Evaluating AI models on Fake News Detection

Dr. Vinicius Woloszyn
April, 2021

Introduction

Outline

- Who am I?
- Context
- Goal
- Projects

What I’ve done about fake news:

- Founder of Untrue.News
- Greenwashing detection in a collaboration with a startup.
- Competition for:
  - Fake news detection
  - Hate speech detection
- Scientific papers regarding
  - Automatic Geenration of ClaimReviews - https://schema.org/ClaimReview
  - Distrust-Rank
  - Hate speech detection

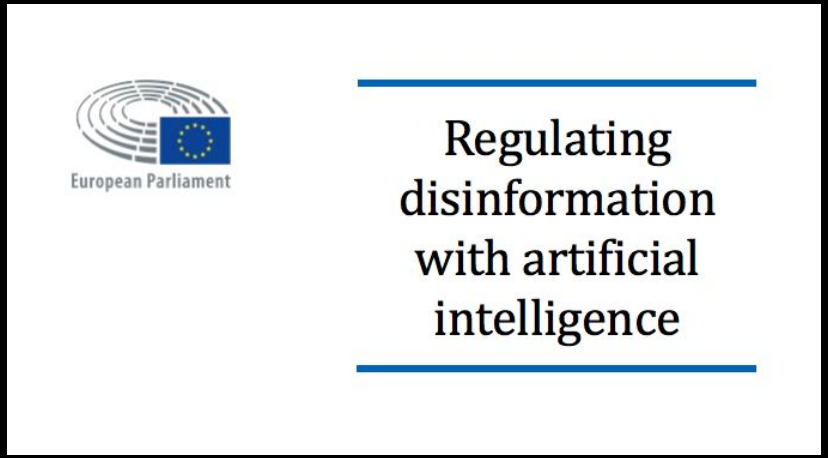
What are the implications of an AI for automatic moderation of content on social media?

Regulating disinformation with artificial intelligence (2019)

Goal: analyze the **implications of artificial intelligence (AI) disinformation initiatives on freedom of expression**, media pluralism and democracy.

The authors **warn against technocentric optimism as a solution to disinformation online**, that proposes use of automated detection, (de)prioritisation, blocking and removal by online intermediaries without human intervention.

When AI is used, it is argued that far more **independent, transparent and effective appeal** and oversight mechanisms are necessary in order to minimise inevitable inaccuracies.



https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf

Alan Turing Institute debate on AI and Fake News (2020)

“Automatic moderation of content on social networks could accelerate a race where AI will be created to counter-attack AI”

EU proposes the first legal framework on AI (2021)

“To promote the development of AI and address the **potential high risks it poses** to safety and fundamental rights equally, the Commission is presenting both a proposal for a regulatory framework on AI and a revised coordinated plan on AI”

EUROPEAN
COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Consensus: there is a fear of using AI in sensible domains without a better understanding (regulation) of implications / downsides.

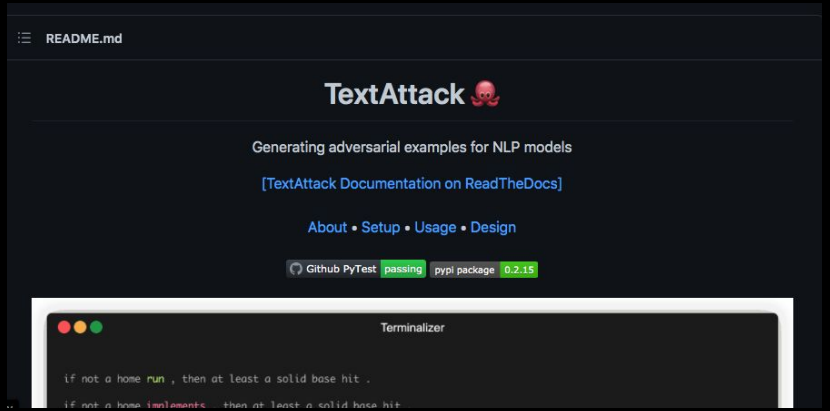
Our goal: understand more about the vulnerabilities and limitations of automatic fact-checking detection for supporting development of a better technology / regulations.

Let’s explore the vulnerabilities and limitations of AI 🤖 💣

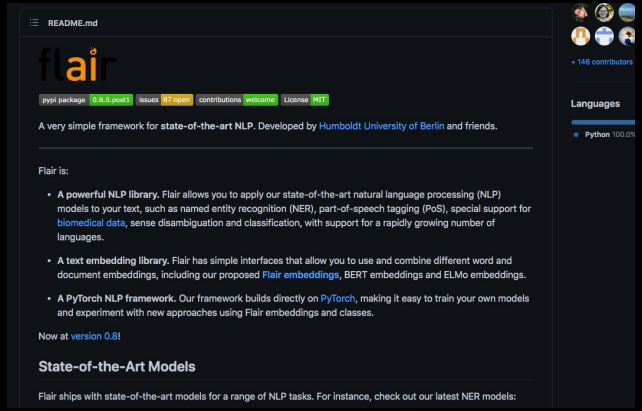
a) How vulnerable is fake news detection to adversarial attacks?
b) How fast models become outdated / generalisation across languages?

HOW VULNERABLE IS FAKE NEWS DETECTION TO ADVERSARIAL ATTACKS?

1) Train a model using flair
   a) Python / Google colab
2) Attack the model using the receipts available at TextAttack
3) Write a scientific paper
   a) Latex / Overleaf
4) Publish at https://arxiv.org/



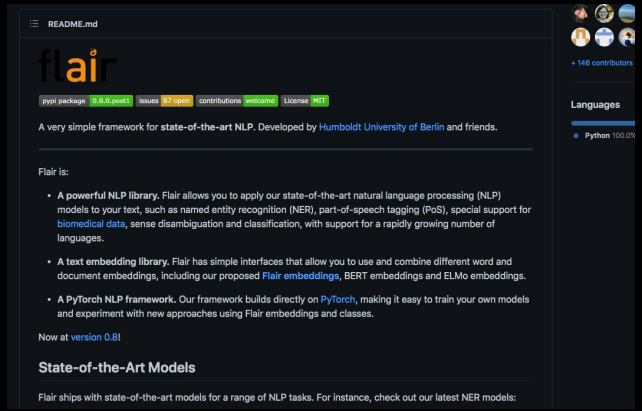
https://github.com/QData/TextAttack



https://github.com/flairNLP/flair

HOW FAST MODELS BECOME OUTDATED / GENERALISATION ACROSS LANGUAGES?

1) Train a model using flair
   a) Python / Google colab
2) Evaluate the model across different domains and timeframes
3) Write a scientific paper
   a) Latex / Overleaf
4) Publish at https://arxiv.org/



https://github.com/flairNLP/flair

FIRST TASK: Train the model

- I’ll provide an example (and data) on how to use flair for training a model.
- You’ll have 2 weeks to create your models (3) and evaluation
- Each groups will present their implementation,
- We will discuss (together) about the process of training an AI

Any question?