in terms of Krippendorff's α-scores between the different automated sentiment analysis tools for both the headlines (above) and full-text bodies (below); Appendix A shows that the results are very similar when using pair-wise correlation analysis. One will notice that the agreement between automated approaches, overall, is surprisingly low – yet, they all (except for "recession" on full texts) point in the positive direction. It is remarkable that most automatic approaches yield largely unrelated measurements of tone. The overlap *between* these methods is weak in all instances. The only methods that relatively agree the most are *LIWC, Polyglot*, and *DANEW* (.28 ≤ *α* ≤ .34 for the headline; .30 ≤ *α* ≤ .36 for the full text), whereas the other combinations mostly yield completely unrelated tone assessments between each other. This implies that sentiment tools often assess the tone of articles differently, most likely due to non-overlapping lexicons that underlie their classification, and different contexts and applications for which they were originally developed. Looking at correlations instead of α-values, we find very comparable results (see Appendix A).

Remarkable is the lack of agreement between most approaches. For example, *SentiStrength* and *Pattern* (α = .09 for headline, and α = .07 for full text) find almost completely unrelated sentiment values, and similarly so for *SentiStrength* and *LIWC* (α = .19 for headline, and α = .17 for full text). Interesting, moreover, is the comparably high (but still low) agreement between *LIWC* and the 65-

Table 4. Agreement between sentiment scores using different methods for headlines (above) and full texts (below).

	Headline						
	Manual coding	Recession	D & B	LIWC	SentiStrength	Pattern	Polyglot
Recession	−0.01 [−0.02, −0.01]						
Damstra & Boukes	0.07 [0.06, 0.09]	−0.00 [−0.0, −0.00]					
LIWC	0.23 [0.20, 0.25]	−0.00 [−0.01, 0.00]	0.18 [0.14, 0.21]				
SentiStrength	0.18 [0.15, 0.21]	−0.05 [−0.06, −0.04]	0.02 [−0.01, 0.04]	0.19 [0.15, 0.22]			
Pattern	0.17 [0.14, 0.20]	−0.00 [−0.01, 0.00]	0.02 [−0.00, 0.04]	0.27 [0.24, 0.30]	0.09 [0.06, 0.12]		
Polyglot	0.25 [0.22, 0.28]	0.00 [−0.00, 0.00]	0.09 [0.07, 0.10]	0.28 [0.26, 0.31]	0.25 [0.23, 0.28]	0.23 [0.20, 0.26]	
DANEW	0.22 [0.19, 0.24]	−0.0 [−0.01, 0.01]	0.05 [0.03, 0.07]	0.33 [0.31, 0.36]	0.23 [0.19, 0.26]	0.23 [0.20, 0.26]	0.34 [0.32, 0.37]
	Full text						
	Manual coding	Recession	D & B	LIWC	SentiStrength	Pattern	Polyglot
Recession	−0.02 [−0.05, −0.00]						
Damstra & Boukes	0.25 [0.20, 0.29]	−0.05 [−0.09, −0.00]					
LIWC	0.35 [0.30, 0.39]	0.01 [−0.01, 0.03]	0.27 [0.22, 0.32]				
SentiStrength	0.15 [0.10, 0.20]	−0.07 [−0.10, −0.05]	0.14 [0.09, 0.19]	0.17 [0.12, 0.22]			
Pattern	0.09 [0.04, 0.14]	−0.01 [−0.03, 0.02]	0.03 [−0.02, 0.08]	0.24 [0.19, 0.28]	0.07 [0.02, 0.13]		
Polyglot	0.26 [0.21, 0.30]	0.03 [0.01, 0.05]	0.20 [0.15, 0.24]	0.36 [0.32, 0.41]	0.15 [0.10, 0.20]	0.26 [0.22, 0.32]	
DANEW	0.15 [0.01, 0.20]	0.02 [−0.00, 0.05]	0.08 [0.03, 0.13]	0.33 [0.28, 0.37]	0.15 [0.10, 0.20]	0.28 [0.23, 0.32]	0.30 [0.25, 0.34]

Agreement-based predictions for each tool for negative, neutral or positive categories. Krippendorff's α informed for ordinal variable. The 95% confidence intervals were calculated using *R*-script as proposed by Zapf et al. (2016).

words dictionary of Damstra & Boukes (α = .18 for headline, and α = .27 for full text), whereas the latter is mostly unrelated to the other tools (most notably *Pattern* and *DANEW*, α-values ≤ .09).

Recall, Precision and F_1-scores

In Table 5, we present performance metrics of the sentiment analysis tools regarding recall and precision of the three specific classes within the "true" human annotator labels (i.e., negative, neutral, or positive tone). The weighted F_1-scores express the harmonic mean of precision (proportion of predictions made for this class that was correct) and recall (of all documents known to belong to this class, which proportion was identified), as such F_1 best captures overall performance in a class. These results help the interpretation of why certain automated approaches agree more strongly with the human-coded data than others. Hence, we can compare whether the dictionaries perform better or worse on articles of a positive or negative tone about the economy. We only show the results for the headlines of articles, as these were already manually divided into three classes (i.e., tone of full texts was coded on 5-point scale) and, have a higher intercoder reliability (amongst human coders), which makes these scores the most interesting. Appendix B presents the results for full texts.

It is important to note the strong differences in performance and ranking between tone groups. The best performing measurement is printed in bold in Table 5. Overall (i.e. irrespective of tone), the results show that *Polyglot* and *DANEW* perform the best in terms of its F_1-score (0.43), but the

Table 5. Classification results (weighted F1, precision, recall) for trinary sentiment detection task split by tone on article headlines.

	All tones combined (overall score)				
	F_1		*n (human coding)*	precision	recall
Recession	0.26		4640	0.30	0.43
Damstra and Boukes (2018)	0.32		4640	0.52	0.45
LIWC	0.42		4640	**0.53**	**0.48**
SentiStrength	0.42		4640	0.45	0.45
Pattern	0.41		4640	0.45	0.45
Polyglot	**0.43**		4640	0.44	0.44
DANEW	**0.43**		4640	0.46	0.45
	Negative Tone				
	F_1	***n (predicted)***	***n (human coding)***	**precision**	**recall**
Recession	0.00	6	1524	0.33	0.00
Damstra and Boukes (2018)	0.08	99	1524	**0.62**	0.04
LIWC	0.29	471	1524	**0.62**	0.19
SentiStrength	0.39	1158	1524	0.45	0.34
Pattern	0.30	692	1524	0.48	0.22
Polyglot	**0.42**	1158	1524	0.48	**0.37**
DANEW	0.36	794	1524	0.52	0.27
	Neutral Tone				
	F_1	***n (predicted)***	***n (human coding)***	**precision**	**recall**
Recession	**0.60**	4634	2008	0.43	**1.00**
Damstra and Boukes (2018)	**0.60**	4366	2008	0.44	0.96
LIWC	**0.60**	3750	2008	**0.46**	0.86
SentiStrength	0.55	3103	2008	0.45	0.70
Pattern	0.56	3260	2008	0.45	0.74
Polyglot	0.47	2231	2008	0.45	0.50
DANEW	0.53	2776	2008	**0.46**	0.63
	Positive tone				
	F_1	***n (predicted)***	***n (human coding)***	**precision**	**recall**
Recession	0.00	0	1108	0.00	0.00
Damstra and Boukes (2018)	0.14	175	1108	**0.53**	0.08
LIWC	0.29	419	1108	0.52	0.20
SentiStrength	0.22	379	1108	0.42	0.14
Pattern	0.30	688	1108	0.39	0.24
Polyglot	**0.39**	1251	1108	0.37	**0.42**
DANEW	0.36	1070	1108	0.37	0.35

differences with the other tools are minimal; respectively, *LIWC, SentiStrength*, and *Pattern* – all had and F_1-score above 0.41. The two more basic approaches (recession and Damstra & Boukes) generally performed the least well in terms of overall F_1-scores.

Looking at the results of the sentiment dictionary that had relatively the best performance with the manual coding of headlines, *Polyglot*'s relative success in detecting the tone of economic news can mainly be explained by its *high recall* of negative and positive items. *Polyglot* clearly identified most items with a non-neutral tone and consequently had the highest accuracy for both tones (negative: $F_1 = .42$; positive: $F_1 = .39$). However, its precision is lower than the *LIWC* and Damstra & Boukes dictionaries: Polyglot relatively frequently classifies news of a certain tone unjustly (positive or negative) even when a tone is not present (i.e., false positives). This also makes that *Polyglot* performs worst in detecting a neutral tone in the headlines ($F_1 = .47$), both regarding recall (.50) and accuracy (.45) it scores the lowest within neutral items.

LIWC – relatively, the second-best performing tool in terms of Krippendorff's α – especially had an acceptable performance in the neutral class ($F_1 = .60$), while coming in fourth for positive ($F_1 = .29$) and fifth for the negative ($F_1 = .29$) classes. Apparently, *LIWC* is relatively precise in capturing the tone of economic news *when* it detects something – precision scores for negative (.62) and positive news (.52) are among the highest of all – but it often tends to miss the tonality (low recall) and classifies articles as neutral instead (i.e., false negatives). While the comparably high α-score of *Polyglot*, thus, can be explained by its recall, *LIWC* benefited from its relatively high precision in detecting tone.

Overall, *DANEW* ties in first with *Polyglot* ($F_1 = .43$); however, whereas *Polyglot* excels with high recall of positive and negative headlines, *DANEW* performs average on all three categories. Interesting is that *SentiStrength*'s performance clearly differs on tonalities. Whereas it performs second best on negative headlines with a relatively high recall ($F_1 = .39$), it scores low (i.e., fifth) on positive headlines ($F_1 = .22$). The latter mainly stems from poor ability to recall positive headlines (recall = .14). The opposite is true for *Pattern*, which performs average on positive news, but is especially bad in detecting negative headlines.

The two dictionaries with a limited number of words (recession: 1 word; Damstra & Boukes: 65 words) did hardly detect any headlines of a positive or negative valence, which makes their recall very low for these categories. Because the one-word recession dictionary is undefined for positive class articles (it can only detect negative news), the performance for this category is by definition zero. Unfortunately, the word "recession" even performs badly on negative article detection ($F_1 = 0.00$), as only six articles contain this word in the headline and only 33% of these articles are in fact classified as negative (i.e., "recession" not necessarily indicates negativity). The Damstra & Boukes algorithm performs second worst ($F_1 = .32$) overall. The reason, most likely, is that the words present in this dictionary do not occur frequently in the headlines, which makes that it misses many positive and negative cases (i.e., low recall). The precision of this tailor-made measurement is amongst the highest, though: So, when it indicates that an article is negative or positive, it mostly is correct – but it does so too seldom due to its small dictionary. The recall scores (and therefore also the F_1-scores) of the "recession" and Damstra & Boukes dictionaries are very high for the neutral category due to their small word corpus with only a few (but relevant) words. It, thus, is less susceptible to false positives of the valenced classes.

Compared to the findings of the headlines, results for the full texts are comparable (see Appendix B). For the full texts, *LIWC* not only performed relatively the best in terms of Krippendorffs's α but also had the highest overall F_1-score (.49). Again, its *precision* was the highest for both the negatively (.66) and positively (.72) classified texts. *Polyglot*, which on headlines tied for the first place overall, performs second best for the full texts with a .04 difference ($F_1 = .45$). It is noteworthy that all tools perform badly on neutral full texts ($.14 \leq F_1 \leq .17$) as compared to headlines ($.47 \leq F_1 \leq .60$). Evidenced by the low precision scores (all below .11) on the neutral texts, the extensive vocabularies of these tools are easily led astray when texts are long: They frequently will classify an article as having no tone even in the case of valenced news. Notably, the performance of the Damstra &

Boukes dictionary vis-à-vis the other tools improved considerably for the full texts compared to the headlines.

Combined Approach: Stronger Together Than Apart?

Table 3, which was presented earlier, also introduced the results of the combined approach, which was constructed from the mean scores of the best four off-the-shelf dictionaries (*LIWC, Polyglot, DANEW*, and *SentiStrength*). The table shows that this combined approach outperformed the sentiment analysis tools on all levels using headline data.[8] With α = .30, the combined approach has a .05-points higher agreement with the manual coding than the best individual dictionary, but it still remains far below the generally accepted threshold of α ≥ .67 for reliable coding.

The relatively good performance of this combined approach may be due to the low correlation of errors between models. The different mistakes, thus, may cancel each other out when combined. This suggests that a composite measurement (i.e., average of different tone scores) can reduce the problems of individual tools and thereby approach human-level performance a little more closely. Also, with regards to the full texts, the combined approach performed comparably well (α = .32). Only *LIWC* yielded a (non-significant) higher agreement with the manually coded tone. Moreover, partially overlapping confidence intervals are found with *Polyglot*. Hence, the combined approach did not clearly stand out as relatively *the best* performing measure for the full texts (i.e., one of the best). Altogether, evidence is found that confirms H_1 within the news headlines, but H_1 cannot be confirmed for the full texts. The first hypothesis, therefore, is only partially supported.

Headlines versus Full Texts

The expectation was that automated measurements of sentiment should agree more strongly with human coding for headlines than for full texts. Table 3 compares the correlations between manual and automatic measurements of tone for the headlines versus the full texts. The statistical difference between the two is examined by the overlap in confidence intervals. The last column of Table 3 indicates whether there was a significant difference.

Generally, no pattern can be discovered as to whether automatic methods performed better (or worse) in detecting the tone in headlines compared to full articles. In line with the hypothesis, *DANEW* yielded a higher Krippendorff's α for the headlines than for the full texts (α = .22 vs. α = .15).[9] The same trend is visible for *Pattern*, which performed slightly less bad for the headlines (α = .17) than for the full texts (α = .09). The small corpus of Damstra & Boukes also performed better on the longer texts. However, opposite evidence was found for *LIWC*, which had a higher agreement with manual annotations of article full texts (α = .35) than with the article headlines (α = .23). No significant differences emerged between the accuracy of headlines and full texts for *Polyglot and SentiStrength*, with closely overlapping confidence intervals that included the estimate of the other. No significant difference was either found for the combined approach. In sum, the evidence is very mixed with positive, negative and insignificant results. Hence, H_2 has to be rejected: Automatic tone measurements of headlines did *not* agree more strongly with human coding for headlines than for full texts.

Correlations at Different Granularities

Most scientific studies that investigate newspaper reporting on economic developments are not interested in the tone of individual articles, but rather examine the aggregated sentiment at the daily, weekly or sometimes even monthly level. Higher granularities are better suited for time-series analysis, as reporting may be sparse and noisy at low temporal granularities. In Table 6, we present the correlations of different tools at the article-level, daily-level, and weekly-level, with scores as mean-aggregations of individual article scores. Logically, a greater aggregation comes together with

Table 6. Correlations to human annotations at different temporal granularities.

	Headline			Full text		
	Per article	Per day	Per week	Per article	Per day	Per week
Recession	−0.01	0.04	0.21	−0.04	−0.08	−0.34
Damstra and Boukes (2018)	0.14 ***	0.28 **	0.17	0.26 ***	0.36 ***	0.63 **
LIWC	0.26 ***	0.41 ***	0.60 **	0.37 ***	0.53 ***	0.59 **
SentiStrength	0.20 ***	0.35 ***	0.38	0.17 ***	0.18	0.42
Pattern	0.18 ***	0.17	0.46 *	0.11 ***	0.07	0.09
Polyglot	0.26 ***	0.40 ***	0.40	0.26 ***	0.26 **	0.52 *
DANEW	0.24 ***	0.26 **	0.29	0.18 ***	0.18	−0.15
Combined approach	0.34 ***	0.47 ***	0.52 **	0.36 ***	0.43 ***	0.53 **
n	4640	152	22	1426	152	22

Pearson correlation coefficients are displayed with *** $p < .001$, ** $p < .010$, * $p < .050$.

a loss of statistical power (i.e., fewer units-of-observation). On full-text content, only Damstra & Boukes, *LIWC*, and *Polyglot* have statistically significant correlations at the weekly level ($n = 22$). On headlines, only *LIWC* and *Pattern* achieved this distinction.

Most sentiment analysis tools have a stronger correlation with the manually coded tone on coarser granularities (day and week); exceptions are *Pattern* and *DANEW* for the full texts. Most obvious are the improvements of *LIWC* and the dictionary approach of Damstra & Boukes: Both correlated moderately for the full text on the article level (respectively, $r = .37$ and $r = .26$) and this improves to a considerable $r \approx .60$ for the weekly level. *LIWC* showed a similar increase for the headlines (from $r = .26$ to $r = .60$); this did not occur for the Damstra & Boukes approach, arguably, due to the low agreement with manually coded headlines to begin with (see α-results presented earlier). Thus, the agreement with manual annotations needs to be relatively high, in the first place, for data aggregation to cause a better performance (also see the findings of *Pattern* for the full text). Clear improvements are also observed for the combined approach, *Polyglot*, as well as *Pattern* (headlines only). This shows that the validity of some (i.e., not all) approaches becomes better when aggregating the data. H_3, thus, is largely confirmed.

Discussion

This study provides several important insights regarding the applicability of automated measurement instruments for sentiment analysis. Although a widely accepted practice, there are some noticeable side notes that emerged as conclusions from the current investigation. First, we conclude that there is wide variety in the quality of performance *across* off-the-shelf sentiment tools. However, even the tools that achieved the best performance in detecting the tone of economic news (i.e., *LIWC* and *Polyglot*) still yielded unacceptably low-reliability scores ($\alpha < .35$) when compared to the manual coding. All the reliability scores are far below the threshold of what is commonly understood as acceptable for a content analysis ($\alpha > .67$). This casts serious doubts on the validity of using automated tone measurements to measure how positive or negative (Dutch) news articles are about the economy and also poses important questions about the usage of sentiment analysis tools more broadly. By the end of the manuscript, we provide recommendations when and how to use these off-the-shelf sentiment analysis tools. Interesting to note was that the different tools behaved differently when it came to the recall and precision of specific tonalities (i.e., negative, neutral or positive news). The sensitivity of specific dictionaries to the tone raises important questions with regard to their utilization in the analysis of economic news: Some may perform structurally worse on detecting certain tonalities. In case of economic crises (a lot of negative news coverage) or economic upturn, the performance of different dictionaries may thus structurally vary, although it is unlikely that this would eventually result in acceptable levels of reliability.

Not only did the sentiment analysis tools not agree with the manual tone assessments, a lack of agreement *between* the tools themselves also became obvious in Table 4, which showed how strongly (or rather: weakly) the tone classifications of each tool overlap with another. In general, this only yielded low Krippendorff's alphas ($\alpha < .35$). Thus, scholars could reach substantially different conclusions if they use different off-the-shelf dictionaries when answering the same research question with the same dataset. This urges them to choose (and explain this choice) and validate the right tool concerning their specific genre and domain (see also Van Atteveldt & Peng, 2018) or to customize a dictionary to their specific research question.

Alternatively, we demonstrate that a combined approach in which the scores of four of the five sentiment analysis tools were aggregated, generally, performed slightly better than the individual measurements. This was especially the case for the headlines on which it stood out with the highest α-score. The combined approach also performed better than average for the full texts but followed *LIWC* as second best (although the difference was not significant). What one could learn from this is that relying solely on one automated sentiment analysis tool is likely to result in a less reliable measurement than by adopting a combined approach. However, the improvements achieved with this combined approach were not very large and by far did still not alleviate the reliability scores to an acceptable level.

The current study had to reject the assumption that agreement with the gold standard of human coding (i.e., "most reasonable benchmark", see Rauh, 2018, p. 7) would per se be stronger for headlines than for the full texts. This was surprising given that these tools are trained, validated, and developed for short texts (e.g., social media or product reviews). However, relatively more agreement with manually coded headlines as compared to the manually coded full bodies of text was only found in two instances (i.e., *DANEW* and *Pattern*) with opposing evidence being found for *LIWC* and the combined approach.

Clear improvements – for both headlines and full texts (especially) – were found for most measurement instruments at higher levels or granularity: If one does not assess the correlation at the individual article level, but instead on the aggregated daily or weekly level, measurements correlated more strongly with the human-coded data. Considering the unacceptably low reliability scores for individual articles, this is somewhat reassuring for scholars who are not interested per se in the tone of separate texts, but who want to analyze how tone develops over time, which is often the reason to choose for automated sentiment analysis, because there are too many articles for human coding when one is interested in a long time span (e.g., Shapiro et al., 2019). This brings us to the last conclusion.

Our results show that as the complexity of automatic measurements increases, this does not necessarily lead to a better performance (Gonçalves et al., 2013; Muddiman et al., 2019). Of the different automated measurements that we compared for the full texts of articles, we found that the method of Damstra and Boukes (2018) yielded one of the relative best performances (still not acceptable, though) when compared to a manual coding of tone in economic news (together with *LIWC*). This tailor-made operationalization, however, only consisted of counting the presence of 30 positive words minus the presence of 35 negative words, whereas the other approaches follow dictionaries with thousands of words. Apparently, more elaborate dictionaries introduce considerable amounts of noise in the measurement of full texts (González-Bailón & Paltoglou, 2015). We did not find the similar pattern for headlines; arguably, because headlines are more focused on the topic of the respective article and all (valenced) words, thus, matter. The smaller dictionary could under such circumstances (short texts) be confronted by a neutrality bias (Rauh, 2018). Consciously choosing the right keywords for the appropriate context and developing a tailor-made sentiment dictionary, thus, could, in some instances, be more effective (Shapiro et al., 2019) than simply choosing the most advanced automatic method of measuring a text's tone (see also Puschmann & Powell, 2018). In sum, lexicon size is no valid proxy for detection validity when it comes to specific domains.

Recommendation

The concrete recommendation that follows from this study is twofold. First, scholars should be conscious of the weak performance of the off-the-shelf sentiment analysis tools, at least in the current context. Whereas human coders agreed with each other in categorizing how positive or negative a news article was about the economy, the overlap of the automated tone measurements with manually coded economic news was below any threshold that normally would be considered acceptable (e.g., Krippendorff, 2004). More importantly, the levels of agreement of the off-the-shelf sentiment analysis tools *among themselves* were strikingly low when categorizing the same news articles, which highlights the importance of carefully deciding *whether* and, if so, *which* tool to use in the first place. Here, one needs to especially consider the context and type of content for which the tool was originally developed.

Second, and aligned to the first recommendation, should researchers – aware of the risks of low levels of reliability – decide to use automated measurement(s) of sentiment, they should not simply choose *one* automatic measurement of sentiment, but instead first compare the results of the different available approaches with a set of manually coded items (see Ribeiro et al., 2016). Without a clear benchmark of human-coded texts, it will simply be impossible to evaluate whether any automated measurement instrument for sentiment classification works acceptably well within a specific domain and genre, and if so which one. Simply choosing one automated dictionary without any validation could result in heavily biased and invalid results. Concerning economic news, using one way (e.g., *SentiStrength*) of measuring the tone automatically could yield completely different and almost unrelated results compared to using another method (e.g., *Pattern*). Eventually, this could lead to contradictory results when answering concrete research questions (Jongeling et al., 2017).

If one, first, creates a manually coded set of items to use as a *golden standard* and, subsequently, compare the different methods available in one's language, researchers can test which tool(s) function(s) most accurately. Although automatic measures are often employed to reduce the costs of data collection, crafting a small dataset of hand-coded items is not expensive and gives the researcher the necessary insights to make a cautious decision about which tool (or combination of tools) would be able to automatically measure the tone of texts. Following this approach and reporting it step-by-step would alleviate the rather nontransparent and arbitrary decision-making process regarding which automatic measurement is applied in a study; it even provides more clarity as to which operationalization the researcher has in mind for the notion of measuring tone of a text. In our view, reviewers and editors should be very skeptical about unverified automatic measurements.

Tone is still difficult to measure automatically, even in the context of relatively straightforward economic news. More generally, across genres and domains, no single method will consistently perform the best – but some off-the-shelf tool(s) within any domain will be relatively more accurate than others (Ribeiro et al., 2016). However, even the best performing dictionary in the current context (i.e., *LIWC*) still yielded unacceptably low-reliability estimates that restrict its use in the context of Dutch economic news. Supervised machine learning might offer a viable alternative when a considerable set of manually coded training data is available (as is in fact the case here, but for many studies is not). When using machine learning, adopting a word-embeddings approach, which is able to take similarities in word meanings into account, might also help to improve the results (Rudkowsky et al., 2018). However, such approaches fall outside of the current paper's scope of comparing the popular off-the-shelf packages that require a little less skills and knowledge about computational methods. Our findings, nevertheless, show that measurement can be significantly improved by moving analyses to higher levels of granularity or by combining the best performing measurement instruments (to know which performs best, one still needs to compare with human coding, though).

Altogether, none of the automatic methods achieved an acceptable agreement with the manually coded tone of economic news. One reason is obvious. Whereas the research assistants for the manual content analysis were carefully instructed to code for explicit assessments in a news article about *how the (Dutch) economy* would develop, automatic dictionaries would base their tone score on the valence of words in the *overall article*. Whereas human coding may thus particularly focus on one sentence from which the tone with regards to the economy could be inferred, automatic methods also consider the sentiment of the (noisy) context in which this sentence would be located. Developed to measure the tone in rather short user-generated texts of an informal nature, automatic dictionaries cannot compete with human coding of news items written by professionals in a more formal style (González-Bailón & Paltoglou, 2015). But, still, on higher levels of granularity (i.e., aggregating observations over time) some of the measurements achieved reasonable results.

Although the current study was employed with a dataset of Dutch news items about the economy, our recommendations are not restricted to that particular genre, domain, or even language. The conclusion that the performance of automatic sentiment analysis tools to analyze the tone of news coverage differs widely (and is mostly unacceptable) and that researchers, ideally, first make a comparison with manually coded items, before they decide which automatic method serves their purpose best, will hold in any context. The process to reveal the most suitable dictionary in any particular setting could follow the same procedure as presented in this paper, irrespective of the research question. A blind application of off-the-shelf tools, especially in contexts for which they were not initially developed, is in general not recommended.

Notes

1. We use "sentiment analysis tool" as a general term to describe different dictionary-based approaches (Boumans & Trilling, 2016) used for sentiment analysis. The word "tool" is used because each approach not only uses a dictionary of words, but also has an internal logic to determine its score for sentiment (e.g., the usage or not of modifier terms such as "very", "extremely", to increase the amplitude, ability to consider negations such as "not good", how negative and positive words are scored against each other, *et cetera*).
2. In Dutch: *Hope* (hoop* hope* gehoopt uitzien uitgezien vooruitzien* vooruitzicht* vooruitblik blikte vertrouw* geestdrift* belofte beloof* inspiratie inspireer* ontplooi* passie gepassioneerd geïnspireerd opgelucht toekomstvisie toekomstgericht* toekomstbestendig* houvast toevlucht toeverlaat perspectief redd* gered herstel*) and not (hopeloos OR hopeloze) (i.e., 30 words).
3. http://sentistrength.wlv.ac.uk/.
4. https://www.clips.uantwerpen.be/pattern.
5. https://github.com/aboSamoor/polyglot.
6. (ontslagen! AND (medewerkers OR werknemers OR banen OR werkpl!)) OR werkgelegen! OR werklo! OR huizenv! OR huizenb! OR huizenk! OR huizenp! OR huizenm! OR inflatie OR deflatie OR Consumentenbesteding! OR TTIP OR (HLEAD(economi!) OR HLEAD(financi!) OR HLEAD(monetai!) OR HLEAD(beroepsbevolk!) OR HLEAD(conjunctu!) OR HLEAD(Centrale Bank) OR HLEAD(Nederlandsche Bank) OR HLEAD(export) OR HLEAD(import) OR HLEAD(nationaal inkomen) OR HLEAD(nationaal product) OR HLEAD(nationaal bruto) OR HLEAD(overheidsuitga!) OR HLEAD(overheidsbested!) OR HLEAD(bezuinig!) OR HLEAD(arbeidspartici!) OR HLEAD(recessie) OR HLEAD(spaargeld!) OR HLEAD(vacatures) OR HLEAD(arbeidsplaatsen) OR HLEAD(spaarrente) OR HLEAD(hypotheekrente) OR HLEAD(rente!)).
7. Ideally, one could also "learn" the optimal cutoff-points inductively, but this seems to be rarely done in practice, as users of off-the-shelf sentiment analysis tools seem to mostly use them "as a black-box" without "changes and adaptations and with none or almost none parameter setting" (Ribeiro et al., 2016, p. 2). The 0.5 thresholds were chosen to best reflect the actual usage off-the-shelf sentiment analysis tools in research contexts with little-to-no manually coded (ground truth) data that can be used for comparison or for determining the cutoff-points inductively.
8. The confidence interval for a small part overlaps with the one of *Polyglot*, but the manually calculated confidence interval of their difference does not include 0. Moreover, the tone scores of the combined approach and *Polyglot* will correlate with each other because the first is partly constructed based on the latter's scores, which results in a conservative estimate. Hence, we can conclude that the difference is significant (see Schenker & Gentleman, 2001).

9. Confidence intervals partly overlap, but the headlines' α point-estimate is not in the full text's confidence interval (and vice versa). Solely relying on the overlap of confidence intervals results in overly conservative and "mistaken conclusions" (i.e., null hypothesis not rejected frequently enough, see Schenker & Gentleman, 2001, p. 182). Especially when both estimates are likely to positively correlate with each other (will be the case here), comparisons of confidence intervals are underpowered. The manually calculated confidence interval of the difference between the two α-estimates does not contain 0, which implies that the difference between headline and full text is significant.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research (NWO) with a VIDI grant under project number: 016.145.369. Data collection of the online news was carried out via INCA on the Dutch national e-infrastructure with support of the SURF Cooperative. An earlier version of this manuscript has been presented at the International Communication Association 68th Annual Conference, 24-28 May 2018, Prague (Czech Republic) as "*The good and bad in economic news: Comparing (automatic) measurements of sentiment in Dutch economic news*."

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study has been made possible by a VIDI grant (project number 016.145.369) from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research (NWO)). Data collection of the online news was carried out via INCA on the Dutch national e-infrastructure with support of the SURF Cooperative.

ORCID

Mark Boukes http://orcid.org/0000-0002-3377-6281
Theo Araujo http://orcid.org/0000-0002-4633-9339

References

Blood, D. J., & Phillips, P. C. (1995). Recession headline news, consumer sentiment, the state of the economy and presidential popularity: A time series analysis 1989–1993. *International Journal of Public Opinion Research*, *7*(1), 2–22. doi:10.1093/ijpor/7.1.2

Boomgaarden, H. G., Van Spanje, J., Vliegenthart, R., & De Vreese, C. H. (2011). Covering the crisis: Media coverage of the economic crisis and citizens' economic expectations. *Acta Politica*, *46*(4), 353–379. doi:10.1057/ap.2011.18

Boukes, M., Damstra, A., & Vliegenthart, R. (2019). Media effects across time and subject: How news coverage affects two out of four attributes of consumer confidence. *Communication Research*, Advance online publication. doi: 10.1177/0093650219870087.

Boukes, M., & Vliegenthart, R. (2017). A general pattern in the construction of economic newsworthiness? Analyzing news factors in popular, quality, regional, and financial newspapers. *Journalism, Advance Online Publication*, 1–22. doi:10.1177/1464884917725989

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism*, *4*(1), 8–23. doi:10.1080/21670811.2015.1096598

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical Report C-1). Florida: The Center for Research in Psychophysiology. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.306.3881&rep=rep1&type=pdf

Burnap, P., & Williams, M. L. (2014). *Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making*. Paper presented at the Internet, Policy & Politics, Oxford, UK. Retrieved from http://orca.cf.ac.uk/65227/

Chen, Y., & Skiena, S. (2014). *Building sentiment lexicons for all major languages*. Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA (pp. 383–389).

Chinchor, N. (1992). *MUC-4 evaluation metrics*. Paper presented at the Proceedings of the 4th Conference on Message Understanding Conference (MUC), Stroudsburg, PA (pp. 22–29).

Damstra, A., & Boukes, M. (2018). The economy, the news, and the public: A longitudinal study of the impact of economic news on economic evaluations and expectations. *Communication Research, Advance Online Publication*, 1–25. doi:10.1177/0093650217750971

De Boef, S., & Kellstedt, P. M. (2004). The political (and economic) origins of consumer confidence. *American Journal of Political Science*, *48*(4), 633–649. doi:10.1111/j.0092-5853.2004.00092.x

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, *13*(June), 2063–2067.

Dietterich, T. G. (2000). *Ensemble methods in machine learning*. Paper presented at the International Workshop on Multiple Classifier Systems (MCS), Berlin, Germany. doi:10.1007/3-540-45014-9_1

Fogarty, B. J. (2005). Determining economic news coverage. *International Journal of Public Opinion Research*, *17*(2), 149–172. doi:10.1093/ijpor/edh051

Goidel, R. K., & Langley, R. E. (1995). Media coverage of the economy and aggregate economic evaluations: Uncovering evidence of indirect media effects. *Political Research Quarterly*, *48*(2), 313–328. doi:10.1177/106591299504800205

Goidel, R. K., Procopio, S., Terrell, D., & Wu, H. D. (2010). Sources of economic news and economic expectations. *American Politics Research*, *38*(4), 759–777. doi:10.1177/1532673X09355671

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). *Comparing and combining sentiment analysis methods*. Paper presented at the Proceedings of the First ACM Conference on Online Social Networks (pp. 27–38), Boston, MA.

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The Annals of the American Academy of Political and Social Science*, *659*(1), 95–107. doi:10.1177/0002716215569192

Günther, E., & Domahidi, E. (2016). *The changing research landscape of our field: A topic model of 80 years in communication science journals*. Paper presented at the 66th Annual ICA Conference, Fukuoka, Japan.

Hester, J. B., & Gibson, R. (2003). The economy and second-level agenda setting: A time-series analysis of economic news and public opinion about the economy. *Journalism & Mass Communication Quarterly*, *80*(1), 73–90. doi:10.1177/107769900308000106

Hollanders, D., & Vliegenthart, R. (2011). The influence of negative newspaper coverage on consumer confidence: The Dutch case. *Journal of Economic Psychology*, *32*(3), 367–373. doi:10.1016/j.joep.2011.01.003

Jongeling, R., Sarkar, P., Datta, S., & Serebrenik, A. (2017). On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering*, *22*(5), 2543–2584. doi:10.1007/s10664-016-9493-x

Kalogeropoulos, A., Svensson, H. M., Van Dalen, A., de Vreese, C., & Albæk, E. (2015). Are watchdogs doing their business? Media coverage of economic news. *Journalism*, *16*(8), 993–1009. doi:10.1177/1464884914554167

Kleinnijenhuis, J., Schultz, F., Oegema, D., & Van Atteveldt, W. (2013). Financial news and market panics in the age of high-frequency sentiment trading algorithms. *Journalism*, *14*(2), 271–291. doi:10.1177/1464884912468375

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65. doi:10.1111/j.1540-6261.2010.01625.x

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. doi:10.1016/j.asej.2014.04.011

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169–177. doi:10.3758/s13428-012-0243-8

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226. doi:10.1080/10584609.2018.1517843

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychological properties of LIWC2007*. Retrieved from http://www.liwc.net/LIWC2007LanguageManual.pdf

Puschmann, C., & Powell, A. (2018). Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. *Social Media + Society*, *4*(3), 2056305118797724. doi:10.1177/2056305118797724

Rauh, C. (2018). Validating a sentiment dictionary for german political language – A workbench note. *Journal of Information Technology & Politics*, *15*(4), 319–343. doi:10.1080/19331681.2018.1485608

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench – A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(23), 1–29. doi:10.1140/epjds/s13688-016-0085-1

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. doi:10.1080/19312458.2018.1455817

Scharkow, M. (2017). Content analysis, automatic. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods* Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781118901731.iecrm0043

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186. doi:10.1198/000313001317097960

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2019). *Measuring news sentiment* (Working Paper 2017-01). Federal Reserve Bank of San Francisco. doi:10.24148/wp2017-01

Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics*, *68*(2), 372–385. doi:10.1111/j.1468-2508.2006.00413.x

Soroka, S. N. (2012). The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, *74*(2), 514–528. doi:10.1017/s002238161100171x

Soroka, S. N., Young, L., & Balmas, M. (2015). Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The Annals of the American Academy of Political and Social Science*, *659*(1), 108–121. doi:10.1177/0002716215569217

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. doi:10.1111/j.1540-6261.2007.01232.x

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with SentiStrength. In J. Holyst (Ed.), *Cyberemotions: Understanding complex systems* (pp. 119–134). Cham, Switzerland: Springer. doi:10.1007/978-3-319-43639-5_7

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. doi:10.1002/asi.21416

Trilling, D., & Boumans, J. (2018). Automatische inhoudsanalyse van Nederlandstalige data: Een overzicht en onderzoeksagenda. *Tijdschrift Voor Communicatiewetenschap*, *46*(1), 5–24.

Trilling, D., Van de Velde, B., Kroon, A. C., Löcherbach, F., Araujo, T., Strychar, J., … Jonkman, J. G. F. (2018). *INCA: Infrastructure for content analysis*. Paper presented at the IEEE 14th International Conference on E-Science (pp. 329–330), Amsterdam, the Netherlands. doi:10.5625/lar.2018.34.4.329

Tuarob, S., Tucker, C. S., Salathe, M., & Ram, N. (2014). An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of Biomedical Informatics*, *49*(June), 255–268. doi:10.1016/j.jbi.2014.03.005

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (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*(1), 73–94. doi:10.1080/19331680802154145

Van Atteveldt, W., & Peng, T. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. doi:10.1080/19312458.2018.1458084

Van Atteveldt, W., Ruigrok, N., Takens, J., & Jacobi, C. (2014). *Inhoudsanalyse met AmCAT*. Retrieved from http://vanatteveldt.com/wp-content/uploads/amcatbook.pdf

Van Dalen, A., de Vreese, C., & Albæk, E. (2017). Economic news through the magnifying glass. *Journalism Studies*, *18*(7), 890–909. doi:10.1080/1461670X.2015.1089183

Van der Wees, M., Bisazza, A., Weerkamp, W., & Monz, C. (2015). *What's in a domain? Analyzing genre and topic differences in statistical machine translation*. Paper presented at the 53rd Annual Meeting of the Association for Computational Linguistics (pp. 560–566), Beijing, China.

Wettstein, M. (2018). *Nogrod 1.1 (beta): Quick tutorial*. Retrieved from https://www.ikmz.uzh.ch/dam/jcr:fd26fbd3-2b3f-4421-96f8-4c2750417551/Nogrod_1–1.pdf

Wu, H. D., Stevenson, R. L., Chen, H., & Güner, Z. N. (2002). The conditioned impact of recession news: A time-series analysis of economic communication in the United States, 1987–1996. *International Journal of Public Opinion Research*, *14*(1), 19–36. doi:10.1093/ijpor/14.1.19

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. doi:10.1080/10584609.2012.671234

Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – Which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*(93), 1–10. doi:10.1186/s12874-016-0200-9

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond*, *32*(4), 271–281.

Appendix A. Correlations between automated sentiment analysis tools

Table A1. Correlations between sentiment scores using different methods for headlines (above) and full texts (below).

	Headline							
	Manual coding	Recession	D & B	LIWC	SentiStrength	Pattern	Polyglot	DANEW
Manual coding	1.00 ***							
Recession	-	-						
Damstra and Boukes (2018)	0.16 ***	-	1.00 ***					
LIWC	**0.30 *** **	-	0.16 ***	1.00 ***				
SentiStrength	0.24 ***	-	0.08 **	0.26 ***	1.00 ***			
Pattern	0.22 ***	-	0.00	0.30 ***	0.22 ***	1.00 ***		
Polyglot	0.30 ***	-	0.19 ***	0.32 ***	0.37 ***	0.26 ***	1.00 ***	
DANEW	0.24 ***	-	0.04	**0.43 *** **	0.33 ***	0.23 ***	0.32 ***	1.00 ***
	Full text							
	Manual coding	Recession	D & B	LIWC	SentiStrength	Pattern	Polyglot	DANEW
Manual coding	1.00 ***							
Recession	−0.06 *	1.00 ***						
Damstra and Boukes (2018)	0.27 ***	−0.16 ***	1.00 ***					
LIWC	0.39 ***	0.02	0.27 ***	1.00 ***				
SentiStrength	0.17 ***	−0.01	0.10 ***	0.18 ***	1.00 ***			
Pattern	0.13 ***	−0.02	0.04	0.28 ***	0.12 ***	1.00 ***		
Polyglot	0.26 ***	0.05	0.17 ***	0.41 ***	0.21 ***	0.30 ***	1.00 ***	
DANEW	0.15 ***	0.06 *	0.05	0.36 ***	0.18 ***	0.29 ***	0.37 ***	1.00 ***

The word "recession" did not occur in headlines of our sample, as such, no correlation coefficient is available for the recession classifier; *** $p < .001$, ** $p < .010$, * $p < .05$.

Appendix B. Precision, recall and F_1-scores for full texts

Table B1. Classification results for trinary sentiment detection task split by tone on article full texts.

	All tones combined (overall score)				
	F_1		n (human coding)	precision	recall
Recession	0.04		1426	0.13	0.10
Damstra and Boukes (2018	0.36		1426	0.62	0.29
LIWC	0.49		1426	0.64	0.43
SentiStrength	0.41		1426	0.52	0.37
Pattern	0.37		1426	0.48	0.33
Polyglot	0.45		1426	0.58	0.40
DANEW	0.41		1426	0.52	0.37
	Negative Tone				
	F_1	n (predicted)	n (human coding)	precision	recall
Recession	0.05	56	594	0.30	0.03
Damstra and Boukes (2018)	0.32	213	594	0.62	0.22
LIWC	0.54	409	594	0.66	0.46
SentiStrength	0.36	270	594	0.57	0.26
Pattern	0.36	383	594	0.46	0.30
Polyglot	0.47	408	594	0.58	0.40
DANEW	0.44	433	594	0.52	0.38
	Neutral Tone				
	F_1	n (predicted)	n (human coding)	precision	recall
Recession	0.16	1370	130	0.09	0.95
Damstra and Boukes (2018)	0.15	923	130	0.09	0.61
LIWC	0.17	641	130	0.10	0.51
SentiStrength	0.14	574	130	0.09	0.39
Pattern	0.16	631	130	0.10	0.48
Polyglot	0.16	616	130	0.10	0.47
DANEW	0.17	609	130	0.11	0.49
	Positive tone				
	F_1	*n (predicted)*	*n (human coding)*	precision	recall
Recession	0.00	0	702	0.00	0.00
Damstra and Boukes (2018)	0.42	290	702	0.72	0.30
LIWC	0.50	376	702	0.72	0.39
SentiStrength	0.51	582	702	0.56	0.47
Pattern	0.42	412	702	0.57	0.33
Polyglot	0.49	402	702	0.68	0.39
DANEW	0.43	384	702	0.61	0.33