chist-era CHIST-ERA Call Development Conference 2021 - Topic 2 - OSNEM #4

Practice area

Practice adding notes to the MURAL board here

Click and drag the post it to move it

Double click to produce a post it note and type on it

Drop the post it note here

Nikos

Sonia

Georgios Hi

Hi!

Double click tag to type your name and affiliation, drag to your location on the map

Tag your name and affiliation

Name, Location

Name, Location

Name, Location

Name, Location

Name, Location

Name, Location

Name, Location

Free Parking

Zoom in and add post it notes

This space is for any ideas you may have that there is no place for, or any feedback you would like to give the organisers

Session I

Double click where you would like to write and strta typing

Brainstorming

Well Sorted Outputs

In the mid-term, new synthetic media generation approaches may be able to create synthetic media streams that consistently represent whole alternate realities. This will pose new risks and call for new detection approaches.

Deciding whether user-generated content is violating certain ToS or community norms is a highly subjective/complex task. For this, human-in-the-loop approaches will be needed for detecting harmful online behavior without putting freedom of speech at risk.

Zoom in and double click to produce a post it note

Long term benefits

Once autonomous agent and media synthesis algorithms reach new capabilities, we expect complete services and suites for creating and managing online armies for different goals (propaganda). Detecting such activities will be crucial for online safety.

Misbehaviors are made possible and sometimes favored by the way the information is structured: attention is sought on the platforms. These dynamics can only change by chaging the platform design, which is consequence of economic incentives.

Resources and Skills

Availability of data from social media providers allowing for the development of methods to detect malicious content

Requirement for interdisciplinary training/education (AI, Computer/ Social Sciences etc)

Methodologies

The issue with fake news is not much their existence, but their (excessive) circulation. How can we leverage an interdisciplinary approach to study these dynamics, including mathematicians, computer scientists, social scientists, engineers?

It is of interest to detect the use of AI-based techniques for the automatic creation and manipulation of visual data (e.g., GANs for image/video generation, motion transfer, video translation, style transfer, automated inpainting)

It is of interest to detect re-sharing of visual media content that has not been produced/acquired natively by the user but has been already shared through the same or other web services.

OSNEM

Gaps

As artificially created intelligent agents are becoming more easy to create it is becoming crucial to be able to recognise such activities as they can easily be used for malicious purposes

Synthetic media generation techniques are getting so good that it will soon be difficult to differentiate it from authentic content. Methods and workflows are urgently needed for detecting inauthentic media content

The simple textual format allow bots to disguise as with simple & empirical rules. Even with very sophisticated models & deep fakes, video-based misbehavior will be possible. OSNEM beyond twitter. How do we detect & address such attacks?

As online content has already exploded intelligent methods are becoming essential for the automatic identification of harmful content, ensuring that this is done reliably so that freedom of speech is not harmed

Poorly false content is clearly a bad thing, but at least we can try to identify it. Instead, there is a broad grey area of low-quality content that are hard to classify. Hence the need to understand and mitigate the circulation of these "fakish" news.

Overarching Themes

Double click where you would like to write and start typing

Overarching Theme

From the points made in Part 1, list the overarching themes that can be identified

1.
2.
3.
4.
5.

Double click the grey space to start writing

Explore the Stregnths, Weaknesses, Opportunities and Threats for the Overarching Themes from the previous session

Session II

Double click the grey space to start writing

Ranking Overarching Themes

From the points made in Session I, rank and add a description to the Theme

1. Need for technological improvement and systematic capturing methods for misbehaviour/fake news/ AI generated text

Specific aspects to be addressed:
- detecting bias/toxicity/hate speech/radical speech/radicalisation
- tools for automated fact checking (fact verification, incl. identifying relevant evidence from multiple sources)
- multilingual dimension (esp. if some languages are less supported - transferring information between different languages)
-Systems level issues

2. Detection of manipulated/ generated content (of various types, including image/video and multimodal media)

- Detection of entirely generated content
- Detection of falsified/manipulated content (as opposed to entirely computer-generated content - different techniques involved)
- Detecting entirely generated text
- Combinations of the above (e.g. generated text + real image or vice versa, or combinations of e.g. genuine text and video which combine to create a false narrative)
- Detecting authenticity based on metadata

3. Understanding the propagation dynamics of harmful/manipulated content

- Interdisciplinary aspects in this research challenge - including social science
- ML to capture differences between propagation of fake news and genuine content
- Data storage aspects (graph processing systems, networks and complexity)

-Influence analysis

Policies on the management of manipulated content without jeopardising freedom of speech

4. Linking/interpreting misbehaviour in the physical and cyber worlds

- Understanding the role / motivation of a certain actor in the cyber world
- classifying online actors based on their behaviour
- Data/behaviour visualisation and interpretation
- Decision support system
- explainable machine learning systems
- blockchain systems that track the authenticity of content

- Policy aspects relevant to all 4 other overarching themes identified
- Regulatory issues

SWOT Analysis

Zoom in and add post it notes

Overarching Theme #1

Availability of open datasets

Semantic issues (e.g. encompassing slang/ figures of speech in different languages)

New techniques for text analysis (e.g. transformers, deep RNNs)

Strength | Weakness

Opportunity | Threat

Availability of open datasets in languages other than English

threat: reluctance of adopting this technology due to lack of explainability

opportunity: creating explainable processes

The development of appropriate Copmputational and open science platforms

unwillingness of data owners- corporations and public organisations to share their data

Overarching Theme

Strength | Weakness

Opportunity | Threat

Overarching themes 2+ 3

existing techniques to identify generated text/ images

Detecting anonymity -

distinguishing genuine from fake actors

Availability of metadata that can be used

Strength | Weakness

Opportunity | Threat

producing synthetic data containing falsified and manipulated contents (e.g., images tied to unrelated text) is doable

Opportunity for interdisciplinarity

Infodemic

opportunity: creating explainable AI

Overarching Theme #4

Strength | Weakness

Opportunity | Threat

opportunity: creating explainable AI

new social norms created by grand challenges (e.g. COVID)

Need to draw insights and predict beyond the expected

Session III

Zoom in and double click to produce a post it note

Maximising Success.....

How can Weaknesses and Threats be addressed

Refer back to the SWOT analysis and discuss how the weaknesses and threats could be mitigated

What is deliverable in a 2-3 year project

What could a CHIST-ERA Project potentially deliver?

Explainable decision-support/ automated fact checking systems, encompassing multilingual and multimodal data aspects. Would be multidisiplinary (e.g. research would include journalists)

Multi-domain fact verification and fake news detection

Utilising blockchain + graph systems analysing metadata of images to assess veracity

Transnational difficulties

What challenges could be faced by a BCI CHIST-ERA project due to its transnational nature