Language Models for Fact-Checking and Claim Assessment

Othman EL HOUFI
Pr. D. KOTZINOS – Research supervisor

M2 Research in Data Science & Machine Learning

6/7/22

Fake news & NLP

This tweet caused $130 billion value drop in stock market!

This article affected COVID-19 vaccination decisions

Fake news & NLP

Fake news

false, often sensational, information disseminated under the guise of news reporting.

Collins English Dictionary

Humans

have been proven irrational and vulnerable when differentiating between real and fake news. Typical accuracy ranges between 55% and 58%.

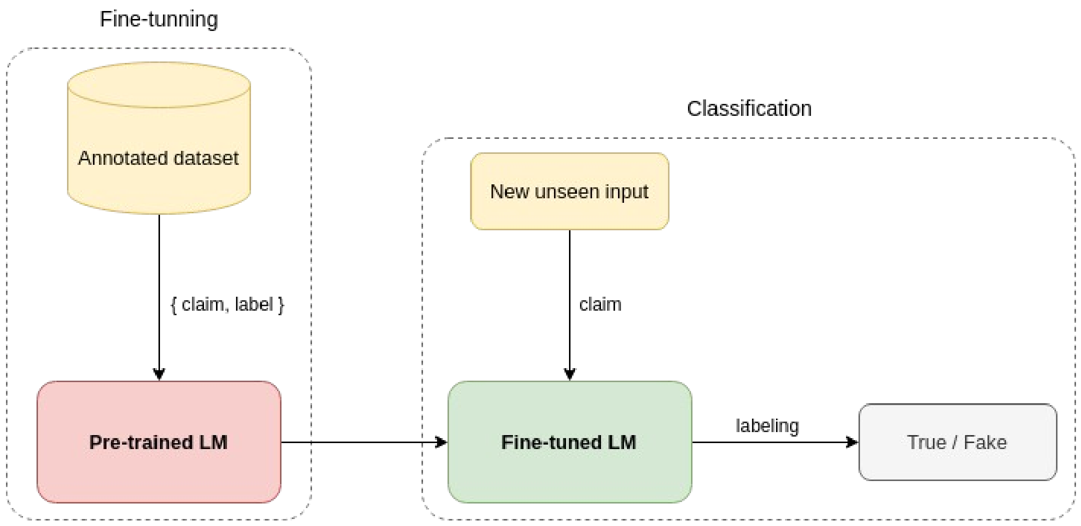
Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges.

Language Models

Datasets

Proposed solution

- Pre-process each dataset:
  - FEVER, Liar, MultiFC, COVID-19, ANTi-Vax
- Fine-tune a set of LMs:
  - BERT, RoBerta, ALBERT, XLNET, DistilBERT, BigBird, ConvBERT
- Deploy the best LM to assess the validity of new input claims



Results & Discussion

Dataset	Metric	2-labels	3-labels	5-labels	6-labels
FEVER	**accuracy**	0.81	0.64	-	
	macro f1	0.81	0.63	-	
MultiFC	**accuracy**	0.72	-	0.50	-
	macro f1	0.64	-	0.40	-
Liar	**accuracy**	0.69	-		0.31
	macro f1	0.61	-		0.30
Covid19	**accuracy**	0.98	-		
	macro f1	0.98	-		
ANTiVax	**accuracy**	0.99	-		
	macro f1	0.99	-		

Conclusion & Perspective

- LMs have a great potential to solve different NLP problems,
- LMs for fact-checking is good but not great,
    - Does not beat state-of-art traditional models.
- Different paths can be explored
    - We still have much to learn about LM.

Fake news & NLP

<u>Automatic</u> fake news <u>detection</u> is a practical <u>NLP problem</u> useful to all online content providers.

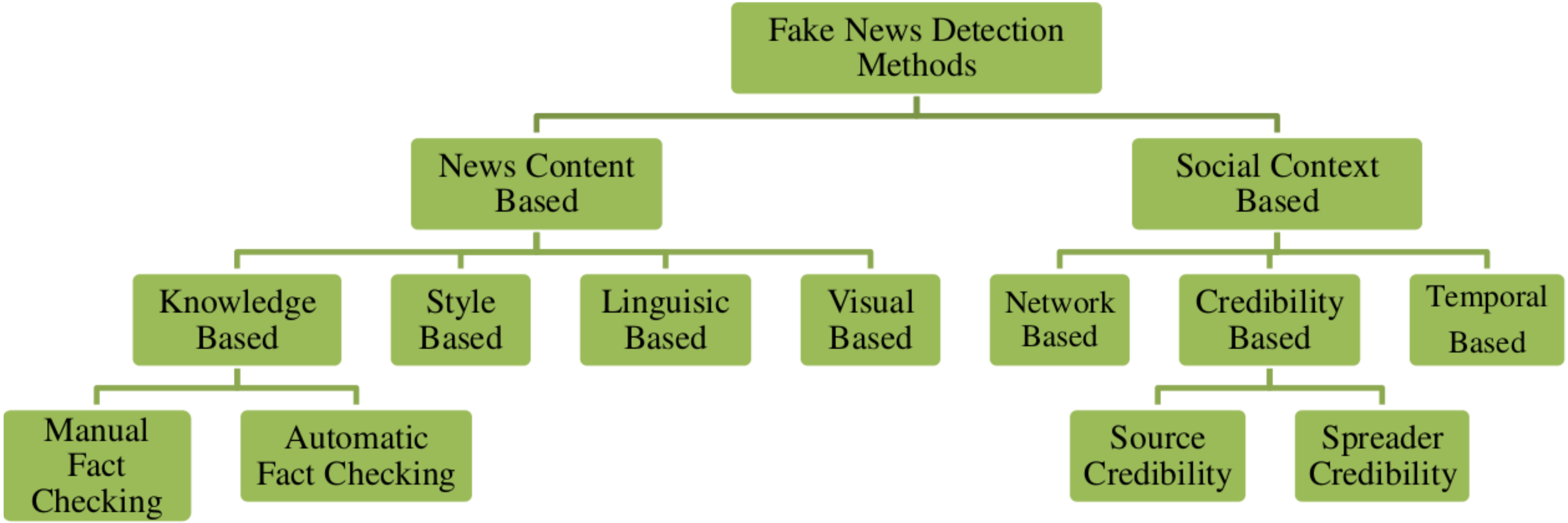
➔ Reduce the human time and effort to detect fake news,

➔ Can sweep through huge data streams,

➔ Capable of ceasing the spreading much faster.

- How can we differentiate fake news from real news?
- At what level of confidence can we do so?
- What are the existing methods that solves this problem?

Related work



S Hangloo, B Arora. Fake News Detection Tools and Methods – A Review

Related work

Knowledge-based Fake News Detection

aims to assess news authenticity by comparing the knowledge extracted from to-be verified news content with known facts, also called fact-checking.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Whatthewikifact: Fact-checking claims against wikipedia.

Style-based Fake News Detection

focuses on the style of writing, i.e. the form of a text rather than its meaning.

P. Przybyla. Capturing the style of fake news. In Proceedings of the AAAI Conference on Artificial Intelligence.

Related work

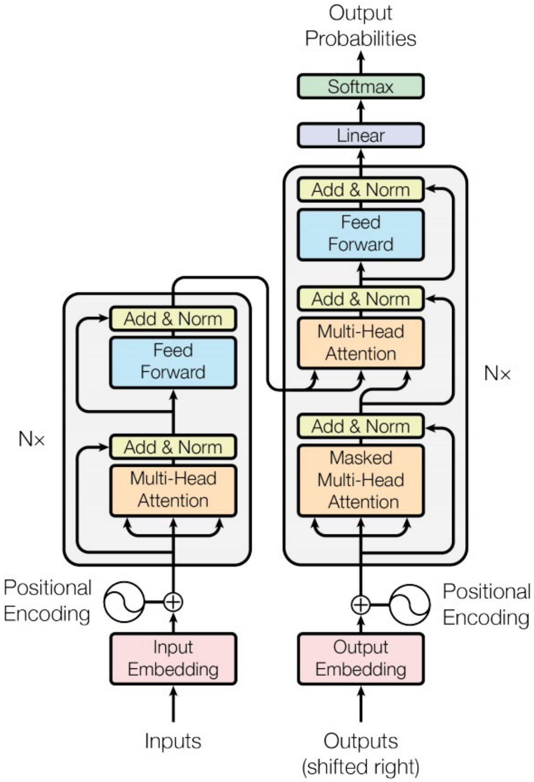
Language Model based Fact-Checking

a new approach that relies on fine-tuning state-of-art LMs like BERT that were pre-trained on Wikipedia's articles in order to solve the claim classification problem.

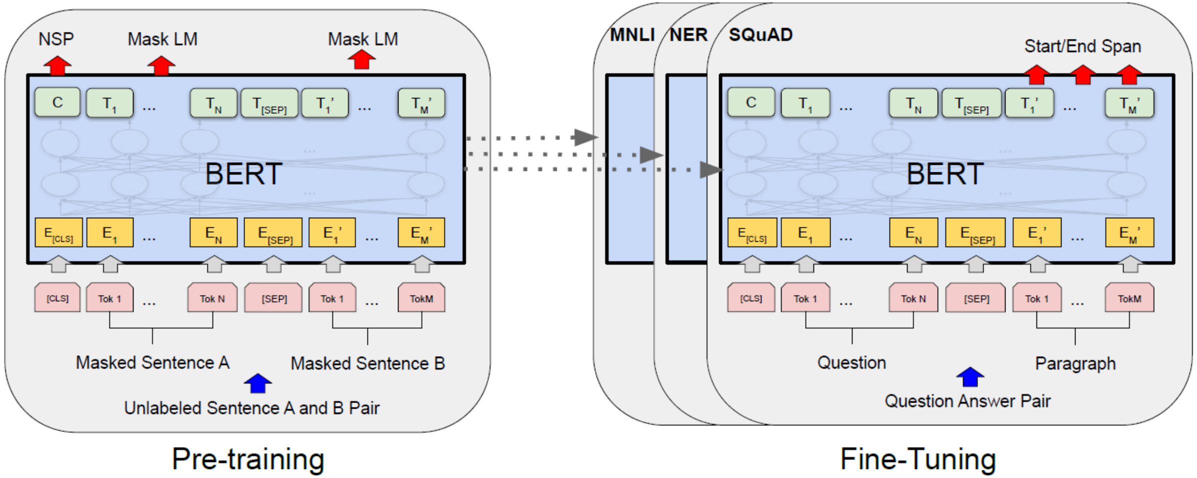
Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact checkers?

- What are Language Models?
- How can they be employed as fact-checkers?

Language Models



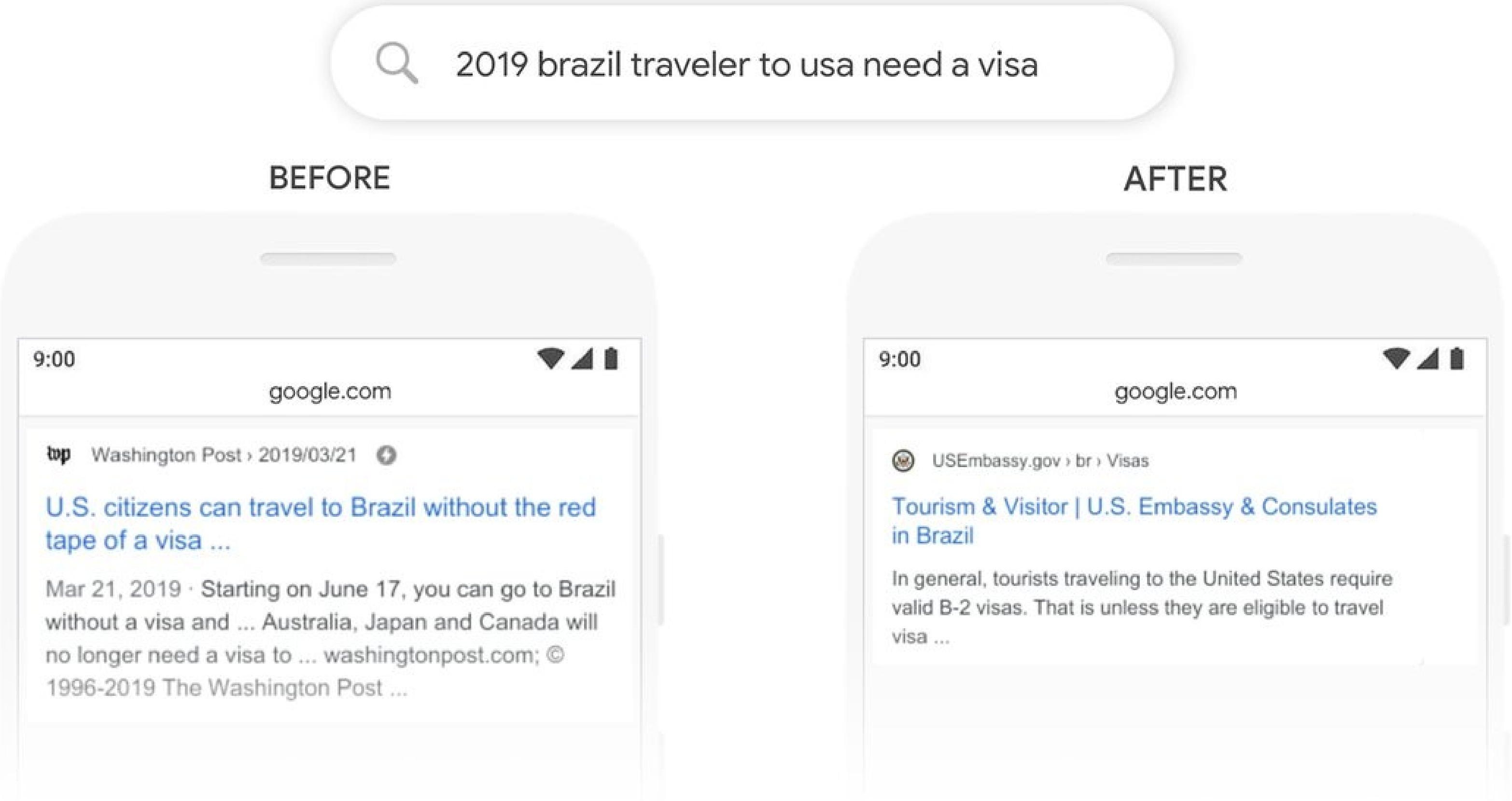
Attention is all you need.



Bert: Pre-training of deep bidirectional transformers for language understanding.

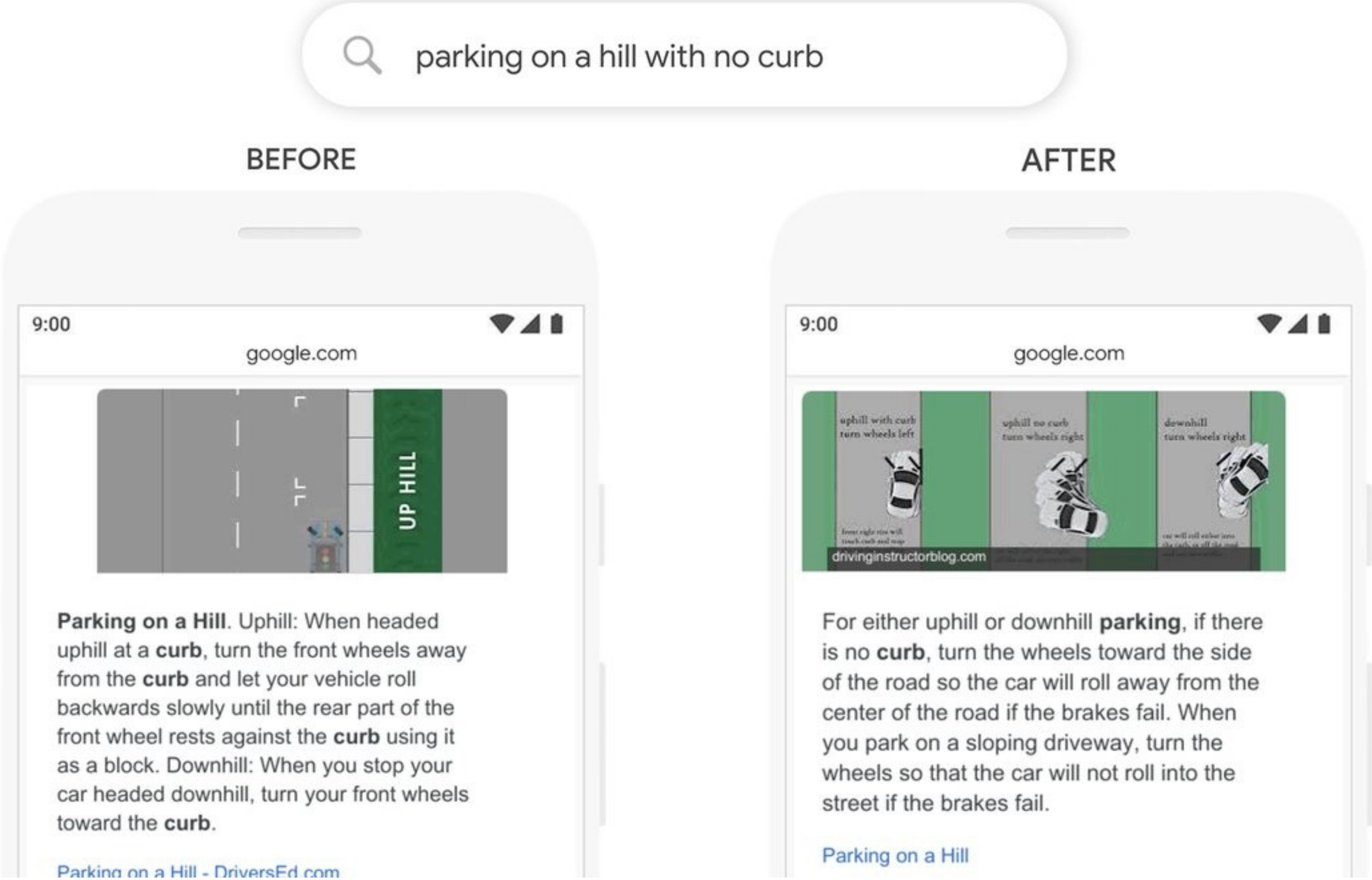
Language Models

BERT



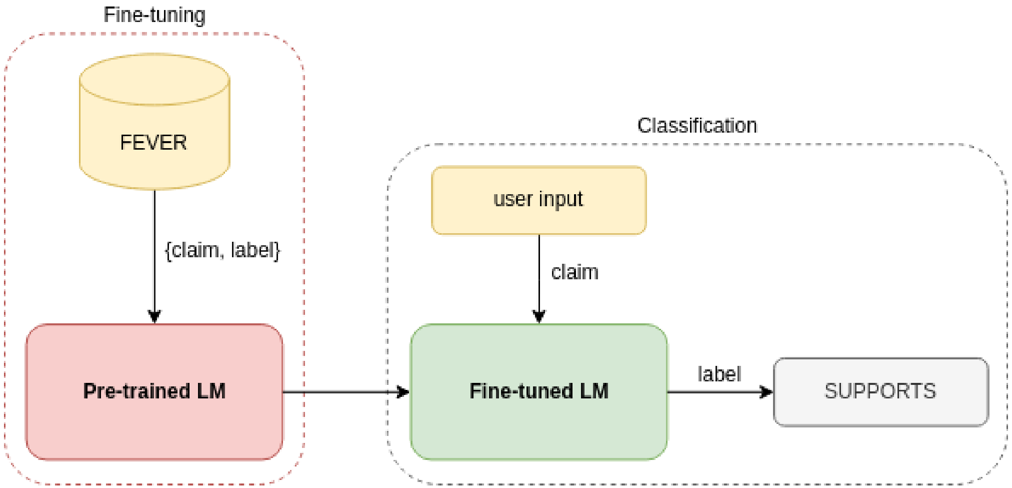
Language Models

BERT



Proposed architecture

- We start by fine-tuning a set of LMs for the downstream task that is claim classification using the FEVER dataset,
- Then each model is employed to assess the validity of new input claims.



Proposed architecture

FEVER dataset

FEVER (Fact Extraction and VERification)

consists of 185,445 claims generated by altering sentences extracted from Wikipedia. The claims are classified as Supported, Refuted or NotEnoughInfo.

Claim: The Rodney King riots took place in the most populous county in the USA.

`[wiki/Los_Angeles_Riots]`

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

`[wiki/Los_Angeles_County]`

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

ID	Claim	Label
79044	The Apple Store first opened in 2001.	1
117129	Adventure Time won an Oscar.	0
55061	Yamaha Corporation produces hardware.	2

EXAMPLES OF FEVER CLAIMS AND LABELS .

Split	SUPPORTS	REFUTES	NEI	Total
Train	80,035	29,775	35,639	145,449
Val	3,333	3,333	3,333	9,999
Test	3,333	3,333	3,333	9,999

DATASET SPLIT SIZES FOR SUPPORTS, REFUTES AND NOTENOUGHINFO (NEI) CLASSES .

Proposed architecture

Fine-tuning Language Models

LMs used in this experiment:

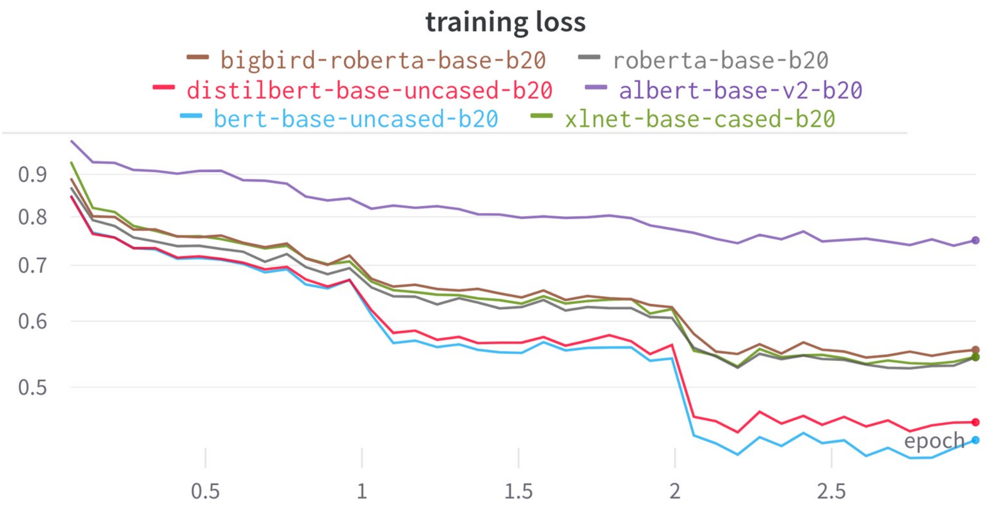
- BERT-base-uncased
- RoBERTa-base
- DistilBERT-base-uncased
- XLNET-base-cased
- ALBERT-base-v2
- BigBird-RoBERTa-base

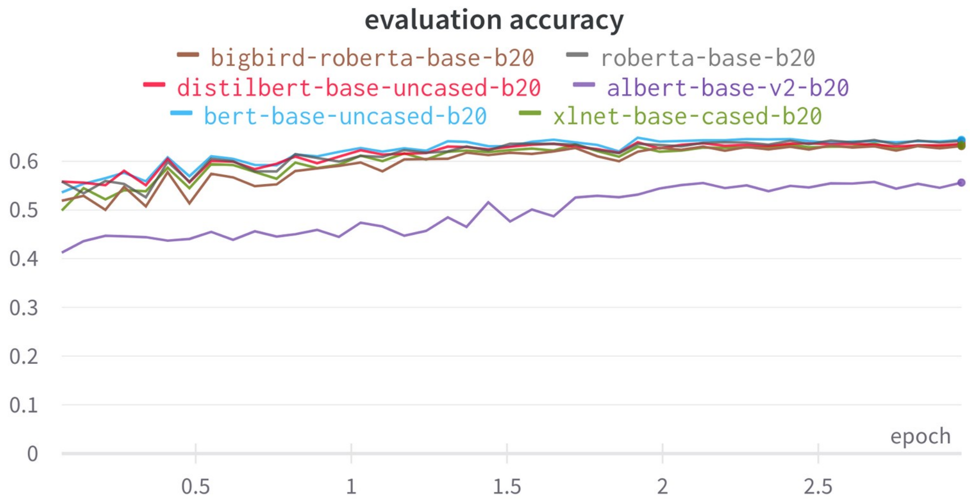
Hyperparameters:

- Tokenizer max sequence length: 128
- Output layer size: 3
- Activation function: GeLU
- Learning rate: 3e-5
- Optimization: Adam with linear decay
- Loss function: Cross-Entropy
- Epochs: 3
- Training batch size: 20
- Validation batch size: 20

Proposed architecture

Learning & Validation





Results & Discussion

Fine-tuned model	Label	prec	recall	f1	accuracy	macro prec	macro recall	macro f1
BERT-base-uncased	SUPPORTS REFUTES NEI	0.55 0.75 0.61	0.78 0.59 0.47	0.64 0.66 0.53	**0.62**	0.63	**0.62**	**0.61**
ALBERT-base-v2	SUPPORTS REFUTES NEI	0.46 0.77 0.50	0.81 0.46 0.33	0.59 0.58 0.40	0.53	0.58	0.53	0.52
DistilBERT-base-uncased	SUPPORTS REFUTES NEI	0.54 0.75 0.60	0.78 0.58 0.47	0.64 0.65 0.53	0.61	0.63	0.61	**0.61**
RoBERTa-base	SUPPORTS REFUTES NEI	0.54 0.75 0.63	0.81 0.59 0.45	0.65 0.66 0.53	**0.62**	**0.64**	**0.62**	**0.61**
BigBird-RoBERTa-base	SUPPORTS REFUTES NEI	0.53 0.75 0.63	0.81 0.58 0.44	0.64 0.66 0.52	0.61	**0.64**	0.61	0.60
XLNET-base-cased	SUPPORTS REFUTES NEI	0.53 0.74 0.63	0.81 0.59 0.43	0.64 0.65 0.51	0.61	0.63	0.61	0.60
Related work	**Label**	**prec**	**recall**	**f1**	**accuracy**	**macro prec**	**macro recall**	**macro f1**
BERT-large [7]	SUPPORTS REFUTES NEI	0.54 0.62 0.57	0.67 0.55 0.49	0.59 0.58 0.53	0.57	0.57	0.57	0.57
FEVER Baseline [19]	-	-	-	-	0.49	-	-	-
Ohio State University [19]	-	-	-	-	0.50	-	-	-
Columbia NLP [19]	-	-	-	-	0.58	-	-	-
Papelo [19]	-	-	-	-	0.61	-	-	-
UNC-NLP [19]	-	-	-	-	0.68	-	-	-
DREAM [20]	-	-	-	-	**0.77**	-	-	-

Results & Discussion

- Pre-trained LMs can classify claims,
- LMs surpasses most of the existing fact-checking methods,
- We don’t need bigger models for better results
    - BERT-large Vs. BERT-base Vs RoBERTa-base
- Using LMs does not require an external source of knowledge.

Conclusion & Perspective

- LMs have a great potential to solve different NLP problems,
- LMs for fact-checking is good but not great,
    - Does not beat state-of-art traditional models.
- Different paths can be explored
    - We still have much to learn about LM.

Overview

- Fake news & NLP
- Related work
- Language Models
- Proposed architecture
    - FEVER dataset
    - Fine-tuning Language Models
    - Learning & Validation
- Results & Discussion
- Conclusion & Perspective

Proposed Method

Language Models for classification

Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 125 **Large:** 355	**Base:** 66	**Base:** 12M **Large:** 18M
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	**Base:** 8 x V100 x 3.5d (4 times less than BERT)	[not given] **Large:** 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Trans-former, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced para-meters & SOP (not NSP)

https://humboldt-wi.github.io/blog/research/information_systems_1920/uncertainty_identification_transformers/