Fake News and Propaganda Detection for Russia's War in Ukraine

Nitin Kulkarni, Chunming Qiao, and Alina Vereshchaka

University at Buffalo
Department of Computer Science and Engineering

University at Buffalo
The State University of New York

Overview

With the world being connected, news can spread around the globe instantly. Often, people believe what they read and propagate the news within their social circles without fact-checking first. Many governments also use news agencies to spread propaganda. E.g., Russia is waging not only a classic war against Ukraine and democracy, but also an information war: using propaganda to spread lies and sow doubts to paint a distorted reality of its war. Thus, it is paramount to identify these cases of disinformation.

Our goal

Distinguish between real and fake news.

What is disinformation?

False information deliberately and often covertly spread (as by the planting of rumors) in order to influence public opinion or obscure the truth. It is the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist.

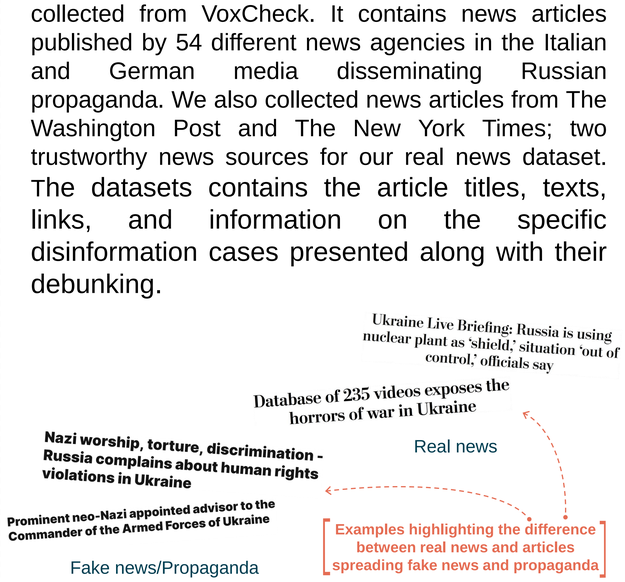
Our approaches

We addressed the problem of fake news and propaganda detection using two Machine Learning (ML) approaches:

1. Word2Vec. It uses word embeddings to transform individual words into a numerical representation. It can capture the semantic relationships, definitions and contexts of words in the text.
2. Language Models. We use Bidirectional Encoder Representations from Transformers (BERT); a language model developed by Google. BERT can accurately capture the context representation in sentences and produces state-of-the-art results on several natural language processing tasks.

Dataset

We use a dataset of news articles pertaining to Russia's war in Ukraine. The fake news dataset was collected from VoxCheck. It contains news articles published by 54 different news agencies in the Italian and German media disseminating Russian propaganda. We also collected news articles from The Washington Post and The New York Times; two trustworthy news sources for our real news dataset. The datasets contains the article titles, texts, links, and information on the specific disinformation cases presented along with their debunking.



- 409 articles
- 983 words per article
- Sources: 54 different news agencies

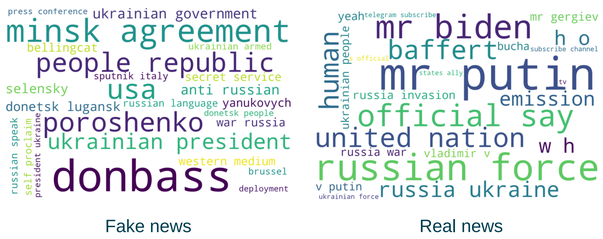
Fake news/Propaganda

- 422 articles
- 964 words per article
- Sources: The Washington Post and The New York Times

Real news

Data Analysis

WORDCLOUDS

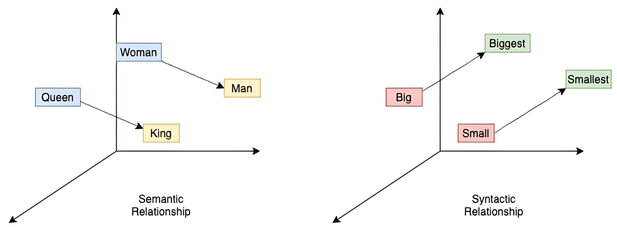


Fake news | Real news

In fake news the words "donbass" and "minsk agreement" are more prevalent, while in real news the words "putin", "united nation" and "Russian force" are prevalent.

Methodologies

Word2Vec



Trained Word2Vec Vectors with Semantic and Syntactic relationship

The effectiveness of Word2Vec comes from its ability to group together vectors of similar words. Given a large enough dataset, Word2Vec can make strong estimates about a word's meaning based on their occurrences in the text. These estimates yield word associations with other words in the corpus. For example, words like "King" and "Queen" would be very similar to one another.

Language Models

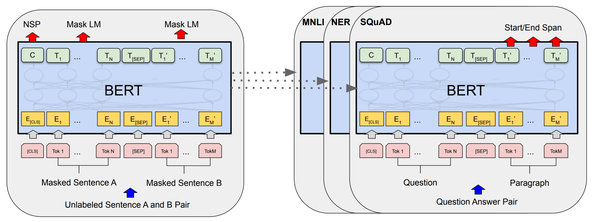


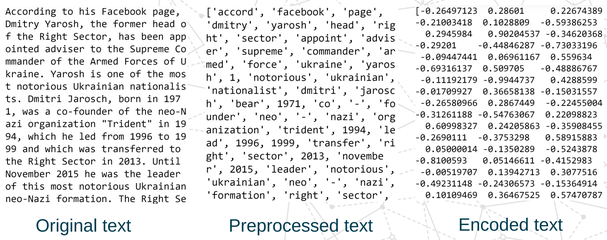
Figure: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training.

Deep Learning

WORD EMBEDDING

We preprocess the data to remove extra whitespaces, any HTML tags, stop words, punctuations, and special characters. We also expand contractions, convert text to lowercase, and do lemmatization and tokenization.

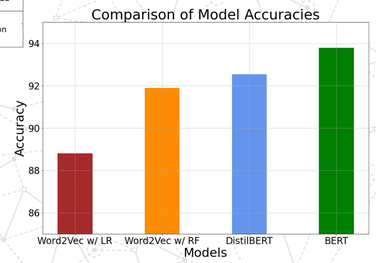


DEEP LEARNING MODELS

We have applied DistilBERT and the full BERT models to do feature extraction and perform binary classification.

Comparison	BERT October 11, 2018	DistilBERT October 2, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 66
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	**Base:** 8 x V100 x 3.5d (4 times less than BERT)
Performance	Outperforming SOTA in Oct 2018	97% of BERT-base's performance on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Transformer, MLM & NSP	BERT Distillation

Hyperparameters for training the models



RESULTS

Test Accuracy

Word2Vec w/ LR: 88.8%
Word2Vec w/ RF: 91.9%
DistilBERT: 92.55%
BERT: 93.79%

Conclusion

We can notice one of characteristics of disinformation is its ideological context. Machine learning methods can help social media companies identify these cases of disinformation and curb their spread. Language models produce state-of-the-art results.

References

Alina Vereshchaka, Seth Cosimini, and Wen Dong. Novel Approaches to Analyzing and Distinguishing Fake and Real News to Mitigate the Problem of Disinformation, 2019

Contact us: {nitinvis, qiao, averesch}@buffalo.edu