Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 7.1 (2025) 1–32
HTTPS://DOI.ORG/10.5117/CCR2025.1.1.NEVE

Boosting Transformers: Recognizing Textual Entailment for Classification of Vaccine News Coverage

Luiz Neves
Federal University of Goiás
Brazilian Institute of Public Communication of Science and Technology

Chico Q. Camargo
University of Exeter
Ewha Womans University, South Korea

Luisa Massarani
Brazilian Institute of Public Communication of Science and Technology
Oswaldo Cruz Foundation

Abstract
The introduction of Transformers neural network revolutionized Natural Language Processing by effectively handling long-range dependencies and context. Models like BERT and GPT are at the forefront of Large Language Models and have been used in text classification. Despite their benchmark performance, real-world applications pose challenges, including the requirement for substantial labeled data and class balance. Few-shot learning approaches, like the Recognizing Textual Entailment framework, have emerged to address these issues. RTE identifies relationships between a text T and a hypothesis H. T entails H if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T. This study explores an RTE framework for classifying vaccine-related headlines using 1,000 labeled data points distributed unevenly across 10 classes. We evaluate eight models and procedures, including both open-source and closed-source, as well as paid and free options. They were tested from four perspectives. The results highlight that deep transfer learning, combining language and task knowledge, like Transformers and RTE, enables the development of text classification models with superior performance, addressing data scarcity and class imbalance. This approach provides a valuable protocol for creating classification models and delivers an automated model for classifying vaccine-related content.

Keywords: Natural Language Processing, Transformers, Recognizing Textual Entailment, BERT, GPT

© The authors. This is an open access article distributed under the CC BY 4.0 license

Introduction

Just over a decade ago, the study of social phenomena through large scale data—the then emerging computational social science—was much closer to the corporate world than to the academic domains (Lazer et al., 2009). The confluence of at least three developments—the growing availability of digital data, the improvement of tools to analyze them, and the emergence of powerful and accessible processing capacity (van Atteveldt & Peng, 2018)—contributed to the following years being considered as the "new era" in this field (Wallach, 2016). Currently, computational social science continues to expand as an academic discipline, evident in the increasing interest from researchers (Metzler et al., 2016) and the rising number of publications each year (Liu et al., 2022b). This has led to the establishment of an international and interdisciplinary community of scholars (Watts, 2016) and a wide range of methods, process, and tools able to address old and new research questions and challenges (Chen at al., 2021).

In this context, significant advancements have been achieved in Natural Language Processing (NLP), an area of Artificial Intelligence dedicated to enabling computers to understand, interpret, and generate human language (Chowdhary, 2003). This development has provided researchers the opportunity to delve into extensive textual databases, which is an advantage not limited to the possibilities of deeper analysis, but also in terms of validation, generalization, and reproducibility of the studies (Widemann, 2013). One crucial aspect of NLP is the vector representation of text, where words, phrases, or sentences are encoded as numerical vectors to capture their semantic relationships (Turian et al., 2010). This concept is rooted in the distributional hypothesis, which posits that words appearing in similar contexts have similar meanings. For many years, some popular ways to do this included using neural networks like Word2Vec (Mikolov et al., 2013), matrix factorization such as GloVe (Pennington et al., 2014), and prediction models like fastText (Joulin et al., 2017). The main goal of these methods is utilizing large text corpora to learn dense representations of words, effectively capturing semantic nuances.

A turning point occurred with the introduction of Transformers (Vaswani et al., 2017), a type of neural network that revolutionized NLP and other tasks in machine learning. Transformers rely on a mechanism called self-attention to process input data in parallel, giving models the ability to capture relationships between words regardless of their positions in a sequence. This

attention mechanism enables Transformers to excel in tasks such as language translation, text generation, and sentiment analysis. The architecture's success is attributed to its capacity to handle long-range dependencies and capture context effectively, making it a cornerstone in modern AI advancements and models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPT (Generative Pre-trained Transformer) (Radford et al., 2018).

GPT models are one of the most famous classes of Large Language Models (LLM), trained on massive amounts of text data to acquire a deep understanding of language structures and semantics. These models consist of millions to billions of parameters to generate coherent and contextually relevant text, perform complex language-related tasks, and exhibit human-like language abilities (Wei et al., 2022). These pre-trained models can be fine-tuned for various downstream tasks, such as text classification, which is why they have been widely used in communication studies (Ziems et al., 2023). In text classification, the models' contextualized embeddings are fed into a classification layer of the neural network, and fine-tuned on a specific dataset. This approach falls within the category of supervised machine learning (Burkart & Huber, 2021) and has allowed models to learn task-specific features and achieve state-of-the-art results in various text classification challenges, benefiting from its contextual understanding of language (Sun et al., 2019). It is important to note that, although GPT models are accessible through API services, their underlying code and model weights are not publicly available, and there is a cost associated with their use.

Although the superior performance of these models is being attested in numerous NLP benchmarks, their application in real-world tasks is not without challenges. One of the biggest obstacles for many social scientists is having available not only a sufficient amount of labeled data to fine-tune the model, but also ensuring that the classes or topics of this training dataset are balanced (Wilkerson & Casas, 2017). This has led scholars to explore the capabilities of current language models to be few-shot learners, that is, to perform NLP tasks without the need for large training datasets (Brown et al., 2020). One alternative has been to use the Recognizing Textual Entailment (RTE) framework to boost Transformers' contextual and semantic representations (e.g., Conneau et al., 2017; Laurer et al., 2023). In this approach, the task is to recognize, given two text fragments, whether the meaning of one text can be inferred (entailed) from another text (Dagan et al., 2006). Initially used as a generic task for evaluating and comparing applied seman-

tic inference models (for a historical overview, see Poliak, 2020), in recent years RTE-based models have been used in text classification, showing the potential to function as few-shot learners (Yin et al., 2019; Wang et al., 2021).

Our study aligns with this framework to address a real-world challenge: to develop an effective automated model to classify vaccine-related news headlines from a dataset with 1,000 labeled data points, unequally distributed in 10 classes. We detail eight different models and procedures—including open-source and closed-source, paid and free options—which were tested and evaluated from four perspectives:

1. **Fine-tuned Transformers-based models:** we fine-tuned pre-trained models derived from BERT, using only the class labels;
2. **Transformers-prompt-based models:** instead of using the class labels, we employed their descriptions to prompt a GPT model with instructions for text classification;
3. **Off-the-shelf RTE-based models:** also using the label descriptions, we evaluated ready-to-use models trained on RTE task;
4. **Fine-tuned Transformers-RTE-based models:** we fine-tuned pre-trained models derived from BERT for the RTE task.

The models were compared from several perspectives: as zero or few-shot learners, we analyzed both open and closed-source options; as fine-tuned models, we evaluated options with and without RTE; and as zero or few-shot learners, we contrasted fine-tuned models with closed-source models.

Our experiments demonstrate that deep transfer learning (Pan & Yang, 2010; Ruder, 2019)—wherein a "language knowledge" like Transformers embeddings is combined with "task knowledge" such as RTE—enabled the development of a text classification model exhibiting superior performance compared to the alternatives, precisely tackling issues of data scarcity, minority classes, and unbalanced data, as well as accessibility in terms of open-source availability and costs. In this sense, our study validates and provides protocols that have been developed for the adoption of this approach in creating new and effective text classification models (Yin et al., 2019; Wang et al., 2021; Laurer et al., 2023). Additionally, by applying this method to our corpus, we produced a cutting-edge automated model for classifying vaccine-related content with comprehensive classes based on the science communication literature, also allowing for the incorporation of new categories by other researchers.

Unsupervised vs. Supervised Learning: An Overview

The classic definition of unsupervised learning refers to the ability of a model to learn to represent particular patterns in a way that reflects the statistical structure of the overall collection of input data (Ghahramani, 2003). In practical terms, this means that the model learns these patterns from unlabeled data. This differentiates it from the supervised paradigm, where the labeled data is used to train the model and thus develop a predictive model capable of classifying new instances (Yom-Tov, 2004). One of the main applications of these models in the social sciences, and more specifically in communication studies, is text analysis through topic modeling. Topic modeling is a class of NLP models which leverage statistical patterns in language to extract underlying topics from large corpora (for a review of topic modeling methods, see Vayansky & Kumar, 2020).

The vectorization of textual data through the popularization of word embeddings has led to more sophisticated text classification and topic modeling techniques. Algorithms like Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022) have incorporated pre-trained embedding models and have overcome the limitations of models primarily based on word co-occurrence (for a comparison between LDA, NMF, Top2Vec, and BERTopic, see Egger & Yu, 2022). One of the major advantages of these techniques is their plug-and-play nature and the relatively easy interpretability of the results.

Although extensively employed for the analysis of diverse textual data (e.g., Maier et al., 2018; Obadimu et al., 2019; Hendry et al., 2021; Adewunmi et al., 2021), unsupervised methods, despite not relying on labeled data, encounter problems such as the difficulty of tailoring them to specific tasks and validating the results (Denny & Spirling, 2018). Taking our corpus of news headlines related to vaccines as an example, topic modeling with BERTopic would categorize the headlines "COVID-19 vaccine shows good results in clinical trials," "COVID-19 vaccines face distribution issues," and "Anti-vaccine groups target COVID-19" under the same topic (e.g., *COVID-19 vaccine*). However, communication researchers interested in a more nuanced analysis could classify these headlines into distinct aspects, such as *vaccine testing*, *logistical issues*, and *anti-vaccine groups*. Cutting-edge LLMs, such as GPT and its conversational interface, ChatGPT, have shown promising abilities (Wei et al., 2022) mainly because they operate via elementary prompts (e.g., "Please extract the main topics from this set of texts"). However, the model's lack of transparency and the challenge in reducing the randomness of the outputs (Borji, 2023) lead to issues of reproducibility,

validation, and scalability.

In this context, supervised learning methods prove to be much more suitable for theory-based hypothesis analyses, involving the operationalization of pre-established categories of interest (Edwards et al., 2020). This brings it closer to the standard research paradigm in social science (Markus et al., 2023). In supervised computational content analysis, researchers begin by defining and operationalizing the categories of interest. They then provide the learning algorithm with samples that exemplify these concepts, aiming to create models that are specifically trained to identify these classes within the unlabeled target corpus. Markus and colleagues (2023) enumerate other advantages of this approach, such as the flexibility in determining the unit of analysis (e.g., word, sentence, paragraph, book), the nature of the classification scheme, the interrelation between categories, as well the model's architecture. Furthermore, in supervised learning, researchers can choose to focus on semantics, style, and other structural features of the text, as exemplified by the vaccine news headlines above.

Within this framework, the training dataset plays a fundamental role in the performance of a supervised classification model (Kavzoğlu, 2009). It needs to be defined in a way that is typical and representative of each individual class. Therefore, quality and size are key elements. Generally, a small sample does not provide the model with the capacity to recognize all classes and determine their boundaries precisely (Kavzoğlu, 2009). However, acquiring a robust training dataset may require a significant investment of both time and financial resources, which is not the reality for many researchers (Wilkerson & Casas, 2017). Furthermore, as pointed out by Laurer and colleagues (2023), the lack of data is particularly problematic in the social sciences because many research questions start from different tasks, and sometimes the researcher's concepts of interest are present in a small portion of the corpus.

There have been many attempts to mitigate the problem of scarce training data and unbalanced classes (Krawczyk, 2016). Alternatives include procedures and strategies to be adopted at different times in the process—during or after training dataset annotation—, and in different aspects—at a data level or an algorithmic level (e.g., Kaur et al., 2019; Tyagi & Mittal, 2020; Markus et al., 2023). Our study focuses on the algorithmic level when testing and fine-tuning supervised text classification models in a real-world scenario, but from the perspective of zero or few-shot learning—with little or no labeled data. We started from theory-driven categories defined a priori, but with a considerably reduced and unbalanced annotated dataset. To tackle

T **= The cat is sleeping on the chair**
H_1 = The feline is resting
H_2 = The animal is awake
H_3 = The chair is made of wood

Table 1: Example of an RTE task

this challenge, we sought to combine the advantages in terms of contextual and semantic representation offered by Transformers with the potentialities of the Recognizing Textual Entailment (RTE) approach, which we will detail below.

Recognizing Textual Entailment

Recognizing Textual Entailment (RTE)[1] is one of the basics of Natural Language Understanding (NLU), which in turn is a subclass of Natural Language Processing (NLP) (Shajalal et al. 2023). Textual entailment is defined as a relationship between a coherent text T and a language expression, which is considered as a hypothesis H. We say that T entails H (H is a consequence of T, denoted by $T \rightarrow H$) if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T (Dagan & Glickman, 2004).

Consider the example in Table 1. In that case, the relation $T \rightarrow H$ can be considered true for the text T and the hypothesis H_1, false for the text T and the hypothesis H_2, and neutral for the text T and the hypothesis H_3. Within certain contexts, this relationship can be classified in terms of *entailment*, *contradiction*, and *neutrality* (Williams et al., 2018). In the previous example, the relationship between the text and hypotheses would be described as entailment for the text T and the hypothesis H_1, contradiction for the text T and the hypothesis H_2 (since the meaning of the hypothesis is clearly opposed to that of the text), and neutral for the text T and hypothesis H_3 (as there are no sufficient elements to infer the meaning of the hypothesis from the meaning of the text).

RTE was developed to address one of the major challenges of NLP, which is dealing with the semantic variability of human language—the same meaning can be expressed by, or inferred from, different words and phrases (Dagan et al., 2006). Going back to the example above, a machine learning

[1]Recognizing Textual Entailment is also known as Natural Language Inference (NLI), although Poliak (2020) notes that the correspondence between terms is not a consensus among scholars. We adopted the first one because it is the form used in most of the literature used for this study.

T **= Cheaper and more effective vaccine against pneumonia is tested in humans**
H_1 = This headline is about health
H_2 = This headline is about sports
H_3 = This headline has a positive sentiment
H_4 = This headline has a negative sentiment
H_5 = This headline addresses a scientific novelty
H_6 = This headline addresses the climate crisis

Table 2: RTE applied to text classification, with example hypotheses for a news headline

model would have to be able to approximate (or oppose) the meanings of the words related to the subject of the sentence (*cat*, *feline* and *animal*) and to the action (*sleeping* and *resting*), in addition to the context in which they are inserted. In this sense, this approach began to be used as a generic task to evaluate and compare applied semantic inference models, such as Question Answering, Information Extraction, Summarization, and Machine Translation (Dagan et al., 2006; Tătar et al., 2009; Poliak, 2020). In summary, RTE allows measuring the degree to which the statistical processing of NLP models is effective in capturing the meanings of the text.

More recently, RTE has been used for zero and few-shot text classification (Yin et al., 2019; Wang et al., 2021). In this approach, the text T is the textual unit that one wants to classify (e.g., a social media post, a news headline, a news story), and statements are made to it that will function as hypotheses H_n. Potentially, these statements can address different aspects of the text, such as topics, sentiments, or situations (Yin et al., 2019). In Table 2, the text is a news headline, and the hypotheses are statements formulated by the researcher.

From an RTE perspective, entailment holds true only between the text T and the hypotheses H_1, H_3, and H_5. When integrated into a text classification model, these hypotheses could be operationalized as topic (*health*), sentiment (*positive*), and stance (*scientific novelty*). Expanding on this concept, Wang and colleagues (2021) take a step further by suggesting the conversion of class labels into comprehensible natural language sentences that provide sufficient explanation. This approach allows for a more detailed representation of a class, encompassing various facets of a given text. For instance, a class labeled as "Vaccination Rollout" would be transformed into the hypothesis: "The headline addresses vaccination rollout, campaigns, priority groups, vaccine deals, and/or the monitoring of vaccination rates." Note that the description covers several specific aspects within the broad

concept of "Vaccination Rollout." Consequently, whenever the model identifies entailment between a headline and some element of this descriptive sentence, the headline would be categorized under that class. This possibility of verbalizing a label is one of the advantages of this method (Laurer et al., 2023).

Evidently, the ability to correctly capture the relationship between a text and a hypothesis is due to the knowledge background, which refers to the internal representation that a model has of human language (Dang et al., 2006). This representation has been improved with the advancement of language models trained on massive textual corpora, such as BERT and its derivatives, which has bolstered the ability of RTE-based models as zero or few-shot learners (Wang et al., 2021). In addition, there are currently over a million unique hypothesis-text pairs, as well as off-the-shelf datasets and models to be fine-tuned available in various repositories (Williams et al., 2018; Nie et al., 2019, 2020; Parrish et al., 2021; Liu et al., 2022a). Our study explores different models and procedures in these two perspectives to classify vaccine-related news headlines into 10 distinct classes, despite having a limited and imbalanced training dataset.

Material and Methods

The challenge of developing a model for classifying texts related to vaccines originated from a real-world scenario: we needed to categorize a corpus of 13,387 headlines on the subject, published by elite newspapers in the United States, United Kingdom, China and Brazil, between 1st January, 2020, and December 31, 2021. These headlines were collected using Python's web scraping tools (BeautifulSoup, Newspaper, and Newsplease), directly from the websites of four newspapers—*The New York Times*, *The Guardian*, *China Daily*, and *Folha de S.Paulo*[2]. The search criteria included the presence of at least one of the specified keywords in the title *(vaccine*, *vaccines*, *vaccination*, *vaccinations*, *vaccinating*, *vaccinated*). All headlines meeting this criterion were included in the corpus. The classification aimed to support a longitudinal and comparative analysis of media coverage on vaccines in the context of the COVID-19 pandemic in countries that, at different times, were severely affected by the health crisis.

[2]As all pre-trained models used in this study were primarily based on English texts, we chose to translate the headlines from the Brazilian newspaper (the selected Chinese newspaper is published in English). While translation for content analysis purposes is a common practice in some social science studies (de Melo & Figueiredo, 2021), this limitation will be discussed further.

Several studies in different countries have demonstrated that the pandemic has broadened the topics covered by the media in their reporting on science and health, including social, political, and economic issues (Liu et al., 2020; Hart et al., 2020; Crabu et al., 2021). Regarding vaccines, evidence points to a shift from the traditional episodic coverage, primarily focused on vaccination campaign announcements, towards a more comprehensive coverage that explores additional dimensions and actors, including the rigorous process of safety and efficacy trials, institutional and geopolitical disputes, logistical challenges in vaccine distribution, and the economic impacts of vaccination (Christensen et al., 2022; Neves & Massarani, 2022). As previously mentioned, such nuances elude conventional unsupervised topic modeling, and in our case, this method functioned more as an initial exploration of the corpus. Consequently, these limitations led us to seek supervised methods capable of capturing the subtleties of the textual data.

To define the classes, a random sample was initiated to be labeled by a science communication researcher with a background in science journalism. In addition to the researcher's expertise, this process drew upon a series of studies regarding media coverage of the COVID-19 pandemic and the vaccine (in addition to the previously mentioned studies, see also Massarani & Neves, 2021; Ju et al., 2022; Schwarz et al., 2023). These studies point to a diversity of topics that were also found in our corpus, such as scientific aspects, vaccine geopolitics, disinformation, public policies, and economic consequences. The researcher had the autonomy to adapt and create labels specific to the corpus, providing a description for each label. Following team discussions, 10 classes and their respective descriptions were defined. It was decided that the classes would be mutually exclusive. In case of ambiguity, the annotator should consider the primary focus of the text. The researcher then proceeded to annotate the headlines, reaching 1,000 data points. Table 3 presents an overview of each label, its description, and illustrative headlines.

ID	Topic Label	Description	Examples
0	**Global Access to Vaccine**	The headline addresses initiatives for equal access to covid-19 vaccines, such as COVAX, and/or the need to combat inequity, ensure global distribution, and contain the disease, especially to low- and middle-income countries	UN chief calls for intl solidarity to find vaccine accessible to all (CHI, 16/09/2020) G20 leaders pledge to distribute Covid vaccines fairly around world (GUA, 22/11/2020)
1	**Science and Technology**	The headline addresses the science behind vaccine development, including research, studies, safety and efficacy trials, emergency approval, and side effects	US starts first human testing for coronavirus vaccine (CHI, 19/03/2020) Covid-19 vaccine: what have we learned from Oxford phase one trial? (GUA, 16/07/2020)
2	**Public Health Policies**	The headline addresses vaccine mandates, vaccine passports, and/or other public health policies such as social distancing, quarantine, lockdowns, and mask mandates	New York City imposes vaccine mandate for many high school athletes and coaches (NYT, 20/08/2021) Israel will allow entry of vaccinated tourists starting in May (FOL, 14/04/2021)
3	**Vaccination Roll-out/Campaign**	The headline addresses vaccination rollout, campaigns, priority groups, vaccine deals, and/or the monitoring of vaccination rates	Britain begins coronavirus vaccine rollout (NYT, 08/12/2020) Drive-thru for the vaccination of the elderly starts this Monday in São Paulo (FOL, 07/02/2021)

Continued

ID	Topic Label	Description	Examples
4	**Vaccine Hesitancy and Mis-/ Disinformation**	The headline addresses the circulation of disinformation and misinformation regarding vaccines, which contribute to vaccine hesitancy and reluctance	Anti-vax group mounts legal blitz to sow disinformation against vaccinations (GUA, 22/06/2021) Twitter will begin removing vaccine misinformation (NYT, 16/12/2020)
5	**Public Endorsement to Vaccine**	The headline addresses public endorsement and incentive for vaccines, such as celebrities and artists getting vaccinated	Jane Fonda is vaccinated against Covid-19 in the United States: 'It doesn't hurt' (FOL, 01/02/2021) Elton John and Michael Caine star in video to promote Covid vaccination (GUA, 10/02/2021)
6	**Institutional Affairs**	The headline addresses institutional and governmental issues, political disputes, conflicts, political authorities, and export bans	Bolsonaro denies interference in Anvisa and says he is in a hurry for a vaccine (FOL, 27/12/2020) Biden criticises Trump over slow Covid-19 vaccine rollout (NYT, 29/12/2020)
7	**Problems in Vaccination**	The headline addresses problems in vaccination, such as delays, fraud, disparities, vaccine shortages	Hours of scrolling, endless refreshing: US tech woes make scheduling vaccine a nightmare (GUA, 28/01/2021) European countries turning to East for vaccines amid supply shortage (CHI, 28/02/2021)

Continued

ID	Topic Label	Description	Examples
8	**Public Perception of Vaccine**	The headline addresses opinion polls and surveys on the public perception of vaccines and/or willingness to get vaccinated	Intention to get vaccinated against Covid-19 grows in Brazil, according to Datafolha (FOL, 20/03/2021) Two-thirds of Australians 'definitely' want Covid vaccine, while 27% are unsure (GUA, 16/02/2021)
9	**Economic Consequences**	The headline addresses the impacts and benefits of vaccination on the economy	COVID-19 vaccination key to economic recovery (CHI, 25/04/2021) O.E.C.D. Raises Global Growth Forecast Sharply, Citing Vaccines (NYT, 31/05/2021)

Table 3: Topic labels, description, and illustrative headlines

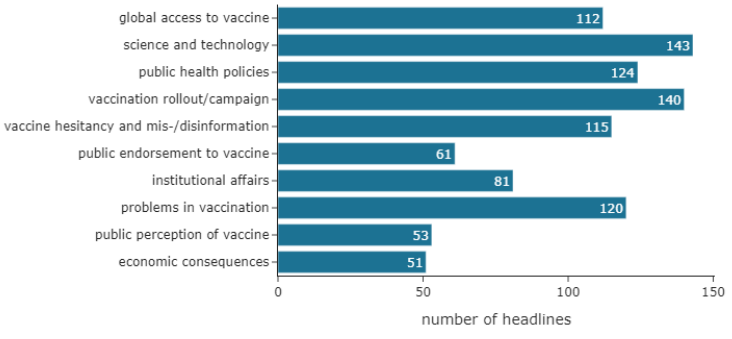


Figure 1: Distribution of classes in the annotated dataset (N=1,000)

To validate the data quality, a second researcher, using the classes and their descriptions, annotated the same 1,000 headlines. The intercoder reliability was 0.825 (Cohen's Kappa and Krippendorff's Alpha), which is considered a strong agreement rate (McHugh, 2012).

In Figure 1, it can be observed that the classes are unevenly distributed within the annotated dataset, a situation not uncommon in social science research. Six of the 10 classes gather more than 110 headlines, comprising over 75% of the dataset (the class with the most headlines is *Science and Technology*). The remaining headlines are distributed across the remaining four classes, with *Economic Consequences* having the lowest amount of text—only 51 headlines, equivalent to 5% of the dataset. From the 1,000 headlines, a stratified sample of 20% was extracted to serve as the test set, while the remaining headlines were used to create stratified training sets with 200, 400, 600, and 800 data points.

We opted to address the challenges posed by scarce data and imbalanced classes focusing on testing and fine-tuning supervised text classification models grounded on deep contextual and semantic representation of language. These models show potential in handling zero or few-shot learning scenarios. We tested eight models based on the four approaches explained below (see Appendix A for the technical details of each procedure)[3].

1. **Fine-tuned Transformers-based models:** In this approach, our objective was to assess the ability of pre-trained language models on

[3]The codes are available in our GitHub repository: https://github.com/lffernandes08/ccr_codes

extensive textual data to capture patterns solely based on class labels, rather than their descriptions. We utilized two derivatives of the BERT architecture: DeBERTaV3base[4] and DeBERTaV3large[5] (He et al., 2021a, 2021b). Both models improve upon the original one through disentangled attention and an enhanced mask decoder. The primary distinction between the two lies in their total number of parameters (435M for large and 184M for base) and the number of parameters in the embedding layer (131M vs. 98M). It is expected that the large version outperforms in tasks that require understanding intricate relationships, long-range dependencies, and fine-grained distinctions in the input data. Conversely, the base model demands fewer computational resources for training and offers faster inference. Our training datasets were used for fine-tuning both models. For simplicity, we shall refer to these models as *Transformers-base* and *Transformers-large.*

2. **Prompt-based models:** In the second approach, we evaluated the zero or few-shot learning capacity of one of the most popular Large Language Models. Utilizing the OpenAI API, we employed the gpt-3.5-turbo model to classify the test set of 200 headlines. The API leveraged the model's conversational nature using a prompt with the designated instruction. Our prompt was formulated to encompass the label and description of each class, as depicted in Table 3. We instructed the model to assign each headline to a specific class (the full prompt is available in Appendix A). In a subsequent experiment, we supplemented two headlines as examples for each label and description (resulting in a total of 20 illustrative headlines). These models will be referred to as *GPT-zero* and *GPT-few*.

3. **Off-the-shelf RTE-based models:** Here we began the testing of the actual textual entailment framework. As with the previous approach, we used classes descriptions. Initially, our aim was to evaluate ready-to-use models tailored for the RTE task, with no specific training in vaccine-related content. We utilized two of the most downloaded models for zero-shot text classification available on the Hugging Face Hub, a collaborative repository for developing and sharing NLP models, datasets, and training pipelines. The first model is a version of the pre-trained BART model[6] (Lewis et al., 2020), fine-tuned on the

[4]Available at: https://huggingface.co/microsoft/deberta-v3-base
[5]Available at: https://huggingface.co/microsoft/deberta-v3-large
[6]Available at: https://huggingface.co/facebook/bart-large-mnli

Multi-Genre Natural Language Inference (MultiNLI) corpus, which consists of a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information (Williams et al., 2018). The second model employs a version of DeBERTaV3large[7], fine-tuned on five datasets of textual entailment (MultiNLI, Fever-NLI, ANLI, LingNLI, and WANLI), totaling more than 885k text-hypothesis pairs (Laurer et al., 2023). In presenting the results, we will refer to these models as *BART* and *RTE-general*. Both models implement the same procedure: each headline functions as the text T, and each of the 10 label descriptions is considered as a hypothesis H_n (as shown in Table 3). The models then compute the percentage of entailment between the headline and each description, selecting the one with the highest value. Subsequently, the headline is assigned the label corresponding to the description that achieved the highest entailment score. The classification was performed on the test set of 200 headlines.

4. **Fine-tuned Transformers-RTE-based models:** Finally, we leveraged the deep transfer learning approach (Pan & Yang, 2010) to combine the language knowledge of Transformers with the task knowledge of RTE, followed by a fine-tuning step using our annotated data. We followed the protocol and models proposed by Laurer and colleagues (2023)—one model based on DeBERTaV3large[8] and the other on DeBERTaV3base[9]. The former utilizes five training datasets for textual entailment, comprising over 885k text-hypothesis pairs. The latter model is trained on more than 1.2M text-hypothesis pairs derived from eight text entailment datasets (MultiNLI, Fever-NLI, and DocNLI, which includes ANLI, QNLI, DUC, CNN/DailyMail, and Curation). Our training datasets (200, 400, 600 and 800 data points) were applied for fine-tuning both models. These models are herein referred to as *Transformers-RTE-base* and *Transformers-RTE-large*.

To select evaluation scores and ensure comparability with other studies, we also followed the methodological approach of Laurer and colleagues (2023). The authors posit that Accuracy (counting of the overall fraction of correct predictions, equivalent to F1 Micro) might overestimate the performance of classifiers that overpredict majority classes while neglecting

[7]Available at: https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

[8]Available at: https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

[9]Available at: https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c

Model	Accuracy/F1 Micro	Balanced Accuracy/Recall Macro	F1 Macro
GPT-zero	0.815	0.825	0.823
GPT-few	0.855	0.864	0.858
BART	0.575	0.598	0.557
RTE-general	0.305	0.346	0.336

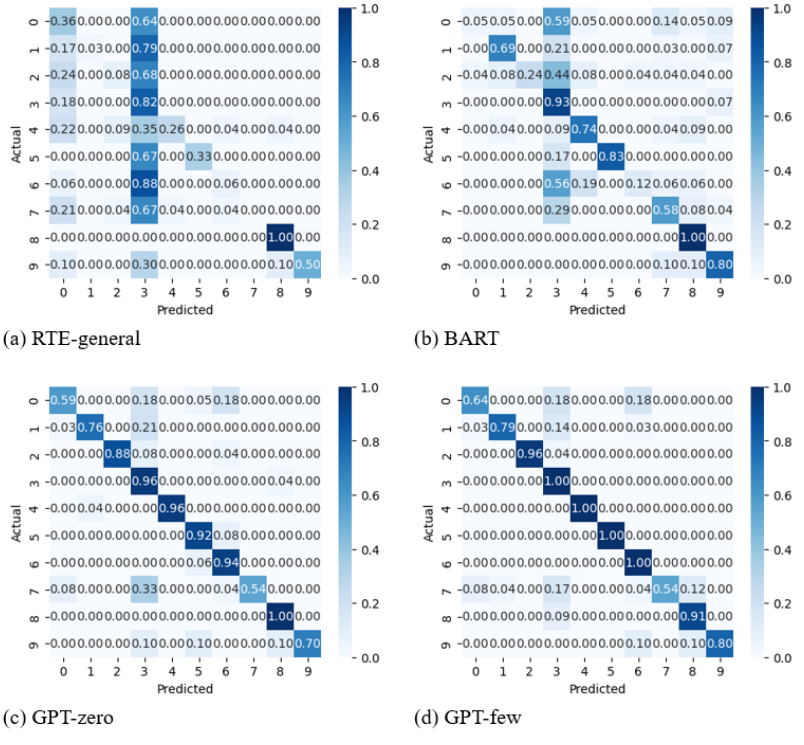
Table 4: Performance metrics of Transformers-prompt-based models and off-the-shelf RTE-based models on text classification of vaccine-related headlines

minority classes. An alternative is Balanced Accuracy (the average of Accuracy for each class calculated separately, equivalent to Recall Macro). However, Balanced Accuracy could favor models that predict numerous minority classes well but perform less effectively on a few majority classes. On the other hand, F1 Macro (the harmonic mean of Precision and Recall) assigns equal weight to all classes, regardless of their size, and is deemed suitable for research in social science (Laurer et al., 2023). All results will be presented considering these three metrics—Accuracy/F1 Micro, Balanced Accuracy/Recall Macro, and F1 Macro.

Results

The initial comparison focuses on the performance of text classification models that have not undergone fine-tuning, wherein no (or very little) labeled data was employed for their training (closed-source prompt-based models and off-the-shelf open-source RTE-based models). For each of these models, descriptions for every class were provided. As depicted in Table4, proprietary Large Language Models capable of processing conversational inputs (*GPT-zero* and *GPT-few*) exhibited notably superior performance (above 80% across all three metrics) when compared to open-source models trained on textual entailment datasets (*BART* and *RTE-general*). The experiment also demonstrated that supplying two example headlines per class led to a modest improvement between the two GPT models.

On the other hand, text entailment-based models that have not undergone fine-tuning have shown limited performance within our highly specific corpus focused on vaccines. Given the 10-class classification scenario, *BART*'s scores of around 55% is somewhat acceptable. However, notable discrepancies emerged across classes (3 and 8 performed well, while 0, 2, and 6 performed poorly) (Figure 2b). Conversely, the confusion matrix for the *RTE-general* model (Figure 2a) reveals that a significant portion of the head-



(a) RTE-general (b) BART

(c) GPT-zero (d) GPT-few

Figure 2: Confusion matrix of Transformers-prompt-based models and off-the-shelf RTE-based models on text classification of vaccine-related headlines. *Note.* The classes indices correspond to those shown in Table 3.

lines were classified into class 3 (*Vaccination Rollout/Campaign*), notably the most general category, reflecting the lack of specificity in the original datasets used for the model's training. Remarkably, the model performed well in class 8 (*Public Perception of Vaccine*), pertaining to headlines concerning vaccine-related public opinion surveys. We posit that this outcome arises from the distinct and consistent structure of headlines on this class, setting them apart from the rest of the corpus (e.g., "Just half of Americans plan on getting the COVID-19 vaccine, poll shows").

Oppositely, the performance of open-source models fine-tuned with portions of the labeled data was notably superior. In both Transformers models, where training was conducted exclusively on class labels (rather than their description), results were definitely linked to the quantity of training data (Figure 3). With only 200 training data points, *Transformers-*

base and *Transformers-large* exhibited very low scores, with some classes remaining unidentified by the models. The performance of the two models improves significantly with 400 and 600 training data points, with the large version presenting higher accuracy. With 800 training data points, the two models performed similarly across all evaluation scores—above 87%. This shows that even smaller pre-trained models, if exposed to an ample amount of data, can be competitive with their larger counterparts, but with the advantage of being faster and using less computational resources.

Models that are built upon the Transformers architecture and have received specific training for the text entailment task, coupled with a fine-tuning step using our annotated data (using class descriptions as hypotheses), exhibited an interesting ability: they performed notably well even with a limited amount of training data. Notably, the scores with only 200 training data points exceeded 74% for *Transformers-RTE-base* and surpassed 84% for *Transformers-RTE-large* (Figure 3). This result was achieved even on classes with very few labeled headlines (e.g., *Economic Consequences*, 10 data points; *Public Perception of Vaccine*, 11 data points; *Public Endorsement of Vaccine*, 12 data points). Consequently, both models, while certainly benefiting from additional data points, are less affected by the quantity of training data and handle imbalanced classes more adeptly. Interestingly, these models exhibited a slight performance decline when trained with 600 data points, suggesting an increase in ambiguity of some classes with this training set configuration.

Similarly, the pre-trained models' size—in both its large and base variants—also affected classification performance, though less decisively than models that did not undergo the fine-tuning RTE phase. The best performance was reached by the fine-tuned version from the RTE perspective, using the large pre-trained model: with 800 labeled headlines employed for training, *Transformers-RTE-large* achieved scores of 92%.

Figure3 also presents a comparative analysis between closed-source GPT models (on the left) and fine-tuned open-source models (on the right), highlighting the capability of the former to achieve substantial classification accuracy even with minimal training data (zero or two headlines per class, totalling 20 headlines). *GPT-zero* and *GPT-few* models outperformed some of the Transformers and Transformers-RTE models, particularly those fine-tuned with little training data and built upon base versions of pre-trained language models. However, the nature of the two approaches and how they perform the classification task raise important concerns, which will be discussed next. Table5 provides a summary of performance results for all

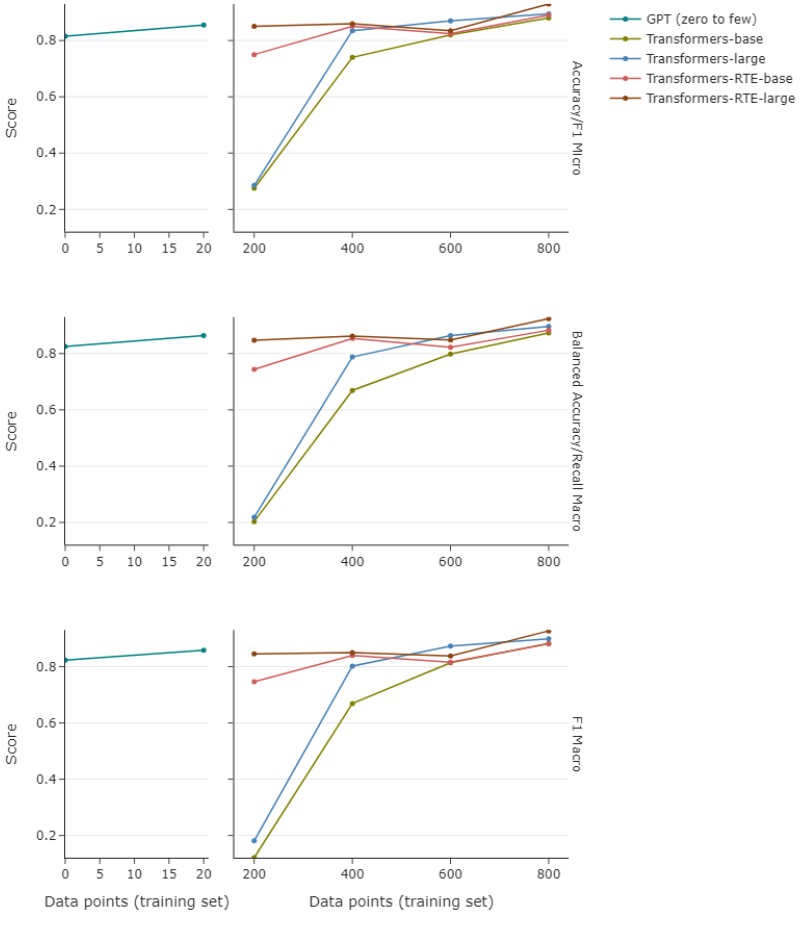


Figure 3: Performance of GPT models, and fine-tuned Transformers-based and fine-tuned Transformers-RTE-based models with different training data. *Note.* Due to the difference in intervals on the x-axis, we present the results of the GPT models (on the left) and the fine-tuned Transformers models (on the right) in separate charts to facilitate comparative visualisation.

Model	Training set	Accuracy/ F1 Micro	Balanced Accuracy/ Recall Macro	F1 Macro
BART	0	0.575	0.598	0.557
RTE-general	0	0.305	0.346	0.336
GPT-zero	0	0.815	0.825	0.823
GPT-few	20	0.855	0.864	0.858
Transformers-base	200	0.275	0.203	0.121
Transformers-base	400	0.740	0.669	0.669
Transformers-base	600	0.820	0.798	0.814
Transformers-base	800	0.880	0.873	0.882
Transformers-large	200	0.285	0.218	0.181
Transformers-large	400	0.835	0.788	0.802
Transformers-large	600	0.870	0.864	0.873
Transformers-large	800	0.895	0.896	0.899
Transformers-RTE-base	200	0.750	0.744	0.746
Transformers-RTE-base	400	0.850	0.854	0.840
Transformers-RTE-base	600	0.825	0.822	0.816
Transformers-RTE-base	800	0.890	0.883	0.881
Transformers-RTE-large	200	0.850	0.847	0.845
Transformers-RTE-large	400	0.860	0.862	0.850
Transformers-RTE-large	600	0.835	0.848	0.838
Transformers-RTE-large	800	0.930	0.924	0.927

Table 5: Performance scores of closed-source models and fine-tuned open-source models

tested models across different training dataset sizes.

Discussion

Much has been discussed about the abilities of modern language models to perform intricate NLP tasks using unsupervised methods (Wei et al., 2022; Liu et al., 2023). Despite notable advancements, our comparative study demonstrates that some degree of fine-tuning remains essential for achieving superior results in text classification tasks, particularly within domain-specific corpus. In the framework of text entailment as a zero or few-shot learner, even off-the-shelf models (such as *BART* and *RTE-general*) trained on thousands or millions of text-hypothesis pairs proved inadequate in capturing the nuances of vaccine-related content. We argue that the

diversity of classes—some highly generic, such as *Problems in Vaccination*, and others very narrowed, like *Economic Consequences*—might have further complicated the classification task. While the latter pertains exclusively to economic issues, the former includes a range of subjects, including delays, fraud, disparities, and shortages.

In this regard, GPT-based models performed notably better (above 80%). When examined individually, several classes achieved over 90% correct headline classification, indicating heightened disambiguation capability. Nevertheless, discerning certain categories posed challenges. An example is the *Problems in Vaccination* class, where some of its headlines were classified as *Vaccination Rollout/Campaign*. Although the description clearly outlines the issues addressed by the class (delays, fraud, disparities, and shortages), the model was unable to fully disambiguate all the headlines.

Evidently, this type of model benefits from the ability to generate prompts in natural language and to request tasks without any training text, given the extensive pre-trained data. However, their drawbacks are accentuated by several factors. Firstly, scaling a research project using this approach can be technically and cost-wise unfeasible. Classifying only 200 headlines using *GPT-zero* and *GPT-few* took about 20 minutes, with some attempts being interrupted for unknown reasons (classification with the other models took less than five minutes, and even the fine-tuning process took under 15 minutes). This delay is not solely due to processing time, but also because of the model's API rate limit, which restricts the number of calls within a given time period. In terms of costs, using the provided prompt and classifying the 200 headlines (approximately 270 tokens per call) incurred a charge of roughly USD 0.07. While this may seem low, it is important to consider that classifying a corpus involves multiple tests (thus, running the model numerous times) and that scientific work takes place within diverse economic contexts.

Another issue was the model generating variations and hallucinations in class labels. For example, the class *Economic Consequences* was sometimes classified as Economy, which is not necessarily an error but indicates a deviation from the standardization that does not occur in the other models. However, more problematically, the model occasionally created the class Vaccine Tourism, which was not specified in the prompt. These issues require a more rigorous post-classification validation. Furthermore, the inability to eliminate randomness prevents exact result replication, making these closed-source tools unreliable for application in academic research.

Many of the problems of the previous approaches were mitigated by

fine-tuning with labeled data. Additionally, the textual entailment approach clearly improved performance, proving to be effective even with limited training data. Thus, our findings empirically corroborate the advantages of deep transfer learning: initially, by leveraging Transformers-based models, pre-trained across diverse domains, to facilitate the acquisition of broad, generalized statistical knowledge of language patterns; subsequently, fine-tuning on annotated data, refining the learning process for a very specific task (Laurer et al., 2023). In this context, the accumulated knowledge of Transformers is used not only to identify patterns of similarity within texts of the same class, but also to associate (entail) a text more to a hypothesis (a class expressed in natural language) than the other. Effectively, it is giving Transformers a boost. However, it is crucial to acknowledge that the size of the general pre-trained model (base or large) plays a key role for the success of the text entailment strategy in achieving optimal classifier performance, particularly with smaller training datasets. As a result, the *Transformers-RTE-large* model outperformed all other alternatives, with the advantage of being open-source, scalable, and reproducible.

The inclusion of explicit class verbalization in RTE offers an additional advantage. In the process of fine-tuning, researchers can modify label descriptions to make them more accurate (Wang et al., 2021). Such refinements are not feasible when relying solely on class labels for fine-tuning. Moreover, the notable performance of this approach even on classes with very few training texts can be attributed to a data augmentation effect of the text-hypothesis strategy (Wang et al., 2021). In the protocol outlined by Laurer and colleagues (2023), which we followed in this study, each true text-hypothesis pair is accompanied by a randomly false pair during training. This pre-processing method allows the model to learn to differentiate between true and false stances, in addition to doubling the size of the training dataset.

As evidence of the methodology's validation, the implementation of the top-performing model (*Transformers-RTE-large* trained with 800 data points) to categorize the entire corpus of 13,387 vaccine-related news headlines yielded results consistent with recent science communication literature, particularly studies on media coverage of vaccines (Liu et al., 2020; Hart et al., 2020; Crabu et al., 2021; Massarani & Neves, 2021; Ju et al., 2022; Schwarz et al., 2023). Notably, it highlighted the prominence of specific topics in media coverage before and after the initial rollout of COVID-19 vaccination: *Science and Technology* in 2020 and *Vaccination Rollout/Campaign* in 2021. Furthermore, the classification indicated an increase in headlines

concerning public health policies, reflecting the public debate post-COVID-19 vaccination rollout on vaccine mandates and passports, as well as the lifting of measures such as lockdown, quarantine, social distancing, and mask mandates. Additionally, the prevalence of the *Institutional Affairs* topic in U.S. and Brazilian newspapers coverage suggests a significant media focus on the political conflicts involving the then presidents of the two countries during that period—Donald Trump and Jair Bolsonaro.

Limitations

Although our study demonstrates that the text entailment approach has proven to be successful in classifying text, even with limited training data and unbalanced classes, the reduced size of the annotated dataset consequently limits validation, as the test dataset is also small. In this sense, we encourage researchers to assess our final classification model (*Transformers-RTE-large*)[10]. This model can be readily implemented with minimal code in vaccine-related content, either by using our predefined classes or by exploiting its flexibility to capture the meaning of new classes.

Another limitation not only in this study but also in the application of pre-trained language models in a general context is their predominant training in English. Significant efforts have produced versions in other languages (e.g., CamemBERT for French, BETO for Spanish, BERTimbau for Portuguese) as well as multilingual versions (e.g., mDeBERTa). In our case, as the Portuguese headlines constituted a minor portion of the corpus, we opted to translate them into English. Considering that the 10 classes pertain more to thematic aspects rather than semantic ones, we believe this did not result in significant meaning loss. It is important to remember that any pre-trained model carries biases from the data on which they were trained. Therefore, we recommend caution when conducting any form of automated content analysis, highlighting the researcher's role in identifying reproductions of stereotypes and biases (for a discussion on bias and LLMs, see Lucy & Bamman, 2021; Kaneko et al., 2022; Abid et al., 2021).

Conclusions

In this study, we systematically tested machine learning models to classify vaccine-related news headlines. Our investigation originated from a real-world research challenge, utilizing a curated dataset of 1,000 texts distributed across 10 imbalanced classes. Our aim was to provide social science

[10]Available at: https://huggingface.co/LuizNeves/BERT-NLI-vaccine-v2

researchers, particularly in communication, with alternative approaches to tackle similar issues. We evaluated eight distinct models from four perspectives, including both open-source and closed-source, as well as paid and free options: fine-tuned Transformers-based models, Transformers-prompt-based models, off-the-shelf RTE-based models, and fine-tuned Transformers-RTE-based models. Consequently, we assessed the performance of various methods in the context of zero or few-shot learning, with a particular emphasis on harnessing the potentialities of the text entailment approach.

Our findings suggest that applying the RTE task to text classification can enhance model performance, particularly when fine-tuned using a large pre-trained language model such as DeBERTaV3large. Deep transfer learning also emerges as a potent strategy for tackling limited data and imbalanced classes, as evidenced by the models achieving high Accuracy and F1 scores even with scarce training data. In our comparative analysis, the conventional approach of fine-tuning Transformers-based models (excluding label descriptions) remains effective, albeit being more responsive to the inclusion of additional training data. Additional advantages include the open-source nature of these models, which can make them cost-effective and flexible, particularly for long-term, large-scale projects with established technical expertise. Ultimately, off-the-shelf models reveal their inherent generality and limited scope. While the proprietary GPT-based model showed satisfactory performance, it is important to acknowledge that issues related to accessibility, scalability, and reproducibility somewhat undermine its advantages. However, it's worth noting that in certain cases, commercial APIs may offer greater ease of use and cost-effectiveness—especially for smaller-scale applications, such as processing limited datasets of short text, where the streamlined setup and recent price reductions can offset the potential savings of open-source models.

Further research can explore the potential of the text entailment method in text classification, broadening its scope to encompass subjective elements such as emotions, semantic nuances, and discursive tones. More studies are also needed to determine optimal protocols and establish the best practices. For instance, this could involve refining label descriptions to ensure accuracy and comprehensiveness, while maintaining ease of interpretation by the model. Ultimately, these efforts will contribute to the advancement of our understanding and application of text classification methodologies.

Acknowledgments

This study was carried out in the scope of the Brazilian Institute of Public Communication of Science and Technology, with support of the National Council for Scientific and Technological Development (CNPq) and the Rio de Janeiro State Research Support Foundation (FAPERJ). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Luisa Massarani thanks CNPq for the Productivity Grant 1B and FAPERJ for the Scientist of Our State grant.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. https://doi.org/10.1145/3461702.3462624

Adewunmi, M., Sharma, S. K., Sharma, N., Sushma, N. S., & Mounmo, B. (2022). Cancer Health Disparities Drivers with BERTopic Modelling and Pycaret Evaluation. *Cancer Health Disparities*, *6*. Retrieved October 18, 2024, from https://www.companyofscientists.com

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics [Version Number: 1]. https://doi.org/10.48550/ARXIV.2008.09470

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borji, A. (2023). A Categorical Archive of ChatGPT Failures [Version Number: 8]. https://doi.org/10.48550/ARXIV.2302.03494

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners [Version Number: 4]. https://doi.org/10.48550/ARXIV.2005.14165

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Chen, H., Yang, C., Zhang, X., Liu, Z., Sun, M., & Jin, J. (2021). From Symbols to Embeddings: A Tale of Two Representations in Computational Social Science. *Journal of Social Computing*, *2*(2), 103–156. https://doi.org/10.23919/JSC.2021.0011

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Christensen, B., Laydon, D., Chelkowski, T., Jemielniak, D., Vollmer, M., Bhatt, S., & Krawczyk, K. (2022). Quantifying Changes in Vaccine Coverage in Mainstream Media as a Result of the COVID-19 Outbreak: Text Mining Study. *JMIR Infodemiology*, *2*(2), e35121. https://doi.org/10.2196/35121

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670–680. https://doi.org/10.18653/v1/D17-1070

Crabu, S., Giardullo, P., Sciandra, A., & Neresini, F. (2021). Politics overwhelms science in the Covid-19 pandemic: Evidence from the whole coverage of the Italian quality newspapers (F. Zollo, Ed.). *PLOS ONE*, *16*(5), e0252034. https://doi.org/10.1371/journal.pone.0252034

Dagan, I., & Glickman, O. (2004). Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, (26-29), 2–5.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge [Series Title: Lecture Notes in Computer Science]. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché-Buc (Eds.), *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment* (pp. 177–190, Vol. 3944). Springer Berlin Heidelberg. https://doi.org/10.1007/11736790_9

De Melo, T., & Figueiredo, C. M. S. (2021). Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. *JMIR Public Health and Surveillance*, *7*(2), e24585. https://doi.org/10.2196/24585

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]]. Retrieved October 18, 2024, from http://arxiv.org/abs/1810.04805

Edwards, A., Camacho-Collados, J., De Ribaupierre, H., & Preece, A. (2020). Go Simple and Pre-Train on Domain-Specific Corpora: On the Role of Training Data for Text Classification. *Proceedings of the 28th International Conference on Computational Linguistics*, 5522–5529. https://doi.org/10.18653/v1/2020.coling-main.481

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Ghahramani, Z. (2004). Unsupervised Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (pp. 72–112). Springer. https://doi.org/10.1007/978-3-540-28650-9_5

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Version Number: 1]. https://doi.org/10.48550/ARXIV.2203.05794

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and Polarization in COVID-19 News Coverage. *Science Communication*, *42*(5), 679–697. https://doi.org/10.1177/1075547020950735

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing [Version Number: 4]. https://doi.org/10.48550/ARXIV.2111.09543

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention [Version Number: 6]. https://doi.org/10.48550/ARXIV.2006.03654

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021). Topic Modeling for Customer Service Chats. *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 1–6. https://doi.org/10.1109/ICACSIS53237.2021.9631322

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. https://doi.org/10.18653/v1/E17-2068

Ju, W., Sannusi, S. N., & Mohamad, E. (2023). "Public goods" or "diplomatic tools": A framing research on Chinese and American media reports regarding Chinese COVID-19 vaccine. *Media Asia*, *50*(1), 43–81. https://doi.org/10.1080/01296612.2022.2081651

Kaneko, M., Imankulova, A., Bollegala, D., & Okazaki, N. (2022). Gender Bias in Masked Language Models for Multiple Languages. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2740–2750. https://doi.org/10.18653/v1/2022.naacl-main.197

Kaur, H., Pannu, H. S., & Malhi, A. K. (2020). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Computing Surveys*, *52*(4), 1–36. https://doi.org/10.1145/3343440

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software*, *24*(7), 850–858. https://doi.org/10.1016/j.envsoft.2008.11.012

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, *32*(1), 84–100. https://doi.org/10.1017/pan.2023.20

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Liu, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2022). WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation [Version Number: 5]. https://doi.org/10.48550/ARXIV.2201.05955

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4 [Version Number: 3]. https://doi.org/10.48550/ARXIV.2304.03439

Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C. J. P., & Ming, W.-K. (2020). Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. *Journal of Medical Internet Research*, *22*(4), e19118. https://doi.org/10.2196/19118

Liu, Y., Feng, X., Zhang, Y., Kong, Y., & Yang, R. (2022). Paths Study on Knowledge Convergence and Development in Computational Social Science: Data Metric Analysis Based on Web of Science (F. Lai, Ed.). *Complexity*, *2022*(1), 3200371. https://doi.org/10.1155/2022/3200371

Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2-3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Markus, D. K., Mor-Lan, G., Sheafer, T., & Shenhav, S. R. (2023). Leveraging Researcher Domain Expertise to Annotate Concepts Within Imbalanced Data. *Communication Methods and Measures*, *17*(3), 250–271. https://doi.org/10.1080/19312458.2023.2182278

Massarani, L., & Neves, L. F. F. (2021). Communicating the "Race" for the COVID-19 Vaccine: An Exploratory Study in Newspapers in the United States, the United Kingdom, and Brazil. *Frontiers in Communication*, *6*, 643895. https://doi.org/10.3389/fcomm.2021.643895

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016, September). *Who Is Doing Computational Social Science? Trends in Big Data Research* (tech. rep.). SAGE Publishing. https://doi.org/10.4135/wp160926

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September). Efficient Estimation of Word Representations in Vector Space [arXiv:1301.3781 [cs]]. Retrieved October 18, 2024, from http://arxiv.org/abs/1301.3781

Neves, L. F. F., & Massarani, L. (2022). A vacina em dois jornais brasileiros antes e durante a covid-19. *MATRIZes*, *16*(2), 191–216. https://doi.org/10.11606/issn.1982-8160.v16i2p191-216

Nie, Y., Chen, H., & Bansal, M. (2019). Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A New Benchmark for Natural Language Understanding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

Obadimu, A., Mead, E., & Agarwal, N. (2019). Identifying Latent Toxic Features on YouTube Using Non-negative Matrix Factorization.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Parrish, A., Huang, W., Agha, O., Lee, S.-H., Nangia, N., Warstadt, A., Aggarwal, K., Allaway, E., Linzen, T., & Bowman, S. R. (2021). Does Putting a Linguist in the Loop Improve NLU Data Collection? [Version Number: 1]. https://doi.org/10.48550/ARXIV.2104.07179

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Poliak, A. (2020). A survey on Recognizing Textual Entailment as an NLP Evaluation. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 92–109. https://doi.org/10.18653/v1/2020.eval4nlp-1.10

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). Improving Language Understanding by Generative Pre-Training. Retrieved October 20, 2024, from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Ruder, S. (n.d.). *Neural Transfer Learning for Natural Language Processing* [Doctoral dissertation, National University of Ireland, Galway].

Schwarz, A., Alpers, F., Wagner-Olfermann, E., & Diers-Lawson, A. (2024). The Global Study of COVID News: Scope, Findings, and Implications of Quantitative Content Analyses of the COVID-19 News Coverage in the First Two Years of the Pandemic. *Health Communication*, *39*(8), 1568–1581. https://doi.org/10.1080/10410236.2023.2226932

Shajalal, M., Atabuzzaman, M., Baby, M. B., Karim, M. R., & Boden, A. (2023). Textual Entailment Recognition with Semantic Features from Empirical Text Represen-

tation. In A. K. M, B. R. Chakravarthi, B. B, C. O'Riordan, H. Murthy, T. Durairaj, & T. Mandl (Eds.), *Speech and Language Technologies for Low-Resource Languages* (pp. 183–195). Springer International Publishing. https://doi.org/10.1007/978-3-031-33231-9_12

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics* (pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16

Turian, J., Ratinov, L.-A., & Bengio, Y. (2010, July). Word Representations: A Simple and General Method for Semi-Supervised Learning. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384–394). Association for Computational Linguistics. Retrieved October 18, 2024, from https://aclanthology.org/P10-1040

Tyagi, S., & Mittal, S. (2020). Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. In P. K. Singh, A. K. Kar, Y. Singh, M. H. Kolekar, & S. Tanwar (Eds.), *Proceedings of ICRIC 2019* (pp. 209–221). Springer International Publishing. https://doi.org/10.1007/978-3-030-29407-6_17

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August). Attention Is All You Need [arXiv:1706.03762 [cs]]. Retrieved October 18, 2024, from http://arxiv.org/abs/1706.03762

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Wallach, H. (2016). Computational Social Science: Toward a Collaborative Future. In R. M. Alvarez (Ed.), *Computational Social Science: Discovery and Prediction* (pp. 307–316). Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.014

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as Few-Shot Learner [Version Number: 1]. https://doi.org/10.48550/ARXIV.2104.14690

Wang, Y.-X., & Zhang, Y.-J. (2013). Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1336–1353. https://doi.org/10.1109/TKDE.2012.51

Watts, D. (2016). Computational Social Science: Exciting Progress and Future Challenges. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 419. https://doi.org/10.1145/2939672.2945366

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models [Version Number: 2]. https://doi.org/10.48550/ARXIV.2206.07682

Wiedemann, G. (2013). Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 14*, No 2 (2013). https://doi.org/10.17169/FQS-14.2.1949

Wilkerson, J., & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges [Publisher: Annual Reviews]. *Annual Review of Political Science, 20*(Volume 20, 2017), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Yin, W., Hay, J., & Roth, D. (2019, November). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3914–3923). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1404

Yom-Tov, E. (2004). An Introduction to Pattern Classification. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (pp. 1–20). Springer. https://doi.org/10.1007/978-3-540-28650-9_1

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can Large Language Models Transform Computational Social Science? [Version Number: 3]. https://doi.org/10.48550/ARXIV.2305.03514