Too Many Claims to Fact-Check: Prioritizing Political Claims Based on Check-Worthiness

Yavuz Selim Kartal, Mucahid Kutlu, and Busra Guvenen
Department of Computer Engineering
TOBB University of Economics and Technology
Ankara, Turkey
{ykartal, m.kutlu, bguvenen}@etu.edu.tr

Abstract

The massive amount of misinformation spreading on the Internet on a daily basis has enormous negative impacts on societies. Therefore, we need automated systems helping fact-checkers in the combat against misinformation. In this paper, we propose a model prioritizing the claims based on their check-worthiness. We use BERT model with additional features including domain-specific controversial topics, word embeddings, and others. In our experiments, we show that our proposed model outperforms all state-of-the-art models in both test collections of CLEF Check That! Lab in 2018 and 2019. We also conduct a qualitative analysis to shed light detecting check-worthy claims. We suggest requesting rationales behind judgments are needed to understand subjective nature of the task and problematic labels.

1 Introduction

The World Economic Forum (WEF) has ranked massive digital misinformation as one of the top global risks in 2013[1]. Unfortunately, the foresight of WEF seems right as we encountered many unpleasant incidents due to the misinformation spread on the Internet since 2013 such as the gunfight due to "Pizzagate" fake news[2] and increased mistrust towards vaccines[3].

In order to combat against misinformation and its negative outcomes, fact-checking websites (e.g., Snopes[4]) detect the veracity of claims spread over the Internet and share their findings with their readers [5]. However, fact-checking is an extremely time-consuming process, taking around one day for a single claim [12]. While these invaluable journalistic efforts help to reduce the spread of misinformation, Vosoughi et al. [22] report that false news spread eight times faster than true news. Therefore, systems helping fact-checkers are urgently needed in the combat against misinformation.

As human fact-checkers are not able detect the veracity of all claims spread on the Internet, it is vital to spend their precious time in fact-checking the most important claims. Therefore, an automatic system monitoring social media posts, news articles and statements of politicians, and detecting the *check-worthy* claims is needed. A number of researchers focused on this important problem (e.g., [12, 19, 13]). Furthermore, Conference and Labs of Evaluation Forum (CLEF) Check That! Lab (CTL) has been organizing shared-tasks on detecting check-worthy claims since 2018 [18, 2, 4]. In CTL tasks, a political debate or a transcribed speech is separated by sentences and participants are asked to rank the sentences according to their priority to be fact-checked. In CTL'20 [3], tweets have also been used for this task.

In this paper, we propose a ranking model that prioritizes claims based on their check-worthiness. We propose a BERT-based hybrid system in which we first

Title of the Proceedings: "Proceedings of the CIKM 2020 Workshops October 19-20, Galway, Ireland"
Editors of the Proceedings: Stefan Conrad, Ilaria Tiddi

[1]http://reports.weforum.org/global-risks-2013

[2]www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

[3]www.washingtonpost.com/news/wonk/wp/2014/10/13/the-inevitable-rise-of-ebola-conspiracy-theories

[4]https://www.snopes.com/

fine tune a BERT [6] model for this task, and then use its prediction and other features we define in a logistic regression model to prioritize the claims. The features we use include word-embeddings, presence of comparative and superlative adjectives, domain-specific controversial topics, and others. Our model achieves 0.255 and 0.176 mean average precision (MAP) scores in CTL'18 and CTL'19 datasets, respectively, outperforming all state-of-the-art models including participants of the corresponding shared-tasks, ClaimBuster [12], BERT, XLNET [24], and Lespagnol et al.[15]'s model. We share our code for the reproducibility of our results[5].

2 Related Work

As the US presidential election in 2016 is one of the main motivating reasons for fact-checking studies, prior work mostly used debates and other speeches of US politicians as their datasets (e.g., [12, 15]). Therefore, the majority of studies focused on English. The Arabic datasets used in prior work ([13, 18]) are just translations of English datasets.

ClaimBuster [12] is one of the first studies about check-worthiness. ClaimBuster is a supervised model using many features including part-of-speech (POS) tags, named entities, sentiment, and TF-IDF representations of claims. TATHYA [19] uses topics, POS tuples, entity history, and bag-of-words as features. The topics are detected by LDA model trained on transcripts of all presidential debates from 1976 to 2016.

Gencheva et al. [8] propose a neural network model with a long list of sentence level and contextual features including sentiment, named entities, word embeddings, topics, contradictions, and others. Jaradat et al. [13] use roughly the same features with Gencheva et al., but extend the model for Arabic. In its followup work, Vasileva et al. [21] propose a multi-task learning model to detect whether a claim will be fact-checked by at least five (out of nine) pre-selected reputable fact-checking organizations.

CLEF has been organizing Check That! Labs (CTL) since 2018. Seven teams participated in check-worthiness task of CTL'18. The participant teams used various learning models such as recurrent neural network (RNN) [10], multilayer perceptron [26], random forest (RF) [1], k-nearest neighbor (kNN) [9] and Support Vector Machine (SVM) [25] with different sets of features such as bag-of-words [26], character n-gram [9], POS tags [26, 10, 25], verbal forms [26], named entities [26, 25], syntactic dependencies [26, 10], and word embeddings [26, 10, 25]. On English dataset, Prise de Fer [26] team achieved the best MAP scores using almost every feature mentioned before with SVM-Multilayer perceptron learning.

In 2019, 11 teams participated in check-worthiness task of CTL'19. Participants used varying models such as LSTM, SVM, naive bayes, and logistic regression (LR) with many features including readability of sentences and their context [2]. Copenhagen team [11] achieved the best overall performance using syntactic dependency and word embeddings with weakly supervised LSTM model.

Lespagnol et al. [15] investigated using various learning models such as SVM, LR, and Random Forests, with a long list of features including word-embeddings, POS tags, syntactic dependency tags, entities, and "information nutritional" features which represent factuality, emotion, controversy, credibility, and technicality of statements. In our experiments we show that our model outperforms Lespagnol et al. on both test collections.

Our proposed an approach distinguishes from the existing studies as follows. 1) We propose a BERT-based hybrid model which uses fine-tuned BERT's output with many other features. 2) As the topic might be a strong indicator for check-worthiness, many studies used various types of topics such as general topics [25], globally controversial topics [15], and topics discussed in old US presindential debates [19]. However, we believe that check-worthiness of a claim depends on local and present controversial topics. Thus, we use a list of hand-crafted controversial topics related to US elections. 3) We also use two different sets of features including a hand-crafted list of words and presence of comparative and superlative adjectives and adverbs.

3 Proposed Approach

We propose a supervised model with a number of features described below. We investigate various learning models including LR, SVM, random forest, MART [7], and LambdaMART [23]. Now we explain the features we use.

BERT: We first fine tune BERT using respective training data. Next, we use its prediction value as one of our features.

Word Embeddings (WE): Words that are semantically and syntactically similar tends to be close in the embedding space, allowing us to capture similarities between claims. We represent a sentence as the average vector of its words excluding the out-of-vocabulary ones. Word embedding vectors are extracted from the pre-trained word2vec model [17] which has a feature vector size of 300.

Controversial Topics (CT): Sentences about controversial topics might include check-worthy claims. Lespagnol et al. [15] use a list of

[5]https://github.com/YSKartal/political-claims-checkworthiness

controversial issues compiled from Wikipedia article "Wikipedia:List_of_controversial_issues". However, the list they use covers many controversial issues which have very limited coverage in current US media such as "Lebanon", "Chernobyl", and "Spanish Civil War" while the data we use are about recent US politics. We believe that controversy of a topic depends on the society. For instance, US politicians propose different policies for immigrants, yielding heated discussions among them and their supporters. On the other hand, US domestic politics are much less interested in refugee crisis in Mediterranean sea than European countries. Therefore, a claim about Mexican immigrants might be check-worthy for people living in US while they might find claims about refugees taking a dangerous path to reach Europe not-check-worthy. In contrast, people living in Europe might consider the latter case as check-worthy and the former one as not-check-worthy. In addition, controversy of a topic might change over time. For instance, Cold War (which also exists in that Wikipedia list) might be one of the most discussed topics in US politics before the collapse of the Soviet Union in 1991. However, nowadays it is rarely covered by US media. Therefore, we propose using controversial issues related to the data we use, instead of any controversial issue around the globe and in the history.

Firstly, we identified 11 major topics in current US politics including immigration, gun policy, racism, education, Islam, climate change, health policy, abortion, LGBT, terror, and wars in Afghanistan and Iraq. For each topic, we identified related words and calculate the average of these words using their word embedding vectors. For instance, for the immigration topic, we used words "immigrants", "illegal", "borders", "Mexican", "Latino" and "Hispanic".

In this feature set of size 11, we calculate cosine similarity between sentences and each topic by using their vector presentation. We use the average of word embeddings for sentences excluding stopwords with NLTK [16].

Comparative & Superlative (CS): Politicians frequently use sentences comparing themselves with others because each candidate tries to convince the public that s/he is better than his/her opponent. Therefore, the comparisons in political speeches might impact people's voting decision and, thereby, it might be important to check their veracity. Thus, in this feature, we use the number of comparative and superlative adjectives and adverbs in sentences.

Handcrafted Word List (HW): Particular words convey important information about check-worthiness because 1) it might be related to an important topic (e.g., "unemployment"), 2) it represents a numerical value, increasing the factuality of the sentence (e.g., "percent") and 3) its semantic represents a comparison between two cases (e.g., "increase" and "decrease"). Thus, we first identified 66 words analyzing training datasets of CTL'18 and CTL'19. In this feature, we check whether there is an overlap between lemmas of selected words and lemmas of words in the respective sentence.

Verbe Tense (VT): We cannot detect the veracity of claims about future while we can only verify claims about the present or past. Thus, the verbe tense of sentences might be an effective indicator for check-worthiness of claims. This feature vector represents the existence or absence of each tense in the predicate of the claims.

Part-of-speech (POS) Tags: If a sentence does not contain any informative words, then it is less likely to be check-worthy. To represent the information load of a claim, we use the number of nouns, verbs, adverbs and adjectives, separately.

4 Experiments

4.1 Experimental Setup

Implementation: We use ktrain library[6] to fine-tune BERT model with 1 cycle learning rate policy and maximum learning rate of 2e-5 [20]. We use SpaCy[7] for all syntactic and semantic analyses. We use Scikit toolkit[8] for the implementations of SVM, Random Forest (RF), and LR. The parameter settings of the learning algorithms are as follows. We use default parameters for SVM. We set the number of trees to 50 and the maximum depth to 5 for RF. We use multinomial and lbfgs settings for LR. For MART and LambdaMART models, we use RankLib[9] library, and set the number of trees and leaves to 50 and 2, respectively.
Data: We evaluate the performance of our system with two datasets used in CTL'18 and CTL'19. The details about them are given in **Table 1**. CTL'18 consists of transcripts of debates and speeches while CTL'19 contains also press conferences and posts.

Table 1: Details about CTL'18 and CTL'19 datasets.

		CTL'18	**CTL'19**
Train	**# Docs**	3	19
	# Sentence	4,064	16,421
	# CW Claims	90 (2,2%)	433 (2,6%)
Test	**# Docs**	7	7
	# Sentence	4,882	7,079
	# CW Claims	192 (3,9%)	110 (1,6%)

[6] https://pypi.org/project/ktrain/
[7] https://spacy.io/
[8] https://scikit-learn.org
[9] https://sourceforge.net/p/lemur/wiki/RankLib/

Baselines: We compare our model against the following models.

- *Lespagnol et al. [15]*: Lespagnol et al. report the best results on CTL'18 so far. Therefore, we use it as one of our baselines. In order to get its results for CTL'19, we contacted with the authors to get their own code. The authors provide us the values of "information nutrition" features and instructions about how to generate WE embeddings. We implemented their method using the values they shared and following their instructions[10].

- *ClaimBuster*: We use the popular pretrained ClaimBuster API[11] [12] which is trained on a dataset covering different debates that do not exist on CTL'18 and CTL'19.

- *BERT*: As it is reported that BERT based models outperform state-of-the-art models in various NLP tasks, we compare our model against using only BERT. We fine tune BERT model using the respective training dataset and predict the check-worthiness of claims using the fine-tuned model.

- *XLNET*: It is reported that XLNet outperfroms BERT in various NLP tasks [24]. Thus, we use XL-NET for this task by fine-tuning with the respective training dataset.

- *Best of CTL'18 and CTL'19*: For each dataset, we also report the performance of best systems participated in the shared-tasks, i.e., Prise de Fer team [26] and Copenhagen team [11] for CTL'18 and CTL'19, respectively.

Training & Testing: We use the same setup with CTL'18 and CTL'19 to maintain a fair comparison with the baselines. We follow the evaluation method used on CTL'18 and CTL'19: We calculate average precision (AP), R-precision (RP), precision@5 (P@5) and precision@10 (P@10) for each file (i.e., debate, speech) and then report the average performance.

4.2 Experimental Results

In this section, we present experimental results on test data using different sets of features and varying learning algorithms.

Comparison of Learning Algorithms. In our first set of experiments, we evaluate logistic regression (LR), SVM, random forest (RF), MART and LambdaMART models using all features defined in Section 3. **Table 2** shows MAP scores of each model. Interestingly, LR outperforms all other models. In a similar experiment Lespagnol et al.[15] conducted, they also report that LR yields higher results than other models they used. Nevertheless, we use LR in our following experiments.

Table 2: **MAP Score for Varying Models Using All Features**

Learning Model	CTL'18	CTL'19
LR	**.2303**	**.1775**
RF	.1468	.1542
SVM	.1716	.1346
MART	.1764	.1732
Lambda MART	.0671	.0564

Feature Ablation. In order to analyze the effectiveness of features we use, we apply two techniques: 1) *Leave-one-out methodology* in which we exclude one type of feature group and calculate the model's performance without it, and 2) *Use-only-one methodology* in which only a single feature group is used for prediction. The results are shown in **Table 3**.

From the results in Table 3, we see that features have different effects on each dataset. BERT is the most effective feature on CTL'19. However, in contrast to our expectations, WE seems more effective feature than BERT on CTL'18. On CTL'18, the performance decreases by nearly 25% when WE is excluded. In addition, we achieve the highest MAP score when we use only WE. On CTL'19, we achieve 0.1356 MAP score using only WE, showing that it is more effective than other features except BERT. However, the performance of our model increases when we exclude WE (0.1775 vs. 0.1786 in Table 3), suggesting that the information it contributes is covered by other features on CTL'19.

Excluding hand-crafted word list (HW) features causes performance decrease in both test collections. In addition, using only HW features outperforms all participants of CTL'18 (0.153 vs 0.1332 in Table 3). These promising results suggest that expanding this list might lead further performance increases.

Our results also suggest that Controversial Topics (CT) are effective features. Excluding them decreases the performance of the model in both collections while using only CT features yield high scores, slightly outperforming the best performing system on CTL'18 (0.1363 vs. 0.1332 in Table 3).

Excluding CS and POS features also slightly decrease the performance of the model in both test collections. Regarding time tense features, our results are

[10] It is noteworthy that we obtain 0.2115 MAP score on CTL'18 with our implementation of their method while they report 0.23 MAP score in their paper. We are not aware of any bug in our code but the performance difference might be because of different versions of the same library. Nevertheless, the results we present for their method on CTL'19 should be taken with a grain of salt.

[11] https://idir.uta.edu/claimbuster/

Table 3: **MAP Scores for Varying Feature Sets**

Leave-One-Out			Use-Only-One		
Features	**CTL18**	**CTL19**	**Features**	**CTL18**	**CTL19**
All	.2303	.1775			
All-CS	.2239	.1765	**CS**	.751	.604
All-BERT	.2211	.1580	**BERT**	.1850	**.1701**
All-VT	**.2547**	.1761	**VT**	.1007	.598
All-HW	.2126	.1727	**HW**	.1530	.1043
All-WE	.1756	**.1786**	**WE**	**.2068**	.1356
All-CT	.2170	.1739	**CT**	.1363	.1046
All-POS	.2283	.1767	**POS**	.1048	.631

Table 4: **Comparison with Competing Models. * sign indicates the results obtained from our implementation of the respective competing model.**

	CTL'18				**CTL'19**			
Model	**MAP**	**RP**	**P@5**	**P@10**	**MAP**	**RP**	**P@5**	**P@10**
BERT	.1850	.2218	.3142	.2857	.1701	.1945	**.2571**	**.2429**
XLNET	.1974	.2393	.2857	.2571	.0932	.0770	.1429	.1143
Lespagnol et al. [15]	.230	.254	.314	.2857*	.1292*	.1347*	.1714*	.2000*
Prise de Fer Team	.1332	.1352	.2000	.1429	-	-	-	-
Copenhagen Team	-	-	-	-	.1660	**.4176**	**.2571**	.2286
ClaimBuster	.2003	.2162	.2571	.2429	.1329	.1555	.1714	.2000
Our Model	**.2547**	**.2579**	**.4000**	**.3429**	**.1761**	.2028	**.2571**	.2143

mix. Excluding time tense feature causes a slight performance decrease on CTL'19, but yields higher performance score on CTL'18.

Comparison Against Baselines. We pick the model that includes all features except VT as our primary model because it achieves the highest MAP score on average. We compare our primary model with the baselines. The results are presented in **Table 4**.

Our proposed model outperforms all other models based on all evaluation metrics on CTL'18. On CTL'19, our proposed model achieves the highest MAP score, which is the official metric used in CTL. BERT model outperforms other models based on P@10 on CTL'19. Regarding P@5 metric, our model, BERT and Copenhagen Team achieve the same highest scores with 0.2571. Regarding RP, Copenhagen Team achieves the highest score. Overall, our model outperforms all other models based on the official evalution metric of CTL while BERT and Copenhagen Team [10] also achieve comparable performance on CTL'19.

5 Qualitative Analysis

In this section, we present our qualitative analysis for the output of our primary model. For each input file, we rank the claims based on their check-worthiness and then detect not-check-worthy claim with the highest rank. **Table 5** shows these not-check-worthy statements for each file with our system's ranking and speaker of the statement.

The statement in Row 1 is a claim about the future. Our model with verb tense could rank this statement at lower ranks but our primary model does not use verb tense features because it yields lower performance on average. In Row 2, the statement is very complex with many relative clauses, in perhaps decreasing the performance of BERT model and WE features in representing the statement. In Row 3, our model makes an obvious mistake and ranks a statement which does not have even any predicate, at very high ranks. Perhaps our model falls short because the word "jobs" indicates that the statement is about unemployment, which is one of the controversial topics we defined.

As reported by Vasileva et al. [21] fact-checking organizations investigate different claims with very minimal overlaps between selected claims. We observe this subjective nature of annotations in Rows 4-14 because all statements are actually factual claims and some of them might also be considered as check-worthy. For instance, statements in Row 8, 11 and 13 are clearly said to change people's voting decision. In addition, almost all statements are about economics which is an important factor on people's votes. Therefore, checking their veracity might be also important not to misinform public. Nevertheless, these examples show the

Table 5: Highest ranked non check-worthy statements from each test document by our primary model

Row	Rank	File Name	Speaker	Statement
1	4	task1-en-file1	CLINTON	The plan he has will cost us jobs and possibly lead to another Great Recession.
2	1	task1-en-file2	CLINTON	Then he doubled down on that in the New York Daily News interview, when asked whether he would support the Sandy Hook parents suing to try to do something to rein in the advertising of the AR-15, which is advertised to young people as being a combat weapon, killing on the battlefield.
3	1	task1-en-file3	TRUMP	Jobs, jobs, jobs.
4	2	task1-en-file4	TRUMP	Before that, Democrat President John F. Kennedy championed tax cuts that surged the economy and massively reduced unemployment.
5	3	task1-en-file5	TRUMP	The world's largest company, Apple, announced plans to bring $245 billion in overseas profits home to America.
6	1	task1-en-file6	TRUMP	America has lost nearly-one third of its manufacturing jobs since 1997, following the enactment of disastrous trade deals supported by Bill and Hillary Clinton.
7	1	task1-en-file7	TRUMP	Our trade deficit in goods with the world last year was nearly $800 billion dollars.
8	1	20151219_3_dem	O'MALLEY	We increased education funding by 37 percent.
9	1	20160129_7_gop	KASICH	We're up 400,000 jobs.
10	1	20160311_12 _gop	TAPPER	Critics say these deals are great for corporate America's bottom line, but have cost the U.S. at least 1 million jobs.
11	3	20180131_state _union	TRUMP	Unemployment claims have hit a 45-year low.
12	1	20181015_60_min	TRUMP	–if you think about it, so far, I put 25% tariffs on steel dumping, and aluminum dumping 10%.
13	3	20190205_trump _state	TRUMP	Unemployment for Americans with disabilities has also reached an all-time low.
14	1	20190215_trump _emergency	TRUMP	They have the largest number of murders that they've ever had in their history - almost 40,000 murders.

the subjective nature of check-worthiness annotations.

In addition to subjective judgments, we also noticed inconsistencies within the annotations. For instance, the statement in Row 9 ("We are up 400,000 jobs") also exists in "20160311_12_gop" file but annotated as "check-worthy". In addition, there exists semantically very similar statements with different labels. For instance, Donald Trump's statement "I did not support the war in Iraq" in 1079^{th} line of 20160926_1pres file is labeled as "not-check-worthy" while his statement in 1086^{th} line of the same file "I was against the war in Iraq" is labeled as "check-worthy". Both statements have similar meanings and exists in the same context (i.e., their position in file are very close). Therefore, both might have the same labels. As a counter argument, "being against" suggests an action while "not supporting" does not require any action to be taken. Thus, different annotations for similar statements might also be again due to the subjective nature of check-worthiness judgments.

Furthermore, there are also annotations that we strongly disagree with the label. For instance, in 20170315_nashville file (training data on CTL'19), Donald Trump's statement "We're going to put our auto industry back to work" is labeled as check-worthy. However, the statement is about future and cannot be verified.

Overall, our qualitative analysis suggests that annotating check-worthiness of claims is a subjective task and the annotations might be noisy. Kutlu et al. [14] show that using text excerpts within documents as rationales help understanding disagreements in relevance judging. Similarly, we might request rationales behind check-worthiness annotations to understand if the label is due to a human judging error or the subjective nature of the annotation task. Furthermore, rationales behind these annotations might help us develop effective solutions for this challenging problem.

6 Conclusion

In this paper, we presented a supervised method which prioritize claims based on check-worthiness. We use logistic regression classifier with features including state-of-the-art language model BERT, domain-specific controversial topics, pretrained word embeddings, handcrafted word list, POS tags and comparative-superlative clauses. In our experiments on CTL'18 and CTL'19, we show that our proposed model outperforms all state-of-the-art models in both collections. We show that BERT's performance can be increased by using additional features for this task. In our feature ablation study, BERT model and word embeddings appear to be the most effective features while handcrafted word list and domain-specific controversial topics also seem effective. Based on our qualitative analysis, we believe requesting rationales for the check-worthiness annotations would further help in developing effective systems.

In the future, we plan to work on weak supervision techniques to extend the training dataset. With the increased data, we will be able explore using deep learning techniques for this task. In addition, we plan to extend our study to detect check-worthy claims in social media platforms because it is the channel where most of the people affected by misinformation. Moreover, working on different languages and building a multilingual model is an important research direction in the combat against misinformation.

References

[1] R. Agez, C. Bosc, C. Lespagnol, N. Petitcol, and J. Mothe. IRIT at checkthat! 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.

[2] P. Atanasova, P. Nakov, G. Karadzhov, M. Mohtarami, and G. Da San Martino. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. In *CEUR Workshop Proceedings*, 2019.

[3] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, and Z. S. Ali. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 215–236, Cham, 2020. Springer International Publishing.

[4] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. D. S. Martino, M. Hasanain, R. Suwaileh, and F. Haouari. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. *Advances in Information Retrieval*, 12036:499 – 507, 2020.

[5] F. Cherubini and L. Graves. The rise of fact-checking sites in europe. *Reuters Institute for the Study of Journalism, University of Oxford*, 2016.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, 2019.

[7] J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[8] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, 2017.

[9] B. Ghanem, M. Montes-y-Gómez, F. M. R. Pardo, and P. Rosso. UPV-INAOE - check that: Preliminary approach for checking worthiness of claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.

[10] C. Hansen, C. Hansen, J. G. Simonsen, and C. Lioma. The copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the clef-2018 checkthat! lab. In *CLEF*, 2018.

[11] C. Hansen, C. Hansen, J. G. Simonsen, and C. Lioma. Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, 2019.

[12] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, and M. Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10:1945–1948, 2017.

[13] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, and P. Nakov. Claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Demonstrations, pages 26–30, 2018.

[14] M. Kutlu, T. McDonnell, Y. Barkallah, T. Elsayed, and M. Lease. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement? In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 805–814. ACM, 2018.

[15] C. Lespagnol, J. Mothe, and M. Z. Ullah. Information nutritional label and word embedding to estimate information check-worthiness. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 941–944. ACM, 2019.

[16] E. Loper and S. Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[18] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, and G. Da San Martino. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387, 2018.

[19] A. Patwari, D. Goldwasser, and S. Bagchi. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262. ACM, 2017.

[20] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *ArXiv*, abs/1803.09820, 2018.

[21] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, and P. Nakov. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2019.

[22] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[23] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, June 2010.

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[25] K. Yasser, M. Kutlu, and T. Elsayed. bigir at CLEF 2018: Detection and verification of check-worthy political claims. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2018.

[26] C. Zuo, A. Karakas, and R. Banerjee. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In *CLEF*, 2018.