Advancing Automated Content Analysis for a New Era of Media Effects Research

The Key Role of Transfer Learning

Kroon, A.; Welbers, K.; Trilling, D.; van Atteveldt, W.

DOI
10.1080/19312458.2023.2261372

Publication date
2024

Document Version
Final published version

Published in
Communication Methods and Measures

COMMUNICATION METHODS AND MEASURES
2024, VOL. 18, NO. 2, 142–162
https://doi.org/10.1080/19312458.2023.2261372

Routledge
Taylor & Francis Group

Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning

Anne Kroon [a*], Kasper Welbers [b*], Damian Trilling [a], and Wouter van Atteveldt [b]

[a]Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, Netherlands; [b]Department of Communication Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

ABSTRACT

The availability of individual-level digital trace data offers exciting new ways to study media uses and effects based on the actual content that people encountered. In this article, we argue that to really reap the benefits of this data, we need to update our methodology for automated text analysis. We review challenges for the automatic identification of theoretically relevant concepts in texts along three dimensions: *format/style*, *language*, and *modality*. These dimensions unveil a significantly higher level of diversity and complexity in individual-level digital trace data, as opposed to the content traditionally examined through automated text analysis in our field. Consequently, they provide a valuable perspective for exploring the limitations of traditional approaches. We argue that recent developments within the field of Natural Language Processing, in particular, *transfer learning* using *transformer-based models*, have the potential to aid the development, application, and performance of various computational tools. These tools can contribute to the meaningful categorization of the content of social (and other) media.

In an increasingly digitized and fragmented media landscape, the notions of mass media and mass media effects are losing their usefulness for communication research. For example, next to traditional entertainment shows on television shows we now have not only a long tail of shows available through streaming services, but also atomized clips on YouTube. This makes it very hard to estimate what media a given person has used – let alone what effects this has had. People can craft their own media diets by self-selecting within and between platforms that best fit their political preferences or social interests (Steppat et al., 2022). At the same time, platforms are increasingly using algorithmic curation to cater content to individual interests— in political contexts (e.g., Jürgens & Stark, 2022; Wojcieszak et al., 2021) as well as in entertainment contexts. Indeed, the core selling point of platforms like TikTok is their ability to create highly customized experiences. If the variety of content increases so much and if in addition, the mass audience is replaced by personalized media diets, then this poses challenges for media effects theory (Bennett & Iyengar, 2008).

The fragmentation and increased diversity of content creates two interlinked challenges for performing *linkage analysis*, a key methodological approach used to measure media effects by combining a content analysis of major media outlets with a (panel) survey to create an individual measure of exposure (De Vreese et al., 2017; Kleinnijenhuis et al., 2007; Scharkow & Bachl, 2017). First, with regard to *exposure measures*, retrospective self-reports tapping into media consumption are

CONTACT Anne Kroon a.c.kroon@uva.nl University of Amsterdam, Amsterdam School of Communication Research (ASCoR), Amsterdam, Netherlands

*These authors contributed equally to this work.

traditionally used. Yet, it is challenging for individuals to recall the full spectrum of media platforms and outlets they were exposed to, let alone how much time they have spent there (Burnell et al., 2021; Deng et al., 2019; Prior, 2009; Sewall et al., 2020). The resulting exposure measure is rather coarse at best, and systematically biased and inaccurate at worst (see for a discussion Otto et al., 2022; Scharkow & Bachl, 2017).

Second, *content measures* typically employed in a traditional linkage design no longer cater to the defining features of our modern-day media environments. Specifically, the increasing fragmentation of (social) media has made the analysis of a few mainstream media outlets less meaningful (Otto et al., 2022). With the long tail of news consumption and media landscape fragmentation, more audience members now rely on niche sources for information (Loecherbach et al., 2022). Consequently, focusing primarily on large-scale legacy outlets in a content analysis overlooks the diversity found in local and fringe content, as well as user-generated content and interactivity in social media. These defining characteristics of media uses are, however, crucial in grasping their effects in our current digital societies.

In this contribution, we focus on the problem of the quantitative analysis of this vast and diverse collection of content into theoretically relevant categories and measures. In particular, we discuss the challenges that digital trace data poses for automated text analysis techniques traditionally employed within the field of communication science and the social sciences. Additionally, we review key opportunities and future directions based on advancements in the field of natural language processing, with a particular focus on the potential of *transfer-learning* using *transformer-based models*. Notably, we argue for the relevance of using models such as *BERT* (Devlin et al., 2019): and *GPT* (Radford et al., 2019) for addressing problems related to the content analysis of digital trace data in a media effects paradigm. Transformers have revolutionized the full spectrum of tools available to computational social scientists by explicitly acknowledging the contextual meaning of language (Acheampong et al., 2021), resulting in a real breakthrough in terms of performance within the field of (political) communication science (Bestvater & Monroe, 2022; Laurer et al., 2023; Licht, 2023; Lin et al., 2023; Viehmann et al., 2022; Widmann & Wich, 2022) and beyond (Devlin et al., 2019).

Our focus will be on the *"top-down"* classification of *a-priori-defined* content measures (see Boumans & Trilling, 2016). This means that we do not focus on "bottom-up" techniques such as topic models, where the aim is to inductively extract patterns (or "topics") from the data; our focus is rather on designs where the categories of interest are defined in advance, typically based on theoretical considerations. These could be news categories (e.g., politics vs economy vs sports), but also whether or not some message contains a threat, an insult, emotional support, etc. Automatic top-down classification is a valuable technique for media effects research because it enables researchers to test hypotheses regarding the effects of exposure to specific content characteristics, such as *frames* or *tonality*, at a scale that facilitates longitudinal and comparative analysis. Until very recently, however, automated approaches were often only of limited use in this regard. Their validity was substantially lower compared to manual approaches, especially for more latent and abstract constructs such as sentiment and hate speech (see e.g., Arango et al., 2019; Van Atteveldt et al., 2021).

This contribution will review challenges and opportunities for automated content analysis along the lines of the diversity of content across individual-level digital traces in terms of *format/style* (1), *language* (2), and *modality* (3). These challenges tend to be especially pronounced when analyzing digital traces and social media data.

First, individuals may visit several legacy news sites and/or read their social media posts, come across and engage with social media posts of friends, colleagues, and strangers, post content online, follow links to niche platforms, or read content written by members from fringe communities (Loecherbach et al., 2022; Vermeer et al., 2020). Obviously, all of these genres will differ considerably in terms of *format/style* (e.g., Lin et al., 2023).

Second, given the high prevalence of English-*language* content on the internet, in many countries, many of the traces may be in English next to the local language — a problem that is even more pronounced in countries with multiple official languages. For instance, Lin et al. (2022) estimate based

on data donations that roughly half of news-related YouTube videos that young Dutch adults watch are in English, and a bit less than the other half in Dutch. This poses an additional burden on the researcher to develop models for different languages, and furthermore requires careful error analysis to see and if possible limit biases across languages. The argument can be extended to different dialects or sociolects - an issue that frequently occurs when studying adolescents.

Third and last, digital traces and social media data are generally also multi-modal, in the sense of containing a combination of text, image, video and/or audio. On platforms such as TikTok or Instagram, the challenges of multi-modality are especially pronounced as images, videos, and texts (comments) go hand-in-hand (e.g. van Driel et al., 2022). Even news from traditional sources, however, is now often consumed in form of a video or podcast (e.g., Lin et al., 2022), highlighting the need to consider *modality* even outside of social media analysis.

These diversity dimensions will serve as the lens to discuss the challenges of classifying digital trace data into useful concepts. The high volume, variety, and unstructured nature logically call for (at least some form of) computational techniques. While manual content analysis may very well remain (an essential) part of the analytical workflow, a fully manual content analysis quickly becomes too labor- and cost-intensive. And as we will argue below, the "classical" automatic content analysis methods that are often used in communication research perform poorly in the face of multiple formats, languages, and modalities.

In contrast, we argue that transformer-based models have the potential to overcome challenges in each of these domains, and therefore aid the development, application, and performance of various computational tools to meaningfully categorize personalized media diets in a media effects paradigm. Furthermore, we will address how transformer-based models enable us to analyze more latent and abstract constructs that before were only possible through manual content analysis. The techniques that will be discussed are more broadly relevant for researchers engaged in any type of content analysis (with or without linkage). Together, we aim to contribute to the literature by highlighting the contemporary challenges faced by communication scholars and providing an updated overview of computational tools available to them (cf. Boumans & Trilling, 2016; Breuer et al., 2020; Elliott et al., 2009; Flew et al., 2012; van Atteveldt & Peng, 2018; Wilkerson & Casas, 2017).

The article progresses as follows. We begin with an overview of traditional content analysis approaches. In particular, we focus on how these methods represent texts as a *bag of words*, which renders these approaches domain-specific but restricts their generalizability. Next, we discuss how the field of natural language processing has evolved from bag of words representations to models that can learn much richer and transferable text representations. This transfer-learning paradigm has enabled groundbreaking performance on a variety of language tasks. After establishing this current state-of-the -art, the potential of transfer learning using large language models for classifying context across different formats, styles, languages, and modalities is discussed, followed by a consideration of the limitations of these models in terms of biases.

Limitations of Traditional Content Analysis Techniques

In this section, we discuss techniques for automated content analysis that are commonly used in the field of communication science. We argue that these techniques pose limitations for the challenges that we addressed above, which we attribute primarily to their pronounced *domain-specificity* and restricted *generalizability*. Domain-specific tools perform well in particular tasks or within specific domains, whereas generalizability refers to their ability to perform effectively across linguistic contexts or domains. We will evaluate these tools and their performance by considering the trade-off between these two factors:

Bag of Words (BoW) models

The maturing field of *Computational Communication Science* (CCS) offers a solid toolkit to analyze media content (for a review, consult Boumans & Trilling, 2016; van Atteveldt & Peng, 2018). In order to identify and classify *predefined* theoretical concepts in textual data— which is part of a deductive, top-down approach- communication scholars have traditionally relied on *Bag-of-Words* (BoW) approaches. In this approach, word order is ignored, and therefore information about the grammatical structure, semantic and syntactic contextual word meanings is lost. This is not without consequence, as contextual linguistic information is highly informative when it comes to understanding language in different domains and contexts (e.g., Grimmer & Stewart, 2013). To mitigate this limitation, BoW approaches sometimes employ the use of brief word sequences of *n* words, known as *n-grams*, to handle short-distance dependencies like *"not good."*. However, this simple approach has various limitations, particularly for *long-distance dependencies*.[1]

While the BoW model is beneficial in terms of speed, explainability, and computational costs, its simplified assumptions challenge the automatic classification of texts especially in the context of high-variance individual-level digital trace data. We discuss these challenges next.

Dictionary-based approaches

The first group of studies that draw on BoW representations uses either "off-the-shelf" or custom-made lexical-based dictionaries or similar rule-based approaches[2] for text classification. These dictionaries are typically created for a specific linguistic and content domain. For instance, there are dictionaries optimized for financial contexts (e.g., Loughran & McDonald, 2011), media domains like social media texts (e.g., Thelwall, 2017), or online discussions (e.g., Stoll et al., 2023). Their validity is closely tied to the context for which a dictionary is developed, as the meanings of wordsand the variety of linguistic expressions vary across domains, making it context-specific. Consequently, dictionaries' performance is typically limited in terms of generalizability. To illustrate, words such as *dope* or *cool* carry a significantly different meaning when used as slang in a social media context than when coined professionally by a professional journalist in an investigative piece or a politician during a parliamentary debate. At the same time, scholars working with digital traces of social media use could anticipate all these different genres in a single data set.

An additional limitation is that the inclusion (and exclusion) of the particular vocabulary terms is not free from *researcher bias* as individual knowledge of and perspective on the issue plays a role in its composition (Burscher et al., 2014; Guo et al., 2016) — although recent work focuses on the automatic creation of dictionaries, their expansion or adaptation to specific domains (e.g., Beigi & Moattar, 2021; Wijayanti & Arisal, 2021). As a consequence, it is highly challenging, if not impossible, to create a unique word list that adequately reflects the *entire* data set under scrutiny, and can validly capture latent or abstract concepts of interest in different communicative contexts (González-Bailón & Paltoglou, 2015; Ribeiro et al., 2016). Indeed, several studies show that the out-of-context application of off-the-shelf dictionaries leads to poor results (Boukes et al., 2020; Loughran & McDonald, 2011).

When working with digital traces of social media use, researchers are per definition confronted with diverse topical domains and media genres, requiring a type of "one-size-fits-all" dictionary that is tailored to and validly generalizes across various linguistic and content domains. It is questionable whether the creation of such a dictionary is feasible given the inherent domain-specific nature of dictionaries, *especially* when the aim is to identify latent or abstract constructs. In any case, when using a dictionary-based approach, careful customization and validation are warranted, but even then, its

[1]For instance, consider the sentence *Zoë, who had always dreamt of becoming a computer scientist, achieved her goals after years of dedication*. Here, the clause *who had always dreamt of becoming a computer scientist* forms a long-distance dependency with the noun *Zoë*, providing important information about Zoë's ambitions and offering important contextual details.

[2]One may argue that rule-based approaches like "find *A* within *n* words distance of *B*" are not based on a strict BoW representation. However, like pure BoW approaches, they only to an extremely limited extent exploit the order and structure of language.

out-of-context performance may remain unsatisfactory in the high-variance context of personalized social media use (see e.g, Boukes et al., 2020; González-Bailón & Paltoglou, 2015; Grimmer & Stewart, 2013; Kroon et al., 2022; Loughran & McDonald, 2011; Soroka & McAdams, 2015; Van Atteveldt et al., 2021; Widmann & Wich, 2022).

Supervised machine learning based on BoW representations

The second group of published empirical work in the field of communication research that utilizes BoW representations employs supervised machine learning. These models, however, have important limitations in terms of generalizability, which we explain in the next paragraph. Drawing on BoW representations, typically, input features are computed as simple word counts, either with or without an additional weighting step. More in particular, in deductive (top-down) classification tasks, supervised classification algorithms are typically trained on the BoW-representations of labeled data, to learn how to automatically classify data into categories. By letting the algorithm estimate which input features (e.g., words or entities) map to specific output labels (e.g., topics, sentiment) using exemplar input-output pairs, this approach has some advantages over dictionary-based approaches. Particularly, one can better account for the particularities of the textual data under investigation and reduce the influence of researcher bias (Burscher et al., 2015a; Van Atteveldt et al., 2021). Supervised machine learning applications that leverage BoW-models have been shown to outperform dictionaries in different contexts (Van Atteveldt et al., 2021; Vermeer et al., 2019; Widmann & Wich, 2022).

At the same time, also machine learning applications that leverage BoW representations do not necessarily scale well across languages, domains, and genres. Some have argued that machine translation in combination with BoW representations works well enough (De Vries et al., 2018), and others have compared different approaches to topic modeling in multilingual corpora (Lind et al., 2022). Still, it seems fair to say that the problem of the automated analysis of multilingual or otherwise diverse corpora is far from solved. We can state that in general and beyond translation issues, this problem tends to originate from a scarcity of labeled data, enabling supervised learning in the exclusive context of the provided training set. Studies published in the field of communication science generally employ task- and domain-specific data sets with generally hundreds or thousands of training examples from a specific source annotated by humans. The models trained on these data sets typically do well (enough) for one domain-specific task, but likely perform poorly beyond that particular context (see Burscher et al., 2015b; Rudkowsky et al., 2018). Simply put, a machine learning model trained on newspaper articles will work substantially worse when predicting news articles written by journalists from outlets not included in its training set.It becomes even more problematic when one tries to use the same classifier to classify political texts (Burscher et al., 2015), tweets, or niche content derived from online fringe communities (see Loecherbach et al., 2022).

To improve the performance of machine learning applications that use BoW representations, obtaining high-quality training data is necessary. In particular, a large and representative data set is needed to come closer to the state-of-the-art performance of modern deep-learning models that leverage millions or even billions of training examples (cf. Laurer et al., 2023; Wilkerson & Casas, 2017). Manually annotating a representative and diverse sample of documents is a time-consuming and costly process, which is an important feasibility consideration to take into account. Due to these challenges, the collection of a large-scale, high-quality manually annotated dataset may be out of reach of those communication scholars that lack extensive financial resources, time and/or strong collaborative networks.

In conclusion, BoW approaches, including dictionary-based and machine-learning applications, are susceptible to variations in genre, format, and linguistic style in the data under investigation, limiting their performance outside of the original context. This is particularly problematic in light of

the diversified nature of digital trace data. For example, a BoW-based classifier may successfully recognize positive affect in legacy news content, but systematically miss expressions of positive sentiment embedded in a person's social media feed. As these mistakes are systematic, they can introduce misleading biases, such as underestimating sentiment in social media content while overestimating sentiment in legacy news content.

From BoW to Large Language Models: Revolutions in text representations

Language is complicated. When people construct sentences, they implicitly use in-depth knowledge of syntax, semantics, morphology and pragmatics to turn the meaning they wish to convey into a sequence of words. It follows that this level of language knowledge is also required to properly infer this meaning from the sentence. To build models that can perform complicated language tasks, such as classifying abstract concepts, we somehow need these models to acquire some of this *language knowledge*.[3]

In this section we outline important developments in the field of natural language processing that have made it possible to create text representations that not only encode much more language knowledge compared to BoW representations, but that are also transferable across tasks. This has given rise to new state-of-the-art models that vastly outperform classic approaches.

Word embeddings

To overcome the limitations associated with the simplistic notions of BoW approaches, and to leverage the benefits from large amounts of training data, researchers have developed a variety of techniques to better model the meaning and structure of the text. Particularly, a large body of work has been devoted to *language modeling*, which refers to the "task of predicting the likelihood of a string given a sequence of preceding or surrounding context words – at its simplest, guess the next word in a sentence" (Gasparetto et al., 2022, p. 8). Nearly a decade ago, this has resulted in the inception of distributed representations of words, also called "word embeddings," vectorial numeric representations that encode part of the meaning of words (Mikolov et al., 2013, 2013).

The first innovation posed by word embeddings is how a word is represented. In a BoW model, words are discrete categories, which can be represented numerically as *one-hot vectors*. This representation is similar to what is more commonly known in our field as a *dummy variable*. It only tells us whether a word encountered in a text is (1) or is not (0) a specific word in our vocabulary.[4] The problem with this representation is that the categories are treated as completely orthogonal, which is obviously not accurate for how words are used in language. Consider the words *money*, *finance*, and *panda*. For a human, *panda* is clearly the odd one out, but for a BoW all three words are equally uncorrelated. Word embeddings can, to some extent, capture this "knowledge" about the semantic relations between words by representing each word as a continuous vector. This is a great boon for training supervised classifiers because they can now learn more from the words in the training data. Even if a model has never seen the word *finance* in the training data, it can still represent this word based on its similarity to words that it has seen (e.g. the word *money*). Consequently, feeding supervised classifiers continuous word representations rather than one-hot encodings has the potential to boost performance in the field of communication science (e.g., Rheault & Cochrane, 2020; Rudkowsky et al., 2018).

The second innovation of word embeddings lies in the discovery that these semantic representations can be learned from any set of texts, without requiring any (manually coded) labels. The key

[3]We use "knowledge" loosely here to indicate only that the model has information that allows it to make better inferences from text. For example, that the word "not" indicates a negation. Whether or not a complex mathematical model that processes negations correctly can be said to possess language knowledge in an ontological sense is of no concern to this article.

[4]For reference, a common BoW representation of a corpus is a document-term matrix, which indicates how often each term occurred in each document. The rows in this matrix (the document vectors) are the one-hot vectors of the terms summed together.

insight is that words with similar syntactic and semantic information tend to occur in similar contexts. Given a large enough collection of texts, we can thus obtain representations of this information by learning to predict words based on the words that surround them. This means that we can learn word embeddings from a huge corpus, such as the entire Wikipedia archive and web crawling data (Mikolov et al., 2017), to obtain representations of words that capture a sense of general semantic meaning across languages. Using so-called Multilingual Sentence Embeddings, this even works with multilingual datasets, and can outperform the more traditional combination of machine translation with BoW machine learning, especially when the training set is small (Licht, 2023). In essence, word embeddings can be seen as an early step toward the idea of transfer learning, where the fruits of the labor from one general task (learning the embeddings) are used to enhance performance on various specific tasks (domain-specific classifiers).

Making models context-aware

A limitation of the first generation of word embeddings is that they are non-contextual. This implies that each vocabulary word receives a single vector, and thus each word carries *one* dominant meaning. In reality, many words that are morphologically similar or identical can carry a different meaning depending on the context in which they are used. To illustrate this, let us consider the sentences: "she put a *date* in his lunchbox" (1); "they went on a *date*" (2); and "what's the *date* today?" (3). The meaning of the word *date* is different in all three sentences, so having a single vector representation can be misleading. Fixed (i.e. non-contextualized) word embeddings are thus limited in their ability to deal with lexical ambiguity, and even simple syntactic modifiers like "not good" (see Loureiro et al., 2021). This limitation recalls a famous quote from the field of linguistics: "The complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously" (Firth, 1935, p. 37)

It follows that to build better models for text analysis, these models need to be context aware. Great strides toward building more context aware models have been made with the introduction of neural networks, in particular the *Convolutional Neural Network* (CNN) and the *Recursive Neural Network* (RNN). It is not crucial for researchers to possess a thorough understanding of the underlying mechanisms of these techniques in order to implement them, similar to the scenario in advanced statistical models where many scholars are unaware of their inner workings. However, we will provide a more abstract explanation to aid a more general understanding of what these techniques are meant to achieve, because this can help to better understand the essence behind transfer learning.

Where n-grams can *represent* words within specific contexts of short sequences, RNNs and CNNs are able to *learn* such contextual representations from the data. CNNs achieve this with convolutional layers that apply filters to transform collections of elements in the input data (e.g., a sequence of words, pixels in an image) into a representation of these collections. This process is easier to understand for images than for text. For a computer, an image is just a 2D matrix where each value represents the color of a single pixel. It does not understand that a row of black pixels on a white background forms a line. To let the computer learn such simple shapes, we can let it look at a small square of pixels (e.g., a 3 by 3 grid) and assign a score for how strongly this square represents a line. Convolutional neural networks provide a way for computers to learn to recognize patterns like lines, or more complicated patterns, that are useful for making good predictions. For text analysis, this same method can be applied to let computers learn to recognize patterns in sequences of words. For example, for a sentiment classification task, it could learn to recognize patterns such as "not good" and "very good." CNNs therefore proved very powerful for text analysis tasks such as sentiment classification (Van Atteveldt et al., 2021).

Recursive Neural Networks take a very different approach, that is specifically geared toward dealing with sequences. This makes them very suitable for texts, because a sentence can be seen as just a sequence of words. A RNN can recursively loop over a sequence such as a sentence to propagate information from the start to the end. This can capture short-distance dependencies like "not good,"

but also dependencies over longer distances. This ability to model words in the context of short and long-distance contexts made them very effective at tasks such as auto-completion and translation.

Regardless of how these techniques work specifically, let us take a step back to look more generally at what makes them so powerful: the ability to *learn* representations of language. Instead of hand-picking what n-grams to use, we need models that can acquire a general "knowledge" of how language works by training on large amounts of data. The next step is to store this information in a way that it can also be transferred to be used in other tasks. This is the essence of transfer-learning, and there is an ongoing research program for building models that learn from more data to acquire more useful "knowledge."

The rise of the transformer

The current state-of-the-art architecture for creating transferable representations of language is the transformer. It is not necessary to understand how a transformer works underneath the hood in order to use it, but given its central position in recent revolutions in NLP, we provide a basic introduction to the key components. Transformers show excellent performance on classification tasks, with BERT (Bidirectional Encoder Representations from Transformers) being the most well known example (Devlin et al., 2019). But transformers can also perform text generation tasks, with GPT (Generative Pre-trained Transformer) being an example that has garnered much attention following the advent of ChatGPT. In the current article we focus primarily on applications of transformers for classification, but this broader picture illustrates just how disruptive transformers have been in the field of NLP.

The transformer is composed of two key components: the *encoder* and *decoder*. Classification models like BERT mainly use the encoder component, and the decoder is the driving mechanism behind tools like ChatGPT and Google Translate. Figure 1 shows a simplified representation of this encoder-decoder architecture, that focuses on the main elements to understand the flexibility of the transformer.

The encoder component takes a text as input and creates a numerical vector representation (or *encoding*) for each word. Models like BERT add a classification head to the encoder to perform classification tasks. But you could also use an encoder simply to obtain the encoding vectors, and use these instead of a BOW or 'regular' word embedding representation in downstream tasks.

The decoder component can take encodings as input and *decode* them into text. As illustrated in Figure 1, this works by first inputting the <bos>; (begin of sentence) token, based on which the decoder generates a token using the language knowledge as encoded in the representation. This token is then also put into the decoder to generate the next token, and this process is repeated until a <eos>; (end of sentence) token is generated. This process can be used for all sorts of text generation tasks,

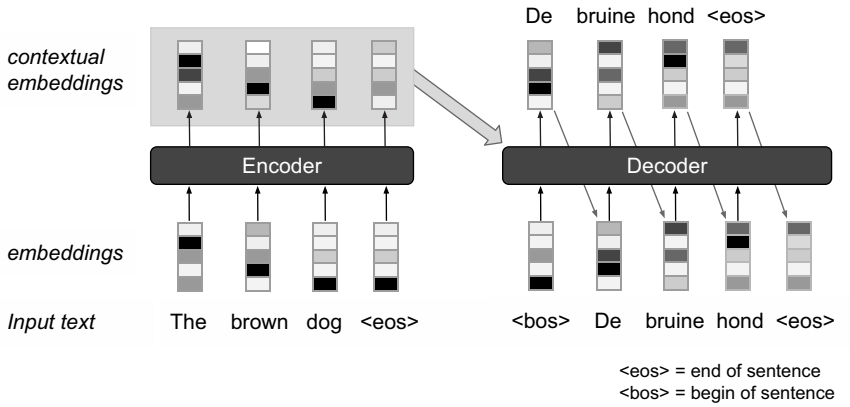

Figure 1. Illustration of the encoder-decoder architecture used in transformers.

ranging from summarization and translation to answering questions. In Figure 1, we illustrate how the sequence "The brown dog" is translated into the Dutch "De bruine hond."

The encoder-decoder architecture in itself had already been used before, and to better understand the break-through innovation behind transformers one needs to look a bit closer underneath the hood of the encoder and decoder to see how the language representation used in transformers differs from the representations discussed above. Vaswani et al. (2017) came up with a new way to use the encoder-decoder architecture to generate powerful representations of language without using convolutions or recurrence, and instead relying entirely on *self-attention*. In the article, aptly titled "Attention is all you need," Vaswani et al. (2017) argue for the benefits of self-attention compared to recurrence and convolution, and demonstrate how combined with the encoder-decoder architecture this produces stellar results.

For an abstract but intuitive understanding of the self-attention mechanism, let us revisit the sentence "she put a *date* in his lunchbox." To understand what the meaning of the word "date" is in the context of this sentence, we primarily need to pay *attention* to the word "lunchbox." In the context of the sentence, it is this specific word that informs us that we should interpret "date" as a piece of fruit. The self-attention mechanism offers a way to calculate how much attention we should pay to "lunchbox," and then to represent the word "date" in a way that incorporates this contextual meaning. It first computes the relations from a word to all other words in the same sentence (i.e. self), and attributes a stronger relation if interpreting the word requires attention for the other word. It then uses this information to update the representation of the word (i.e. the word embedding). Through this process, the self-attention mechanism can transform normal word embeddings, that are the same for every occurrence of the word, into contextualized word embeddings, that allow the same word to have different meanings depending on it's context.[5]

Transformers can learn the parameters for self-attention mechanisms without requiring supervision. Like the process for training word-embeddings, a huge number of training texts is required, but these texts do not need to be labeled. The classic BERT model, for instance, is pre-trained using two tasks: masking random words and then predicting them, and predicting the next sentence (Devlin et al., 2019). This training can be performed on any natural language document, and for the classic BERT model they used a corpus of books and Wikipedia articles with a combined size of 3,300 million words.

The training of these deep neural network models on vast amounts of data necessitates a correspondingly large amount of computational resources. Because of this, they are often referred to as *large* or *deep* Gasparetto et al. (2022), referring to their architecture of multiple layers of neural networks. In this article, we refer to these models as *Large Language Models* (LLMs).

Implementing large language models in research

The key takeaway from the previous section is that the value of LLMs lies in their ability to obtain transferable language knowledge by looking at vast amounts of unstructured, generic textual data. Without the need for laborious labeled texts, they can learn to represent semantic relations between words like "money" and "finance," and even use contextual information to indicate that "a romantic date" and "date of birth" are very different things. This knowledge can then be transferred and used in downstream tasks, such as supervised machine learning. By relying on existing, prior knowledge, we can build more accurate classifiers using less training data (Laurer et al., 2023). Due to this role of LLMs as a foundation that can be adapted to various tasks, they have also been referred to as *foundation models* (Bommasani et al., 2021)

[5]In practice the contextualized embeddings are not created with a single self-attention calculation. In state-of-the-art Transformers like BERT (Devlin et al., 2019), multiple self-attention *heads* are performed in parallel in a process called multi-head attention, and multiple multi-head attention layers are stacked together. In the "normal" *BERT base* model there were no less than 12 layers with 12 attention heads. Each of these 144 self-attention mechanisms featured close to 150,000 parameters.

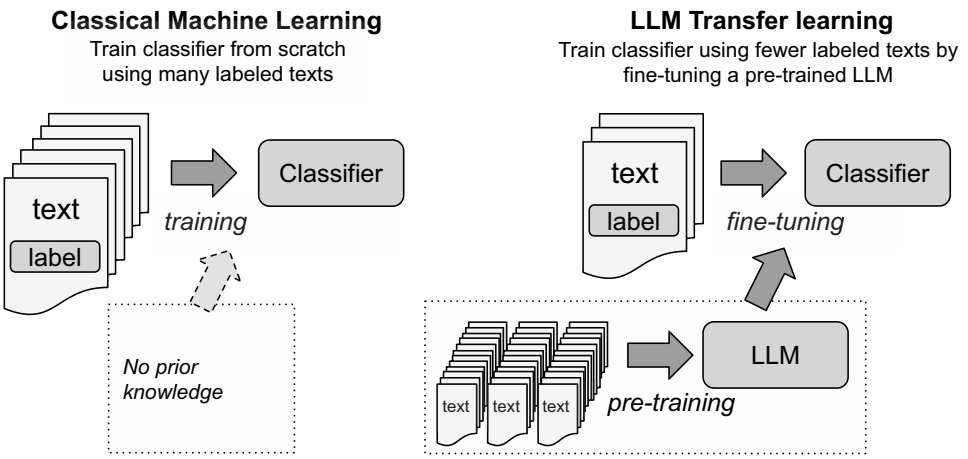

Figure 2. Classical machine learning versus LLM transfer learning.

Figure 2 illustrates how this use of LLMs compares to classical machine learning approaches. In classical machine learning a model is typically *trained* from scratch, using only labeled texts. The key difference when using LLMs is that the training process is conducted in two steps. First, the LLM is *pre-trained* on a huge amount of unlabeled texts. At this stage the model acquires general language knowledge, but has not yet been trained to perform specific classification tasks.[6] In the next step, we can *fine-tune* the LLM to perform a classification task by providing it with labeled examples. Because the pre-trained model already comes equipped with general language knowledge, the amount of data needed to fine-tune it is generally much lower compared to the amount of data needed to train a classic machine learning model. This is particularly relevant in the field of communication science, where we often have limited annotated data about the constructs that we want to investigate.

For the researcher, the practice of using LLMs is not very different from using classical machine learning. The most laborious part would be to pre-train the LLM, which is a very expensive process in time and computational resources. However, researcher do not have to perform the pre-training themselves. Powerful pre-trained models have been made public by large digital platforms (e.g., Google) and research groups. Many of these models are freely available online, for example via the *Hugging Face* library.[7]

In practice, this means that using LLMs as a researcher is not much harder compared to using classical machine learning, and if the challenge of achieving high accuracy is taken into account it might even be considered to be easier. There is one major bottleneck: fine-tuning an LLM is much more computationally expensive and slower, and there are some technical challenges in setting up the hardware. Still, while this makes it more difficult for researchers to adopt the technique, it should not pose a critical barrier. The required hardware is affordable at the price of a high-end personal computer, and it is also possible to hire resources in online computing environments. There are also many tutorials available for getting started that do not require extensive prior computational knowledge. Once a researcher has set-up an environment, leveraging the power of LLMs can be as simple as running a few lines of code to download and fine-tune an existing pre-trained model.

[6]Notably, some LLMs can already perform decent zero-shot classification at this stage, meaning that they can to some extent recognize classes that they haven't yet seen any labeled examples of.
[7]https://huggingface.co/bert-large-cased

Table 1. Overview of the advantages and disadvantages, domain-specificity, and generalizability of various automated techniques.

Method	Advantages	Disadvantages	Domain-Specificity	Generalizability
Dictionary-based classification	Easy to understand, computationally cheap, high degree of interpretability	Limited ability to capture semantic meaning.	Typically tailored to a specific domain	Generally limited performance of 'off-the-shelf' dictionaries to new domains.
Supervised machine learning based on Bag of Words representations	Generally improved performance over dictionary-based/rule-based classifiers, relatively computationally cheap	Limited ability to capture semantic meaning, performance is hindered by out-of-vocabulary words	Models tend to be domain-specific: Performance is best when training and test data are from the same domain.	Performance decreases in out-of-context domains.
Supervised machine learning based on static word embeddings	Better representation of semantic meaning and improved performance of out-of-vocabulary terms. If available, embedding models can easily be pre-trained on large amounts of unlabeled text data	The ability to capture context-dependent meaning is limited.	High domain specificity can be achieved if (1) word embeddings are of high quality and (2) trained on data representative of the application data	Word embeddings capture *static* semantic meaning that may be useful across domains, but generalizability is restricted due to a lack of contextual information.
Transfer learning (using contextualized embeddings from Large Language Models)	Capture context-dependent meaning, improving performance on unseen data.	Computationally expensive, limited interpretability, and the potential to encode structural and societal biases inherent in human communication.	When contextualized embeddings are trained on a data representative of the application domain (in terms of language and domain), domain-specificity tends to be high.	Potential for high generalizability as the pre-trained model can be fine-tuned on the domain-specific target data using *transfer learning*.

Leveraging transfer learning for the analysis of digital trace data

In this section we revisit the key challenges for analyzing individual-level digital trace data to argue how LLMs and transfer learning can help us address them. Table 1 summarizes the key benefits and limitations of the techniques discussed above.

Format and styles

The ability of LLMs to transfer knowledge of language to *various* target applications makes them particularly useful to understand and classify the large variety that exists in digital trace data. As a starting point, researchers could leverage a generic pre-trained LLM and fine-tune it on a representative sample of digital trace data.

For example, to trace the effects of Public Service Announcements (PSA) about mental health on social media, one could collect a representative sample of social media platforms from which digital traces are collected (e.g., YouTube, Twitter, Instagram), and manually code that data set for the presence of mental health-related PSA's. Next, one could leverage a pre-trained LLM and fine-tune this on the annotated dataset, which will result in a social media-based model that can classify mental health PSA content in the remaining social media messages in the full digital trace data set. By leveraging a transfer learning framework, one can train such a model with significantly lower time, financial resources, and environmental costs. The resulting *fine-tuned* model will likely perform better compared to training a model from scratch (Devlin et al., 2019).

Recent empirical work corroborates these claims in the context of political communication. Particularly, Laurer et al. (2023) find that transfer learning outperformed classical supervised machine learning on a broad set of eight different classification problems common within political communication, such as the classification of *topics* and *sentiment*. Importantly, they found that using transfer learning with only 500 data points achieved similar results to traditional algorithms trained with 5,000 data points. Similarly, Widmann and Wich (2022) reported that a transformer-based approach using transfer learning outperformed machine learning based on static word embeddings in classifying discrete emotions. However, it is important to note that just because one approach outperforms others does not necessarily mean that it performs adequately. In this case, although the transformer-based approach did outperform all other methods, the F1 scores for 6 out of 8 emotions were below .7. This raises questions about the acceptability of these scores, and it is certainly debatable whether they are seen as adequate.

Transfer learning represents a promising strategy to boost performance in any type of automated content analysis, but especially when working with the high-dimensional and voluminous nature of digital trace data. At the same time, and although less training data might be needed compared to training a model from scratch, it is still necessary to provide the classifier with sufficient training examples of the different domains — as training the model with a single domain dataset could result in overfitting for that specific context causing performance degradation (e.g. Ryu et al., 2022). When labeled data is highly skewed or otherwise not representative of the different domains or genres present in the digital trace data set, it becomes problematic to fine-tune LLMs to the undersampled contexts, degrading its performance. To overcome this problem, several promising applications focus on increasing the *domain* and *genre* awareness of transformer-based models. In particular, a LLM can be adapted for a specific domain by using additional pre-training on relevant data. For instance, in the study by Sun et al. (2021) on automatically classifying incivility on Reddit content, the researchers used BERT and further pre-trained it on 3 million unlabeled Reddit comments to make the model more sensitive to the language and context of the research question. This approach can be effective when there is a large amount of unlabeled data available from the target domain. This additional pre-training step may help to customize the language model for the specific application domain the research is interested. Subsequently, the customized LLM can fine-tuned for the specific classification task. Additionally, previous works shows that combining transfer-learning with a classifier that enables the model to learn domain-independent features improves the detection of fake news across domains (Shu et al., 2022). Finally, transformer-based methods combined with domain adaptation frameworks have been shown to boost cross-domain sentiment analysis (Du et al., 2020) and author profiling (Barlas & Stamatatos, 2020; Delmondes Neto & Paraboni, 2022).

Multilingual content

A particular benefit of using transfer-learning in the context of digital trace data is that it allows scholars to measure concepts in textual data that covers multiple nationalities and languages. So far, most research in the field of communication science has used automated text analyses geared toward one language only, among which English stands out (Boumans & Trilling, 2016; Guo et al., 2022; Lind et al., 2022). More in particular, dictionaries are often available in a single language, while supervised algorithms using BoW models are generally not well equipped to deal with multilingual content — unless researchers have recruited and trained coders with a particular understanding of a given language to perform to label the content. This is, however, challenging due to high costs and difficulties in achieving inter-coder reliability.

Transformer-based models can facilitate the analysis of non-English or multilingual content in digital trace data. When working with digital trace data from a particular national setting, *monolingual* pre-trained models might be useful. These language models are available for a broad spectrum of languages, but do not *directly* account for the multilingual nature of digital trace data.

When confronted with multilingual content in datasets of individual's social media use, researchers can roughly choose from two routes. First, a simple approach, in which the application data is fed through a translation model (such as Google Cloud), which transforms all text to the reference language, after which a generic language model (such as BERT) is applied (see e.g., Bach et al., 2022). Second, *multilingual* pre-trained models, such as *mBert* or *XLM-R* are available for a wide range of cross-lingual classification tasks. XLM-R is a commonly used model that extends BERT by using internet content data of 100 languages, and obtains state-of-the-art performance in cross-lingual classification (Conneau et al., 2019). Such multilingual models that can scale across languages are arguable especially useful in the multidimensional and multilingual context of digital trace data.

Multi-modal content

Computational communication research has for a long time focused primarily on the analysis of natural language (van Atteveldt & Peng, 2018). This was easily justified given that much work was concerned with news content, which was for a large part consumed in textual format. From a methodological point of view this was a blessing of convenience. The computational analysis of natural language was already a well-established field, whereas automatic analysis of image, audio, and video content for a long time seemed nigh impossible.

At present this defense is not holding strong. Image, video, and audio content is becoming increasingly more common, especially on popular social media platforms such as Instagram and TikTok. This makes the analysis of multi-modal content a critical next challenge to overcome (Poria et al., 2016). This challenge is particularly (though not exclusively) relevant for the analysis of personal digital trace data, as can for instance be collected via data donation methods. In this case, data is collected with the person at the center, and a selective focus on textual content would impose a substantial blind spot.

But the field of computational methods has not been idle. Great strides have been made toward adopting methods for visual content analyzes into our toolkit (Casas & Williams, 2019, 2022). This not only enables the analysis of images at scale but also lays down the first stones toward the analysis of multi-modal content. Indeed, a recent special issue in *Computational Communication Research* on the topic of visual analysis shows that many of the authors in this field "are already explicitly working in a multi-modal space, combining images with text, or video with text" (Casas & Williams, 2022, p. 7).

The challenge of analyzing multi-modal content (e.g., words and images) is that their meaning is not simply the sum of the individual modalities. The picture next to a news story or social media post cannot be separated from the words, and even if we can separately analyze the words and image, it is not always evident how to put this information back together. An alternative, therefore, is to combine multiple modalities in the same model. For example, a machine learning model for predicting hate speech in social media posts could take complete social media posts as input, combining the image and natural language data. The goal of multi-modal models is that the image and natural language information are then not just processed separately, but in context of each other.

The combination of image and natural language data in particular is already showing great promise. This type of multi-modal data is also very common (e.g., newspaper articles, social media posts, images with captions, memes), which not only means that these models are useful, but also that sufficient training data is available. Models that jointly learn from both image and words have been shown to be capable of state-of-the-art results on various downstream tasks (Huang et al., 2020; Majumdar et al., 2020). Li et al. (2019) demonstrate that their VisualBERT model performs well on complicated tasks like "Visual Commonsense Reasoning," in which models are asked to "infer people's actions, goals and mental states" from images (Zellers et al., 2019, p. 6720). Perhaps most relevant to our field is that this level of visual-and-language understanding seems to improve performance on classification tasks that require complex interactions between visual and language cues to be taken into account. Maheshwari

and Nangi (2022) for instance used a multi-modal BERT model that uses both visual and language information to better detect misogynistic content in memes.

The development of models that can accurately extract useful inferences from such complicated multi-modal texts is still in its infancy. As much as there is to hope for, there is also reason to fear how these ever more complicated models can hold harmful and difficult to detect biases. Both the hopes and fears are reasons why communication scientists should pay heed to the development of these models. The step toward multi-modal content analysis seems inevitable to study present day media consumption, and we need to inventorize and validate our toolkit accordingly.

Limitations: biases in LLMs

The introduction of transfer learning to leveraging large-scale language models to the field of communication science comes with its own risks and limitations. Importantly, LLMs are generally trained on massive amounts of textual data scraped from the web. The environmental impact of training models that are ever-increasing in size is substantial and warrants careful consideration (Bender et al., 2021). Additionally, the data used for training LLMs is by no means neutral (Founta et al., 2018; Hutchinson et al., 2020), but inherits structural and societal biases ingrained in human communication (Bender et al., 2021). As a consequence, large language models have the potential to embed racial, political, gender, or other harmful societal biases (e.g., Abid et al., 2021; Guo & Caliskan, 2021; Nadeem et al., 2020; Silva et al., 2021). It is difficult to avoid such biases during the training phase, as even models that are trained on arguably more "neutral" or selective content, such as Wikipedia or journalism content (i.e., online and print articles) exhibit societal biases (Huang et al., 2019). Importantly, simply increasing the size of the training data does not lead to greater diversity and equity in these models. Instead, it's crucial to focus on carefully curating and documenting the training data (Bender et al., 2021).

It follows that, when fine-tuning a LLM for a particular classification task, it is important to recognize that the transfer of knowledge may encompass social biases as well (Bender et al., 2021). The act of fine-tuning a model does not magically remove such intrinsic bias of these language models, rather, such biases may seep through to the target application (see for example Urman & Makhortykh, 2022; Zhang et al., 2019). For example, a sentiment classifier that leverages a pre-trained language model containing bias may be more likely to return the label *negative* when particular countries, occupations, or societal groups are mentioned.

As a consequence, an emerging body of work focuses on the question of how to remove such biases from LLMs – often under the banner of *de-biasing* language models. Yet, despite continuous efforts in this domain, it remains notoriously difficult to truly "neutralize" LLMs (e.g, Liu et al., 2021; Qian et al., 2019; Silva et al., 2021). As biases are often implicit in nature and context-dependent, it is difficult to fully account for the wide scope of societal biases ingrained in these models.

For social scientists, the potential impact of biases on the classification accuracy of latent concepts is a significant concern. To address this issue, it is important to carefully evaluate the validity and quality of the dataset used for fine-tuning in transfer learning. Recent work suggests that biases often originate from the fine-tuning dataset (Steed et al., 2022). Hence, it is important to ensure that this dataset is diverse and representative of the intended target population. To validate the results of their research, scholars may also evaluate classifiers trained using transfer learning by inspecting correlations of performance indicators with possible extra-textual independent variables. For example, one could compute correlation coefficients to measure the strength and direction of the relationship between the errors made by the classifier and the extra-textual variables, such as party or source.

To further reduce the potential impact of biases in their research, social scientists may want to consider using *debiasing strategies* such as counterfactual data augmentation or adversarial learning. First, *counterfactual data augmentation* involves increasing the balance of sensitive categories in the training corpus, by swapping them out with alternative sentences (such as by swapping *he* and *she* to yield a more balanced representation) (Lu et al., 2019; Zmigrod et al., 2020). Second, *adversarial learning*, involves adding an additional layer to the training process to weaken the model's ability to

learn from protected classes (such as gender, age, and ethnicity) to predict the target. In this manner, the model learns to optimally predict the target category while minimizing the adversary's predictive power (Zhang et al., 2018). These and other debiasing techniques may help scholars in reducing the potential for bias in their automated classifications.

Discussion

This contribution discussed the challenges and opportunities for the automated content analysis of digital trace data in a media effects paradigm. Digitalization has steadily broken up the shared media buffet served by mass media channels. With digital algorithmic curation and selective exposure serving highly personalized digital experiences based on user profiles and behavior, traditional approaches for measuring media effects have started to lose their meaning. In particular, it has reduced the reliability and validity of traditional linkage analysis combining content analytical data of mainstream outlets with self-reported media.

The combination of digital trace data with self-reports can overcome some of these limitations, and advance our understanding of individual-level media effects. Such an approach warrants the sensible grouping, categorizing, and analyzing of digital trace data. The particularities of digital trace data, especially their high volume, variety, and unstructured nature, logically call for (at least some form of) computational techniques. The particularities of this data, however, also introduce a range of new challenges for communication scholars.

The article provided a discussion of some of the limitations of commonly used deductive or "top-down" techniques in the field of automated content analyses, namely dictionary-based analyses and supervised machine learning using BoW representations. We have discussed that the simplified assumptions of dictionary-based and BoW approaches regarding the contextualized meaning of language hinders these models to perform well in contexts, topical domains, or time frames for which they were not directly designed. The denial of contextual knowledge in BoW representations leads to a loss of information; hindering accuracy (Grimmer & Stewart, 2013). As a consequence, these techniques are typically limited in the extent to which they can validly and reliably scale up analyses across substantive domains and media genres – which is needed when working with digital trace data.

To better account for the full spectrum of diversity present in personalized media diets, the article has argued for the use of *transfer learning* using state-of-the-art large language models. Such models help communication scholars to reap the benefits from much larger amounts of training data and introduce contextual knowledge to classification tasks. We argue that this will help us deal with the problematic diversity of content across individual media diets in terms of format/style (1), language (2), and modality (3). The application of pre-trained language models is particularly useful to understand and classify the large variety that exists in digital trace data.

In this article, we have explicated the potential benefits of using transfer learning in the realm of digital trace data. Despite a growing body of literature suggesting the effectiveness of this approach (Bestvater & Monroe, 2022; Laurer et al., 2023; Licht, 2023; Lin et al., 2023; Widmann & Wich, 2022); empirical evidence remains relatively limited due to its novelty. Consequently, more work is necessary to evaluate the potential for transfer learning to address the particular challenges of analyzing digital traces of human behavior. Specifically, future research should further investigate the conditions that enable transfer learning to enhance our understanding of the impact of media content on media effects, including its influence on cognition, emotions, and behavior. The integration of these state-of-the-art computational methodologies within the field of media effects research holds great promise to deepen our understanding of media effects of current digitalized and personalized media diets.

To conclude, we emphasize that the still recent rise of transformers in the field of NLP is not merely a hype, but offers stunning improvements in supervised machine learning performance that can help us address the complexities of today's digital content. Although there are definitely hurdles involved in the adoption of these innovations, there are both short and long-term merits in investing in the

required resources and training. Platforms such as *Hugging Face* provide accessible tutorials and resources, and developments in computing power make it possible to employ powerful pre-trained models using local devices and affordable cloud computing. R packages like grafzahl (Chan, 2023) make it easy to fine-tune transformer for a wider audience of communication researchers. Making transformers "smaller, faster, cheaper, and lighter" (Sanh et al., 2019) is an active field of research that will further simplify their use (Tay et al., 2020). On the short term, great boosts in performance can be gained by fine-tuning a pre-trained model on a manually coded sample of digital trace data. The use of multilingual models is already within reach, and we should invest in validation efforts to determine their value to our field. In the longer term, these models hold realistic promise of moving beyond the realm of text, toward the inclusion of the multi-modal content that is increasingly prominent in the online experiences of many.

In short, there are many reasons to invest in adopting these innovations in our field sooner rather than later. We expect that these techniques will help the field gain a deeper understanding of the content that individuals use, create, and engage with, moving to a nuanced understanding of media effects across various contexts, languages, and modalities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Anne Kroon is an associate professor at the Amsterdam School of Communication Research at the University of Amsterdam. Her research centers on employing computational techniques to explore the causes and consequences of bias in algorithms in the domain of digital job markets.

Kasper Welbers is an assistant professor at the Department of Communication Science at the Vrije Universiteit Amsterdam. His research focuses primarily on how the gatekeeping process of news messages has changed due to the rise of new media technologies, and how we can study this using computational methods.

Damian Trilling is associate professor of political communication and journalism at the University of Amsterdam. He is interested in news use and dissemination and in the adoption and development of computational methods.

Wouter van Atteveldt is professor of Computational Communication Science and Political Communication at the Vrije Universiteit Amsterdam. He focuses on automatic analysis of (political) communication, including both traditional and social media, and the methods and data required for studying this.

ORCID

Anne Kroon http://orcid.org/0000-0001-7600-7979
Kasper Welbers http://orcid.org/0000-0003-2929-3815
Damian Trilling http://orcid.org/0000-0002-2586-0352
Wouter van Atteveldt http://orcid.org/0000-0003-1237-538X

References

Abid, A., Farooqi, M., & Zou, J. (2021, May 19 - 21). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 aaai/acm conference on ai, ethics, and society*, Virtual Event USA (pp. 298–306).

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021, 12). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, Paris France (pp. 45–54).

Bach, R. L., Kern, C., Bonnay, D., & Kalaora, L. (2022). Understanding political news media consumption with digital trace data and natural language processing. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185* (Supplement_2), S246–S269. https://doi.org/10.1111/rssa.12846

Barlas, G., & Stamatatos, E. (2020). Cross-domain authorship attribution using pre-trained language models. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 255–266). New York: Springer International Publishing.

Beigi, O. M., & Moattar, M. H. (2021). Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. *Knowledge-Based Systems*, *213*, 106423. https://doi.org/10.1016/j.knosys.2020.106423

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3 - 10). On the dangers of stochastic parrots: Can language models be too big?[U+FFFD][U+FFFD]. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, Canada (pp. 610–623).

Bennett, W. L., & Iyengar, S. (2008). A new era of minimal effects? the changing foundations of political communication. *Journal of Communication*, *58*(4), 707–731. https://doi.org/10.1111/j.1460-2466.2008.00410.x

Bestvater, S. E., & Monroe, B. L. (2022, April). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, *31*(2), 235–256. https://doi.org/10.1017/pan.2022.10

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & others. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14*(2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Breuer, J., Wulf, T., & Mohseni, M. R. (2020, August). New formats, new methods: Computational approaches as a way forward for media entertainment research. *Media and Communication*, *8*(3), 147–152. https://doi.org/10.17645/mac.v8i3.3530 https://www.cogitatiopress.com/mediaandcommunication/article/view/3530

Burnell, K., George, M. J., Kurup, A. R., Underwood, M. K., & Ackerman, R. A. (2021). Associations between self-reports and device-reports of social networking site use: An application of the truth and bias model. *Communication Methods and Measures*, *15*(2), 156–163. https://doi.org/10.1080/19312458.2021.1918654

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190–206. https://doi.org/10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015a). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015b, May). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, *659* (1), 122–131. https://doi.org/10.1177/0002716215569441

Burscher, B., Vliegenthart, R., & Vreese, C. H. D. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The Annals of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Casas, A., & Williams, N. W. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, *72*(2), 360–375. https://doi.org/10.1177/1065912918786805

Casas, A., & Williams, N. W. (2022). Introduction to the special issue on images as data. *Computational Communication Research*, *4*(1), 1–10. https://doi.org/10.5117/CCR2022.1.000.CASA

Chan, C.-H. (2023). Grafzahl: Fine-tuning transformers for text data from within R. *Computational Communication Research*, 5(1), 76–84. https://github.com/chainsawriot/grafzahl

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Delmondes Neto, J. P., & Paraboni, I. (2022). Multi-source BERT stack ensemble for cross-domain author profiling. *Expert Systems*, *39*(3), e12869. https://doi.org/10.1111/exsy.12869

Deng, T., Kanthawala, S., Meng, J., Peng, W., Kononova, A., Hao, Q., & David, P. (2019). Measuring smartphone usage and task switching with log tracking and self-reports. *Mobile Media & Communication*, *7*(1), 3–23. https://doi.org/10.1177/2050157918761491

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of deep bidirectional transformers for language understanding. (arXiv:1810.04805[cs]). http://arxiv.org/abs/1810.04805

De Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). Linking survey and media content data: Opportunities, considerations, and pitfalls. *Communication Methods and Measures*, *11*(4), 221–244. https://doi.org/10.1080/19312458.2017.1380175

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Du, C., Sun, H., Wang, J., Qi, Q., & Liao, J. (2020). Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4019–4028). Association for Computational Linguistics.

Elliott, E. W., Ho, K., & Holmes, J. S. (2009, May). Political science computing: A review of trends in computer evolution and political science research. *Journal of Information Technology & Politics*, *6*(2), 166–175. https://doi.org/10.1080/19331680902821569

Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society*, *34*(1), 36–73. https://doi.org/10.1111/j.1467-968X.1935.tb01254.x

Flew, T., Spurgeon, C., Daniel, A., & Swift, A. (2012, April). The promise of computational journalism. *Journalism Practice*, *6*(2), 157–171. https://doi.org/10.1080/17512786.2011.616655

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., & Kourtellis, N. (2018, April). *Large scale crowdsourcing and characterization of twitter abusive behavior*. arXiv. Retrieved September 7, 2023, from (arXiv:1802.00393[cs]) http://arxiv.org/abs/1802.00393

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, *13*(2), 83. https://doi.org/10.3390/info13020083

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. https://doi.org/10.1177/0002716215569192

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guo, W., & Caliskan, A. (2021, July). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 122–133). Virtual Event USA: ACM. Retrieved 2022-12-28, from. https://doi.org/10.1145/3461702.3462536

Guo, L., Su, C., Paik, S., Bhatia, V., Akavoor, V. P., Gao, G., & Wijaya, D. (2022). Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*, *11*(2), 276–297. https://doi.org/10.1080/21670811.2022.2031241

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., & Kohli, P. (2019). Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5491–5501). Online: Association for Computational Linguistics. Retrieved 2023-09-07, from https://doi.org/10.18653/v1/2020.acl-main.487

Jürgens, P., & Stark, B. (2022). Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication*, *72*(3), 322–344. https://doi.org/10.1093/joc/jqac009

Kleinnijenhuis, J., Van Hoof, A. M., Oegema, D., & De Ridder, J. A. (2007). A test of rivaling approaches to explain news effects: News on issue positions of parties, real-world developments, support and criticism, and success and failure. *Journal of Communication*, *57*(2), 366–384. https://doi.org/10.1111/j.1460-2466.2007.00347.x

Kroon, A. C., van der Meer, T., & Vliegenthart, R. (2022, October). Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, *4*(2), 528–570. https://doi.org/10.5117/CCR2022.2.006.KROO

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 1–33. https://doi.org/10.1017/pan.2023.20

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, *31*(3), 366–379. https://doi.org/10.1017/pan.2022.29

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2022). Building the bridge: Topic modeling for comparative research. *Communication Methods and Measures*, *16*(2), 96–114. https://doi.org/10.1080/19312458.2021.1965973

Lin, Z., Welbers, K., Vermeer, S., & Trilling, D. (2022, October 20–22). Who is watching what? Exploring news consumption on YouTube through data donation. In *European Communication Conference (ECREA)*. Aarhus, Denmark.

Lin, Z., Welbers, K., Vermeer, S., & Trilling, D. (2023). Beyond discrete genres: Mapping news items onto a multidimensional framework of genre cues. In *International Conference on the Web and Social Media (ICWSM)*. (https://arxiv.org/abs/2212.04185)

Liu, R., Jia, C., Wei, J., Xu, G., Wang, L., & Vosoughi, S. (2021, 2-9 February). Mitigating political bias in language models through reinforced calibration. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 14857–14866).

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2022). *Don't miss the long tail: Website classification to identify local and niche news*. Paris: ICA Conference 2022.

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. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, *47*(2), 1–57. https://doi.org/10.1162/coli_a_00405

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2019, May). *Gender bias in neural natural language processing*. arXiv. (arXiv:1807.11714 [cs]). Retrieved December 28, 2022, from http://arxiv.org/abs/1807.11714

Maheshwari, P., & Nangi, S. R. (2022). Teamotter at semeval-2022 task 5: Detecting misogynistic content in multimodal memes. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)*, Seattle, United States (pp. 642–647).

Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., & Batra, D. (2020). Improving vision-and-language navigation with image-text pairs from the web. In *European conference on computer vision*, Glasgow, United Kingdom (pp. 259–274).

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Nadeem, M. L. A. I. L., Bethke, A., & Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Otto, L. P., Thomas, F., Glogger, I., & De Vreese, C. H. (2022). Linking media content and survey data in a dynamic and digital media environment–mobile longitudinal linkage analysis. *Digital Journalism*, *10*(1), 200–215. https://doi.org/10.1080/21670811.2021.1890169

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, *174*, 50–59. https://doi.org/10.1016/j.neucom.2015.01.095

Prior, M. (2009). The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly*, *73*(1), 130–143. https://doi.org/10.1093/poq/nfp002

Qian, Y., Muaz, U., Zhang, B., & Hyun, J. W. (2019). Reducing gender bias in word-level language models with a gender-equalizing loss function. *arXiv preprint arXiv:1905.12801*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAi Blog*, *1*(8), 9.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133. https://doi.org/10.1017/pan.2019.26

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*(1), 1–29. https://doi.org/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. https://doi.org/10.1080/19312458.2018.1455817

Ryu, M., Lee, G., & Lee, K. (2022). Knowledge distillation for BERT unsupervised domain adaptation. *Knowledge and Information Systems*, *64*(11), 3113–3128. https://doi.org/10.1007/s10115-022-01736-y

Sanh, V. L. A. I. L., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* .

Scharkow, M., & Bachl, M. (2017, July). How measurement error in content analysis and self-reported media use leads to minimal media effect findings in linkage analyses: A simulation study. *Political Communication*, *34*(3), 323–343. https://doi.org/10.1080/10584609.2016.1235640

Sewall, C. J., Bear, T. M., Merranko, J., & Rosen, D. (2020). How psychosocial well-being and usage amount predict inaccuracies in retrospective estimates of digital technology use. *Mobile Media & Communication*, *8*(3), 379–399. https://doi.org/10.1177/2050157920902830

Shu, K., Mosallanezhad, A., & Liu, H. (2022). Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection* (pp. 215–232). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1524-6_9

Silva, A., Tambwekar, P., & Gombolay, M. (2021). Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2383–2389).

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, *32*(1), 1–22. https://doi.org/10.1080/10584609.2014.881942

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3524–3542). Dublin, Ireland: Association for Computational Linguistics. Retrieved 2022-12-28, from. https://doi.org/10.18653/v1/2022.acl-long.247

Steppat, D., Castro Herrero, L., & Esser, F. (2022, February). Selective exposure in different political information environments – how media fragmentation and polarization shape congruent news use. *European Journal of Communication*, *37*(1), 82–102. https://doi.org/10.1177/02673231211012141

Stoll, A., Wilms, L., & Ziegele, M. (2023). Developing an incivility dictionary for German online discussions–a semi-automated approach combining human and artificial knowledge. *Communication Methods and Measures*, *17*(2), 131–149. https://doi.org/10.1080/19312458.2023.2166028

Sun, Q., Wojcieszak, M., & Davidson, S. (2021, November). Over-time trends in incivility on social media: Evidence from political, non-political, and mixed sub-reddits over eleven years. *Frontiers in Political Science*, *3*: 741605. https://doi.org/10.3389/fpos.2021.741605

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 55(6), 1–28. https://doi.org/10.1145/3530811

Thelwall, M. (2017). The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions* (pp. 119–134). Springer International Publishing. https://doi.org/10.1007/978-3-319-43639-5_7

Urman, A., & Makhortykh, M. (2022). "Foreign beauties want to meet you": The sexualization of women in google's organic and sponsored text search results. *New Media & Society*, *14614448221099536*, 146144482210995. https://doi.org/10.1177/14614448221099536

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

Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures, Online First*, *16*(4), 266–282. https://doi.org/10.1080/19312458.2022.2109608

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vermeer, S. A., Araujo, T., Bernritter, S. F., & van Noort, G. (2019). Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. *International Journal of Research in Marketing*, *36*(3), 492–508. https://doi.org/10.1016/j.ijresmar.2019.01.010

Vermeer, S. A., Trilling, D., Kruikemeier, S., & de Vreese, C. (2020). Online news user journeys: The role of social media, news websites, and topics. *Digital Journalism*, *8*(9), 1114–1141. https://doi.org/10.1080/21670811.2020.1767509

Viehmann, C., Beck, T., Maurer, M., Quiring, O., & Gurevych, I. (2022). Investigating opinions on public policies in digital media: Setting up a supervised machine learning tool for stance classification. *Communication Methods and Measures*, *17*(2), 150–184. https://doi.org/10.1080/19312458.2022.2151579

Widmann, T., & Wich, M. (2022, June 29). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 1–16. https://doi.org/10.2139/ssrn.4127133

Wijayanti, R., & Arisal, A. (2021). Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *20*(1), 1–16. https://doi.org/10.1145/3425632

Wilkerson, J., & Casas, A. (2017, May). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F. F., & Weeks, B. (2021, May). Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. *The International Journal of Press/politics*, *27*(4), 860–886. https://doi.org/10.1177/19401612211009160

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, Long Beach, CA, USA (pp. 6720–6731).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340). New Orleans LA USA: ACM. Retrieved 2022-12-28, from. https://doi.org/10.1145/3278721.3278779

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., & Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2020, May). *Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology*. arXiv. (arXiv:1906.04571 [cs]). Retrieved December 28, 2022, from http://arxiv.org/abs/1906.04571