Ofcom – Machine Learning Researcher, Automatic Detection of Memes
TIN-OFC-002

About the Organisation

Whether connecting with loved ones, accessing trusted information, working and educating remotely, or viewing entertaining content, almost every part of our lives is at least now partly online. At the same time there are real risks posed online by harmful or illegal activity - from hate speech to terrorist content, and from bullying and harassment to child sexual exploitation and abuse.

In light of the scale of these challenges, the Government has appointed Ofcom as the regulator for online safety in the UK. Ofcom’s role will be to hold online services to account for the steps they take to protect their users from harmful content, in order to ensure a safer life online for UK internet users.

Role Description and Responsibilities

Background

The rise of social media has led to the spread of memes, which are images or videos that are widely shared and can have a significant impact on public opinion. While memes can be humorous and entertaining, they can also be used to spread harmful or false information, such as hate speech, propaganda, and fake news. In recent years, there has been growing concern about the impact of harmful memes on society, and there is a need for effective tools to detect and combat them.

Although current methods of moderation rely on user reports and content moderation, the sheer volume of content makes it challenging to identify and remove harmful memes in a timely manner. Therefore, artificial intelligence (AI) has the potential to help address this issue by automatically identifying harmful memes and providing insights into their origins and dissemination.

Project: Investigating the Use of AI to Detect and Combat Harmful Memes

The goal of this research is to investigate the use of AI to detect and combat harmful memes and to (if time permits) develop new methods and techniques to improve their effectiveness. The research will be carried out through a combination of literature review and experimental research. Some of the objectives of this research are:

- To conduct a systematic review of existing literature on the use of AI for detecting and combating harmful memes, including the state-of-the-art techniques and their limitations.
• To investigate commercial memes moderating technologies/APIs, evaluate their performance with a view to identifying if they can be used reliably.
• To evaluate the performance of existing solutions on a large dataset of memes, with a focus on accuracy, precision, recall, and generalization ability.
• To explore the potential of AI to provide insights into the spread and impact of harmful memes, including their origin, target audience, and diffusion patterns.

Methodology

The research will be conducted in several stages, including:

1. Literature review: A comprehensive review of existing literature on the use of AI for detecting and combating harmful memes will be conducted, including the latest state-of-the-art techniques and their limitations. The review will evaluate the current state of the art in AI technology for meme detection and analysis. The ethical and social implications of using AI to combat harmful memes will also be evaluated. Additionally, this review will consider the limitations and challenges of using AI in this context, such as issues related to bias, accuracy, and scalability. The review will also investigate whether there is off-the-shelf (commercial) moderation software that platforms are currently using. The outcome of the review will be presented in the form of a written document.

2. Dataset preparation: To investigate the use of AI to detect and combat harmful memes, an essential step is to prepare a suitable dataset. There are several open datasets available for meme detection, such as the Hateful Memes Dataset, MemeTracker Dataset, and Dank Learning Dataset. The dataset will then be prepared to a suitable state. This will start with pre-processing the images, and then splitting them into training, validation, and test sets.

3. Experimental research: The experimental research will be carried out in two phases. First, the effectiveness and reliability of existing AI technologies that can be used to transcribe and understand memes, this will include commercial technologies identified during literature review and other open-source technologies such as pytesseract, easyocr, and KERAS-ocr, will be evaluated. Second, if time permits, an AI model for meme detection and analysis will be developed and tested. The model will be trained on a dataset of known harmful memes and compared to a dataset of benign memes. The accuracy and effectiveness of the model will be assessed through a series of experiments, including both quantitative and qualitative analyses. Furthermore, the limitations and challenges of implementing the model in real-world scenarios, such as the scalability and adaptability of the model to different types of memes and contexts, will be investigated.

4. Insights and impact analysis: The potential of AI to provide insights into the spread and impact of harmful memes will be explored, including their origin, target audience, and diffusion patterns.

The intern, therefore, will be responsible for:

- Conducting research and literature review on state-of-the-art methods for meme detection
- Gathering and pre-process large-scale datasets of memes
- Developing and testing algorithms for classification and clustering of memes
- Communicating fundings and results through a report and presentation to colleagues.

**Expected Outcomes**

The investigation is expected to provide an understanding of the technologies currently available for understanding and classifying memes, their effectiveness, and whether social media platforms are using them to moderate content. Secondly, the research will contribute to the development of AI technology for detecting and combating harmful memes. Additionally, it will increase the understanding of the ethical and social implications of using AI in this context. The research will identify limitations and challenges associated with using AI in meme detection and analysis and provide recommendations for future research and development.

**Supervision and Mentorship**

The role holder will have a Supervisory Manager assigned who will oversee the day-to-day work. Ofcom also offers a voluntary coaching and mentoring programme.

**Person Specification**

We are looking for an enthusiastic Machine Learning researcher to research into automatic detection of memes. The successful candidate will work independently and collaboratively with a team of experts in the field of computer vision and natural language processing.

The ideal intern will possess:

**Essential**

- Proven experience in developing and validating machine learning models for image and text analysis;
- Knowledge of deep learning techniques and algorithms;
- Experience with popular machine learning libraries such as Tensorflow, PyTorch, and Keras;
- Familiarity with computer vision and natural language processing techniques;
- Problem-solving and analytical skills;
- Strong written and verbal communication skills;
- Ability to work independently and as part of a team;
- Excellent collaboration skills, with the ability to work effectively in a team environment.

**Preferred**

- Experience with meme culture and understanding of humour.

If you are a researcher having ML experience with a passion for developing innovative solutions and want to work on a challenging project in the field of computer vision and natural language processing, we encourage you to apply!
Internship Logistics

Start Date: September 2023

Salary: £30,000 per annum pro rata

Internship Duration: 6 months, full time or part time (International students on a Student Visa are welcome to apply). Flexible working hours are also available.

Location: The job is based on a hybrid remote setup, with the expectation of visiting Manchester at least once per week.