"To trust a LIAR": Does Machine Learning Really Classify Fine-grained, Fake News Statements?

Mark Mifsud[1], Colin Layfield[1], Joel Azzopardi[2] and John Abela[1]

[1]Dept of Computer Information Systems, Faculty of ICT, University of Malta, Msida Malta
[2]Dept of Artificial Intelligence, Faculty of ICT, University of Malta, Msida Malta

Abstract

Fake news refers to deceptive online content and is a problem which causes social harm [1]. Early detection of fake news is therefore a critical but challenging problem. In this paper we attempt to determine if state-of-the-art models, trained on the LIAR dataset [2] can be leveraged to reliably classify short claims according to 6 levels of veracity that range from "True" to "Pants on Fire" (absolute lies). We investigate the application of transformer models BERT [3], RoBERTa [4] and ALBERT [5] that have previously performed significantly well on several natural language processing tasks including text classification. A simple neural network (FcNN) was also used to enhance each model's result by utilising the sources' reputation scores[1]. We achieved higher accuracy than previous studies that used more data or more complex models. Yet, after evaluating the models' behaviour, numerous flaws appeared. These include bias and the fact that they do not really model veracity which makes them prone to adversarial attacks. We also consider the possibility that language-based, fake news classification, on such short statements is an ill-posed problem.

Keywords

Fake News, Natural Language Processing, Artificial intelligence, Deep Learning, Transformer Models

1. Introduction

Social media has made the creation and spreading of information easier, quicker and cheaper than ever before. This scenario has resulted in an epidemic of what is termed as 'fake news' - content that deliberately gives false information to deceive and manipulate, often with negative results [1].

1.1. Fake News Classification

Since fake news spread fast, early detection is necessary in order to limit the spread and, consequently, the harm. The use of Machine Learning (ML) techniques for Natural Language Processing (NLP) is one way to build classifiers that could serve as potential early detectors. Two distinct approaches that use NLP-based models [6] are:

[1]All the source code used is available at: https://github.com/MarkMifsud/To-Tust-A-Liar

OHARS'21: Second Workshop on Online Misinformation- and Harm-Aware Recommender Systems, October 2, 2021, Amsterdam, Netherlands

mark.mifsud.16@um.edu.mt (M. Mifsud); colin.layfield@um.edu.mt (C. Layfield); joel.azzopardi@um.edu.mt (J. Azzopardi); john.abela@um.edu.mt (J. Abela)

1. Model-Based approaches: A ML model is used to find some reliable features in a dataset that correlate to the classification label.
2. Feature-Based approaches: The ML model relies on pre-defined, linguistic or textual feature(s) assumed to indicate deception.

Our approach is model-based, since transformer models are used to learn features that correlate short statements with their respective class labels.

1.2. Transformers

Transformers are deep learning architectures that have changed the face of NLP research in recent years. A transformer is initially trained on a large corpus of text to build a 'language model'. This model represents words as vectors, yet unlike traditional word embeddings (like Word2Vec), the representation of each word is sensitive to the context within which the word occurs. The trained transformer can then be used for various NLP tasks, such as Question Answering, Named Entity Recognition, Text Classification and others [3].

Google's BERT [3] was the first to gain popularity since it performed very well on a number of NLP benchmark-tasks. Facebook later released RoBERTa which, although sharing many common features, was pre-trained using different algorithms on an English corpus 10 times larger and also supported a larger vocabulary [4]. RoBERTa outperformed BERT in multiple instances. ALBERT, by Google and Toyota is an optimised version of the original BERT that is highly scalable thus making larger architectures possible. ALBERT too performed better at many NLP tasks than previous attempts [5].

1.3. The LIAR Dataset

The LIAR dataset [2] contains 12,836 short claims by prominent players in US politics, extracted from politifact.com. Statements are labelled as either True, Mostly-True, Half True, Barely-True (mostly false), False or Pants-on-fire (6 classes). This makes the measure for veracity more finely grained than a binary (real or fake) label [2]; which is appropriate since statements can have a mix of true and false claims.

Each statement in LIAR also comes with textual metadata including the job title of the speaker, the speaker's affiliation, state of origin and the context in which the claim was uttered, (namely an interview, a debate or another event) [2]. The speaker's reputation is a numeric value that represents the total number of claims under each category of truthfulness uttered by that speaker. These values are important in some studies, and referred to as the speakers' history, credibility or reputation.

1.4. Previous Works on LIAR

Among the studies leading to this one, two are very relevant to our approach. Kirilin & Strube (2018) [7] have the best accuracy score (45.7%) on the 6-way classification, while Liu et al (2019) [8] were among the first to use BERT to classify LIAR entries.

Kirilin & Strube represented the statements and the textual metadata as FastText [9] word embeddings and subsequently used LSTMs to carry out the classification. The reputation score

and the classification result were then combined using an attention mechanism. A dense neural network performs a final classification [7].

Liu et al used BERT-base to classify the statements and textual metadata (except for the speaker's names). BERT's output-vector and the reputation were then utilized in an attention network followed by a simple neural network to carry out the classification. This entire layout was repeated twice. The first one produced a coarse grained (true or false) classification. This output was passed to the second segment, together with the initial input, to derive a final 6-way classification [8].

1.5. Ill-posed & Ill-conditioned Problems

The mathematical definition of a well-posed problem is attributed to mathematician Jacques Hadamard [10]. Pattern classification problems can be viewed as well-posed or ill-posed problems in the sense of Hadamard [11].

A problem is well-posed if it satisfies the following 3 criteria [12]:

1. It has a solution
2. It has only one, uniquely defined solution; and
3. The solution's behaviour changes continuously with the initial conditions

Thus, an ill-posed problem is a problem that fails one or more of the above criteria [12]. Such problems require some modification to be solved or approximated which may include additional data, measurements or boundaries [12].

When considering a few of the short statements from the LIAR dataset, it becomes apparent that many statements can have multiple possible truth-levels based on context, time or who the speaker is. For instance, the veracity of the statement: "*I am pro-life, he is not*" will depend on the speaker, the subject and their opinions at a given time.

These multiple, possible solutions violate the second rule for a well posed problem, giving a strong reason to believe that purely language-based classification of fake news is an ill-posed problem. Another class of problems is called ill-conditioned. Such problems may not satisfy the definition of ill-posed problems but are considered similarly unstable for practical purposes since a small change in the input results in a large change in the output [12].

2. Aims and Objectives

The principal aim of this research was to build a Fake-News Classifier, that matches, or exceeds, previous classifiers' accuracy scores on the LIAR dataset. Objectives to reach this aim included:

1. To analyse the effectiveness of transformers, given that they are the current state-of-the-art NLP models, for the classification task.
2. To investigate whether better trained or larger transformers can achieve a higher accuracy.
3. To evaluate whether the overall classification result can be enhanced by adding neural network layers that use both the transformer's output and the source's reputation score.

Two other objectives were added with the aim of determining the reliability of the resulting models.

4. To investigate the bias, behaviour and learning of the models achieved.
5. To investigate the argument that fake news classification, performed using only NLP (without the classifier having knowledge of the real world), is an ill-posed or ill-conditioned problem.

3. Design & Implementation

3.1. The Classification Models

Three BERT variants were used in order to determine if the differences in transformer size, pre-training or optimisations matter (as initially hypothesised). These were:

1. BERT-base, the smallest of the models used.
2. RoBERTa-Large: a larger model pre-trained on a larger amount of data.
3. ALBERT-Large-V2: comparable to a larger version of BERT that was pre-trained for a longer time.

6-way classification for different levels of truthfulness was performed on the statements in LIAR without the use of any metadata.

Fine tuning the transformers was performed manually for reasons of limited disk space. 80% of LIAR's data was used for fine tuning (training), 10% was used for validation and 10% for testing. LIAR comes already split into these segments allowing for a fair comparison with results reported in different studies.

BERT and RoBERTa converged after two epochs at learning rates of 1.8^{-5} and 2.2^{-5} respectively. ALBERT took 4 epochs at a learning rate of 2.2^{-5}. The batch size for training all models was 64.

3.1.1. Using the Reputation Score

To use both the statement and the reputation scores for classification we created FcNN (Fully Connected Neural Network). This was necessary because NLP-transformers are not applicable for classification with numerical data such as the reputation score [8]. The utilised FcNN has 24 nodes in its first hidden layer, 12 in the second hidden layer and an output of 6 nodes each corresponding to one of our classifications. All layers use the tanh activation function since the values of the transformers' output vector vary from -1 to 1 (or close) just like the upper and lower limits of the hyperbolic tangent function.

Each of the transformers produces a classification vector consisting of 6 values in its final layer, which can be extracted programmatically. These are input to the FcNN together with the 6 values of the reputation score (Figure 1) after the latter are normalised (divide by 200, since this is a value close to the largest reputation score). FcNN is then trained at a learning rate of 9-4. To avoid overfitting their result was checked every 500 epochs of training and the best fitting model was used. In every run, the FcNN managed to fit in less than 9000 epochs.

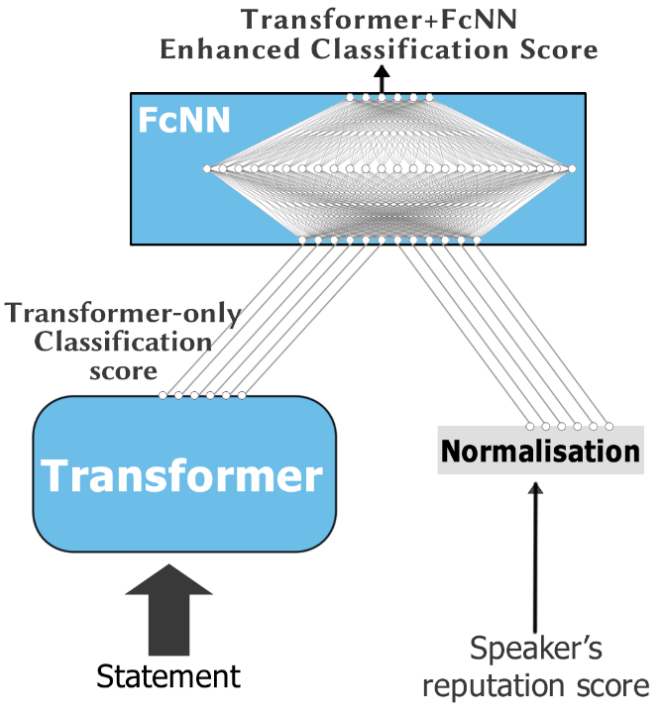

Figure 1: Our Transformer+FcNN architecture.

In a separate attempt, FcNN was applied on its own, for classification using only the reputation vector. This provided a baseline against which other models could be compared to (Table 3).

None of the textual metadata was used because a speaker's name, job or similar details were considered to be unrelated to a statement's truthfulness. Furthermore, these values either repeat frequently, are often null or not normalized (non-atomic and different spelling can be found for the same value). Because of this, we were concerned that it would bias the models unnecessarily.

3.2. Quantifying the Classifiers' Bias

Relying too heavily on the individual's reputation may result in labelling liars instead of lies [7]. To test if this was the case with our models, a small set of 226 statements was used as a test-set for our baseline, FcNN-only model that utilises only the reputation score.

This test-set's 226 statements are truths from liars and lies from mostly-honest speakers. These were chosen by computing each speaker's honesty ratio P, a measure of how honest a speaker is, based on each speaker's classification of his or her claims, such that:

$$P = 1.5(True - PantsOnFire) + (MostlyTrue - False) \quad (1)$$

The numerical difference of a speaker's pants-on-fire statements from true ones was multiplied to give it a higher weighting. Speakers with values close to zero (balanced liars) were ignored. Those with scores less than -15 are considered liars so we take their truthful statements. Speakers

scoring more than 4 are honest ones, for which we take their lies. The reason for these cut-off points was because speakers with honesty ratio between -15 and 4 were ones with relatively fewer claims. The inequality resulted from the fact that the labels of liars are skewed to begin with (3 false, 1 neutral and 2 true labels). Thus, the set of 226 statements was collected.

While only the FcNN-only model (trained normally on LIAR's training set) was used to classify these 226 statements, it was expected that even the FcNN models using the transformer's output will be prone to this same bias, if confirmed.

3.3. Investigating the Effect of Data Quality on Learning

We also trained the same models on datasets with different data quality than that of LIAR and compared the results. This was meant to reveal the effect that the quality of the data has on the models' learning and also if the models are truly able to learn the intended classification task or not.

For this task, two variations of LIAR were created. The first is called Shuffled-LIAR and was obtained by randomly shuffling the spoken claims attribute among all entries in the training set, while leaving every other attribute (column) untouched. By having a dataset with randomised text and all other attributes untouched, we can better determine how much the text really affects the result. If the same results on the actual, unshuffled set are also achievable on a completely random set, then one can conclude that the results are accidental and hence insignificant.

Additionally, the Cleaned-LIAR dataset was created in order to allow training and testing on data of better quality (less errors) [13]. This was done by compensating for flaws found in LIAR[2]. Cleaned-LIAR omits 207 entries that were discovered to not be stated claims at all [13]. For example, some entries are test data, many indicate whether a speaker changed opinion (known as flip-flops on Politifact.com) while others are in Spanish (so the words used would not be in the vocabulary of transformers trained on an English corpus).

The spelling and grammar of the statements were also corrected manually, under the assumption that since the transformers were pre-trained on good quality text and have a limited vocabulary, classification may receive a boost from these corrections. If accuracy is not improved when training on this set, this may suggest that the flawed entries were responsible for the higher accuracy on the original (unchanged) LIAR.

On Cleaned-LIAR, BERT was trained for 2 epochs at a 2^{-5} learning rate. RoBERTa and ALBERT were trained at a learning rate of 1.2^{-5} for 3 and 2 epochs respectively. The models failed to fit for Shuffled-LIAR.

3.4. Testing the Ill-conditioned Property

If instances of the same basic model produce highly varied classifications when they are trained on data that differs gradually (when tested on the same test data); it may indicate that the problem is ill-conditioned (at least in the way the problem is being treated here).

Five copies of the same transformer were trained with training data that varies proportionally each time. The dataset's original training and validation portions were joined and their order

[2]Spelling and other mistakes in LIAR mostly result from how the data was scraped from polifact.com to produce the dataset.

randomised. The resulting set was then stratified, splitting it in 5 folds (parts), such that all folds contain virtually an identical number of statements and variety of labels (truth levels). This keeps the label balance identical for each fold and thus for the 5 folds. For each of the five training runs, a different combination of 4 folds would be used for training, and the fifth would be used for validation. The test set was the same in each of the 5 runs.

This classification was performed with both LIAR and Cleaned-LIAR separately, using the BERT-base model. Then this was all repeated with RoBERTa-Large. For comparison, the same procedure was repeated using the two transformers to carry out a 5-way sentiment analysis on the Stanford Sentiment Treebank (SST-5) dataset [14]. This would offer a baseline. Assuming sentiment analysis is well-conditioned, fake-news classification would give a similar variability to sentiment only if it is well-conditioned too.

The same hyperparameters were used to train the transformers (2^{-5} learning rate, 64 batch size for 2 epochs).

4. Results & Evaluation

In our evaluation, we managed to achieve a higher accuracy than other results reported in literature. However, all other test results suggest that our models are flawed despite their higher accuracy. This is described in more detail below.

4.1. Classifiers Accuracy

The transformer-only classifiers had a performance similar to Wang's previous attempts that utilise statements alone [2], showing they are at least as effective at classification as previous deep learning models (Table 1).

All of our Transformer+FcNN models exceeded accuracy results by Kirilin & Strube (2018) and Liu et al (2019) in spite of these studies using more data. This vindicates our decision to avoid using textual metadata. Furthermore, our BERT model performs better than Liu et al's system, despite it having a far simpler architecture (Figure 1). All transformers produced similar accuracy scores. A bigger or a better trained transformer only marginally improves fake news classification.

4.2. Reputation Bias

When classifying truthful statements from liars and lies from honest speakers, FcNN displays a clear bias caused by utilising reputation. This is clearly visible in the Confusion Matrix found in Table 2.

4.3. Effect of Data Quality on Training

The fact that the transformers did not manage to properly fine-tune for Shuffled-LIAR indicates that the models are correlating some features to the labels, whilst no such correlating feature occurs randomly. However, when trained on the less noisy, Cleaned-LIAR the performance of

Table 1
Comparing to previous work: model, data used and accuracy score achieved on LIAR.

Study	Model/Architecture	Data Used				Accuracy Score
		Statement	Metadata	Reputation	External	
Wang 2017 [2]	SVM	+				25.5%
	CNNs + LSTM	+				27.0%
	CNNs + LSTM	+	+	+		27.4%
Long et al 2017 [15]	LSTM + Attention	+				25.5%
	LSTM + Attention	+	+	+		41.5%
Karimi et al 2018 [16]	CNN +LSTM	+				29.1%
	CNN + LSTM	+	+	+	+	34.8%
Pham 2018 [17]	Dual Attention	+	+			37.3%
	Memory Attention Network	+		+		44.2%
Kirilin & Strube 2018 [7]	LSTM+Attention	+	+			41.5%
	LSTM+Attention	+	+	+		45.7%
Liu et al 2019 [8]	2 stage BERT_base + Attention	+				34.5%
	2 stage BERT_base + Attention	+	+	+		40.6%
Ours (transformer only)	BERT_base	+				27.7%
	RoBERTa_Large	+				27.3%
	ALBERT_Large_V2	+				28.2%
Ours	BERT_base + FcNN	+		+		48.0%
	RoBERTa_Large + FcNN	+		+		47.9%
	ALBERT_Large_V2 + FcNN	+		+		**48.6**%

Table 2
Confusion matrix showing how reputation scores alone biases truths from liars and lies from honest speakers

Actual label	Predicted					
	Pants on fire	Fake	Mostly Fake	Half True	Mostly True	TRUE
Pants	0	2	0	0	23	0
Fake	0	2	0	0	94	0
Mostly Fake	0	2	0	0	65	0
Half True	0	0	0	0	1	0
Mostly True	0	22	0	0	0	0
TRUE	0	14	0	0	0	0

the transformers, without FcNN was generally poorer (Table 3). RoBERTa is the exception in this case, since the cleaned set resulted in marginally better performance.

This unexpected result raised the question of whether the models are really modelling veracity. A test to this effect was done as described below.

4.4. Is Veracity Being Modelled?

Consider the following true statement that was classified correctly:

"One out of every four homeless people on our streets is a veteran."

Table 3
Accuracy scores on data of different quality.

Classifier	Accuracy		
	LIAR	Cleaned-LIAR	Shuffled-LIAR
FcNN only	44.58%	44.78%	44.58%
BERT-base Only	27.67%	25.98%	20.73%
BERT+FcNN	48.17%	48.84%	45.36%
RoBERTa Only	27.28%	27.81%	20.81%
RoBERTa+FcNN	47.31%	49.40%	46.45%
ALBERT Only	28.22%	25.59%	20.34%
ALBERT+FcNN	48.56%	47.81%	44.66%

Table 4
Variance & Overall Mean Square Error on 5 classification tasks.

Transformer	Dataset	Mean Variance	Overall MSE
BERT-base	SST-5	0.12	0.79
RoBERTa-Large	SST-5	0.11	0.54
BERT-base	Cleaned-LIAR	0.48	2.96
RoBERTa-Large	Cleaned-LIAR	0.61	3.07
BERT-base	LIAR	0.64	3.13
RoBERTa-Large	LIAR	0.64	3.03

A change in the fine-grained classification of the statement is expected if any of the following changes is done:

- **Negation of the same statement**: "*One out of every four homeless people on our streets is **not** a veteran.*"
- **Reducing probability of the statement**: "*One out of every four homeless people on our streets is a **friendly** veteran.*"; and
- **Contradiction**: "*One out of every four homeless people on our streets is **not homeless**.*"

Nevertheless, all such modifications are still classified as fully-true, showing that the models are not modelling deception or veracity, thus making them vulnerable to adversarial attacks.

4.5. The Ill-conditioned Property

Fake news classification results varied with gradual changes in input data over 4 times that of Sentiment Analysis on SST-5 (Table 4). Taking the Mean Square Error (MSE) for each run one would measure the difference in classifications from their target label. Taking an Overall MSE for the 5 runs, fake news classification shows considerable changes in output (Table 4).

By the definition of ill-conditioned problems [12], all these are a strong indication that fake news classification of short statements using transformers is an ill-conditioned problem.

4.6. Is NLP-based Fake-news Classification Ill-posed?

Factors supporting the case that NLP-based, fake news classification is an ill-posed problem include:

1. There appears to be no indicator of truthfulness or deception within LIAR's statements unless one has knowledge of the real world. Sentiment Analysis by contrast, can be based on the presence of certain words or expressions.
2. Feature Based detection does not generalise over domains [18].
3. Khan et al 2019, observed that "the performance of models is not dataset invariant" [19].
4. Accurate but non explainable models are not necessarily reliable. Assuming so, is an 'affirming the consequent' fallacy[3].
5. The models produced by this study and at least another previous one (Fakebox) are not modelling veracity [20].
6. Psychology shows that people lie differently. Even the same person's indicators of deception change over time within the same interview and are influenced by numerous factors [21, 22].
7. The models in this study are at least ill-conditioned, as shown [13].

All these point to the likelihood that there cannot be a model that maps a string of text to truth levels without knowledge of the world. This likelihood is strong for the models trained on LIAR and demonstrated for our models despite their relatively higher accuracy.

5. Conclusions & Future Work

5.1. Conclusions

Our best models achieve a higher accuracy on the LIAR dataset utilising the spoken statements and the speakers' reputation alone, outperforming methods that either used more data, more complex models or both. BERT and BERT variants can be leveraged to classify short statements more accurately.

Bigger and better trained transformers yielded only a marginal improvement over the smaller BERT-base transformer. In our case, although ALBERT-Large did perform better than BERT-base, fake news classification accuracy did not scale in proportion to the transformer size.

The most important insights resulted from testing beyond accuracy scores. Flaws were found and these led to questioning the whole idea of language-based classification of content according to deception or veracity. Issues were also identified with the LIAR dataset and these flaws were used to test the effect that data quality has on the models' ability to learn the task. Specifically, we show that although the models' accuracy on LIAR is better than random, the language transformers' contribution to the classification was generally poorer when trained and tested on cleaner data.

[3]Good models give a high accuracy. These models give a high accuracy; therefore, they are good. This is a logical fallacy known as Affirming the Consequent.

Furthermore, when compared with sentiment classification, fake news classification appears to be an unstable problem. We put forward arguments that suggest that purely NLP-based, fake-news classification on short statements, such as those found in LIAR, is not robust since it presents traits of ill-posed and ill-conditioned problems.

A simple test indicates that these models are not really modelling deception or veracity and are thus vulnerable to adversarial attacks.

The models herein, while improving on previous accuracies were thus proven unreliable for classifying arbitrary claims. The biggest contributor to the higher score was the reputation score which was shown to bias the models.

5.2. Future Work

We used only the text and the speaker's reputation in our tests, achieving a better score. An ablation study can also be done to analyse the impact of each attribute on the result.

Future studies can attempt similar investigations with the use of constructed features like part-of-speech tagging or dependency parsing together with those utilised internally by the transformer.

It would also be interesting to establish and standardise a variety of tests and metrics to assess the quality of a fake news classifier, by testing behaviour rather than mere accuracy scores; such as the ability to truly model veracity, stability (whether it is well conditioned or not), its ability to generalise over domains, and tests for bias. The ability to truly model veracity or deception deserves particular attention in future work, since it determines a classifier's robustness against adversarial attacks.

5.3. Recommendations

Researchers need to be aware of the flaws in the LIAR dataset.

Future studies are recommended to treat purely NLP-based, fake news detection as ill-posed, especially those utilising arbitrary or non-explainable features. Using knowledge-graphs to store knowledge about the real world, is likely one potential way to regularise the ill-posed problem.

Lastly, our models stand as examples of why analysis of a model's behaviour should be a better judge of how good the model is, rather than mere accuracy. Going forward, we believe this to be essential for mitigating the fake news problem effectively.

References

[1] A. Coleman, 'Hundreds dead' because of Covid-19 misinformation, 2020. URL: https://www.bbc.com/news/world-53755067.

[2] W. Y. Wang, "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL, Vancouver, Canada, 2017, pp. 422–426.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, Minneapolis, USA, 2019, pp. 4171–4186.
[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019).
[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, 2020. arXiv:1909.11942.
[6] R. Zafarani, X. Zhou, K. Shu, H. Liu, Fake News Research: Fundamental Theories, Detection Strategies & Open Problems, 2019. URL: https://www.fake-news-tutorial.com.
[7] A. Kirilin, M. Strube, Exploiting a speakers credibility to detect fake news, in: Proceedings of Data Science, Journalism & Media workshop at KDD (DSJM18), 2018.
[8] C. Liu, X. Wu, M. Yu, G. Li, J. Jiang, W. Huang, X. Lu, A two-stage model based on BERT for short fake news detection, in: C. Douligeris, D. Karagiannis, D. Apostolou (Eds.), Knowledge Science, Engineering and Management, 2019, pp. 172–183.
[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
[10] J. Hadamard, Sur les problèmes aux derivées partielles et leur signification physique, Princeton University Bulletin (1902) 49–52.
[11] P. Yee, S. Haykin, Pattern classification as an ill-posed, inverse problem: a regularization approach, in: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, 1993, pp. 597–600 vol.1. doi:10.1109/ICASSP.1993.319189.
[12] S. I. Kabanikhin, Definitions and examples of inverse and ill-posed problems, Journal of Inverse and Ill-posed Problems 16 (2008) 317–357.
[13] M. Mifsud, "To Trust a LIAR": Does machine learning really classify fine-grained, fake news statements? (Bachelor's dissertation), 2020. URL: https://www.um.edu.mt/library/oar/handle/123456789/76880.
[14] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, ACL, Seattle, Washington, USA, 2013, pp. 1631–1642.
[15] Y. Long, Q. Lu, R. Xiang, M. Li, C.-R. Huang, Fake news detection through multi-perspective speaker profiles, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 2017, pp. 252–256.
[16] H. Karimi, P. Roy, S. Saba-Sadiya, J. Tang, Multi-source multi-class fake news detection, in: Proceedings of the 27th International Conference on Computational Linguistics, ACL, Santa Fe, New Mexico, USA, 2018, pp. 1546–1557.
[17] T. T. Pham, A study on deep learning for fake news detection, 2018. URL: https://core.ac.uk/download/pdf/156904536.pdf.
[18] T. Gröndahl, N. Asokan, Text analysis in adversarial settings: Does deception leave a stylistic trace?, 2019. arXiv:1902.08939.
[19] J. Y. Khan, M. T. I. Khondaker, A. Iqbal, S. Afroz, A benchmark study on machine learning methods for fake news detection, CoRR abs/1905.04749 (2019). arXiv:1905.04749.

[20] Z. Zhou, H. Guan, M. Bhat, J. Hsu, Fake news detection via NLP is vulnerable to adversarial attacks, Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019) (2019).

[21] D. B. Buller, J. K. Burgoon, Interpersonal deception theory, Communication Theory 6 (1996) 203–242.

[22] J. K. Burgoon, D. B. Buller, C. H. White, W. Afifi, A. L. S. Buslig, The role of conversational involvement in deceptive interpersonal interactions, Personality and Social Psychology Bulletin 25 (1999) 669–686.