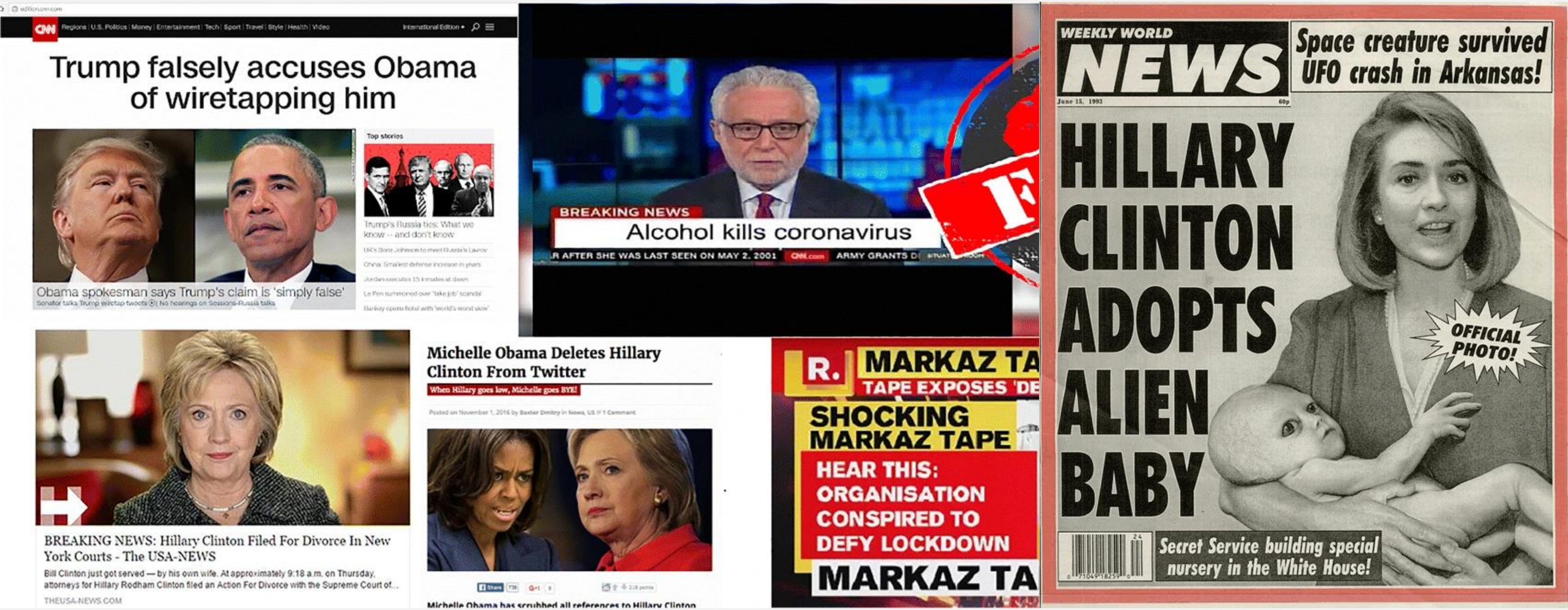
FAKE NEWS DETECTION BY USING LANGUAGE MODELS

Marianna Malaireu

AGENDA

- Some history
- Approaches description
- Linguistic-based methods description
  - Bert
  - Roberta
  - Electra
  - ELMO
- Experiments description

HISTORY


Trump falsely accuses Obama of wiretapping him
Obama spokesman says Trump's claim is 'simply false'
Top stories
Trump's Russia ties: What we know -- and don't know
BREAKING NEWS
Alcohol kills coronavirus
WEEKLY WORLD NEWS
Space creature survived UFO crash in Arkansas!
June 15, 1993
HILLARY CLINTON ADOPTS ALIEN BABY
OFFICIAL PHOTO!
Secret Service building special nursery in the White House!
BREAKING NEWS: Hillary Clinton Filed For Divorce In New York Courts - The USA-NEWS
Bill Clinton just got served — by his own wife. At approximately 9:18 a.m. on Thursday, attorneys for Hillary Rodham Clinton filed an Action For Divorce with the Supreme Court of...
THEUSA-NEWS.COM
Michelle Obama Deletes Hillary Clinton From Twitter
When Hillary goes low, Michelle goes BYE!
MARKAZ TA
TAPE EXPOSES 'DE
SHOCKING MARKAZ TAPE
HEAR THIS: ORGANISATION CONSPIRED TO DEFY LOCKDOWN
MARKAZ TA

FAKE NEWS DETECTION APPROACHES

Network-based - analyze the news source and the propagation pattern of the news in the social network:

- Source credibility analysis;
- User credibility analysis;
- Propagation pattern analysis.

Linguistic-based - analyze the language used in the news article to identify patterns and characteristics that are indicative of fake news:

- Sentiment analysis;
- Linguistic pattern analysis;
- Content-based analysis.

BERT

Bidirectional encoder representations from transformers

- **BERT** was trained on Wikipedia (~2.5B words) and Google's BooksCorpus (~800M words)
- **BERT** is designed to read in both directions at once

We went to the river bank.

I need to go to bank to make a deposit.

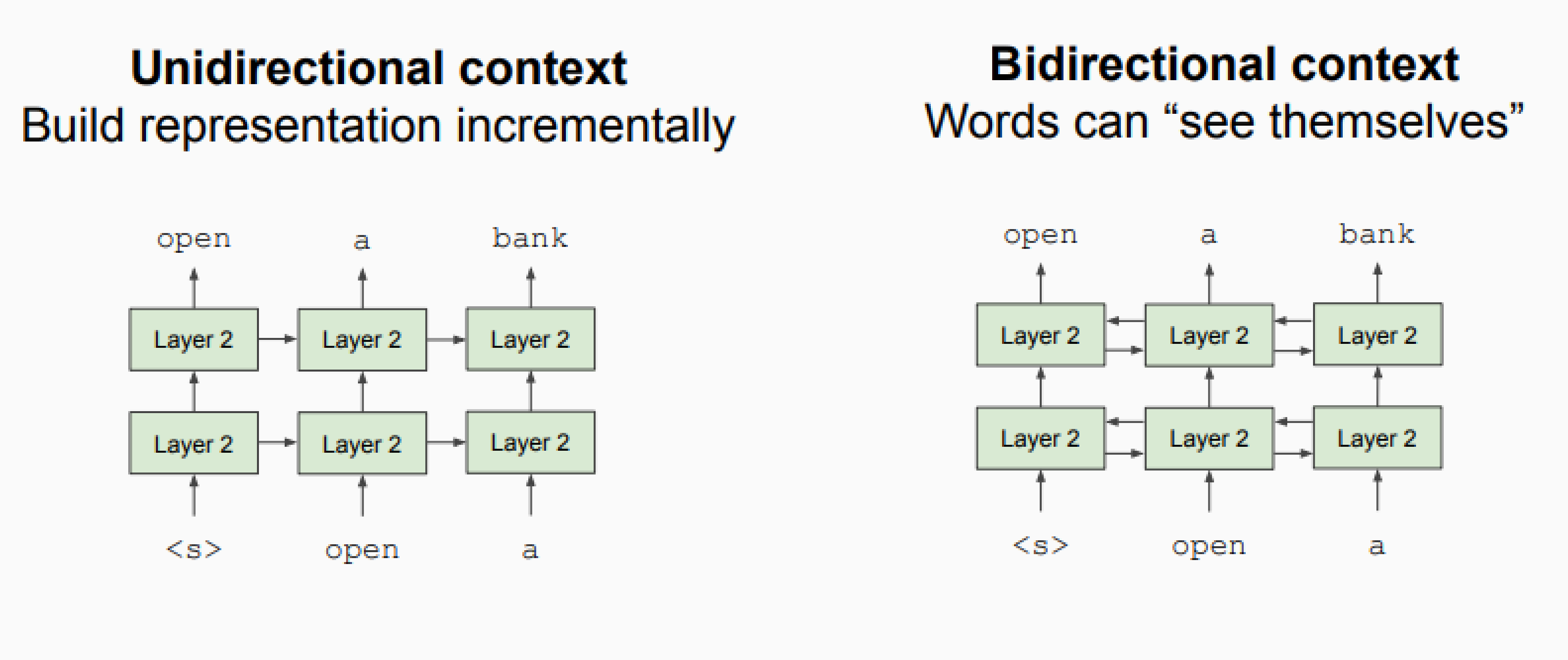


Fig1. Example of bi-directionality

BERT. MASKED LANGUAGE MODEL(1)

Masked Language Model

- MLM enables bidirectional learning from text by masking a word in a sentence and forcing BERT to use the words on either side of the covered word to predict the masked word.
- A random 15% of tokenized words are hidden during training and BERT's job is to correctly predict the hidden words.

"[CLS] my dog [MASK] cute [SEP] he like [MASK] playing [SEP] "

Can you guess the masked words?

Fig2. Example of masking

BERT. MASKED LANGUAGE MODEL(2)

- The model will predict good probabilities for only the [MASK] token.
- During fine-tuning when this model will not get [MASK] as input; the model won't predict good contextual embeddings.

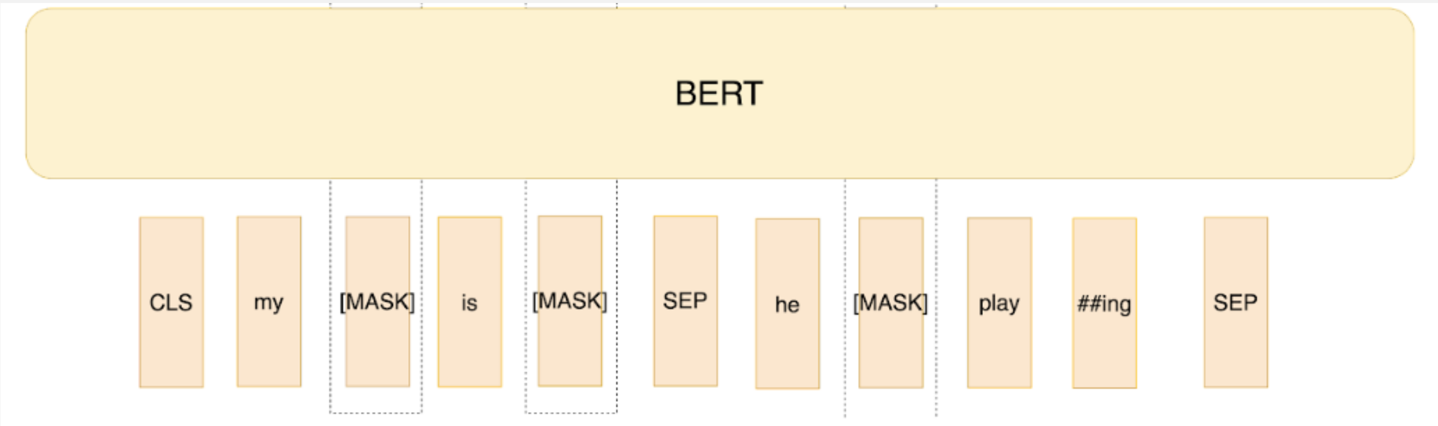


Fig3. Predict only masked words.

- The best setup where model doesn't learn any unnecessary patterns.

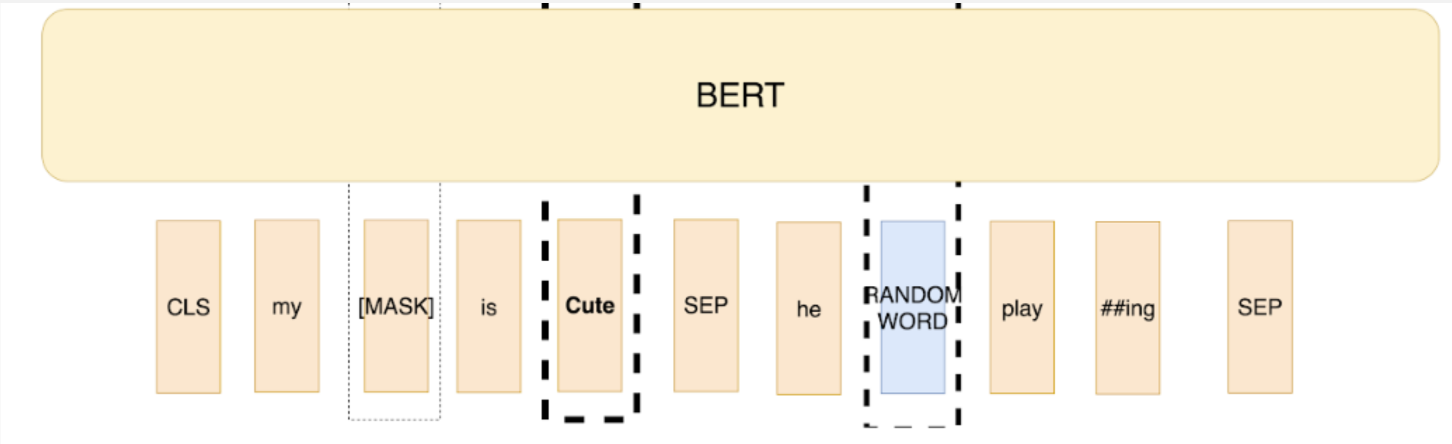


Fig4. Predict masked words, Random Words and Unmasked Words.

BERT. NEXT SENTENCE PREDICTION

Next Sentence Prediction

- NSP (Next Sentence Prediction) is used to help BERT learn about relationships between sentences by predicting if a given sentence follows the previous sentence or not.
- In training, 50% correct sentence pairs are mixed in with 50% random sentence pairs to help BERT increase next sentence prediction accuracy.

BERT is trained on both MLM (50%) and NSP (50%) at the same time.

Input: “[CLS] my dog [MASK] cute [SEP] he like [MASK] playing [SEP] ”

Label: IsNext

Input:“[CLS] my dog [MASK] cute [SEP] he bought a gallon [MASK] milk [SEP] ”

Label: NotNext

Fig5. Next Sentence Prediction Example

BERT LANGUAGE MODEL(I)

The input is processed in the following way before entering the model:

- Insert [CLS] token at the beginning of the first sentence;
- Insert [SEP] token at the end of each sentence;
- A sentence embedding indicating Sentence A or Sentence B is added to each token;
- A positional embedding is added to each token to indicate its position in the sequence;

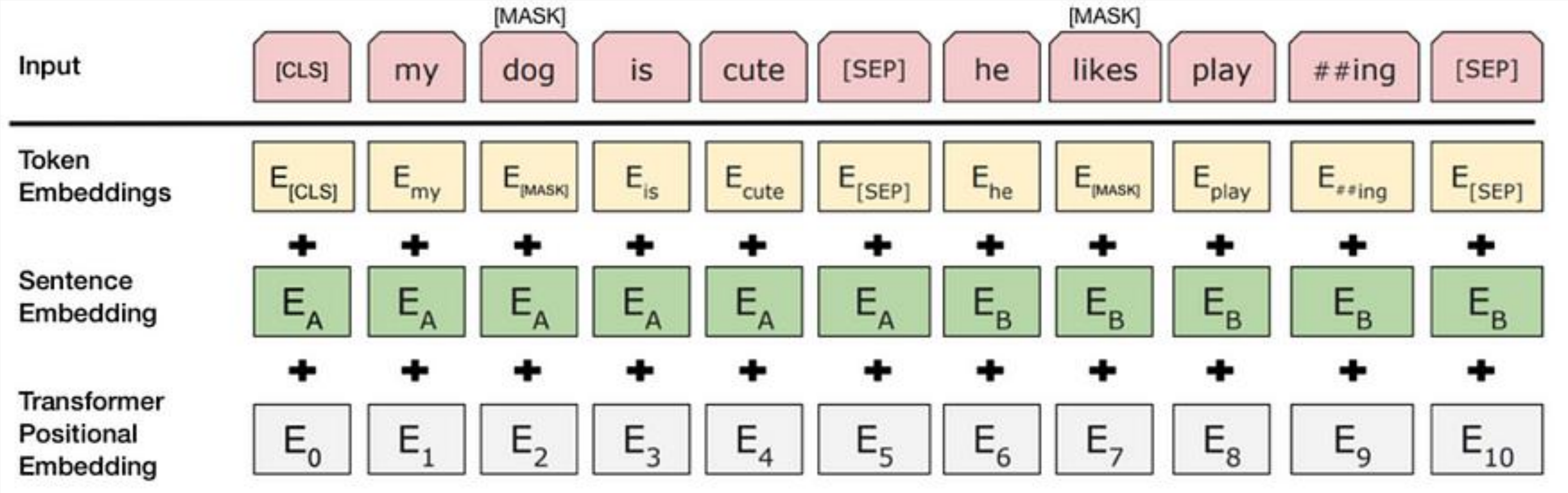


Fig6. BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT LANGUAGE MODEL(2)

- Input - sequence of tokens, embedded into vectors and processed in the neural network;
- The output - sequence of vectors, have same index as input tokens;

- In a well-trained BERT model:
- output vector corresponding to the masked token can show what the original token was
- output of [CLS] token can show if two sentences belong to each other.

- Then, the weights trained in the BERT model can understand the language context well.

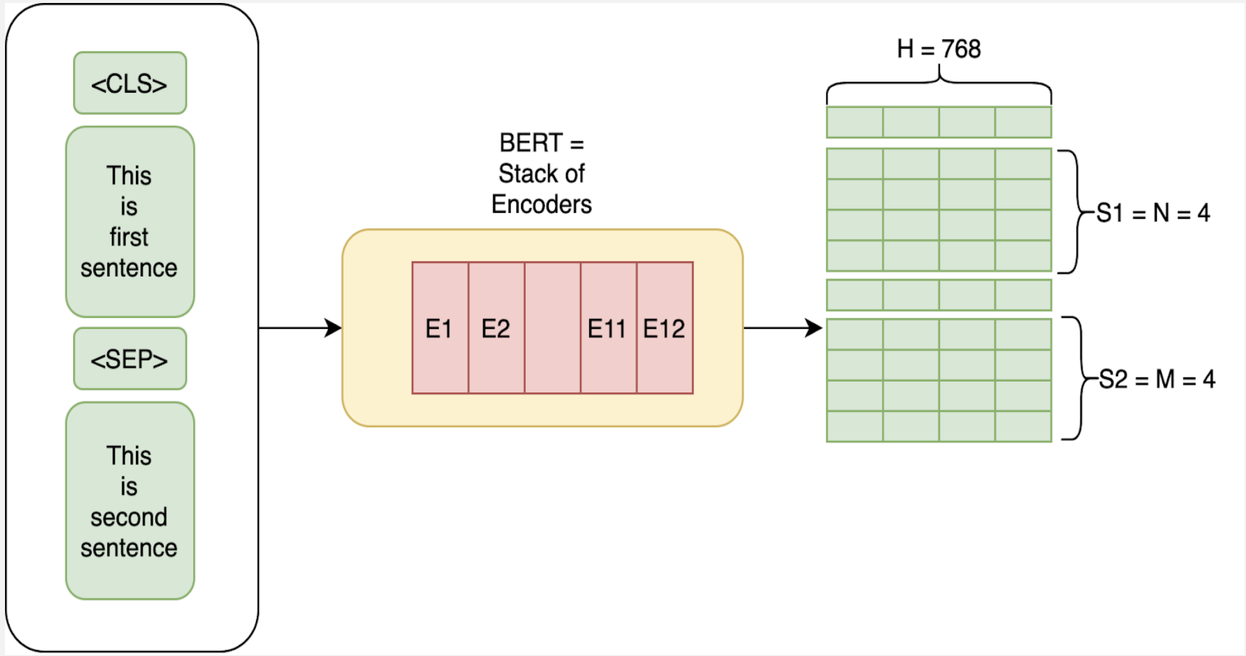


Fig7. High-level description of the Bert encoder.

BERT LANGUAGE MODEL(3)

To predict if the second sentence is connected to the first:

- A simple classification layer on top of encoder output is added in order to classify sentences;
- Calculating the probability of IsNext sentence with softmax.

To detect [MASK] words:

- Classification layer for each encoder layer to detect [MASK] word;
- Transforming vectors into the vocabulary dimension.
- Calculating the probability of each word in the vocabulary with softmax.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Fig8. The softmax formula

z_i - the elements of the input vector for i = 1,…….,K.

ROBERTA

Modifications to BERT:

- **Removing the Next Sentence Prediction (NSP) objective**;
- **Training** on a much larger dataset and using a more effective training procedure;
- **Dynamically changing the masking pattern**.

ELMO. ELECTRA

- **ELECTRA** - instead of masking the input, the approach replaces some input tokens with similar ones.
- The model is trained to predict if token in the input was replaced or is original.

- **ELMo** is a bi-directional LSTM based language model.
- The model is taking into account the entire context of a word in a sentence. It predicts the next word in a sequence given the previous words.

EXPERIMENTAL SETUP

Fine-tuning for the fake news detection task:

- Add classification head on the top of the pre-trained language models;
- Use the respective pre-trained embeddings of the model as the input of the classification head

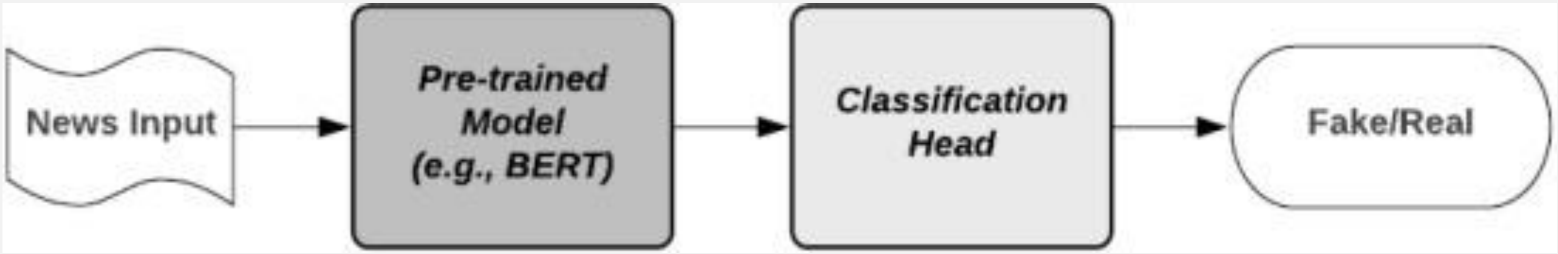


Fig9. Fine-tuning of pre-trained language models.

EVALUATION METRICS

Training and test set for each of the three datasets by splitting it in an 80:20 ratio

- Accuracy - $Accuracy \quad (A) = \frac{TP+TN}{TP+FN+TN+FP}$

- Precision - $P\,(R) = \frac{TP}{TP+FP}, \quad P\,(F) = \frac{TN}{TN+FN}, \quad P = \frac{P(R)+P(F)}{2}$

- Recall – $R\,(R) = \frac{TP}{TP+FN}, \quad R\,(F) = \frac{TN}{TN+FP}, \quad R = \frac{R(R)+R(F)}{2}$

- F1-score – $F1 = \frac{2 \cdot P \cdot R}{P+R}$

R - real news as ‘positive class’, F - fake news as ‘negative class’

Possible concepts of classification: TP - True Positive, FP – False Positive, TN- True Negative, FN - False Negative

STUDIED DATASETS

Dataset	#Total data	#Fake news	#Real news	Avg. length of news articles (in words)	Topic(s)
LIAR	12791	5657	7134	18	Politics
Fake or real news	6335	3164	3171	765	Politics (2016 USA election)
Combined corpus	79548	38859	40689	644	Politics, economy, investigation, health, sports, entertainment

Tab1. Properties of datasets.

DATA PREPROCESSING

Before feeding into the models, texts require some preprocessing:

- Eliminate unnecessary IP and URL addresses from our texts;
- Remove stop words (a, at, , an, another, towards, before);
- Correct the spelling of words;
- Remove suffices from words by stemming them (playing → play + ##ing);
- Convert text data into lowercase letters;
- Remove all symbols from the text data.

STUDIED FEATURES

Used features for traditional machine learning models:

- Lexical - word count, article length, count of parts of speech;
- Sentiment (i.e., positive and negative polarity) of every article;
- Uni-gram and bi-gram features;
- Empath generated features - generate lexical categories from a given text using a small set of seed terms.

EXPERIMENTAL RESULTS

Model type	Model	Rationale for picking	Feature used	Summary of result (Acc.)		
				Liar ~	Fake or real	Combined corpus
Traditional machine learning models	SVM	These traditional models are used in different classification tasks including text classification. Different	Lexical	0.56	0.67	0.71
	SVM		Lexical + Sentiment	0.56	0.66	0.71
	Decision Tree		Lexical + Sentiment	0.51	0.65	0.67
	Naïve Bayes		Unigram	0.60	0.82	0.91
	Naïve Bayes		Bigram	0.60	0.86	0.93
	k-NN		Empath	0.54	0.71	0.71
Advanced pre-trained language models	BERT	These language models are~ pre-trained on large text corpus~ and can be fine-tuned for~ text classification.	BERT embeddings	0.62	0.96	0.95
	RoBERTa		RoBERTa embeddings	0.62	0.98	0.96
	ELECTRA		ELECTRA embeddings	0.61	0.96	0.95
	ELMo		ELMo embeddings	0.61	0.93	0.91

Tab2. Experimental results.

EXPERIMENTAL RESULTS

Model	Datasets											
	Liar				*Fake or real news*				*Combined corpus*			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
BERT	.62	.62	.62	.62	.96	.96	.96	.96	.95	.95	.95	.95
RoBERTa	**.62**	**.63**	**.62**	**.62**	**.98**	**.98**	**.98**	**.98**	**.96**	**.96**	**.96**	**.96**
DistilBERT	.60	.60	.60	.60	.95	.95	.95	.95	.93	.93	.93	.93
ELECTRA	.61	.61	.61	.61	.96	.96	.96	.95	.95	.95	.95	.95
ELMo	.61	.61	.61	.61	.93	.93	.93	.93	.91	.91	.91	.91

Tab3. Experimental results of language models.

THANK YOU FOR ATTENTION