Journal of Universal Computer Science, vol. 29, no. 6 (2023), 569-594
submitted: 23/10/2022, accepted: 6/3/2023, appeared: 28/6/2023 CC BY-ND 4.0

Politically-oriented information inference from text

Samuel Caetano da Silva
(University of São Paulo, Brazil
https://orcid.org/0000-0002-4469-8901, sam.kaetano@gmail.com)

Ivandré Paraboni
(University of São Paulo, Brazil
https://orcid.org/0000-0002-7270-1477, ivandre@usp.br)

Abstract: The inference of politically-oriented information from text data is a popular research topic in Natural Language Processing (NLP) at both text- and author-level. In recent years, studies of this kind have been implemented with the aid of text representations ranging from simple count-based models (e.g., bag-of-words) to sequence-based models built from transformers (e.g., BERT). Despite considerable success, however, we may still ask whether results may be improved further by combining these models with additional text representations. To shed light on this issue, the present work describes a series of experiments to compare a number of strategies for political bias and ideology inference from text data using sequence-based BERT models, syntax- and semantics-driven features, and examines which of these representations (or their combinations) improve overall model accuracy. Results suggest that one particular strategy - namely, the combination of BERT language models with syntactic dependencies - significantly outperforms well-known count- and sequence-based text classifiers alike. In particular, the combined model has been found to improve accuracy across all tasks under consideration, outperforming the SemEval hyperpartisan news detection top-performing system by up to 6%, and outperforming the use of BERT alone by up to 21%, making a potentially strong case for the use of heterogeneous text representations in the present tasks.

Keywords: Natural language processing, Text classification, Politically-oriented inference, Sentiment analysis, Author profiling
Categories: I.2.7
DOI: 10.3897/jucs.96652

1 Introduction

Inferring politically-oriented information from text data is a popular research topic in Natural Language Processing (NLP), with a wide range of applications to predict individuals' political views [dos Santos and Paraboni, 2019, Jiang et al., 2019, Pavan et al., 2020, Pavan and Paraboni, 2022] and behaviour [Potthast et al., 2018, Kiesel et al., 2019, Li and Goldwasser, 2021]. Existing work in the field often draws a distinction between two main tasks, namely, the issues of political bias and political ideology in general.

Following [Kiesel et al., 2019] and others, political bias is presently defined as any extreme one-sided (or hyperpartisan) discourse that, in a political context, clearly leans towards a liberal or conservative agenda, and which may be deliberately produced as a means to convince or gather support. This contrasts with several other more nuanced forms of political expression, hereby called ideology, in which arguments tend to be more balanced, and which do no explicit promote a particular political agenda. These include expressions of political orientation (e.g., being left- or right-leaning) and political stance in general (e.g., being for or against a political party, a principle, or an individual), among others.

From a computational perspective, both bias and ideology inference from text data may be further divided into two problem definitions, hereby called text- and author-level political inference. Text-level inference is more closely related to sentiment analysis [Zhang et al., 2018, Berka, 2020, Singh and Singh, 2021] and stance classification [Mohammad et al., 2017, Siddiqua et al., 2019, Pavan et al., 2020], and it is intended to determine the meaning of an input text (e.g., whether the text expresses a liberal or conservative view). Less-known author-level inference, by contrast, is an instance of computational author profiling [Vijayaraghavan et al., 2017, Preotiuc-Pietro et al., 2017, Takahashi et al., 2018, Pizarro, 2019, Rangel et al., 2020, Polignano et al., 2020, Price and Hodge, 2020], that is, the task of inferring an individual's demographics (e.g., their political leaning) based on samples of text that they have authored, and which may or may not convey politically-oriented information explicitly.

In recent years, both text- and author-level inference tasks have been implemented with the aid of text representations ranging from simple count-based models (e.g., bag-of-words) to sequence-based models built from transformers such as BERT [Devlin et al., 2019] and others. The latter - which may be seen as large, pre-trained language models - are able to capture deep contextual relations between words, and have been shown to significantly improve downstream task results [Lee et al., 2019, Baly et al., 2020].

Despite the benefits afforded by the use of pre-trained language models, however, in this paper we hypothesise that political inference from text may be improved even further with the aid of two additional text representations based on syntax- and semantics-motivated features. At the syntactic level, we notice that dependency relations (such as those computed from dependency graphs in [Sidorov et al., 2014]) indicate, for instance, whether a word is a subject or an object in the sentence - and may help the underlying model to focus on more relevant information. At the semantic level, we notice that psycholinguistics-motivated features (e.g, words associated with negative emotions, money, religion, etc. such as those provided by the LIWC word categories in [Pennebaker et al., 2001]) may also provide valuable domain-dependent information for political bias and ideology inference.

These observations give rise to the questions of how we may combine transformer-based text representations with syntax- and semantics-motivated features, and which of these so-called heterogeneous text representations may have an impact on different political inference tasks. To shed light on these issues, we envisaged a series of experiments in supervised machine learning addressing a number of text- and author-level tasks alike. As in much of the existing work in the field, some of these experiments make use of English text data but, in addition to that, we also address a number of tasks based on Portuguese text data as well, and introduce a novel dataset for this particular language.

Our main contributions are summarised as follows.

i A neural architecture that combines transformer-based language models, syntactic dependencies and psycholinguistics-motivated features for political bias and ideology inference from text.

ii Text- and author-level formulations of the general political bias and ideology inference tasks.

iii Experiments involving both mainstream English and less-studied Portuguese text data.

iv A novel, large dataset of Twitter text data labelled with political stance information.

The reminder of this paper is structured as follows. Section 2 reviews existing work in political inference from text, and Section 3 describes the computational tasks to be addressed in this work. Section 4 presents our main approach to using heterogeneous text representations. Section 5 summaries our experiment results, which are further discussed in Section 6. Finally, Section 7 presents conclusions and future work.

2 Related work

Table 1 presents an overview of recent NLP work on political bias and ideology inference from text data. All selected studies happen to be devoted to the English language. Further details are discussed below.

Study	Category	Labelling	Features	Method	Genre
[Potthast et al., 2018]	bias	t,a	text	U	N
[Jiang et al., 2019]	bias	t,a	text	CNN	N
[Srivastava et al., 2019]	bias	t,a	text	LogReg	N
[Drissi et al., 2019]	bias	t,a	text	BERT	N
[Lee et al., 2019]	bias	t,a	text	BERT	N
[Bestgen, 2019]	bias	t,a	text	SVM, LogReg	N
[Patankar et al., 2019]	bias	t	text	NPOV	N
[Li and Goldwasser, 2021]	bias	t	text	biLSTM	N
[Iyyer et al., 2014]	ideology	t	text	RNN	D
[Bhatia and P, 2018]	ideology	t	text	LogReg	N
[Kulkarni et al., 2018]	ideology	t	text,graph	CNN	N
[Baly et al., 2020]	ideology	t	text	LSTM, BERT	N,T
[Stefanov et al., 2020]	ideology	t	text,graph	FastText	T
[Feng et al., 2021]	ideology	a	graph	RGC	W

Table 1: Related work in political bias (top) and political ideology (bottom) inference from text. For each study, we report task category as either 'bias' (e.g., hyperpartisan news detection) or 'ideology' (political ideology, stance, and alignment), label granularity (t = text-level, a = author-level), main learning features (text or network relations), learning methods (e.g., RNN = recurrent networks, U = stylometry, CNN = convolutional networks, LogReg = logistic regression, SVM = Support Vector Machines, RGC = graph networks, NPOV=neutral point of view), and text genre (D = discourse/debate, N = news, T = Twitter, W = Wikipedia).

2.1 Political bias

Extremely one-sided political bias has largely focused on the issue of hyperpartisan news detection. The work in [Potthast et al., 2018], which is among the first prominent NLP studies in this field, analysed a corpus of extremely one-sided news and presents a stylometry-based approach to distinguish biased and neutral news, fake news and satire. Among other findings, results suggest that stylometry is of limited use in fake news detection, and that left- and right-wing news share a significant amount of stylistic similarities.

Some of the results from [Potthast et al., 2018] were taken as the basis for the influential SemEval-2019 shared task on hyperpartisan news detection in [Kiesel et al., 2019]. The task comprised two formulations of the problem, each of them based on a different dataset. The *b*y_article dataset contained 1,273 manually labelled texts, and the *b*y_publisher dataset contained 754,000 articles labelled via distant supervision (based on the publishing source of each article). Results reported in [Kiesel et al., 2019] suggest that the *b*y_publisher task was generally more challenging than the *b*y_article task.

Among the participant systems at SemEval-2019, the work in [Jiang et al., 2019] reported the overall highest accuracy in the *b*y_article task. The system made use of a pre-trained ELMo [Peters et al., 2017] language model to encode news texts as input features to a CNN classifier. This approach is currently taken as a baseline to our own experiments described in the next sections.

The work in [Srivastava et al., 2019] made use of a logistic regression classifier to detect hyperpartisan news with the aid of hand crafted features (e.g., bias scores obtained from a bias lexicon, article- and sentence-level polarity, subjectivity and modality scores, etc.) and Universal Sentence Encoder embeddings, among other alternatives. The system obtained the overall highest F1 score in the *b*y_article task at SemEval-2019.

The studies in [Drissi et al., 2019] and in [Lee et al., 2019] were among the first to attempt using pre-trained BERT language models for hyperpartisan news detection. In both cases, however, results remained below those obtained by several more traditional approaches in both *b*y_article and *b*y_publisher tasks.

The work in [Bestgen, 2019] reported the overall highest accuracy in the *b*y_publisher task using a Logistic Regression classifier with a bag-of-words text representation, outperforming a wide range of more complex models based on deep neural networks and others.

The work in [Patankar et al., 2019] took a more application-oriented approach to hyperpartisan bias detection by presenting a real-time system that flags political bias on news articles, and then recommending similar articles from alternative sources. To this end, the system made use of a bias lexicon and unsupervised methods for clustering articles by topic similarity.

Finally, the work in [Li and Goldwasser, 2021], although not an original participant system at SemEval-2019, used the SemEval *b*y_article dataset to address the hyperpartisan news detection task as well. The work made use of a heterogeneous text representation consisting of BERT pre-trained model enriched with social and political information and linguistically-motivated features alike, and obtained results comparable to the top-performing systems at SemEval-2019 with the aid of a multihead Bi-LSTM architecture.

2.2 Political ideology

In addition to extremely one-sided bias, natural language text may convey many other more nuanced forms of political view, which are presently labelled as ‘ideology’ for conciseness. The work in [Iyyer et al., 2014] is among the first of this kind, addressing the issue of political ideology (defined as left, right, or neutral leaning) detection in text. In what nowadays may be seen as a standard approach to the task, the work uses a recurrent neural network model and word embeddings built from left- and right-leaning data to detect ideology in US congressman debates. Results suggest that this approach outperforms a range of logistic regression baseline systems using bag-of-words and word embeddings representations.

Also in the US congressman debates domain, the work in [Bhatia and P, 2018] introduces a sentiment-oriented model to identify political ideology in text, the underlying assumption being that sentiment words may be revealing of an individual’s political leaning (e.g., conservatives may arguably express more positive sentiment towards the free-market topic, etc.) To investigate this issue, potentially relevant topics were selected from a debate corpus, including issues related to health care, US military program and others, and topic-specific sentiments were computed as a probability distribution over ordinal polarity classes ranging from strongly positive to strongly negative. Results suggest that a logistic regression classifier based on sentiment features outperforms the use word embeddings and others.

The work in [Kulkarni et al., 2018] addresses the issue of political ideology detection in news text using a so-called multi-view approach. In this approach, document-level features (e.g., the news headline and contents) and a network of links exhibited in the text are regarded as being complementary properties in the sense that authors may refer to links that reinforce their political views. The study is carried out using a corpus of 120k politically-related news in English, and results suggest that a multi-view approach based on convolutional networks outperforms a range of baseline alternatives including the use of logistic regression classifiers and hierarchical attention models, among others.

The work in [Baly et al., 2020] addresses the task of detecting political ideology in news texts from previously unseen media sources, which prevents models from learning the text source rather than the ideology proper. The work makes use of complementary knowledge obtained from Twitter and Wikipedia sources, and uses adversarial adaptation and triplet loss pre-training (TLP) with both LSTM and BERT models. Results suggest that combining TLP with additional Twitter information outperforms a range of alternatives, including the use of pre-trained transformer models alone, and those with access to additional Wikipedia information.

The work in [Stefanov et al., 2020] addresses the task of characterising the general political leaning of online media and influencers by using unsupervised learning to determine the stance of Twitter users towards a polarising topic based on their retweet behaviour, and then performing label propagation to take the resulting user stance information as training data for media political leaning detection. Results suggest that a combination of User-to-Hashtag and User-to-Mention graph embeddings with BERT models built from both article titles and contents outperforms the use of these individual strategies in isolation.

Finally, the work in [Feng et al., 2021] combines socially- and politically-related features to address the issue of entity stance prediction, an which may be regarded as an instance of author-level inference as discussed in Section 1. Examples of entities under consideration include US presidents, parties, and states. The study builds a heterogeneous information network from Wikipedia articles, in which nodes are social entities and edges are relations between them (e.g., party affiliation, home state, etc.) The network is combined with the Wikipedia text summary of each entity using a *R*oBERTa transformer [Liu et al., 2019] in a gated relational graph convolutional network for representation learning. Results suggest that the approach outperforms a wide range of baseline alternatives, including bag-of-words, average word embeddings, transformers and graph-based models alike.

2.3 Considerations

As discussed in the previous sections, the sheer variety of task definitions, target languages and others may suggest that a direct comparison between existing work and our current approach is not entirely possible. These difficulties not withstanding, however, we notice that none of the reviewed studies attempted to use syntactic dependencies information, nor combines text- and psycholinguistics-motivated features as presently considered.

3 Task definitions

The present work addresses two text-level political inference tasks (T1,T2) and three author-level tasks (T3,T4,T5) using data from three sources: the SemEval-2019 Hyperpartisan news corpus [Kiesel et al., 2019], the BRmoral essay corpus [Pavan et al., 2023] and a novel dataset, hereby called GovBR corpus, to be introduced in Section 4.1. These tasks are summarised in Table 2 and discussed individually in the next sections.

Task	Level	Target	Classes	Corpus	Dataset	Lang.
T1	text	hyperpartisan news	hyperpartisan, neutral	SemEval	by_articles	En
T2	text	political orientation	left, (centre), right	BRmoral	by_opinion	Pt
T3	author	hyperpartisan news	hyperpartisan, neutral	SemEval	by_publisher	En
T4	author	political orientation	left, (centre), right	BRmoral	by_author	Pt
T5	author	political stance	for, against	GovBR		Pt

Table 2: Text- and Author-level political inference tasks according to annotation level, target problem, class definition, corpus, dataset, and language (En=English, Pt=Portuguese).

3.1 Text-level tasks (T1,T2)

Text-level tasks concern the inference of politically-oriented information directly associated with the meanings of the input texts, in which case class labels are annotated at the individual text level. As discussed in Section 1, this is analogous to sentiment analysis, stance classification and related tasks.

Task T1 addresses the issue of text-level hyperpartisan news detection based on the SemEval *b*y_articles dataset [Kiesel et al., 2019]. This consists of a binary classification task intended to distinguish ‘hyperpartisan’ from ‘neutral’ information, as in the following examples.

- Hyperpartisan: *'Trump can't get Congress to repeal Obamacare, he's making changes that will penalize low-income people'*

- Neutral: *'Colin Kaepernick told a CBS reporter that he would represent the national anthem if he was hired by an NFL team'*

Task T2 addresses the issue of text-level political orientation detection based on the BRmoral corpus [Pavan et al., 2023] *b*y_opinion dataset following both binary ('left' / 'right') and ternary ('left' / 'centre' / 'right') class definitions (to be further discussed in Section 4.1.2). Examples taken from short essays related to the issue of same-sex marriage are illustrated as follows (translated from the original texts in Portuguese).

- Left-leaning: *'Agreed, same-sex people must have their marriages valid, as they pay taxes and have the same civil obligations, so they should also have the same rights as the other.'*

- Centre: *'It's not up to the State to forbid this. However if a person asks me if I'm in favour, I'll say no, although I respect (their choice).'*

- Right-leaning: *'From the perspective of the civil institution, I do not see problems in marriage, however, the Christian doctrine makes it clear that the family consists of a man and a woman.'*

3.2 Author-level tasks (T3,T4,T5)

Author-level tasks concern the inference of politically-oriented information related to *the individual* who wrote the input texts rather than to the literal meaning of the text. In this case, class labels are annotated at the author (or publisher) level. As discussed in Section 1, this is analogous to NLP author profiling.

Task T3 is the author-level version of previous task T1, addressing the issue of author-level hyperpartisan news detection based on the SemEval *b*y_publisher dataset [Kiesel et al., 2019]. Once again, this consists of a binary classification task intended to distinguish 'hyperpartisan' from neutral' information, but using weakly labelled data determined by the source of information rather than individually annotated texts. Thus, for instance, all texts produced by a publisher deemed to be a hyperpartisan source are labelled as 'hyperpartisan' news regardless of their actual contents.

Task T4 is the author-level version of previous task T2, addressing the issue of author-level political orientation detection based on the BRmoral [Pavan et al., 2023] *b*y_author dataset following both binary ('left' / 'right') and ternary ('left' / 'centre' / 'right') class definitions. To this end, all texts are weakly labelled with the self-reported political orientation of their authors regardless of the actual text contents. Thus, for instance, opinions that are annotated (at text-level) as 'left-leaning' for the purpose of previous Task T2 may nevertheless be assigned a (author-level) 'right-leaning' label if their authors happen to identify themselves as right-leaning individuals.

Finally, task T5 addresses the issue of author-level stance classification based on the GovBR corpus to be described in the next section. This consists of a binary classification task intended to distinguish Twitter users who are 'for' or 'against' the former Brazilian president Jair Bolsonaro. In this case, tweets are weakly labelled with for/against stance information derived from popular hashtags. Thus, for instance, all tweets accompanied by a *#RespectThePresident* hashtag are assumed to be favourable to the president (the actual hashtags are not included in the data). Examples of both classes are as follows.

- For: *'THIS IS MY PRESIDENT'*
- Against: *'This misgovernment is formed by indecent, immoral, ignorant, stupid and perverse people.'*

4 Materials and method

The goal of the present study is to investigate the use of heterogeneous text representations - based on transformer-based language models, syntactic dependencies, and psycholinguistics-motivated features - for political bias and ideology inference from text as discussed in the previous section. In what follows we describe the data for each task, the classifier models to be investigated, and evaluation procedure.

4.1 Data

The following sections discuss the three corpora used as train and test data for our experiments, and present descriptive statistics.

4.1.1 SemEval-2019 hyperpartisan news corpus (tasks T1 and T3)

For the hyperpartisan news detection tasks T1 and T3, we will make use a subset of the SemEval-2019 Hyperpartisan news corpus [Kiesel et al., 2019] in the English language. The SemEval-2019 corpus consists of political news organised in two datasets called *b*y_articles and *b*y_publisher, both of which annotated with 'hyperpartisan' or 'neutral' labels. Hyperpartisan news convey extreme one-sided information of either liberal or conservative nature alike. The full corpus data originally conveys 1,273 news articles in the *b*y_articles set, and 754,000 articles in the *b*y_publisher set, from which the training subsets are presently used in our experiments.

As discussed in the previous section, *b*y_articles texts are labelled individually according to their contents, and were taken as an input to text-level hyperpartisan news detection (task T1). *b*y_publisher texts, by contrast, are weakly labelled according to their media source, and will be taken as the input for author-level hyperpartisan news detection (task T3). Despite using well-known shared task data, however, we notice that we do not presently seek to outperform the existing SemEval benchmark, but rather compare a number of novel computational strategies among themselves, and show that some of these alternatives may represent an improvement over the existing benchmark for certain tasks.

For the purpose of the present work, the original train portions of both datasets were randomly split into development (80%) and test (20%) sets. Table 3 presents the resulting class distribution.

4.1.2 BRmoral essay corpus (tasks T2 and T4)

For the political orientation detection tasks T2 and T4, we used the BRmoral essay corpus [Pavan et al., 2023] in the Portuguese language. This consists of short essays about eight topics of liberal and conservative nature alike (same-sex marriage, gun possession, abortion, death penalty, drug legalisation, lowering of criminal age, racial quotas, and tax exemptions for churches) labelled with both stance scores (from 'totally against' to

	*b*y_articles			*b*y_publisher		
Set	hyperpartisan	neutral	Total	hyperpartisan	neutral	Total
Dev	332 (64.3%)	184 (35.7%)	516	1235 (49.4%)	1265 (50.6%)	2500
Test	75 (58.1%)	54 (41.9%)	129	492 (49.2%)	508 (50.8%)	1000

Table 3: SemEval-2019 by_articles and by_publisher class distribution in development and test sets.

'totally for' each topic) and authors' demographics, including their self-reported political orientation from 'extreme left' to 'extreme right'. The full corpus data conveys 4080 essays written by 510 crowd-sourced volunteers [Pavan et al., 2023].

The dual labelling scheme in the BRmoral corpus (i.e., either based on individual stances or author's own political orientation) gives rise to two dataset definitions, hereby called *b*y_opinion and *b*y_author for analogy with the SemEval *b*y_articles and *b*y_publisher datasets discussed in the previous section. Both *b*y_opinion and *b*y_author are labelled with 'left', 'right' and, depending on the task under consideration (see below), also with 'centre' information.

*b*y_opinion takes as labels the liberal and conservative stance information available from the corpus to determine, albeit indirectly, a text's likely political leaning. More specifically, texts expressing an opinion against so-called liberal topics (same-sex marriage, abortion, drug legalisation, and racial quotas), or those expressing opinions in favour of so-called conservative topics (death penalty, gun possession, lowering of criminal age, and tax exemptions for churches), are labelled as 'right', and so forth. This dataset will be taken as an input to text-level political orientation detection (task T2).

The *b*y_author dataset, by contrast, takes as labels the actual authors' political orientation information available from the corpus, in what may be seen as an instance of weakly labelling not unlike the SemEval *b*y_publisher labels discussed in the previous section. BRmoral *b*y_author texts will be taken as an input to author-level political orientation detection (task T4).

Both datasets were randomly split into development (80%) and test (20%) sets. Table 4 presents the resulting class distribution.

		*b*y_opinion		
Set	left	centre	right	Total
Dev	1201 (36.8%)	685 (21.0%)	1378 (42.2%)	3264
Test	299 (36.6%)	176 (21.6%)	341 (41.8%)	816
		*b*y_author		
Set	left	centre	right	Total
Dev	1210 (37.1%)	1158 (35.5%)	896 (27.4%)	3264
Test	310 (38.0%)	282 (34.5%)	224 (27.5%)	816

Table 4: BRmoral by_opinion and by_author class distribution in development and test sets.

4.1.3 GovBR political stance corpus (task T5)

Finally, for the political stance task T5, we created a novel language resource based on Twitter data in the Portuguese language, hereby called the GovBR corpus. GovBR comprises a collection of tweets written by users who expressed a clear stance towards the current president of Brazil. The corpus was built by selecting two disjoint sets of users - supporters and opponents of the said president - according to the use of a number of popular politically-oriented hashtags (e.g., *#RespectThePresident* or *#NotHim*. The full corpus data - from which non-political tweets were filtered out as discussed below - conveys 13.5 million tweets written by 5452 unique users.

For each selected user, all their publicly available tweets (i.e., disregarding their retweets) were downloaded. Users who simultaneously promoted supportive and opposing hashtags were discarded, and so were all hashtags and all messages shorter than five words. Finally, tweets that did not convey a minimal level of political content were also discarded. To this end, we computed a TF-IDF representation of the political section of the *Folha de SP* newspaper[1], and kept only the tweets conveying a minimum degree of similarity to the political news texts. After the removal of non-political tweets, we obtained approximately 25 tweets per user.

GovBR politically-related tweets will be taken as the input for author-level political stance detection (task T5). In our current work, we use a balanced subset of this data consisting of 4010 randomly selected tweets. These were randomly split into development (80%) and test (20%) sets as illustrated in Table 5

Set	against	for	Total
Dev	1600 (49.9%)	1608 (50.1%)	3208
Test	405 (50.5%)	397 (49.5%)	802

Table 5: GovBR class distribution in development and test sets.

As Twitter privacy policies prohibit the reproduction of tweet texts, the GovBR corpus is provided as a set of tweet identifiers and corresponding class labels only. From these identifiers, the actual dataset may be recreated by downloading each individual tweet[2], and by removing the existing hashtags.

4.2 Models

As a means to investigate the issue of political inference from text (tasks T1-T5 described in the previous sections), in what follows we propose combining heterogeneous text representations into a convolutional neural network architecture. An overview of this architecture is presented in Section 4.2.1, and its individual components are described in Section 4.2.2.

4.2.1 Architecture

We envisaged[3] a convolutional neural network architecture for political inference that combines three kinds of text representation: (i) pre-trained language models provided

[1] https://www.kaggle.com/marlesson/news-of-the-site-folhauol

[2] https://drive.google.com/file/d/1aJreKP6YJ2lJa865jeTaPx705Py8-IkR/view?usp=sharing

[3] https://github.com/samcaetano/ideological_bias_detector

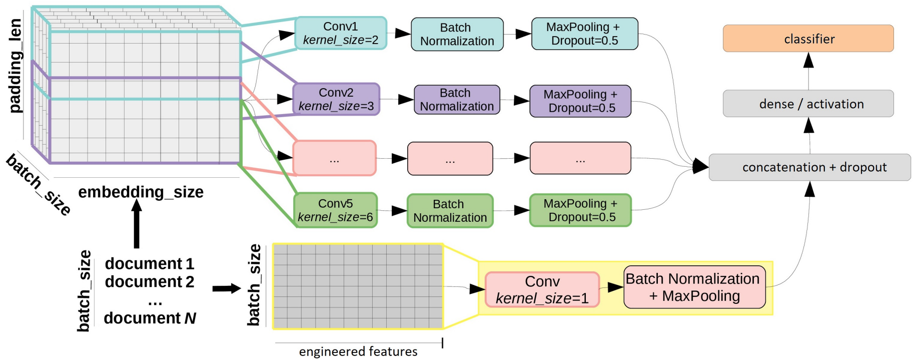

Figure 1: General architecture (best visualised in digital format with zoom tool).

by Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019], hereby called *bert*; (ii) syntactic bigram counts computed from dependency graphs [Sidorov et al., 2014], hereby called *sngram*, and (iii) psycholinguistics-motivated features obtained from Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2001] and from the Medical Research Council (MRC) database [Coltheart, 1981], hereby called *psych*. This architecture is illustrated in Figure 1 and further discussed below[4].

Given a set of input documents (bottom left of the figure), we use a text classifier model that takes as an input both standard text features represented as contextual embeddings *bert* (top left), and engineered features (bottom centre) that combine the alternative text representations based on syntactic dependencies *sngram* and psycholinguistics-motivated features *psych*. In this approach, *bert* embeddings are taken as an input to five convolutional layers (Conv1-Conv5) followed by batch normalisation, a max pooling and a 0.5 dropout layers, whereas engineered features are processed in parallel by a kernel of size 1, also followed by a normalisation layer and max pooling. Finally, the output of the operations Conv1-Conv5 is concatenated with the engineered feature vector followed by a 0.5 dropout, and a softmax activation function produces the class predictions.

The full model architecture, hereby called *b*ert+sngram+psych, will be evaluated against a number of baseline systems, and also against some of its individual components as discussed in Section 4.3.

4.2.2 Text representations

In the *b*ert+sngram+psych model, input texts are to be represented as contextual embeddings (our *bert* component), syntactic bigram counts *sngram*, and psycholinguistics-motivated features (*psych*). These components are described as follows.

Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] are now mainstream in NLP and related fields, and are the basis for our *bert* architectural component as well. This consists of pre-trained BERT language models[5] fine-tuned for each of our classification tasks T1-T5. For the English language models (tasks T1 and T3), we use *base-uncased* BERT, and for the Portuguese language (tasks T2, T4, and T5) we use *multilingual-base-uncased* BERT.

[4] A preliminary version of this approach, originally applied to hate speech prediction, appeared in [da Silva et al, 2020].

[5] https://huggingface.co/models/

The *sngram* component explores the use of text structural information to enrich our classifier models by computing syntactic bigram counts from dependency graphs. To this end, we first compute a syntactic dependency graph from the input text using SpaCy[6], and then generate a TF-IDF bigram model as suggested in [Sidorov et al., 2014]. For the English language models (tasks T1 and T3), we use the en_core_web_sm pipeline, and for the Portuguese language (tasks T2, T4, and T5) we use the pt_core_news_sm pipeline. Finally, we perform univariate feature selection using F1 as a score function in order to keep only the k best features. Optimal k values for each task were obtained through grid search on development data.

The *psych* component makes use of psycholinguistics-motivated features computed with the aid of both LIWC [Pennebaker et al., 2001] and MRC [Coltheart, 1981] lexicons. Examples of LIWC word categories include those related to attention focus (e.g., pronouns and verb tense), affective or emotional processes (positive and negative emotions, anxiety, fear, etc.), social relationships (e.g., family, friends, etc.) and others. Similarly, MRC categories cover lexical features such as concreteness, age of acquisition, and others. For the English language models (tasks T1 and T3), we use the 93-feature LIWC-2015 lexicon [Pennebaker et al., 2015] and the 9-feature MRC database [Coltheart, 1981]. For the Portuguese language (tasks T2, T4, and T5) we use 64-feature LIWC-BR [Balage Filho et al., 2013] and the 6 MRC-like features from [dos Santos et al., 2017].

Both LIWC and MRC text representations consist of word category counts normalised by document size, in which words that belongs to more than one category update all related counts (e.g., 'she' is a pronoun and also a feminine word, etc.) Both representations are concatenated as a single vector of size $93 + 9 = 101$ features for English, or $64 + 6 = 70$ features for Portuguese. As in the case of the syntactic features discussed above, we once again perform univariate feature selection with F1 as a score function to obtain the k best features for each task.

4.3 Procedure

We conducted a series of experiments focused on tasks T1-T5 introduced in Section 3 to assess the use of the *b*ert+sngram+psych model and some of its subcomponents, namely, *bert*, *sngram*, *psych* alone, and also the two BERT-based pairs *b*ert+sngram and *b*ert+psych. More specifically, our goal is to investigate how each of these alternatives compare to two baseline systems, namely, the Bertha von Suttner model described in [Jiang et al., 2019], which was the overall best-performing system at SemEval-2019 Hyperpartisan news detection shared task [Kiesel et al., 2019], and the use of BERT alone as a classifier, hereby called *B*ERT.baseline.

All models were trained in 30 epochs using a development dataset partition, and then evaluated using a previously unseen test data as described in Section 4.1. For BERT-based models, additional pre-processing was performed to remove all non-alphabetic characters, links and HTML tags. All input documents were limited to their first 300 tokens, and shorter documents were completed with the [PAD] token. This representation was taken as the input to a BERT model of size 768, resulting in text embeddings of size 300×768. Depending on the input language of each task, the underlying BERT model was either *base-uncased* or *multilingual-base-uncased* BERT, as discussed in the previous section.

*s*ngram and *p*sych features are concatenated as a single vector, and a z-score function is applied to obtain standardised value ranges. Table 6 summarises the actual number of

[6] https://spacy.io/usage/linguistic-features#dependency-parse

features considered by each models and for each corpus.

Corpus	psych	sngram
*S*emEval by_articles	37	13,107
*S*emEval by_publisher	78	38,220
*B*Rmoral by_opinion	55	13,466
*B*Rmoral by_author	22	13,399
GovBR	25	6,998

Table 6: Number of features used in psych and sngram vectors in each corpus.

Evaluation proper was carried out by (i) measuring Accuracy (Acc), macro F_1 (F_1), Precision (P) and Recall (R) scores, (ii) by assessing statistical significant differences between models, and (iii) by providing model prediction explanations. To this end, statistical significance is assessed by using the McNemar test [McNemar, 1947] in the case of binary classifiers, and by using the Stuart-Maxwell test [Stuart, 1955, Maxwell, 1970] for ternary classifiers. For each task, two kinds of significance tests are conducted. First, we identify those models that are statistically superior to the reference model in [Jiang et al., 2019]. Second, we identify the groups of models that are statistically distinguishable from others. Finally, we performed eli5 prediction explanation[7] to compute the word features more strongly correlated with each class and task, as discussed in Section 6.

5 Results

This section presents results of the full model *b*ert+sngram+psych and its subcomponents (*bert*, *sngram*, *psych*, *b*ert+sngram, and *b*ert+psych) compared with those obtained by the work in [Jiang et al., 2019] and by BERT alone as a classifier (*b*ert.baseline).

Results are reported as Accuracy (Acc), F_1, Precision (P) and Recall (R) scores divided into three groups: (1) on the top row baselines from [Jiang et al., 2019] and *b*ert.baseline, (2) following by model components *bert*, *sngram*, *psych*, *b*ert+sngram, and *b*ert+psych, and (3) on the bottom row the full model *b*ert+sngram+psych. In all scenarios, best accuracy scores are highlighted, and also marked as * when found to be statistically superior to the baseline system in [Jiang et al., 2019]. Finally, in addition to the main results report for each task, models are also depicted in statistically significant clusters ($p < 0.05$) according to their accuracy scores.

5.1 Task T1: Text-level hyperpartisan news detection

Results for task T1 - text-level hyperpartisan news detection - using the SemEval corpus *b*y_articles dataset are summarised in Table 7 and further discussed below.

Based on these results, we notice that the best-performing model is *b*ert+sngram. The difference between this and the baseline in [Jiang et al., 2019] is statistically significant ($\chi = 5.161$, $\alpha = 0.05$, $p < 0.05$). The full model *b*ert+sngram+psych, by contrast, ranks considerably lower. To further illustrate this outcome, the models were clustered into homogeneous groups (A,B,C) by statistical significance according to their accuracy scores as illustrated in Table 8.

[7] https://eli5.readthedocs.io/en/latest/

Model	Acc	F_1	P	R
Jiang et. al.	0.72	0.65	0.69	0.61
BERT.baseline	0.68	0.65	0.61	0.71
bert	0.75	0.69	0.73	0.65
sngram	0.69	0.69	0.70	0.69
psych	0.47	0.46	0.45	0.47
bert+sngram	**0.78***	0.77	0.78	0.78
bert+psych	0.74	0.71	0.66	0.76
bert+sngram+psych	0.67	0.64	0.67	0.67

Table 7: Text-level hyperpartisan news detection results.

Model	Acc	Groups		
bert+sngram	**0.78***	A		
bert	0.75	A		
bert+psych	0.74	A		
Jiang et al.	0.72		B	
sngram	0.69		B	
BERT.baseline	0.68		B	
bert+sngram+psych	0.67		B	
psych	0.47			C

Table 8: Text-level hyperpartisan news detection homogeneous groups.

We notice that, in addition to the best-performing *b*ert+sngram model, both *bert* and *b*ert+psych obtain, to a lesser extent, statistically similar results within group A. The full model and the reference baseline in [Jiang et al., 2019], by contrast, are both members of group B.

5.2 Task T2: Text-level political orientation detection

Results for task T2 - text-level political orientation detection - using the BRmoral corpus *b*y_opinion dataset for both binary (left, right) and ternary (left,centre,right) classification are summarised in Table 9 and further discussed below.

Once again, the best-performing model for both binary and ternary classification is *b*ert+sngram. However, differences between this and others, including the baseline in [Jiang et al., 2019], were not found to be statistically significant. To further illustrate this outcome, homogeneous groups related to the binary classification task are shown in Table 10, and groups related to ternary classification are shown in Table 11.

In both binary and ternary classification tasks, although *b*ert+sngram still obtains the highest accuracy scores, its simpler *bert* sub-component (i.e., the combination of a fine-tuned BERT model with a CNN classifier, as discussed in Section 4) is statistically

	Binary classification				Ternary classification			
Model	Acc	F_1	P	R	Acc	F_1	P	R
Jiang et. al.	0.76	0.76	0.77	0.76	0.62	0.61	0.61	0.62
BERT.baseline	0.57	0.57	0.57	0.57	0.43	0.36	0.43	0.43
bert	0.75	0.75	0.76	0.75	0.60	0.56	0.59	0.60
sngram	0.70	0.70	0.71	0.70	0.55	0.53	0.55	0.55
psych	0.68	0.68	0.68	0.68	0.52	0.50	0.49	0.52
bert+sngram	**0.78**	0.78	0.78	0.78	**0.64**	0.62	0.63	0.64
bert+psych	0.76	0.76	0.77	0.76	0.60	0.59	0.59	0.60
bert+sngram+psych	0.73	0.73	0.74	0.73	0.58	0.58	0.58	0.58

Table 9: Text-level political orientation detection results.

Model	Acc	Groups			
bert+sngram	**0.78**	A			
Jiang et. al.	0.76	A			
bert+psych	0.76	A			
bert	0.75	A			
bert+sngram+psych	0.73		B		
sngram	0.70		B		
psych	0.68			C	
BERT.baseline	0.57				D

Table 10: Binary text-level political orientation detection homogeneous groups.

Model	Acc	Groups			
bert+sngram	**0.64**	A			
Jiang et. al.	0.62	A			
bert	0.60	A			
bert+psych	0.60		B		
bert+sngram+psych	0.58		B		
sngram	0.55		B		
psych	0.52			C	
BERT.baseline	0.43				D

Table 11: Ternary text-level political orientation detection homogeneous groups.

similar. The full model, *b*ert+sngram+psych, composes the group B in both classification tasks, ranking lower than the reference system [Jiang et al., 2019].

5.3 Task T3: Author-level hyperpartisan news detection

Results for task T3 - author-level hyperpartisan news detection - using the SemEval corpus *b*y_publisher dataset are summarised in Table 12 and further discussed below.

Model	Acc	F_1	P	R
Jiang et. al.	0.56	0.62	0.55	0.71
BERT.baseline	0.54	0.60	0.51	0.75
bert	0.58	0.67	0.55	0.85
sngram	0.57	0.55	0.57	0.57
psych	0.53	0.53	0.53	0.53
bert+sngram	0.57	0.54	0.59	0.57
bert+psych	**0.61***	0.67	0.58	0.80
bert+sngram+psych	0.55	0.52	0.50	0.55

Table 12: Author-level hyperpartisan news detection results.

Based on these results, we notice that the best-performing alternative is *b*ert+psych. The difference between this and the baseline in [Jiang et al., 2019] is statistically significant ($\chi = 6.715$, $\alpha = 0.05$, $p < 0.01$). Models were clustered into homogeneous groups by statistical significance according to their accuracy scores as illustrated in Table 13.

Model	Acc	Groups	
bert+psych	**0.61***	A	
bert	0.58	A	
bert+sngram	0.57	A	
sngram	0.57		B
Jiang et al.	0.56		B
bert+sngram+psych	0.55		B
BERT.baseline	0.54		B
psych	0.53		B

Table 13: Author-level hyperpartisan news detection homogeneous groups.

According to Table 13, group A includes most alternatives that are based on BERT language models, and once again the difference between the full model *b*ert+sngram+psych and the reference baseline in [Jiang et al., 2019], both of which in group B, is not statistically significant.

5.4 Task T4: Author-level political orientation detection

Results for task T4 - author-level political orientation detection - using the BRmoral corpus *b*y_author dataset for both binary (left, right) and ternary (left,centre,right) classification are summarised in Table 14 and further discussed below.

	Binary classification				Ternary classification			
Model	Acc	F_1	P	R	Acc	F_1	P	R
Jiang et al.	0.61	0.61	0.61	0.61	0.38	0.37	0.37	0.38
BERT.baseline	0.56	0.55	0.55	0.56	0.37	0.31	0.37	0.37
bert	0.53	0.53	0.58	0.53	0.41	0.37	0.41	0.41
sngram	0.56	0.56	0.57	0.56	0.38	0.37	0.37	0.38
psych	0.59	0.56	0.57	0.59	0.37	0.35	0.35	0.37
bert+sngram	**0.63**	0.63	0.63	0.63	0.41	0.37	0.40	0.41
bert+psych	0.60	0.53	0.59	0.60	0.36	0.30	0.44	0.36
bert+sngram+psych	0.59	0.59	0.61	0.59	**0.42**	0.41	0.41	0.42

Table 14: Author-level political orientation detection results.

The best-performing model for binary classification is *b*ert+sngram, and for ternary classification is *b*ert+sngram+psych. However, the differences between these and the baseline in [Jiang et al., 2019] were not found to be significant. Homogeneous groups for the binary task are illustrated in Table 15, and groups for the ternary task are illustrated in Table 16.

Model	Acc	Groups	
bert+sngram	**0.63**	A	
Jiang et al.	0.61	A	
bert+psych	0.60	A	
psych	0.59	A	
bert+sngram+psych	0.59	A	
BERT.baseline	0.56	A	
sngram	0.56		B
bert	0.53		B

Table 15: Binary author-level political orientation detection homogeneous groups.

In both binary and ternary classification tasks, we notice that several models turned out to obtain statistically equivalent results. This outcome, which is similar to what has been observed in the text-level political orientation task (cf. Section 5.2) based on the same corpus, will be further discussed in Section 6.

5.5 Task T5: Author-level political stance detection

Results for task T5 - author-level political stance detection - using the *G*ovBR corpus are summarised in Table 17 and further discussed below.

Once again, the best-performing model is *b*ert+sngram, but others, including the baseline in [Jiang et al., 2019], were found to be similar. The corresponding homogeneous groups are illustrated in Table 18.

Model	Acc	Groups
bert+sngram+psych	**0.42**	A
bert+sngram	0.41	A
bert	0.41	A
Jiang et al.	0.38	A
sngram	0.38	A
psych	0.37	B
BERT.baseline	0.37	B
bert+psych	0.36	B

Table 16: Ternary author-level political orientation detection homogeneous groups.

Model	Acc	F_1	P	R
Jiang et. al.	0.58	0.55	0.62	0.58
BERT.baseline	0.53	0.52	0.54	0.53
bert	0.61	0.59	0.64	0.61
sngram	0.53	0.53	0.53	0.53
psych	0.52	0.51	0.52	0.52
bert+sngram	**0.62**	0.62	0.62	0.62
bert+psych	0.60	0.60	0.60	0.60
bert+sngram+psych	0.59	0.58	0.59	0.59

Table 17: Author-level political stance detection results.

Model	Acc	Groups
bert+sngram	**0.62**	A
bert	0.61	A
bert+psych	0.60	A
bert+sngram+psych	0.59	A
Jiang et. al.	0.58	A
BERT.baseline	0.53	B
sngram	0.53	B
psych	0.52	B

Table 18: Author-level political stance detection homogeneous groups.

As in some of the previous experiments, although *b*ert+sngram still obtains the highest accuracy among the alternatives, some of its simpler sub-components (as *bert*, in this case) were found to be statistically similar. Moreover, the full model *b*ert+sngram+psych is once again statistically similar to the reference baseline in [Jiang et al., 2019].

6 Discussion

In what follows we summarise our main results (Section 6.1) and report prediction explanations for the main classification tasks.

6.1 Results summary

Table 19 shows the tasks in which each of the models under evaluation ranks among the top-performing alternatives (i.e., belonging to cluster A in each of the homogeneous groups illustrated in the previous section).

		Text-level tasks			Author-level tasks			
Model	Wins	T1	T2.bin	T2.ter	T3	T4.bin	T4.ter	T5
Jiang et al.	5		A	A		A	A	A
BERT.baseline	1					A		
bert	6	A	A	A	A		A	A
sngram	1						A	
psych	1					A		
bert+sngram	7	A	A	A	A	A	A	A
bert+psych	5	A	A		A	A		A
bert+sngram+psych	3					A	A	A

Table 19: Models that rank in the top-performing cluster (A) for each task.

From these results, a number of observations are warranted. First, we notice that there is a large number of statistically-similar models in the author-level tasks (on the right of the table), particularly in the case of the essay (tasks T4) and Twitter (task T5) domains. This was to some extent to be expected as the information to be learned from author-level inference is less explicit in the input texts, making these tasks possibly more challenging than text-level inference in general.

Second, we notice that the full model, comprising the main CNN architecture and *bert*, *sngram* and *psycho* sub-components, is generally not the best choice, being often outperformed by simpler alternatives and/or by the reference baseline system in [Jiang et al., 2019].

Regarding the role of individual sub-components of the main architecture, we notice that using *psych* or *sngram* alone is clearly insufficient, and that even the use of the *bert* component alone generally fails to deliver optimal results. This outcome suggests that using fine-tuned BERT and the CNN architecture alone, as in the present *bert* model (not to be mistaken by the standard *B*ERT.baseline model in the second row from the top, which does not use a CNN classifier) explains much of the best results obtained across our experiments, but not all of them. In fact, it is the combination of BERT and *sngrams* in the CNN architecture (represented by the *b*ert+sngram model) that generally obtains the best results among these alternatives, being the top-performing model in 5 out of 7 tasks. This outcome may be partially explained by the simultaneous use of two representations of the input text (i.e., linearly ordered tokens and count-based syntactic bigram features), but we notice that larger models using this strategy (e.g., including psycholinguistics-motivated features) are not necessarily better.

Finally, we notice that the baseline in [Jiang et al., 2019] remains highly competitive and, although seldom obtaining the highest accuracy, it is often found within the group of top-performing alternatives.

6.2 Feature importance

As a means to illustrate the word features more strongly correlated with each class and task, we performed eli5 model explanation to obtain word weights representing the change (decrease/increase) of the evaluation score when a given feature is shuffled. As it is generally the case in data-driven methods of this kind, the most important information for each task tend to include both highly intuitive features (namely, politically-oriented terms) and others that simply happen to correlate with them in the training data.

Table 20 presents the most important features associated with the hyperpartisan (i.e., non-neutral) class in the SemEval tasks T1 (*b*y_articles) and T3 (*b*y_publisher) datasets, and their positive (top) or negative (bottom) weights.

T1: Text-level		T3: Author-level	
weight	feature	weight	feature
10.467	cnn	6.02	window
10.169	american	5.585	bush
10.069	it	5.542	mr
9.905	women	5.149	that
9.708	hillary	4.833	iraq
9.536	her	4.367	nyse
9.507	now	4.089	billion
...	...	...	...
-6.754	million	-4.699	reuters
-6.756	muslim	-4.829	percent
-6.952	rally	-5.025	globalpost
-7.46	donald	-5.463	mexico
-7.531	isis	-5.629	california
-7.895	<BIAS>	-6.268	albuquerque
-8.17	mr	-6.921	ap

Table 20: Text-based (left) and author-based (right) hyperpartisan news detection most important features, and their positive/negative weights.

Table 21 presents the most important features associated with the left/right classes in the BRmoral corpus tasks T2 and T4 (*b*y_opinion and *b*y_author datasets, respectively), translated from the original Portuguese input, and their positive (top) or negative (bottom) weights.

T1: Text-level				T3: Author-level			
left orientation		right orientation		left orientation		right orientation	
weight	feature	weight	feature	weight	feature	weight	feature
2.533	remove	2.743	black	0.902	women	0.628	favour
2.388	guns	2.526	choice	0.805	about	0.618	mine
2.385	institution	2.493	love	0.739	lay	0.618	be (present)
2.29	am	2.278	marriage	0.668	society	0.6	since
2.234	churches	2.195	wed	0.654	more	0.587	acts
1.86	money	2.174	couple	0.643	body	0.549	should
1.807	god	2.154	abortion	0.625	population	0.54	be (subjunctive)
1.751	youth	1.977	alcohol	0.603	while	0.533	freedom
1.727	life	1.965	traffic	0.596	alcohol	0.481	they
1.696	for	1.944	gender	0.595	think	0.465	I
...	...	...	...	...	...	...	...
-1.677	black	-1.738	reduce	-0.515	defend	-0.483	she
-1.698	child	-1.746	innocent	-0.515	mine	-0.494	country
-1.851	law	-1.817	church	-0.516	they	-0.494	are
-1.857	crimes	-1.953	remove	-0.528	crimes	-0.507	population
-1.984	favour	-2.111	although	-0.546	because	-0.523	this_way
-2.017	consciousness	-2.114	have	-0.572	same	-0.526	women
-2.029	alcohol	-2.138	institution	-0.59	hideous	-0.551	some
-2.083	hideous	-2.394	churches	-0.633	colour	-0.621	or
-2.285	prohibition	-2.505	guns	-0.703	am	-0.621	public
-2.416	freedom	-3.19	perhaps	-0.755	be	-0.658	lay

Table 21: Text-level (left) and author-level (right) political orientation detection most important features, and their positive/negative weights.

Finally, Table 22 presents the most important features associated with political stance in GovBR task T5, once again translated from the original Portuguese input, and their positive (top) or negative (bottom) weights.

7 Final remarks

This paper has addressed the issue of how to combine transformed-based text representations, which are main stream in NLP and related fields, with both syntactic dependency information and psycholinguistics-motivated features for political inference from text data. In doing so, we considered both text- and author-level task definitions in both English and Portuguese languages, and introduced a novel dataset devoted the latter.

As expected in experiments involving a range of tasks, datasets and languages, our present results vary considerably across evaluation settings and, although BERT remains a robust baseline for many of these tasks, in most cases it was possible to obtain significant improvements by making use of additional representations. This is not to say, however, that combining *all* current three text representations (BERT, syntax and psycholinguistics) into a single model is the best option. In particular, it was a subset

for		against	
weight	feature	weight	feature
5.826	supreme_court	-4.087	came
5.709	jair_bolsonaro	-4.138	let's_go
5.159	communists	-4.161	some
4.875	visit	-4.201	poor
4.789	senator	-4.235	human
4.694	criminal	-4.33	(president, derogatory)
4.653	red	-4.399	out
4.589	care	-4.662	sportive
4.561	call	-4.687	damn
4.507	thinks	-4.899	political

Table 22: Political stance detection most important features, and their positive/negative weights.

of our original CNN architecture combining only BERT and the syntactic dependency model that obtained overall best results in most tasks.

The current work leaves a number of opportunities for further research. First, we notice that political bias and ideology are relatively broad terms that may actually include a wide range of distinct politically-related phenomena, and that future NLP studies may benefit from more fine grained task definitions.

We notice also that many other pre-trained language models have been made available in recent years, including ELMo [Peters et al., 2017], RoBERTa [Liu et al., 2019], and GPT-3 [Brown et al., 2020]. Whether any of these may outperform BERT when combined with other text representations (as in the present work) remains an open research question.

Finally, the present use of text representations is only a first step towards more informed models that may ultimately combine BERT with many other text- and author-level features. In particular, the present architecture may be expanded with, for instance, sentiment or emotion-related information, or even with author demographics (e.g., gender, age, personality traits, moral foundations, etc.) obtained with the aid of author profiling classifiers [Rangel et al., 2020, dos Santos et al, 2020]. These and many other related possibilities are also left as future work.

Acknowledgements

The present work received support from the São Paulo Research Foundation (FAPESP grant #2021/08213-0). The first author received support from the Brazilian Foundation CAPES - Coordination for the Improvement of Higher Education Personnel, under grant 88882.378103/2019-01.

References

[Balage Filho et al., 2013] Balage Filho, P. P., Aluísio, S. M., and Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 215–219, Fortaleza, Brazil.

[Baly et al., 2020] Baly, R., Martino, G. D. S., Glass, J., and Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

[Berka, 2020] Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. *Journal of Intelligent Information Systems*, 55:51–66.

[Bestgen, 2019] Bestgen, Y. (2019). Tintin at SemEval-2019 task 4: Detecting hyperpartisan news article with only simple tokens. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1062–1066, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Bhatia and P, 2018] Bhatia, S. and P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

[Brown et al., 2020] 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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

[Coltheart, 1981] Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33(4):497–505.

[da Silva et al, 2020] da Silva, S.C., Ferreira, T.C., Ramos, R.M.S., and Paraboni, I. (2020). Data Driven and Psycholinguistics Motivated Approaches to Hate Speech Detection. *Computación y Systemas*, 24(3):1179–1188. 10.13053/CyS-24-3-3478.

[Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

[dos Santos et al., 2017] dos Santos, L. B., Duran, M. S., Hartmann, N. S., Junior, A. C., Paetzold, G. H., and Aluísio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In *International Conference on Text, Speech, and Dialogue*. Springer.

[dos Santos and Paraboni, 2019] dos Santos, W. R. and Paraboni, I. (2019). Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In *Recents Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria. 10.26615/978-954-452-056-4_123

[dos Santos et al, 2020] dos Santos, W. R. and Ramos, R.M.S., and Paraboni, I. (2020). Computational Personality Recognition from Facebook text: psycholinguistic features, words and facets. In *New Review of Hypermedia and Multimedia*, 24(4) 268–287. 10.1080/13614568.2020.1722761

[Drissi et al., 2019] Drissi, M., Segura, P. S., Ojha, V., and Medero, J. (2019). Harvey mudd college at SemEval-2019 task 4: The clint buchanan hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 962–966, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Feng et al., 2021] Feng, S., Chen, Z., Yu, P., and Luo, M. (2021). Encoding heterogeneous social and political context for entity stance prediction. *arXiv preprint arXiv:2108.03881*.

[Iyyer et al., 2014] Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

[Jiang et al., 2019] Jiang, Y., Petrak, J., Song, X., Bontcheva, K., and Maynard, D. (2019). Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Kiesel et al., 2019] Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Kulkarni et al., 2018] Kulkarni, V., Ye, J., Skiena, S., and Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

[Lee et al., 2019] Lee, N., Liu, Z., and Fung, P. (2019). Team yeon-zi at SemEval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Li and Goldwasser, 2021] Li, C. and Goldwasser, D. (2021). Using social and linguistic information to adapt pretrained representations for political perspective identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4569–4579, Online. Association for Computational Linguistics.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

[Maxwell, 1970] Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *The British Journal of Psychiatry*, 116(535):651–655.

[McNemar, 1947] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

[Mohammad et al., 2017] Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

[Paraboni and de Lima, 1998] Paraboni, I. and de Lima, V.L.S. (1998). Possessive pronominal anaphor resolution in Portuguese written texts. *Proceedings of the 17th international conference on Computational linguistics*, vol.2:1010–1014. Association for Computational Linguistics.

[Patankar et al., 2019] Patankar, A., Bose, J., and Khanna, H. (2019). A bias aware news recommendation system. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 232–238. IEEE.

[Pavan et al., 2020] Pavan, M. C., dos Santos, W. R., and Paraboni, I. (2020). Twitter Moral Stance Classification using Long Short-Term Memory Networks. In *9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319*, pages 636–647. Springer. 10.1007/978-3-030-61377-8_45.

[Pavan et al., 2023] Pavan, M. C., dos Santos, V. G., Lan, A. G. J., Martins, J.T., dos Santos, W. R., Deutsch, C., da Costa, P. B., Hsieh, F. C., and Paraboni, I. (2023). Morality classification in natural language text. *IEEE transactions on Affective Computing*, 14(1):857–863. 10.1109/TAFFC.2020.3034050.

[Pavan and Paraboni, 2022] Pavan, M. C., and Paraboni, I. (2022). Cross-target Stance Classification as Domain Adaptation. In *Advances in Computational Intelligence - MICAI 2022. LNAI 13612*, pages 15–25. Springer. 10.1007/978-3-031-19493-1_2.

[Pennebaker et al., 2015] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas, Austin, Texas, USA.

[Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

[Peters et al., 2017] Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proc. of ACL-2017*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

[Pizarro, 2019] Pizarro, J. (2019). Using N-grams to detect Bots on Twitter. In Cappellato, L., Ferro, N., Losada, D., and Müller, H., editors, *CLEF 2019 Labs and Workshops, Notebook Papers*, page 10. CEUR-WS.org.

[Polignano et al., 2020] Polignano, M., de Gemmis, M., and Semeraro, G. (2020). Contextualized BERT sentence embeddings for author profiling: The cost of performances. In *Computational Science and Its Applications (ICCSA)-2020, LNCS 12252*, pages 135–149, Cham. Springer.

[Potthast et al., 2018] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

[Preotiuc-Pietro et al., 2017] Preotiuc-Pietro, D., Liu, Y., Hopkins, D., and Ungar, L. (2017). Beyond binary labels: Political ideology prediction of twitter users. In *55th Annual Meeting of the Association for Computational Linguistics*, pages 729–740, Vancouver. Association for Computational Linguistics.

[Price and Hodge, 2020] Price, S. and Hodge, A. (2020). Celebrity profiling using twitter follower feeds. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CLEF and CEUR-WS.org.

[Rangel et al., 2020] Rangel, F., Rosso, P., Zaghouani, W., and Charfi, A. (2020). Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, page 1–21.

[Siddiqua et al., 2019] Siddiqua, U. A., Chy, A. N., and Aono, M. (2019). Tweet stance detection using an attention based neural ensemble model. In *NAACL-HLT 2019*, pages 1868–1873, Minneapolis, USA.

[Sidorov et al., 2014] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

[Singh and Singh, 2021] Singh, L. G. and Singh, S. R. (2021). Empirical study of sentiment analysis tools and techniques on societal topics. *Journal of Intelligent Information Systems*, 56:379–407.

[Srivastava et al., 2019] Srivastava, V., Gupta, A., Prakash, D., Sahoo, S. K., R.R, R., and Kim, Y. H. (2019). Vernon-fenwick at SemEval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Stefanov et al., 2020] Stefanov, P., Darwish, K., Atanasov, A., and Nakov, P. (2020). Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

[Stuart, 1955] Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3/4):412–416.

[Takahashi et al., 2018] Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., and Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. In *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, page 12, Avignon, France.

[Vijayaraghavan et al., 2017] Vijayaraghavan, P., Vosoughi, S., and Roy, D. (2017). Twitter demographic classification using deep multi-modal multi-task learning. In *55th Annual Meeting of the Association for Computational Linguistics*, pages 478–483, Vancouver. Association for Computational Linguistics.

[Zhang et al., 2018] Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.