Text Classification of Manifestos and COVID-19 Press Briefings using BERT and Convolutional Neural Networks

KAKIA CHATSIOU*

November 3, 2020

Abstract

We build a sentence-level political discourse classifier using existing human expert annotated corpora of political manifestos from the Manifestos Project (Volkens et al., 2020a) and applying them to a corpus of COVID-19 Press Briefings (Chatsiou, 2020). We use manually annotated political manifestos as training data to train a local topic Convolutional Neural Network (CNN) classifier; then apply it to the COVID-19 Press Briefings Corpus to automatically classify sentences in the test corpus. We report on a series of experiments with CNN trained on top of pre-trained embeddings for sentence-level classification tasks. We show that CNN combined with transformers like BERT outperforms CNN combined with other embeddings (Word2Vec, Glove, ELMo) and that it is possible to use a pre-trained classifier to conduct automatic classification on different political texts without additional training.

1 INTRODUCTION

A substantial share of citizen involvement in politics arises through written discourse especially in the digital space. Through advanced, novel communication strategies, the public can play their part in constructing a political agenda, which has led politicians to increasingly use social media and other types of digital broadcasting to communicate (compared to mainstream press and traditional print media). This is especially pertinent with crisis communication discourse and the recent COVID-19 pandemic has created a great opportunity to study how similar topics get communicated in different countries and the narrative choices made by government and public health officials at different levels of governance (international, national, regional). To aid fellow scholars with the systematic study of such a large and dynamic set of unstructured data, we set out to employ a text categorization classifier trained on similar domains (like existing manually annotated sentences from political manifestos) and use it to classify press briefings about the pandemic in a more effective and scalable way.

The main attraction behind using manually coded political manifestos (Volkens et al., 2020a) as training data is that the political science expert community have been manually collecting and annotating in a systematic way political parties' manifestos for years (since the 1960s) around the world in order to apply content analysis methods and to advance political science. They have subsequently been used as training data in semi-supervised domain-specific classification tasks with good results (Zirn et al., 2016; Nanni et al., 2016; Glavas, Nanni, and Ponzetto, 2017; Bilbao-Jayo and Almeida, 2018a; Bilbao-Jayo and Almeida, 2018b).

In this paper, we build variations of a CNN sentence-level political discourse classifier using existing annotated corpora of political manifestos from the Manifestos Project (Volkens et al., 2020a). We test different CNN and word embedding architectures on the already annotated (english language) sentences of the Manifestos Project Corpus. We then apply them to a corpus of COVID-19 Press Briefings (Chatsiou, 2020), a subset of which was manually annotated by political scholars for the purposes of this work.

The article is organised as follows: we first offer a brief overview of previous related work on the use of human expert annotated political manifestos for discourse classification. We then describe our framework including the training data used, data pre-processing performed and used architecture. We report on a series of experiments with CNN trained on top of pre-trained word vectors for sentence-level classification tasks. We conclude with evaluation of the BERT+CNN architecture against other combinations (Word2Vec+CNN, GloVe+CNN, ELMo+CNN) for both corpora. Experimental results show that a CNN classifier combined with transformers like BERT outperforms CNN combined with other non-context sensitive embeddings (Word2Vec, Glove, ELMo).

2 RELATED WORK

The use of NLP methods to analyse political texts is a well-established field within Political Science and Computational Social science more generally (Lazer et al., 2009; Grimmer and Stewart, 2013; Benoit, Laver, and Mikhaylov, 2009).

Researchers have used NLP methods to acccomplish various classification tasks, such as *political positioning* on a left to right continuum (Slapin and Proksch, 2008; Glavas, Nanni, and Ponzetto, 2017), *identification of political ideology differences from text* (Sim et al., 2013; Menini and Tonelli, 2016), *detection of political events* (Nanni, Ponzetto, and Dietz, 2017), or *detection of opinion and sentiment* (Young and Soroka, 2012).

2.1 Topic Classification of political discourse

A substantial body of recent work has focused on topic classification in political texts (Lauscher et al., 2016; Baturo, Dasandi, and Mikhaylov, 2017) some using supervised models (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Benoit et al., 2016; Glavas, Nanni, and Ponzetto, 2017), others using unsupervised models such as latent semantic analysis (Hofmann, 1999) and latent Dirichlet allocations (LDA) (Blei, 2003) or structural topic modelling (Lindstedt, 2019; Jacobs and Tschotschel, 2019)

Topic classification of domain-specific types of political text, such as *political manifestos* and their use as training data for unsupervised methods is receiving increased attention.

Zirn et al. (2016) independently trained three sentence-level classifiers - one for detecting the topic and two for detecting topic-shifts - and then combined their predictions in a global optimisation setting using a Markov Logic Network. Their experimental results show that the proposed global model achieves high classification performance and significantly outperforms the local sentence-level topic classifier.

Glavas, Nanni, and Ponzetto (2017) propose an approach for cross-lingual topical coding of sentences from electoral manifestos using as training data, manually coded manifestos with a total of 77500 sentences in four languages (English, French, German and Italian) (and CNNs with word embeddings) and inducing a joint multilingual embedding space. They report achieving better results than monolingual classifiers in English, French and Italian but worse results with their multilingual classifier than a monolingual classifier in German.

More recently, Bilbao-Jayo and Almeida (2018a) build a sentence classifier using multi-scale convolutional neural networks trained in seven different languages trained with sentences extracted from annotated parties' election manifestos. They use the full range of the domains defined by the manifestos project and they prove that enhancing the multi-scale convolutional neural networks with context data improves their classification. For a detailed discussion of different deep learning text classification-based models for text classification and their technical contributions, similarities, and strengths (Chatsiou and Mikhaylov, 2020; Minaee et al., 2020, see).

DOMAIN TRANSFER OF POLITICAL MANIFESTOS CLASSIFICATION TO OTHER POLITICAL TEXTS   Using annotated political manifestos as the training dataset for classifying other types of political texts is gaining traction in the literature, especially with the boost in performance of deep learning methods for text.

Nanni et al. (2016) used expert annotated political manifestos in English and speeches to train a local supervised topic classifier (SVM with a bag of words approach) that combines lexical with semantic textual similarity features at a sentence-level. A sub-part of the training set was annotated manually by human experts, and the rest was labelled automatically with the global optimisation step performed via a Markov Logic network presented in Zirn et al. (2016). The advantage of such a domain transfer approach is that no manual topic annotation on the rest of the corpus is needed. They then classify the speeches from the 2008, 2012 and 2016 US presidential campaign into the 7 domains defined by the Manifestos Project, without the need for additional topic annotation.

Bilbao-Jayo and Almeida (2018b) used annotated political manifestos in Spanish and the Regional Manifestos Project taxonomy Alonso, Gomez, and Cabeza (2013), to train a neural network sentence-level classifier (CNN) with Word2Vec word embeddings, also taking account the context of the phrase (like what was previously said and the political affiliation of the transmitter). They used this to analyse social media (twitter) data of the main Spanish political parties during 2015 and 2016 Spanish general elections without the need for additional manual coding of the twitter data.

This paper builds on this area of research presenting a comparison of a CNN classifier trained on the manifestos project annotations for English, but comparing more context-free (Word2Vec, Glove, ELMo) to context-sensitive (BERT) word embeddings. We then apply this to a corpus of daily press-briefings on the COVID-19 status by government and public health authorities.

2.2 Datasets

MANIFESTOS PROJECT CORPUS   The main attraction behind using manually coded political manifestos (Volkens et al., 2020a) as training data is that the political science community has been manually collecting and annotating in a systematic way political parties' manifestos for decades in a combined effort to create a resource for the systematic content analysis and to advance political science. The corpus is based on the work of the Manifesto Research Group (MRG) and the Comparative Manifestos (CMP) projects (Budge et al., 2001). Classification annotations are described in the *Manifesto Coding Handbook* which has evolved over the years, and provides information and instructions to the human annotators on how political parties' manifestos should be coded (latest version in Volkens et al. (2020b)). The handbook also includes a speficic set of policy areas or 'domains' (7) and subareas or 'subdomains' (56) which are available to annotators to use (see Figure 1).

For our training corpus, we use a subset of the corpus contating 115 English Manifestos with 86,500 annotated sentences. Table 1 shows the domain codes distribution in the dataset.

Domain 1: External Relations
101 Foreign Special Relationships: Positive
102 Foreign Special Relationships: Negative
103 Anti-Imperialism: Positive
104 Military: Positive
105 Military: Negative
106 Peace: Positive
107 Internationalism: Positive
108 European Integration: Positive
109 Internationalism: Negative
110 European Integration: Negative
Domain 2: Freedom and Democracy
201 Freedom and Human Rights: Positive
202 Democracy
203 Constitutionalism: Positive
204 Constitutionalism: Negative
Domain 3: Political System
301 Decentralization: Positive
302 Centralization: Positive
303 Governmental and Administrative Efficiency: Positive
304 Political Corruption: Negative
305 Political Authority: Positive
Domain 4: Economy
401 Free-Market Economy: Positive
402 Incentives: Positive
403 Market Regulation: Positive
404 Economic Planning: Positive
405 Corporatism: Positive
406 Protectionism: Positive
407 Protectionism: Negative
408 Economic Goals
409 Keynesian Demand Management: Positive
410 Economic Growth
411 Technology and Infrastructure: Positive
412 Controlled Economy: Positive
413 Nationalization: Positive
414 Economic Orthodoxy: Positive
415 Marxist Analysis: Positive
416 Anti-Growth Economy: Positive

Domain 5: Welfare and Quality of Life
501 Environmental Protection: Positive
502 Culture: Positive
503 Equality: Positive
504 Welfare State Expansion
505 Welfare State Limitation
506 Education Expansion
507 Education Limitation
Domain 6: Fabric of Society
601 National Way of Life: Positive
602 National Way of Life: Negative
603 Traditional Morality: Positive
604 Traditional Morality: Negative
605 Law and Order
606 Civic Mindedness: Positive
607 Multiculturalism: Positive
608 Multiculturalism: Negative
Domain 7: Social Groups
701 Labour Groups: Positive
702 Labour Groups: Negative
703 Agriculture and Farmers
704 Middle Class and Professional Groups: Positive
705 Minority Groups: Positive
706 Non-Economic Demographic Groups: Positive

000 No meaningful category applies

Figure 1: Manifestos Project annotation domains and subdomains used by human expert annotators (Volkens et al., 2020b), taken from Bilbao-Jayo and Almeida (2018a)

Domain 1 (External Relations)	6.5%
Domain 2 (Freedom and Democracy)	4.42%
Domain 3 (Political System)	10.64%
Domain 4 (Economy)	25.45%
Domain 5 (Welfare and Economy of Life)	31.77%
Domain 6 (Fabric of Society)	11.20%
Domain 7 (Social groups)	9.99%

Table 1: Domain Codes' distribution in the English subset of the Manifestos Corpus used for training the CNN classifier.

CORONAVIRUS (COVID–19) PRESS BRIEFINGS CORPUS The Coronavirus (COVID-19) Press Briefings Corpus is a collection of daily briefings on the COVID-19 status and policies from the UK and the World Health Organisation. The corpus is still in development, but we have selected example sentences from the UK and WHO which were the ones available.

During the peak of the pandemic, most countries around the world informed their citizens of the status of the pandemic (usually involving an update on the number of infection cases, number of deaths) and other policy-oriented decisions about dealing with the health crisis, such as advice about what to do to reduce the spread of the epidemic. At the moment the dataset includes briefings covering announcements between March 2020 and August 2020 from the UK (England, Scotland, Wales, Northern Ire-land) and the World Health Organisation (WHO) as follows:

- UK - England: Daily Press Briefings by UK Government between 12 March 2020 – 30 Au-gust 2020 (150 briefings in total, 13,050 sentences)

- UK - Scotland: Daily Press Briefings by Scottish Government between 3 March 2020 - 30 August 2020 (167 briefings in total, 14,529 sentences)

- UK - Wales: Daily Press Briefings by Welsh Government between 23 March 2020 - 30 August 2020 (146 briefings in total, 12,702 sentences)

- UK - Northern Ireland: Daily Press Briefings by N. Ireland Assembly between 23 March 2020 - 30 August 2020 (130 briefings in total, 11,310 sentences)

- World Health Organisation: Press Briefings occurring usually every 2 days between 22 January 2020 - 30 August 2020 (124 briefings in total, 10,788 sentences)

3 NEURAL NETWORK ARCHITECTURE FOR TOPIC CLASSIFICATION

CONTINUOUS SEMANTIC TEXT REPRESENTATIONS (EMBEDDINGS) We obtained pre-trained context-free word embeddings for English (Word2Vec: (Mikolov et al., 2013), GloVe: (Pennington, Socher, and Manning, 2014)). Word2Vec uses a shallow neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence.

Word2Vec uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence.

GloVe is an unsupervised learning model for obtaining vector representations for words. This is achieved by mapping words into a meaningful space where the distance between words is related to semantic similarity. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

We also obtained word embeddings for more context-sensitive word embeddings, namely ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019).

ELMo is a deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). These word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. They can be easily added to existing models and significantly improve the state of the art across a broad range of challenging NLP problems, including question answering, textual entailment and sentiment analysis.

BERT is a deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. It includes a variant that uses the English Wikipedia with 2.5 million words. Unlike previous context-free models, which generate a single word embedding representation for each word in the vocabulary, BERT takes into account the context for each occurrence of a given word, providing a contextualised embedding that is different for each sentence.

CONVOLUTIONAL NEURAL NETWORKS Since Kim (2014)'s paper outlining the idea of using CNNs for text classification (traditionally used for recognising visual patterns from images), CNNs have achieved very good performance in several text classification tasks (Poria, Cambria, and Gelbukh, 2015; Bilbao-Jayo and Almeida, 2018b). CNNs involve convolutional operations of moving frames or windows (filter sizes) which analyse and reduce different overlapping regions in a matrix, to extract different features. The ability to also bootstrap word embeddings in this type of neural network make it an excellent candidate for extracting knowledge and classifying non-annotated texts.

Experiment	Accuracy	F1
M1	65.79%	61.11
M2	68.15%	64.93
M3	72.84%	68.42
M4	87.52%	74.68

Table 2: Domain results of all models using political manifestos

We therefore set up 4 variations of the CNN classifier M1, M2, M3, M4 as follows:

1. Word vectors of the training dataset sentences are created using one of the following word embeddings: Word2Vec (M1), GloVe (M2), ELMo (M3) and BERT (M4). Sentences are fed as sequences of words, then mapped to indexes, then a sequence of word vectors. We have chosen 300 as the word vector size and 60 x d for the space where the convolution operations can be performed.

2. Vectors are fed to the neural network (CNN). we then perform convolution operations with 100 filters and three different filter sizes (2 x d, 3 x d, and 4 x d). We reduce the dimensionality of the feature maps generated by each group of filters using *1-max-pooling*, which are consequently concatenated (Boureau, Ponce, and LeCun, 2010).

3. A dropout rate of 0.5 is applied (Srivastava et al., 2014) as regularisation to prevent overfitting.

4. The layer with *softmax* computes the probability distribution over the labels.

5. We perform optimization using the Adam optimiser with the parameters of the original manuscript (Kingma and Ba, 2017).

Note that this is a sentence-level topic classifier basing its predictions by taking into account only the information local within the sentence.

4 EVALUATION

For our training corpus, we use a subset of the corpus containing 115 English Manifestos with 86,500 annotated sentences. Table 1 shows the domain codes distribution in the dataset. In order to evaluate the different architectures, we divided our training dataset in 2 different subsets: training and validation sets (85%) and test set (15%). Typically, we have used a validation set (or development test set) separate from the test set, to ensure correct evaluation and that our model(s) do not overfit, thus ensuring how each domain is classified and that the evaluation is robust.

We performed 4 experiments, one for each combination of CNN and word embeddings:

- M1: CNN with Word2Vec

- M2: CNN with GloVe

- M3: CNN with ELMo

- M4: CNN with BERT

As shown in Table 2, the performance of the classifier improves when more context-sensitive word embeddings are used. Using BERT with CNN (M4) seems to provide a substantial increase in accuracy and F1, whereas using ELMo performs very well as well.

Domain 1 (External Relations)	0.74%
Domain 2 (Freedom and Democracy)	0.47%
Domain 3 (Political System)	11.58%
Domain 4 (Economy)	33.99%
Domain 5 (Welfare and Economy of Life)	34.62%
Domain 6 (Fabric of Society)	15.02%
Domain 7 (Social groups)	3.58%

Table 3: Manifest Project Domain Codes' distribution in the manually annotated subset of the COVID-19 corpus.

Experiment	Accuracy	F1
M1	50.65%	48.62
M2	54.18%	48.82
M3	60.74%	57.07
M4	68.65%	64.58

Table 4: Domain results of all models using COVID-19 Press briefings corpus

APPLYING THE MODELS ON THE COVID–19 CORPUS   We also tested the performance of the same different pre-trained models on the COVID-19 corpus. We asked two political science scholars to annotate a subset of 20 press briefings (4 of each set), using the 7 domains of the Manifestos Project. This resulting in a dataset of 1740 manually annotated sentences, with domain distrubution as in Table 3. Note that the pre-trained models have been trained using the annotated manifestos from the Manifestos Project, without any additional training on the press briefings corpus sentences.

As shown in Table 4, the performance of the classifier improves when more context-sensitive word embeddings are used in the context of the COVID-19 press briefings corpus as well. Using BERT with CNN (M4) seems to provide a substantial increase in accuracy and F1, whereas using ELMo performs very well as well. As expected there is some loss of accuracy, as we are porting the classifier to a slightly different domain of political text (from manifestos to press briefings).

5   CONCLUSION

In this paper, we built a sentence-level political discourse classifier using existing human expert annotated corpora of English political manifestos from the Manifestos Project (Volkens et al., 2020a). We tested the accuracy and performance of a neural networks classifier (CNN) using different word embeddings as part of the word to vector mapping and we showed that sentence-level CNN classifiers combined with transformers like BERT outperform models with other embeddings (Word2Vec, Glove, ELMo). We then applied the same pre-trained models to a different set of text, the COVID-19 Press Briefings Corpus. We observe similar patterns in the accuracy and F1 scores, and additionally show that it is possible to use a pre-trained classifier to conduct automatic classification on different political texts without additional training

In the future, we aim to conduct similar experiments also considering the 'sub-domain' categories of the Manifesto Corpus Annotations. We also look forward to re-running these experiments for other languages in the Manifestos project, testing the language-agnostic advantage of word embeddings and see if we could obtain different results.

6 ETHICS STATEMENT

This paper follows the AAAI Publications Ethics and Malpractice Statement and the AAAI Code of Professional Conduct. We use publicly available text data to ensure transparency and reproducibility of the research. Additionally, all code will be available as open source code (on github.com) at the end of the submission and reviewing process.

The paper suggests ways to automatically extract topic information from political discourse texts, employing deep learning methods which are usually associated with artificial intelligence and ethical considerations around them. We do not envisage any ethical, social and legal considerations arising from the work outlined in this study, such as impact of AI on humans, on economic growth, on inequality, amplifying bias or undermining political stability or other issues described in recent reports on ethics in AI (see for example (Bird et al., 2020)).

ACKNOWLEDGEMENTS

REFERENCES

[1] Thomas Hofmann. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '99. New York, NY, USA: Association for Computing Machinery, Aug. 1999, pp. 50–57. ISBN: 978-1-58113-096-6. DOI: 10.1145/312624.312649. URL: https://doi.org/10.1145/312624.312649 (visited on 09/10/2020).

[2] Ian Budge et al., eds. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. Oxford, New York: Oxford University Press, Aug. 2001. ISBN: 978-0-19-924400-3.

[3] David M Blei. "Latent Dirichlet Allocation". en. In: *a deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus* 3 (2003), p. 30.

[4] Stephen Purpura and Dustin Hillard. "Automated classification of congressional legislation". In: *Proceedings of the 2006 international conference on Digital government research*. dg.o '06. San Diego, California, USA: Digital Government Society of North America, May 2006, pp. 219–225. DOI: 10.1145/1146598.1146660. URL: https://doi.org/10.1145/1146598.1146660 (visited on 09/10/2020).

[5] Jonathan B. Slapin and Sven-Oliver Proksch. "A Scaling Model for Estimating Time-Series Party Positions from Texts". en. In: *American Journal of Political Science* 52.3 (2008), pp. 705–722. ISSN: 1540-5907. DOI: 10.1111/j.1540-5907.2008.00338.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2008.00338.x (visited on 09/10/2020).

[6] Kenneth Benoit, Michael Laver, and Slava Mikhaylov. "Treating words as data with error: Uncertainty in text statements of policy positions". In: *American Journal of Political Science* 53.2 (2009). Publisher: Wiley Online Library, pp. 495–513. ISSN: 0092-5853.

[7] David Lazer et al. "Life in the network: the coming age of computational social science". In: *Science (New York, N.Y.)* 323.5915 (Feb. 2009), pp. 721–723. ISSN: 0036-8075. DOI: 10.1126/science.1167742. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/ (visited on 09/09/2020).

[8] Brandon M. Stewart and Yuri M. Zhukov. "Use of force and civil–military relations in Russia: an automated content analysis". en. In: *Small Wars & Insurgencies* 20.2 (June 2009), pp. 319–343. ISSN: 0959-2318, 1743-9558. DOI: 10.1080/09592310902975455. URL: http://www.tandfonline.com/doi/abs/10.1080/09592310902975455 (visited on 09/10/2020).

[9] Y.-Lan Boureau, J. Ponce, and Y. LeCun. "A Theoretical Analysis of Feature Pooling in Visual Recognition". In: *ICML*. 2010.

[10] Lori Young and Stuart Soroka. "Affective News: The Automated Coding of Sentiment in Political Texts". In: *Political Communication* 29.2 (Apr. 2012). Publisher: Routledge, pp. 205–231. ISSN: 1058-4609. DOI: 10.1080/10584609.2012.671234. URL: https://doi.org/10.1080/10584609.2012.671234 (visited on 09/10/2020).

[11] Sonia Alonso, Braulio Gomez, and Laura Cabeza. "Measuring Centre-Periphery Preferences: The Regional Manifestos Project". In: *Regional & Federal Studies* 23.2 (May 2013). Publisher: Routledge, pp. 189–211. ISSN: 1359-7566. DOI: 10.1080/13597566.2012.754351. URL: https://doi.org/10.1080/13597566.2012.754351 (visited on 09/10/2020).

[12] Justin Grimmer and Brandon M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". en. In: *Political Analysis* 21.3 (2013). Publisher: Cambridge University Press, pp. 267–297. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mps028. URL: https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20 (visited on 09/09/2020).

[13] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv:1301.3781 [cs]* (Sept. 2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781 (visited on 09/10/2020).

[14] Yanchuan Sim et al. "Measuring Ideological Proportions in Political Speeches". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 91–101. URL: https://www.aclweb.org/anthology/D13-1010 (visited on 09/10/2020).

[15] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *arXiv:1408.5882 [cs]* (Aug. 2014). arXiv: 1408.5882. URL: http://arxiv.org/abs/1408.5882 (visited on 10/25/2018).

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation". In: *EMNLP*. 2014.

[17] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". en. In: *Journal of Machine Learning Research* 15 (2014), p. 30.

[18] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 2539–2544. DOI: 10.18653/v1/D15-1303. URL: https://www.aclweb.org/anthology/D15-1303 (visited on 09/10/2020).

[19] Kenneth Benoit et al. "Crowd-sourced text analysis: Reproducible and agile production of political data". In: *American Political Science Review* 110.2 (2016), pp. 278–295.

[20] Anne Lauscher et al. "Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability". en. In: *IJCoL - Italian Journal of Computational Linguistics* 2.2 (Dec. 2016). URL: https://hal.archives-ouvertes.fr/hal-01483333 (visited on 09/10/2020).

[21] Stefano Menini and Sara Tonelli. "Agreement and Disagreement: Comparison of Points of View in the Political Domain". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2461–2470. URL: https://www.aclweb.org/anthology/C16-1232 (visited on 09/10/2020).

[22] Federico Nanni et al. "TopFish: Topic-Based Analysis of Political Position in US Electoral Campaigns". In: *PolText 2016* 14 (2016), p. 61.

[23] Cacilia Zirn et al. "Classifying Topics and Detecting Topic Shifts in Political Manifestos". en. In: *University of Mannheim working papers* (2016), p. 6. URL: https://madoc.bib.uni-mannheim.de/41552/1/classyman.pdf.

[24] Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. "Understanding state preferences with text as data: Introducing the UN General Debate corpus". en. In: *Research & Politics* 4.2 (Apr. 2017). Publisher: SAGE Publications Ltd, p. 2053168017712821. ISSN: 2053-1680. DOI: 10.1177/2053168017712821. URL: https://doi.org/10.1177/2053168017712821 (visited on 09/10/2020).

[25] Goran Glavas, Federico Nanni, and Simone Paolo Ponzetto. "Cross-Lingual Classification of Topics in Political Texts". en. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 42–46. DOI: 10.18653/v1/W17-2906. URL: http://aclweb.org/anthology/W17-2906 (visited on 09/09/2020).

[26] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs]* (Jan. 2017). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on 09/10/2020).

[27] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. "Building Entity-Centric Event Collections". In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. June 2017, pp. 1–10. DOI: 10.1109/JCDL.2017.7991574.

[28] Aritz Bilbao-Jayo and Aitor Almeida. "Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data". en. In: *International Journal of Distributed Sensor Networks* 14.11 (Nov. 2018), p. 1550147718811827. ISSN: 1550-1477. DOI: 10.1177/1550147718811827. URL: https://doi.org/10.1177/1550147718811827 (visited on 11/30/2018).

[29] Aritz Bilbao-Jayo and Aitor Almeida. "Political discourse classification in social networks using context sensitive convolutional neural networks". en. In: *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 76–85. DOI: 10.18653/v1/W18-3513. URL: http://aclweb.org/anthology/W18-3513 (visited on 09/10/2020).

[30] Matthew E. Peters et al. "Deep contextualized word representations". In: *arXiv:1802.05365 [cs]* (Mar. 2018). arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365 (visited on 09/10/2020).

[31] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (May 2019). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 09/10/2020).

[32] Thomas Jacobs and Robin Tschotschel. "Topic models meet discourse analysis: a quantitative tool for a qualitative approach". In: *International Journal of Social Research Methodology* 22.5 (Sept. 2019). Publisher: Routledge, pp. 469–485. ISSN: 1364-5579. DOI: 10.1080/13645579.2019.1576317. URL: https://doi.org/10.1080/13645579.2019.1576317 (visited on 09/10/2020).

[33] Nathan C. Lindstedt. "Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017". en. In: *Social Currents* 6.4 (Aug. 2019). Publisher: SAGE Publications Inc, pp. 307–318. ISSN: 2329-4965. DOI: 10.1177/2329496519846505. URL: https://doi.org/10.1177/2329496519846505.

[34] Eleanor Bird et al. *The ethics of artificial intelligence: issues and initiatives.* en. Study -European Parliament's Panel for the Future of Science and Technology PE 634.452. LU: Publications Office, 2020. URL: https://data.europa.eu/doi/10.2861/6644 (visited on 09/09/2020).

[35] Kakia Chatsiou. *COVID-19 Press Briefings Corpus*. eng. type: dataset. June 2020. DOI: 10.5281/zenodo.3872417. URL: https://zenodo.org/record/3872417#.X1lTo4vTWUk (visited on 09/09/2020).

[36] Kakia Chatsiou and Slava J. Mikhaylov. "Deep Learning for Political Science". In: *ArXiv* (2020). DOI: 10.4135/9781526486387.n58.

[37] Shervin Minaee et al. "Deep Learning Based Text Classification: A Comprehensive Review". In: *arXiv:2004.03705 [cs, stat]* (Apr. 2020). arXiv: 2004.03705. URL: http://arxiv.org/abs/2004.03705 (visited on 09/09/2020).

[38] Andrea Volkens et al. *Manifesto Project Dataset*. en. Version Number: 2020a type: dataset. 2020. DOI: 10.25522/MANIFESTO.MPDS.2020A. URL: https://manifesto-project.wzb.eu/doi/manifesto.mpds.2020a (visited on 09/10/2020).

[39] Andrea Volkens et al. *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2020a*. 2020. DOI: 10.25522/manifesto.mpds.2020a. URL: https://doi.org/10.25522/manifesto.mpds.2020a.

CONTENTS

LIST OF FIGURES

LIST OF TABLES