Automated Claim Detection for Fact-checking: A Case Study using Norwegian Pre-trained Language Models

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymousest Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

In this work, we investigate to what extent pre-trained language models can be used for automated claim detection for fact-checking in a low resource setting. We explore this idea by fine-tuning four Norwegian pre-trained language models to perform the binary classification task of determining if a claim should be *discarded* or *upheld* to be further processed by human fact-checkers. We conduct a set of experiments to compare the performance of the language models, and provide a simple baseline model using SVM with tf-idf features. Since we are focusing on claim detection, the recall score for the *upheld* class is to be emphasized over other performance measures. Our experiments indicate that the four language models are superior to the baseline system in terms of F1. The baseline model results in the highest precision. However, the two Norwegian models, NorBERT2 and NB-BERT_large, give respectively superior F1 and recall values. We argue that large language models could be successfully employed to solve the automated claim detection problem. The choice of the model depends on the desired end-goal. Moreover, error analysis shows that language models are generally less sensitive to the changes in claim length and source than the SVM model.

1 Introduction

With the growing concerns about misinformation, fact-checking has become an essential part of journalism. To mitigate the time and the human burden of fact-checking and to allow for more fact-checked articles, automated fact-checking (AFC) systems have been developed (Guo et al., 2022; Zeng et al., 2021; Lazarski et al., 2021). To approach automated fact-checking, three basic tasks are defined in the pipeline: claim detection, evidence retrieval, and claim verification. Claim detection refers to monitoring social media and political sources for identifying statements worth checking. The subsequent components retrieve reliable documents for debunking the detected claims and generate a verdict. Several tools have been developed to automate these tasks to meet the expectations of the human fact-checkers[1]. According to the studies on the user needs of fact-checkers, claim detection receives the highest preference among other AFC tools (Graves, 2018; Dierickx et al., 2022). Automated claim detection is a classification problem, where models are trained on sentences parsed from text documents and labelled by humans according to their check-worthiness (Hassan et al., 2017a).

In this work, we explore how well Norwegian pre-trained language models (LMs) perform on the task of automated claim detection. This is, to the best of our knowledge, the first attempt at automated claim detection for Norwegian using LMs. Fine-tuning LMs for the task of automated claim detection is not novel (Cheema et al., 2020; Zhuang et al., 2021; Shaar et al., 2021). However, this has never been done on Norwegian, and we believe that our insights into which errors these models do compared to simple baselines is a valuable contribution. Our research questions are:

- How well do Norwegian LMs perform on the task of automated claim detection compared to a simple SVM baseline?
- Which aspects of claim detection do these LMs still struggle with?

[1] https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html

To address these questions, we first fine-tune each model on a small dataset from a Norwegian non-profit fact-checking organization, comprising claims manually annotated with labels reflecting their check-worthiness. Then we manually analyse the misclassifications of each model and provide an error analysis.

The imperial contributions of this work to the NLP research, has important societal implications, we believe. The case we study here sheds lights on the future directions of claim detection tools for fact-checking based on pre-trained language models for low resource languages. This would contribute to the fight against dis/misinformation by scaling and speeding up the fact-checking process.

The rest of the paper is organized as follows. In Section 2 we give an overview of previous work on automated claim detection. Section 3 describes the dataset and our experimental setup. We present and discuss our results and provide an error analysis in Section 4. Finally, we summarize our main findings, and discuss possible future works in Section 5.

2 Background

Automated claim detection for fact-checking does not have a long history, but it has turned to one of the attractive fields of research in NLP (Hassan et al., 2015; Gencheva et al., 2017; Beltrán et al., 2021; Cheema et al., 2020; Shaar et al., 2021). One of the first studies on claim detection for AFC is initiated as part of the ClaimBuster project Hassan et al. (2017b). Their initial claim detection system was based on a set of features (sentiment, word count, part of speech (PoS) tags and named entity (NE)) followed by a feature selection and a traditional classifier namely Naive Bayes, SVM and Random Forest(Hassan et al., 2015). Claim detection has also been addressed in languages other than English. ClaimRank is a claim detection system that supports both Arabic and English (Gencheva et al., 2017). A comprehensive set of features such as tf-idf, assertiveness, subjectivity, word embeddings, etc. are added to the ClaimBuster features and are fed to a two-layered neural network classifier (Gencheva et al., 2017).

In recent years, employment of pre-trained LM models in automated claim detection has been considered by the researchers (Cheema et al., 2020; Shaar et al., 2021; Beltrán et al., 2021). Several instances of these works are presented in the check-worthiness detection sub-tasks in CLEF CheckThat! editions (introduced in 2018 and ongoing) (Shaar et al., 2021; Nakov et al., 2022). CheckThat! provides data sets in different languages (English, Turkish, Arabic, Bulgarian, and Spanish) for the claim detection task on Twitter and political debates. The teams participated in this task have proposed classifier models mostly based on LMs. For instance, the top-ranked teams in CheckThat! 2020 used BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021) with enhanced generalization capability (Williams et al., 2020) to detect check-worthy Tweets. For the task of detecting claims in political debates, the baseline BiLSTM (Schuster and Paliwal, 1997) model with GloVe embedding outperforms the LM-based systems (Martinez-Rico et al., 2020). ClaimHunter (Beltrán et al., 2021) is another BERT-based claim detection system that leverages XLM-RoBERTa [2], a multilingual version of RoBERTa. It has been proved that the proposed model is superior to the classical baseline models NNLM+LR (Neural-Net Language Models embedding+Logistic Regression) and tf-idf+SVM.

To deal with the problem of small training data for LMs, data augmentation is employed. Claim detection from Twitter has been approached by generating synthetic check-worthy claims with lexical substitutions using BERT-based embeddings (Shaar et al., 2021). This approach improves the performance of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) classification models (Shaar et al., 2021). It has also been shown that BERTweet (Nguyen et al., 2020) model fine-tuned on claims normalized and augmented by substitutions using WordNet surpasses a reference n-gram model (Shaar et al., 2021).

3 Experiments

3.1 Data set

The data set is provided to us by Faktisk [3], a non-profit fact-checking organization and independent newsroom in Norway. The data set comprises 4885 claims in Norwegian collected from social debates and public discourses from 04.03.2018 to 20.05.2022. Each claim in the dataset is provided with its respective source. These cover a selection

[2] https://huggingface.co/xlm-roberta-large

[3] https://www.faktisk.no/om-oss

Class	Data Set Label	Description	#Claims
Discarded	Discarded	The claim has simply been discarded, there is no need for further investigation.	2810
Upheld	Pre-checked and rejected	Some preliminary work has been done to see if the claim is worth fact-checking, with a negative result.	372
	Pre-checking	Preliminary work to see if the claim is worth fact-checking has been started.	336
	Published	The fact-check about the claim has been published.	297
	Suspended	The claim will be taken up for consideration at a later time, and pre-checking or fact-checking will start then.	194
	Checking	A fact-check about the claim is in progress.	107

Table 1: Distribution of class labels and related labels in our dataset.

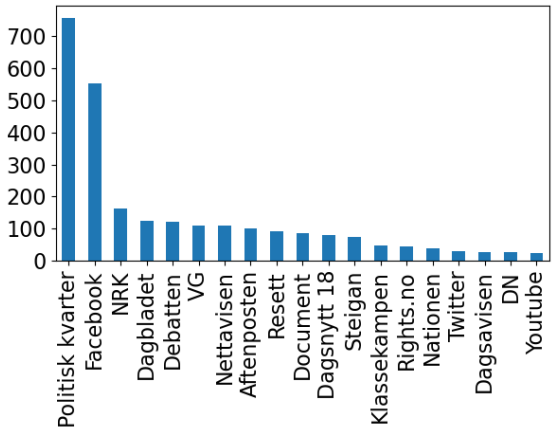


Figure 1: Most frequent sources of claims in our dataset.

of Norwegian newspapers (Dagbladet, VG, Nettavisen, Aftenposten, Klassekampen, Nationen, Dagsavisen, DN), alternative news outlets (Resett, Steigan, Document), think tank (Rights.no), the Norwegian Broadcasting Corporation (NRK), social media (Facebook, Twitter, YouTube), and TV/Radio (news) shows (Dagsnytt18, Politisk kvarter, Debatten). The alternative news outlets and the think tank are generally considered radical and controversial. The distribution of the occurrence of these sources can be seen in Figure 1.

A label is assigned to each claim, which refers to the actions taken by human fact-checkers. This data set has been labelled as part of the daily routine in the organization *Faktisk.no* and is neither hand-crafted nor crowdsourced for training LMs. Thus, it resembles a real world problem. The data set labels are {*Discarded, Checked and rejected, Pre-checked, Published, Suspended, Checked, Facebook*}. After removing the missing values, the rare samples with label 'Facebook' (only nine claims), and the short claims with less than five words, we end up with 4116 claims across six different labels. For our purposes, we aim to focus on class labels specified as whether a claim is worth being considered for further processing or if it should be discarded. We therefore define a binary classification task with the labels *Discarded* and *Upheld*; where the *Discarded* class refers to claims with the same label (Discarded) in the data set, and the *Upheld* class includes the claims originally labelled as *Pre-checked and rejected, Pre-checkeing, Published, Suspended, or Checking*. A brief explanation of these labels as well as the mapping of the original labels to the binary class labels is given in Table 1. The number of claims in each category is also presented. There are 2810 claims in the first class and 1306 claims in the second class. The average and the maximum length of claims in these samples are equal to 16 and 107 words, respectively.

3.2 Experimental setup

Pre-trained language models We fine-tune four Norwegian LMs to perform the binary classification task of claim detection. Norwegian has two official written standards: Bokmål and

Nynorsk, and the four models are trained on data in both written forms. These are:

- **NorBERT** (Kutuzov et al., 2021): trained on the Norwegian newspaper corpus[4], and Norwegian Wikipedia, with a vocabulary of about two billion word tokens.
- **NorBERT2**[5]: trained on the non-copyrighted subset of the Norwegian Colossal Corpus (NCC)[6] and the Norwegian subset of the C4 web-crawled corpus (Xue et al., 2021). The size of the vocabulary is about 15 billion word tokens.
- **NB-BERT_base** (Kummervold et al., 2021): trained on the full NCC, and follows the architecture of the BERT cased multilingual model (Devlin et al., 2019). This model is bigger than the two previous ones, and comprises around 18.5 billion word tokens.
- **NB-BERT_Large**[7]: trained on NCC, and follows the architecture of the BERT-large uncased model. This model is bigger and trained on more data (from the same sources) than it's base-form NB-BERT_base.

Training details The baseline model is an SVM classifier with tf-idf features (Jones, 2004), implemented using Scikit-learn library[8]. To split the data, stratified sampling based on the original data set labels is employed to ensure the distributions of the real world label noise is consistent among the splits. The ratio of the train, validation, and test sets is $70\% - 20\% - 10\%$ respectively. The validation set is employed to tune the hyperparameters of the model. The hyperparameters of the best model are `(C=100, gamma=0.1, kernel='rbf')`. It should be noted that the preliminary experiments revealed that the baseline model performs extremely poor on the minority class, *uphold*. To make a fair comparison between the baseline model and the BERT-based models, we have examined five different random states for splitting the data and chosen the one in favour of the baseline model. Furthermore, we ensured that the distribution of the length of claims in the test split is consistent with the whole data set (See Figure 3 (a)). The same split is used for fine-tuning the pre-trained LMs. The selected split results in the highest F1 for the *upheld* class by the baseline model among the five examined random splits.

Hyperparameter	Value
batch_size	16
init_lr	2e-5
end_lr	0
warmup_proportion	0.1
num_epochs	5
max_seq_length	64

Table 2: Hyperparameter configuration of the four Norwegian language models.

Model	t (s)	Prec.	Rec.	F1
tf-idf+SVM	2	**0.440**	0.168	0.243
NorBERT	44	0.328	0.626	0.430
NorBERT2	45	0.401	0.588	**0.477**
NB-BERT_base	48	0.358	0.336	0.345
NB-BERT_large	103	0.320	**0.740**	0.447

Table 3: Training time and claim detection results for different models.

The claim detection models are fine-tuned using a TensorFlow-based model for sequence classification [9] from the HuggingFace `transformers` library [10]. Bert-based model transformer have a sequence classification head, i.e. a linear layer on top. We use the same train, validation, and test splits as the baseline model and the validation set is deployed to return the best model after five epochs. All experiments are repeated for five times and the best run in terms of F1 is reported. All models are fine-tuned with Adam optimizer (Kingma and Ba, 2014). The other hyperparameter configurations are identical for all the four models, and can be seen in Tabel 2.

[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/

[5] https://huggingface.co/ltgoslo/norbert2

[6] https://github.com/NbAiLab/notram/blob/master/guides/corpus_description.md

[7] https://huggingface.co/NbAiLab/nb-bert-large

[8] https://scikit-learn.org/stable/

[9] TFAutoModelForSequenceClassification

[10] https://huggingface.co/docs/transformers/index

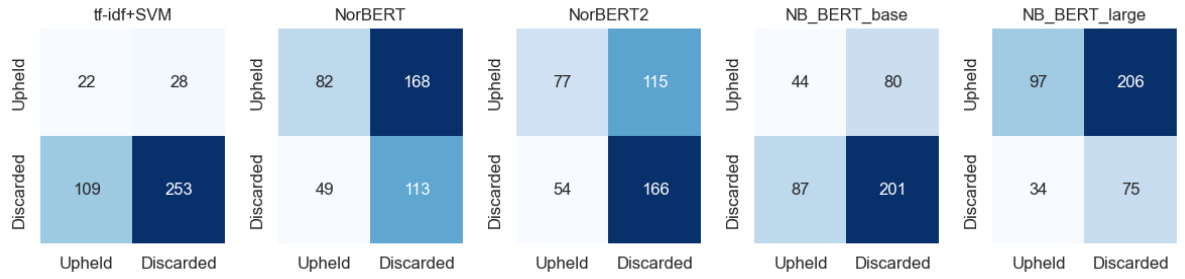
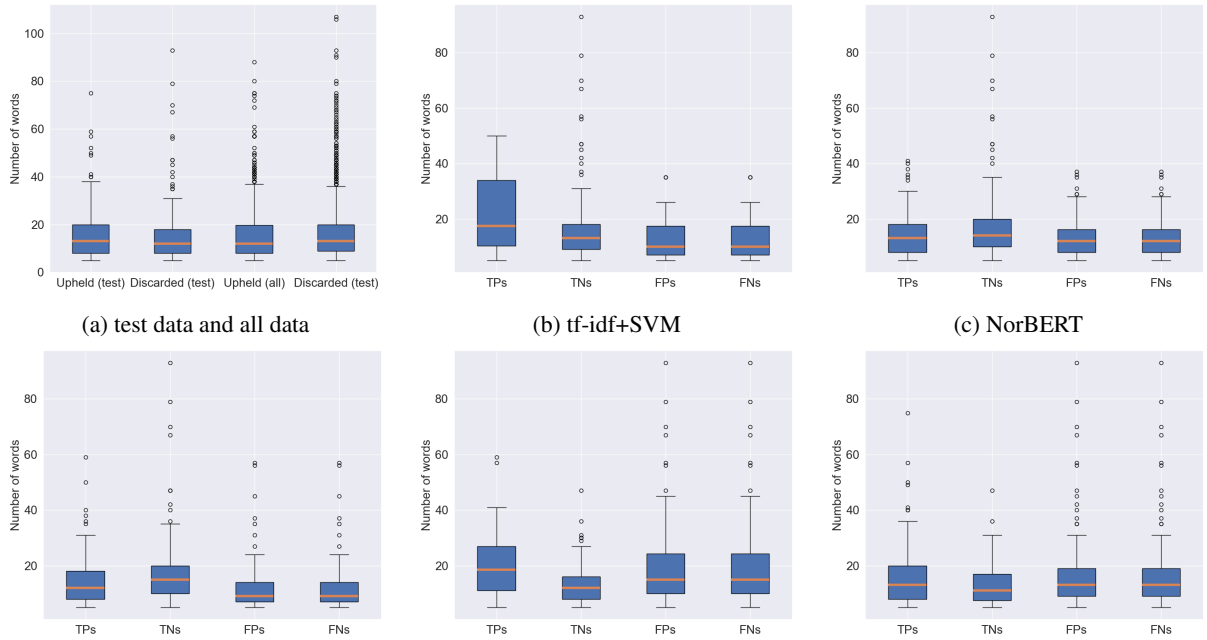


Figure 2: Confusion matrix of model predictions for different models.



(a) test data and all data (b) tf-idf+SVM (c) NorBERT

(d) NorBERT2 (e) NB-BERT_base (f) NB-BERT_large

Figure 3: The number of words in the claims across true and false predictions for different models.

4 Results and discussion

4.1 Classification performance

The performance of the classification models on the test data are measured in terms of precision, recall, and F1. The upheld class is treated as the positive class. It should be noted that in automated claim detection, overlooked important claims have a higher cost than misclassified unimportant claims. In other words, the recall score of the *upheld* class should be given particular emphasis.

Table 3 presents the results for the baseline model and the four fine-tuned language models. The highest score in each column is shown in bold. For the case of precision, the baseline system outperforms the LMs, but recall and F1 are extremely poor. It is noticeable how all the four LMs are superior to the baseline system in terms of F1, with NorBERT standing on the top. Another significant reflection of the results is NB-BERT_large's superior performance in terms of recall. The training time (in seconds) is also given in the table. We run the experiments on a PC with an AMD Ryzen 7 5800X 8 Core Processor, an Nvidia GeForce RTX-3080 GPU with 10 GB graphics memory and 32 GB of RAM. The largest model, NB-BERT_large, requires twice as much training time compared to the other three models.

4.2 Error analysis

To get insights on the errors made by our models, confusion matrices of the predictions are plotted in Figure 2. If we focus on one of the classes in terms of precision-recall, the baseline model and NB-BERT_large are the best models. These mod-

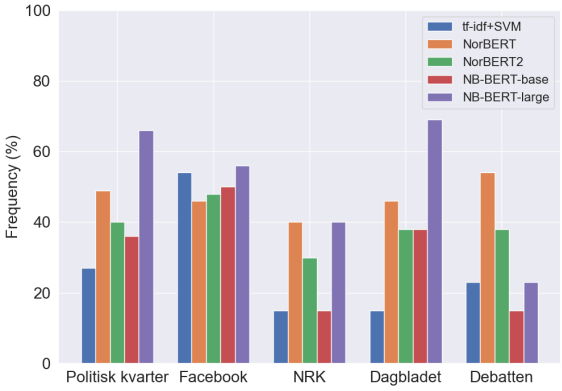


Figure 4: The percentage of the incorrectly classified claims from the five most frequent sources.

els appear to learn one of the classes better, having fewer errors on that class. For example, NB-BERT_large has learnt to correctly classify more instances of the upheld class. But the fact that it also classifies a large proportion of the claims from the discarded class as upheld shows that it simply has overfitted on the upheld class. This observation seems to be true for the SVM model (overfitted to the majority class) as well, and to some extent can be said about NB-BERT_base.

NorBERT and NorBERT2 seem to actually learn a more decent representation of the label distribution. While NorBERT exhibits some similarities with the previous models, by mostly classifying claims as one class rather than the other (in this case the discarded class), NorBERT2 seems to have a more balanced representation between the classes. It is the only model that is able to identify both classes to a certain degree, even if it still confuses many of the upheld claims as discarded claims. If we were to select a model that works fairly good on both classes, NorBERT2 would be the natural choice.

Further analysis is conducted on the length of the claims with respect to the model predictions for true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). Figure 3 illustrates the box and whisker plots of the number of words in each of these groups. In Figure 3 (a), the number of words in the upheld and discarded class are shown for the test set and the whole data set. The length of the claims in the discarded class appears to be slightly larger than the upheld class. However, the quartiles and the median length are very close and thus length is not a significant discriminative feature. For the baseline model, length plays an important role in the model behaviour, though. The SVM model correctly classifies the longer claims from the upheld class and the shorter claims from the discarded class. Among the LMs, NorBERT and NB-BERT_large are less sensitive to the length of the claims, as inferred from the similar statistics for true and false predictions. The figure also indicates that NorBERT2 suffers when predicting shorter claims, while NB-BERT_base deteriorates for longer claims from the discarded class.

We also looked into the sources of the incorrectly classified claims for different models. The five most frequent sources in the data set, namely, 'Politisk kvarter', 'Facebook', 'NRK', 'Dagbladet', and 'Debatten' are considered. The percentage of the claims with false predictions from each source are shown in Figure 1. One interesting observation is claims from 'Facebook' are relatively difficult for all the models, while predicting the ones from 'NRK' seem to be more straightforward. This could be due to the differences in the writing styles in an official broadcasting organization and a social media platform. It is notable that the patterns for NorBERT and NorBERT2 are relatively similar across different sources. NB-BERT_base and NB-BERT_large are more sensitive to the source of the claims.

5 Conclusion

In this work, we conduct a case study using Norwegian pre-trained LMs for the task of automated claim detection. Four existing Norwegian models in addition to an SVM baseline system are examined and compared using a claim detection data set that resembles a real world problem. The results show that language models outperform the baseline system. Different models can be selected for different purposes. If the overall performance is to be prioritized, the NorBERT2 model is the best performing. If the recall is the focus, then the biggest model is to be selected.

Most of our observations can also be due to the differences between the LMs. The behaviour of our models can be due to model architecture, training procedures, and the datasets they were originally trained on. We also show how the length and the source of the claim plays a role in prediction patterns. We believe that there is more that can be uncovered from the behaviour of these models, and we plan to explore this in future works.

Limitations

Our work does have some limitations that might have impacted the outputs of our models. For instance, the behaviour of the models might partly be due to the skewed distribution of classes in the dataset, where the discarded class is the majority class. Another possible issue is how we have decided to split the data. We aimed to achieve the best possible representation of classes across splits, but ended up with a rather small training set and a relatively large validation set. We think that this could have had an effect, and that increasing the size of the training set, while removing coverage from the validation set, could have a positive outcome on our models.

References

Javier Beltrán, Rubén Míguez, and Irene Larraz. 2021. Claimhunter: An unattended tool for automated claim detection on twitter. In *KnOD@WWW*.

Gullal S. Cheema, Sherzod Hakimov, and Ralph Ewerth. 2020. Check_square at checkthat! 2020 claim detection in social media via fusion of transformer and syntactic features. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurence Dierickx, Ghazaal Sheikhi, Duc Tien Dang Nguyen, and Carl-Gustav Lindén. 2022. Report on the user needs of fact-chekers. In *NORDIS Project Report: University of Bergen*, Task 4.2.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cede no, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing*, RANLP '17, Varna, Bulgaria.

Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017b. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Per Egil Kummervold, Javier de la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. In *Proc. of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Eric Lazarski, Mahmood Al-Khassaweneh, and Cynthia Howard. 2021. Using nlp for fact checking: A survey. *Designs*, 5.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Juan R. Martinez-Rico, Lourdes Araujo, and Juan Martínez-Romo. 2020. Nlp&ir@uned at checkthat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Javier Beltrán, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022. Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, CLEF '2022, Bologna, Italy.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Mucahid Kutlu Alex Nikolov, Firoj Alam Yavuz Selim Kartal, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*, CLEF '2021, Bucharest, Romania (online).

Evan M. Williams, Paul Rodrigues, and Valerie Novak. 2020. Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.