Detecting Check-Worthy Claims in Political Debates, Speeches, and Interviews Using Audio Data

Sofia University St. Kliment Ohridski

Petar Ivanov[1] Ivan Koychev[1] Momchil Hardalov[2] Preslav Nakov[3]

[1]Sofia University "St. Kliment Ohridski" [2]AWS AI Labs

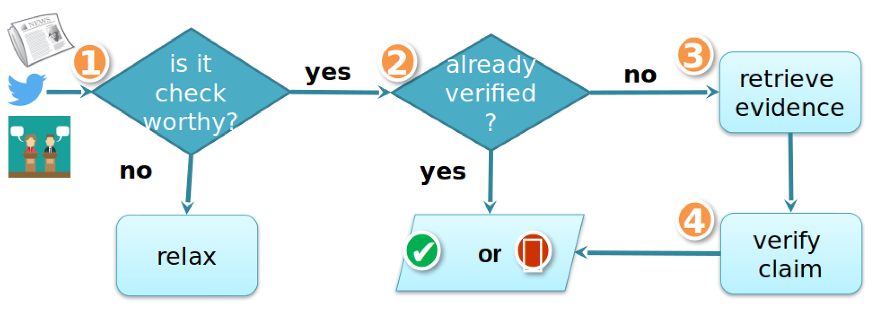
[3]Mohamed bin Zayed University of Artificial Intelligence

peturoi@uni-sofia.bg, koychev@fmi.uni-sofia.bg, momchilh@amazon.com, preslav.nakov@mbzuai.ac.ae

ICASSP 2024 KOREA

Motivation

Fake news and mis/disinformation are booming. While accumulating support, politicians sometimes resort to distorting or hiding the truth, unintentionally or on purpose. Manual fact-checking is the most common and trustworthy way to fight this, but it is tedious and time-consuming. Hence, it is important to prioritize what to fact-check, i.e., to estimate the check-worthiness of the claims.



The general fact-checking flow-chart

The focus of our work is the first step from the fact-checking flowchart: detecting check-worthy claims. Previous work has focused exclusively on the text modality, but here we explore the utility of the audio as an additional input.

Objective

Given a political debate, a speech or an interview, rank the sentences according to their check-worthiness. As this is a ranking task, we use Mean Average Precision (MAP) for evaluation.

Line #	Speaker	Sentence	Check-worthy?
146	Pence	But Hillary Clinton and Tim Kaine want to build on Obamacare.	No
147	Pence	They want to expand it into a single-payer program.	Yes
842	Kaine	The Clinton Foundation is one of the highest-rated charities in the world.	No
843	Kaine	It provides AIDS drugs to about 11.5 million people.	Yes

Political debate transcript

Line #	Speaker	Sentence	Check-worthy?
843	Kaine	It provides AIDS drugs to about 11.5 million people.	Yes
147	Pence	They want to expand it into a single-payer program.	Yes
842	Kaine	The Clinton Foundation is one of the highest-rated charities in the world.	No
146	Pence	But Hillary Clinton and Tim Kaine want to build on Obamacare.	No

Ranked claims from debate

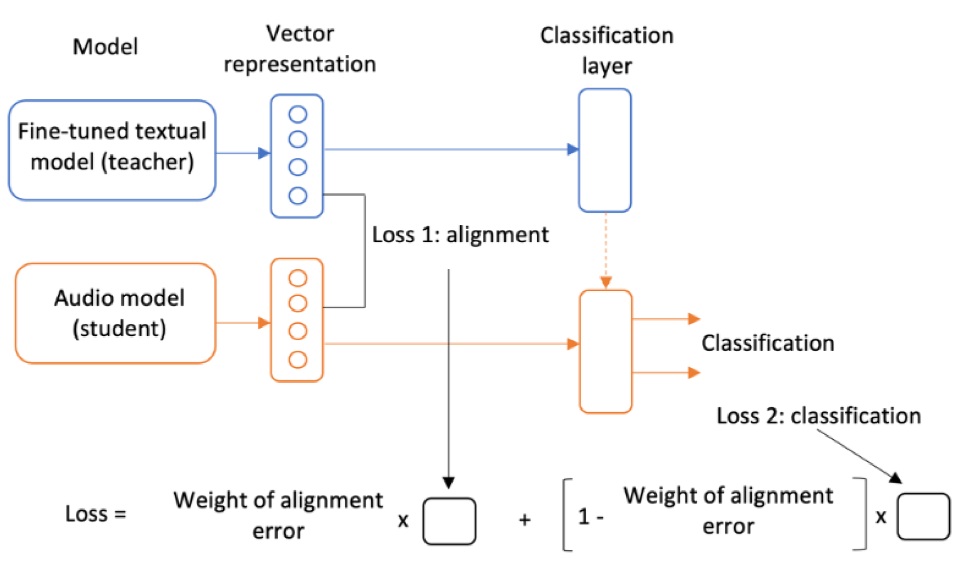
Contributions

- A **multimodal dataset** (text and audio) for detecting check-worthy claims.
- A **novel framework** that combines the text and the speech modalities.
- **Evaluation and comparison** of current state-of-the-art textual and audio models on our multimodal dataset.
- **Positive results - Multiple speakers**: adding the audio modality yields sizable improvements over using the text modality alone.
- **single speaker**: an audio-only model could outperform a strong text-only baseline.

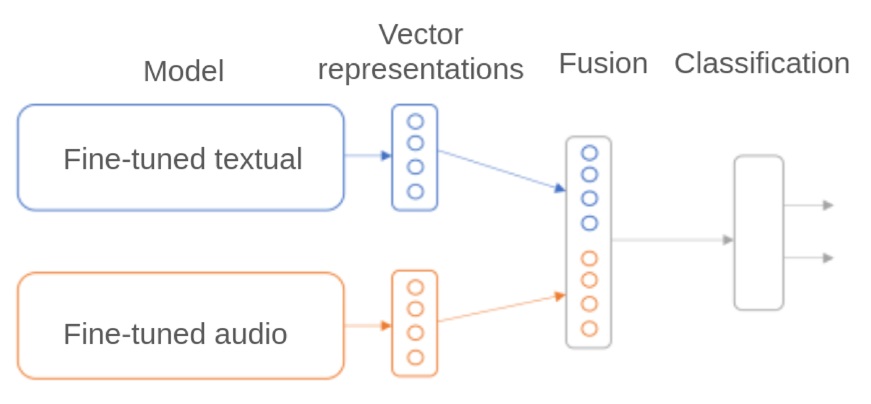
Joint Models

Textual and audio models are combined in several different ways:

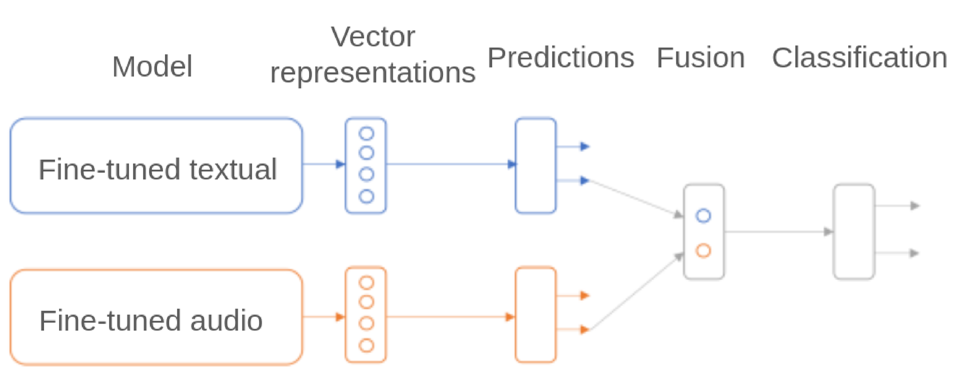
- **Knowledge alignment** - we train an audio model to represent the input it receives in the same way a fine-tuned textual model would represent its input in a teacher-student mode.
- **Early fusion ensemble** - we take fine-tuned models, run the inputs with the respective modalities through them and combine their input representations which in term goes through a classification layer.
- **Late fusion ensemble** - we combine the predictions of the models.



Knowledge alignment



Early fusion



Late fusion

Data

We augment the dataset for the 2021 CheckThat! lab, Task 1b. Our new multimodal dataset (text and audio in English) contains 48 hours of speech in English, comprising 34,489 sentences.

Modality	CheckThat!'21 Text Only	Our Dataset Text + Audio
Train		
# events	40	38
# sentences	42,033	28,715
# check-worthy claims	429	417
Dev		
# events	9	7
# sentences	3,586	1,896
# check-worthy claims	69	40
Test		
# events	8	8
# sentences	5,300	3,878
# check-worthy claims	298	291
All		
# events	57	53
# sentences	50,919	34,489
# check-worthy claims	796	748

Multimodal dataset

	Original	x15	x30	1:1
# non-check-worthy	28,298	28,298	28,298	417
# check-worthy claims	417	6,672	12,927	417
Check-worthy claims	1.5%	19.1%	31.4%	50.0%

Train data with over- and undersampling

	Train	Dev	Test	All
# sentences	8,191	1,650	3,489	13,330
# check-worthy claims	213	39	278	530
Check-worthy claims	2.6%	2.4%	8.0%	4.0%

Single speaker dataset

The check-worthy claims in the multimodal dataset are about 2% of all sentences. Thus, we prepared three variants of the training dataset: upsampling 15 and 30 times, removing random non-check-worthy sentences until their number becomes equal to the number of the check-worthy ones.

We propose a single-speaker setup, leaving aside the speech specifics of the different speakers.

In addition, we prepare a variant of the sentence-level audio segments with reduced background noise.

Models

Text models:

- BERT-base uncased
- SVM with TF.IDF
- Feedforward network focusing on named entities

Audio models (base variants):

- HuBERT
- wav2vec 2.0
- data2vec-audio

Results: Multiple Speakers

Fusing the audio and the textual modalities improves the MAP score.

Row	Model	Train dataset	MAP(test)
1	BERT	1:1	37.15
2	SVM with TF.IDF	x15	23.92
3	Feedforward network with named entities count	x15	22.28

Textual model results

Row	Model	Train dataset	Audio segments	MAP(test)
1	HuBERT	x30	Original	25.26
2	wav2vec 2.0	x15	Original	23.65
3	data2vec-audio	x30	Reduced noise	23.30

Audio model results

Row	Model	Train dataset	Audio segments	MAP(test)
1	data2vec-audio	Without changes	Original	29.99
2	wav2vec 2.0	Without changes	Original	29.96
3	HuBERT	Without changes	Original	27.87

Knowledge alignment results

Row	Ensemble type	Model	Train dataset	Audio segments	MAP(test)
1	Early fusion	BERT & HuBERT	Without changes	Original	38.17
2	Late fusion	BERT & HuBERT	x15	Original	37.58
3	Early fusion	BERT & aligned data2vec	Without changes	Original	37.35
4	Late fusion	BERT & aligned data2vec	x30	Original	37.24

Ensemble results

Results: Single Speaker

Experiments with the single-speaker subset of the dataset. The audio model using audio segments with reduced background noise achieves the highest MAP, outperforming the best textual model.

Row	Model	MAP(test)
1	BERT	32.67
2	SVM with TF.IDF	26.93
3	Feedforward network with named entities count	21.93

Textual model results

Row	Model	Audio segments	MAP(test)
1	wav2vec 2.0	Reduced noise	34.27
2	HuBERT	Original	24.78
3	data2vec-audio	Reduced noise	21.29

Audio model results