HUMANITIES & SOCIAL CHANGE

Centre at the University of Cambridge

UNIVERSITY OF CAMBRIDGE

A report prepared for The New Institute

APRIL 2020

CRASSH

HUMANITIES & SOCIAL CHANGE
Centre at the University of Cambridge

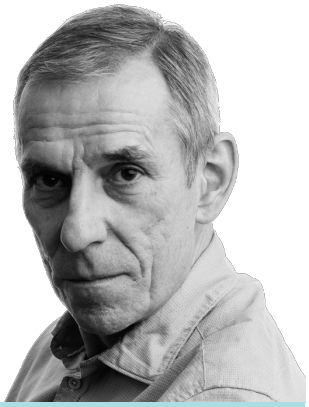
The University of Cambridge extends its sincere thanks to The New Institute for the donation of £2,002,000 in 2017 to support the Humanities and Social Change Centre.

The creation of the Centre has accelerated the ability of our researchers to contribute humanities research to policy and practice at the interface between technology and social change.

The eight sub-projects of Expertise Under Pressure examine the contexts in which expertise is received and utilised, in order to establish a framework for better engagement with expertise. Giving Voice to Digital Democracies is exploring the many ways in which technology can contribute to, and detract from, participation in democracy, from artificial intelligence to fake news. From workshops and presentations, to publications and teaching, the programme leads are ensuring their findings are widely disseminated for maximum impact.

This report contains an update from CRASSH Director Professor Steven Connor, and progress reports from project leads Dr Anna Alexandrova (Expertise Under Pressure) and Dr Marcus Tomalin (Giving Voice to Digital Democracies).

An introduction from **Professor Steven Connor** Director of CRASSH



Professor Steven Connor

Director of CRASSH

Grace 2 Professor of English at the University of Cambridge

Fellow of Peterhouse

I am delighted to be able to present in what follows the evidence of how our two Humanities and Social Change projects have both diversified and gathered impetus over the last 12 months.

Expertise Under Pressure has expanded dramatically from its original three case-studies to eight sub-projects. Beyond its programme of workshops and seminars, Giving Voice to Digital Democracies is developing some remarkable initiatives in the teaching of postgraduate students. Both of these project-clusters have kept engagement and outreach at the forefront of their activity, ensuring that their thinking is brought, in something approaching real-time, to the attention of wide and varied audiences beyond the academic world. Publications by our researchers and postdocs are beginning to appear in high-profile and influential publications. Though the work of both teams has been disrupted by the COVID-19 pandemic, plans are being made for ways in which their work can continue in new and vigorous forms under the social and academic conditions that will emerge over the coming months.

We are very delighted by the rapid progress that has been made: the research of the Centre for Humanities and Social Change has become a beacon for other researchers in CRASSH and the University of Cambridge.

Expertise Under Pressure (EuP)

The project is now in full swing with activities having grown from the original three case studies to eight distinct sub-projects, all concerning areas in which experts and expertise are both necessary and controversial. The subprojects are pursued by overlapping groups of members of EuP and their external partners with regular joint activities involving all and an intense programme of outreach to key stakeholders outside academia.

Dr Anna Alexandrova

Director of Expertise Under Pressure

Reader in Philosophy of Science at the University of Cambridge

Fellow of King's College, Cambridge

Agglomeration Economics (Mike Kenny and Cléo Chassonnery-Zaïgouche)

This project concerns the intellectual history of agglomeration economics and the relationship between economics and urban policy since the 1970s, with the goal of understanding what makes expert knowledge popular and sought-after.

Experts on Trial (Cléo Chassonnery-Zaïgouche and Federico Brandmayr)

This project examines the distinct challenges when courts call on experts to speak on cases they are hearing, especially when these cases involve new areas of knowledge such as discrimination by gender and attribution of responsibility for disasters.

Perception of Heritage (Hannah Baker)

This project builds upon the research Baker completed for her PhD which assessed the decision to demolish or adapt existing buildings on masterplan sites. It explores how experts, such as national heritage bodies, perceive heritage in comparison to laypersons and community groups.

Rapid Decisions Under Risk (Emily So, Hannah Baker, Rob Doubleday)

This project focuses on the use of experts in rapid decisions under risk, such as earthquakes, volcanoes, or topically, potential epidemics. The goal is to query the role of scientific expertise in disaster management under time pressures, high stakes and uncertainty and the potential for inclusion of laypersons in the networks of knowledge.

Sociological Excuses (Federico Brandmayr)

This project examines how sociologists are called on to explain human behaviour, especially crime and terrorism, but their explanations are often taken as excusing action and hence their perceived objectivity is compromised. The project traces the history of this phenomenon in the 20th century and examines high-profile contemporary case studies in France and Italy.

Exceptionalism in Social Sciences (Anna Alexandrova, Federico Brandmayr, with Martin Kusch)

This project examines intellectual and social history of the distinction between social and natural knowledge with the goal of ultimately shedding light on how contemporary social science can inform research into large-scale challenges such as climate change and inequality.

Well-being Policy and Technocracy (Anna Alexandrova)

New models of economy will require new modes of measurement of social progress, and well-being is emerging as a key alternative to traditional indicators such as GDP. However, some methods of measuring well-being and using this knowledge for policy raise the danger of technocracy.

Citizen Science (Rob Doubleday, Anna Alexandrova, Hannah Baker, with Katie Cohen)

Scientists and policymakers increasingly urge public engagement with science, as a way of addressing the crisis of expertise and as a way of making scientific research more relevant to practical concerns. Which of its many models (citizen science, deliberative polling, mini-publics, etc) should we pursue and how?

Events held

1) Workshop "When Does Explaining Become Explaining Away? Compassion, Justification and Exculpation in Social Research", held at CRASSH on 27 September 2019. The programme, photos and a blog post can be viewed at: https://hscif.org/when-does-explaining-become-explaining-away/

2) Workshop 'Disaster Response – Knowledge Domains and Information Flows' - 11 February 2020. For the programme, topics, and participants, see:
https://hscif.org/disaster-response-knowledge-domains-and-information-flows/

3) Regular internal seminar series: Expert Bites. Designed to cover perspectives on expertise from a wide range of disciplines and projects.

- Lisa Stampnitzky (Politics, University of Sheffield), 29 January 2020
- Bill Byrne (Information Engineering, University of Cambridge), 7 November 2019
- Alice Vadrot (Political Science, University of Vienna), 21 June 2019
- Arsenii Khitrov (Sociology, University of Cambridge), 22 May 2019

Participants at the Disaster Response workshop

- Elizabeth Anderson (Philosophy, University of Michigan), 15 May 2019
- Mike Hulme (Geography, University of Cambridge), 26 March 2019
- Alfred Moore (Politics, University of York), 28 November 2018

4) Regular public seminar series:

The Politics of Economics:

- 28 January 2020 / James Forder (University of Oxford) on Theoretical expertise and the weaponising of the Phillips curve, 1970-1977
- 11 February 2020 / Hilary Cooper (Consultant Economist) & Simon Szreter (Cambridge) on Incentivising an ethical economics
- 14 October 2019 / Caitlin Zaloom (New York University) on Indebted: Student finance, social speculation, and the future of the US family
- 29 October 2019 / Diane Coyle (Cambridge) on Economics for the Digital Age?
- 12 November 2019 / Steven Medema (Duke University, USA) & David Gindis (University of Hertfordshire) on The Politics of Law and Economics
  https://hscif.org/portfolio/seminar-the-politics-of-economics/

5) Weekly reading group, Calculating People, meets as part of Brandmayr and Alexandrova's subprojects and is attended by members of sister projects at CRASSH and scholars from all over Cambridge research units: https://www.hps.cam.ac.uk/news-events/seminars-reading-groups/calculating-people

Forthcoming events

Dr Doubleday (with Katie Cohen) is organising a conference "Innovations in Citizen Science for Public Policy" on 24-25 March in Cambridge at the Centre for Science and Policy (CSaP). The event will bring together key scholars and practitioners of public engagement of science all over the world, Alexandrova will offer remarks on how to involve public in measurement of well-being.

The workshop "Economists in the city: exploring the history of urban policy expertise" will take place on 11 May. The workshop will explore when and why did the expertise and knowledge of economists become so highly valued in the world of public policy, especially urban policy in the US, France and the UK. It also examines other social science perspectives upon cities, and evaluates the efforts of a number of leading economists in the last 20 years to develop a more spatially aware body of thinking. The list of speakers and the programme can be seen here:
http://www.crassh.cam.ac.uk/events/28891

International conference: "Are Social Sciences Special?" September 17-18 2020 http://www.crassh.cam.ac.uk/events/28892. This conference is a joint venture between the 'Expertise Under Pressure' project and the 'Philosophy of Science and Epistemology' group at the University of Vienna. The conference will explore the history of 'exceptionalism' (the idea that the social sciences operate by fundamentally different rules to the natural sciences), its contemporary status, and its role in today's social sciences. 11 speakers are confirmed with the closing lecture to be delivered by Rahel Jaeggi (Humboldt University of Berlin) the Head of our sister centre in Berlin.

Publications

- Chassonnery-Zaïgouche, Cléo. 2020. Economists Entered the 'Number Games'. The Early Reception of Wage Decomposition Methods in the U.S. Courtrooms (1971-1989). Forthcoming in *The Journal of the History of Economic Thought*.
- Chassonnery-Zaïgouche, C., 2020. Introduction to the symposium "Economists in Courts". Forthcoming in *The Journal of the History of Economic Thought*.
- Chassonnery-Zaïgouche, C., Cherrier, B., and Singleton, J. 2020. 'Out in the open' controversy: Economists' perspectives on the first gender reckoning in economics. In Lundberg, S., ed, *Women in Economics, CEPR*.

- Singh, R. and Alexandrova, A (forthcoming)"Happiness economics and Technocracy" *Behavioral Science and Policy*. https://www.cambridge.org/core/journals/behavioural-public-policy/article/happiness-economics-as-technocracy/
- Brandmayr, F. 'Public Epistemologies and Intellectual Interventions in Contemporary Italy', *International Journal of Politics, Culture, and Society* (2019): pp. 1-22. (In open access).
- Brandmayr, F. Review of Josh Booth and Patrick Baert, "The Dark Side of Podemos?", *The Sociological Review* (24 June 2019).

External Talks, Lectures, and Presentations

- Chassonnery-Zaïgouche, C., "Programming Expertise: The Political Element in Bergmann's Micro-Simulations (1971-1991)" at The History of Economic Thought Society annual conference, at Goldsmiths University, London, 19 September 2019.
- Chassonnery-Zaïgouche, C., "'Economics is not a men's field': A history of the American Economic Association's Committee on the Status of Women in the Economic Profession", (paper written with Cherrier, B., and Singleton J.) at the Allied Social Science Associations (ASSA) meeting, American Economic Association, San Diego, 3-5 January 2020.
- Brandmayr, F. "Nothing but 'stimulating metaphysical theories'? Cultural expertise in the L'Aquila trial", Centre for Socio-Legal Studies seminar, University of Oxford, 30 January 2020.
- Brandmayr, F. "The Political Epistemology of Explanation in Contemporary French Social Thought", "When does explaining become explaining away?" Workshop, CRASSH, Cambridge, 27 September 2019.
- Brandmayr, F. "Le classement politique des théorie sociologiques de moyenne portée: le cas de la 'prophétie auto réalisatrice' dans les premiers procès états-uniens de déségrégation raciale (1948-1955)", 8th Congress of the Association Française Sociologie, Aix-En-Provence 27-30 June 2019.
- Brandmayr, F. "Entre faits et valeurs : les limites épistémologiques dans l'étude du politique". Guest lecture at EHESS, Paris, 21 February 2019.
- Alexandrova, A. and Mitchell, P., "Pluralism in the science of well-being" The International Society for Quality-of-Life Studies (ISQOLS) 17th Conference (September 2019), Granada, Spain.
- Alexandrova, A., "On the definitions of social science and why they matter" Congress for Logic, Methodology, Philosophy of Science and Technology, Prague, August 2019.
- Alexandrova, A., "Happiness and Technocracy" at:
  - ~ Oxford Internet Institute, July 2019
  - ~ Royal Institute of Philosophy Series, Exeter University, November 2019
  - ~ Copenhagen Philosophy Seminar, December 2019
  - ~ University of Warwick, Leverhulme Bridges Programme Seminar, January 2020.

Impact and public engagement

The project maintains a lively Twitter feed: https://twitter.com/ExpertiseUnder

As do other members of the project:

https://twitter.com/CleoCZ

https://twitter.com/bakes_hannah

https://twitter.com/robdoubleday

https://twitter.com/michaelkenny_

Professor Michael Kenny has been invited to deliver several talks to officials in UK government in his capacity as an expert on issues of place, identity and governance, and also to share some of the findings of our work on the policy and analytical implications of agglomeration economics. He spoke to officials in UK government working in the Departments of Housing, Communities and Local

Government, Business, Economics and Industrial Strategy, and the Foreign and Commonwealth Office.

Professor Kenny also spoke at a private seminar organised by Sir Mark Sedwill, the UK's most senior civil servant, on 'Constitutional Expertise and the Future of the British state', at a symposium on 'British Irish Relations' at University College, Dublin, and at various academic workshops in the UK held on the themes of 'place', left-behind communities, and the UK government's 'levelling up' agenda.

Finally, Professor Kenny has co-authored a series of policy papers analysing the economic trajectories and policy challenges facing the UK's towns, analysing the role of agglomerationist thinking upon decisions about infrastructure funding in particular. These are published within the new 'Townscapes' series at the Bennett Institute for Public Policy:
https://www.bennettinstitute.cam.ac.uk/research/research-projects/townscapes-project/

Dr Chassonnery-Zaïgouche's research on the role of economic expertise on gender (presented at the American Economic Association in January) has been reviewed in *The Economist*: "Economists are discussing their lack of diversity", 9 January 2020.

Dr Brandmayr has written blog posts for the CRASSH and HSCIF websites:

https://hscif.org/when-does-explaining-become-explaining-away/

http://www.crassh.cam.ac.uk/blog/post/when-does-explaining-become-explaining-away

Dr Baker has participated in events hosted by the Centre for Science and Policy (CSaP). During many of these she presented the project which has helped to help create a network of people in policy working in the field of emergency/disaster response. These include: CSaP's annual lecture – Dame Sally Davies (11 February 2020); CSaP's 10 Year Anniversary Reception (17 October 2019); CSaP Workshop: The role of oral histories in understanding science-policy interrelations (25 June 2019); CSaP forum (8 May 2019). She has also participated in several public consultations about building, development, and sustainability projects in Cambridge. Finally she hosted Fodé Beaudet, a Senior Learning Advisor at the Centre for Intercultural Learning with Global Affairs Canada (GAC), who visited the UK as part of a project to better understand how we can design, facilitate and evaluate our work to support behavioural change at the individual, group or system level. A related blog post can be found at: https://hscif.org/expertise-adult-learning-intercultural-effectiveness/

Dr Alexandrova has advised various organisations and charities (including the Institute for Electrical and Electronic Engineers' Ethical Design Initiative) on how well-being outcomes should be incorporated into evaluation of policy and technologies. She recorded a podcast on expertise in social sciences "Philosophy in Three Words with Adrian Currie" for the Royal Institute of Philosophy Podcast, November 2019:
https://www.mixcloud.com/adrian_currie/episode-4-anna-alexandrova-cambridge-social-value-expertise/

Plans for the future

Professor Kenny is working on a paper for a special issue of *Regional Studies* on the economic geography of political disenchantment across central and western Europe, which includes a critique of the impact of agglomerationist policy-making, with Dr Davide Luca (Cambridge, Land Economy). He and Dr Chassonery-Zaigouche are also working on a paper examining the contours and influence of agglomerationist economics. They are also planning a policy paper arising from this research.

Dr Chassonery-Zaigouche is preparing a new paper on experts in courts (on the history of the comparable worth controversy) will be presented in June in two conferences and submitted this summer.

The three postdocs are working to organise an additional workshop in Fall 2020 on experts in courts focused on various case studies of trials that involved a substantial amount of expertise.

Dr Brandmayr is planning with past presidents of all the major social science associations (BSA, RAI, BSC, RHS, AFS in France) about public perceptions of and political challenges to social research. (Excerpts to be published on the HSCIF website.) He is editing a special issue titled "When Does Explaining Become Explaining Away?", the proposal for which was enthusiastically accepted by the *European Journal of Social Theory*. He is involved in the "Scienza in Parlamento" Italian project and in the organisation of a summer school for civil servants, MPs and researchers on the public understanding of science (PUS), science communication and other science and technology studies issues.

Dr Baker will work with CSaP to organise a policy forum in 2020. This will include town planning experts, heritage organisations and local community activists for heritage protection. She will publish a paper based on this, building upon her PhD research and tying in with the theme of expertise and heritage. She is also organising an event, "I Built That!", a celebration of women in construction and real estate" – 18 March 2020. This is part of Cambridge University Library's, *The Rising Tide: Women at Cambridge* exhibition. The event will include a panel discussion with women working in construction, followed by a poster and networking session. There will be approximately 80 attendees and Hannah will use the opportunity to continue raising awareness of the project and build upon her network.

The organisation of the Disaster Response workshop and attendance at other events has raised awareness about the project and helped to form collaborations. One of the speakers, Professor Dorothea Hilhorst, plans to visit Cambridge in September 2020 to collaborate with the project team to jointly update her paper: Hilhorst, D., "Responding to Disasters: Diversity of Bureaucrats, Technocrats and Local People", *International Journal of Mass Emergencies and Disasters*, 21 (1) p.37-55 (2003). Other collaborations have been formed with (but are not limited to): Dr Emma Doyle, Massey University, New Zealand, a key academic in science advice and disaster response; Dame Sally Davies, the former Chief Medical Officer for England; and Josh Macabua, a search and rescue specialist and consultant to the World Bank.

Drs So and Baker are putting together a paper cross-comparing scientific advisory groups in different countries. The aim of this is to gauge what is best practice and form a foundation of information about what already happens in terms of scientific advice and disaster response, in order to make a critique later in the project. They also plan to host a policy forum in collaboration with CSaP with a focus on the Coronavirus and the role of experts in coordinating the government response.

Dr Alexandrova will be a keynote speaker at the international conference "Measurement at the Crossroads 2020" that will be held at the Catholic University of Milan, Italy (July 2020), and at the Third Biannual Conference of the Eastern European Network for Philosophy of Science 2020 that will take place in Belgrade next year on 10-12 June 2020. Both addresses will report on the sub-projects of EuP led by Alexandrova, promoting the work of the Cambridge Centre to two large audiences.

Giving Voice to Digital Democracies

This section does not provide a completely exhaustive list of all the group's activities, but it gives information about the most significant project-related work.

Dr Marcus Tomalin

Project Manager of Giving Voice to Digital Democracies

Senior Research Associate

Director of Studies for Philosophy at Downing College

1. Personnel

The latest member of the Giving Voice to Digital Democracies (GVDD) team, Dr Shauna Concannon, joined the group in early June 2019. She is an interdisciplinary researcher with interests in experimental psychology, computer science, and linguistics. She completed her PhD, 'Taking a Stance: Experimenting with Disagreement', in the Computational Linguistics Lab and Cognitive Science Research Group at Queen Mary University of London. Her research interests focus on interactional accounts of how informational and opinionated content is communicated. In particular, she is interested in how epistemicity and evidentiality are linguistically encoded, and uses this to understand how information is processed and circulated in an increasingly technologically mediated society.

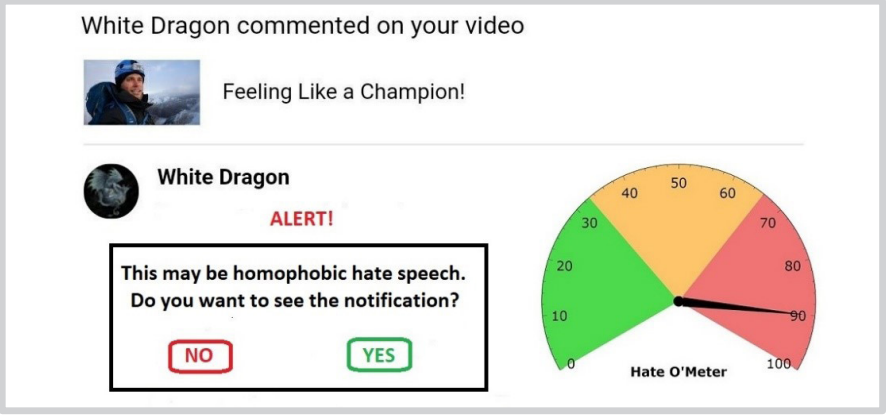
2. Publications

The research undertaken by members of the GVDD team has started to appear in influential publications. For instance, Marcus Tomalin and Stefanie Ullmann wrote an opinion piece for *The Conversation* in October 2019 that focused on the problem of AI systems reinforcing existing social biases:

https://theconversation.com/ai-could-be-a-force-for-good-but-were-currently-heading-for-a-darker-future-124941

More importantly, though, a journal article, "Quarantining online hate speech: technical and ethical perspectives", authored by Stefanie Ullmann and Marcus Tomalin, was published in *Ethics and Information Technology* in November 2019: https://link.springer.com/article/10.1007/s10676-019-09516-z

This article had an immediate impact and has already been downloaded more than 4,000 times. It introduces an innovative framework for handling hate speech online that enables potentially offensive material to be quarantined. More specifically, it describes how an automated hate speech detection system can analyse each post and trigger a warning if the confidence score for the hate speech category exceeds a specified threshold:



While preserving freedom of expression, this framework allows potential victims to decide whether or not to read offensive messages. It thereby achieves a balance between libertarian and authoritarian ideological approaches to censorship.

This research was featured in an article that appeared on the Cambridge University website:

https://www.cam.ac.uk/research/news/online-hate-speech-could-be-contained-like-a-computer-virus-say-researchers

It also received considerable attention from IT-related specialist publications, and from the mainstream media both in the UK and in Europe:

https://www.campusreform.org/?ID=14149

https://www.technologynetworks.com/informatics/news/could-we-contain-hate-speech-like-a-computer-virus-328713

https://www.eurekalert.org/pub_releases/2019-12/uoc-ohs121719.php

https://eandt.theiet.org/content/articles/2019/12/online-hate-speech-could-be-contained-like-a-computer-virus/

https://www.breitbart.com/tech/2020/01/02/cambridge-researchers-launch-hate-ometer-software-to-treat-hate-speech-as-malware/

https://www.scitecheuropa.eu/can-ai-help-quarantine-hate-speech/98996/

https://neurosciencenews.com/ai-contain-hate-speech-15397/

https://inews.co.uk/news/technology/online-hate-speech-should-be-quarantined-like-pc-viruses-experts-argue-1344651

https://eandt.theiet.org/content/articles/2019/12/online-hate-speech-could-be-contained-like-a-computer-virus

Stefanie Ullmann was interviewed about this research for the BBC world service programme 'Digital Planet' and the German radio programme 'Computer und Kommunikation':

https://www.bbc.co.uk/sounds/play/w3csy676

https://www.deutschlandfunk.de/hass-im-internet-es-koennte-eine-quarantaene-fuer.684.de.html?dram:article_id=467684

Needless to say, we are delighted that our work has had such a significant impact so swiftly, and we are confident that some of our forthcoming publications will be equally well received. In particular, a journal article entitled 'Rethinking Bias Reduction in AI Systems: The Practical Ethics of Adaptation' has been submitted to *Ethics and Information Technology*. If accepted, it should appear later this year.

3. Talks and Events

There have been many opportunities for the members of the GVDD team to disseminate their research. In particular, in May 2019, the second in our series of workshops took place in Cambridge, and the topic was 'The Future of Artificial Intelligence: Language, Gender, Technology'. This event considered the social impact of Artificially Intelligent Communications Technology (AICT), and, as the title suggests, the talks and discussions focused on different aspects of the complex relationships between language, gender, and technology. These issues are of particular relevance in an age when:

- virtual personal assistants such as Siri, Cortana, and Alexa present themselves as submissive females
- most language-based technologies manifest glaring gender biases
- 78% of the experts developing AI systems are male
- sexist hate speech online is a widely recognised problem
- many Western cultures and societies are increasingly recognising the significance of non-binary gender identities

Like the first workshop, this one was massively oversubscribed. The structure of the event consisted of short talks by various speakers, including a generous amount of time for Q&A sessions so that the audience members could explore some of the ideas introduced. The speakers were:

- Professor Alison Adam (Sheffield Hallam University)
- Dr Heather Burnett (CNRS-Université Diderot Paris)
- Dr Dirk Hovy (Bocconi University)

- Dr Dong Nguyen (Alan Turing Institute/University of Utrecht)
- Dr Ruth Page (University of Birmingham)
- Dr Stefanie Ullmann (University of Cambridge)

The feedback from speakers and attendees alike was uniformly positive, and further information about the day (including videos of the main talks) can be found here:

http://www.crassh.cam.ac.uk/events/28480

In June 2019, Marcus Tomalin chaired a panel discussion about 'Technology and Society' at the Royal Society in London. This prestigious event formed part of the Centre for Science and Policy's annual conference. The session was well attended and the panellists were Dr Leonie Tanczer (Department of Science, Technology, Engineering and Public Policy, UCL), David Knight (Department for Digital, Culture, Media and Sport), and Vinous Ali (Tech UK). The discussion ranged widely over numerous AI-related topics, and there was a Q&A session with the audience. An audio recording of the event is available online.

The third GVDD workshop took place in September 2019, and the topic addressed was 'The Future of Artificial Intelligence: Language, Society, Technology'. More specifically, the event focused on the impact of artificial intelligence on society, particularly on language-based technologies at the intersection of AI and ICT (henceforth 'Artificially Intelligent Communications Technologies' or 'AICT') – namely, speech technology, natural language processing, smart telecommunications and social media. The social impact of these technologies is already becoming apparent. Intelligent conversational agents such as Siri (Apple), Cortana (Microsoft) and Alexa (Amazon) are already widely used, and, in the next 5 to 10 years, a new generation of virtual personal assistants will emerge that will increasingly influence all aspects of our lives, from relatively mundane tasks such as turning the heating on and off, to highly significant activities such as influencing how we vote in national elections. Crucially, our interactions with these devices will be predominantly language-based. The speakers at the event came from academia, government, and industry, and therefore they explored these themes from a range of contrasting perspectives. The speakers were:

- Maria Luciana Axente (PricewaterhouseCoopers)
- Dr Shauna Concannon (University of Cambridge)

Dr Marcus Tomalin at the second GVDD workshop in May 2019

- Sarah Connolly (UK Department for Digital, Culture, Media & Sport)
- Dr Ella McPherson (University of Cambridge)
- Dr Trisha Meyer (Free University of Brussels – VUB)
- Jonnie Penn (University of Cambridge)

Like the first two workshops, this one attracted a large audience, and the feedback from speakers and attendees was extremely positive. It is clear to us that these events serve an extremely useful purpose, and help to publicise our project-related activities.

In addition, the GVDD project contributed two talks to the Festival of Ideas organised by Cambridge University in October 2019. The topics of the talks were:

- Artificial Intelligence and Social Change (19 October)
- Disempowering Hate Speech: How to Make Social Media Less Harmful (19 October)

The festival organisers anticipated significant interest in both talks, so they published a special 'speaker spotlight' feature in advance which consisted of an interview with Dr Marcus Tomalin:

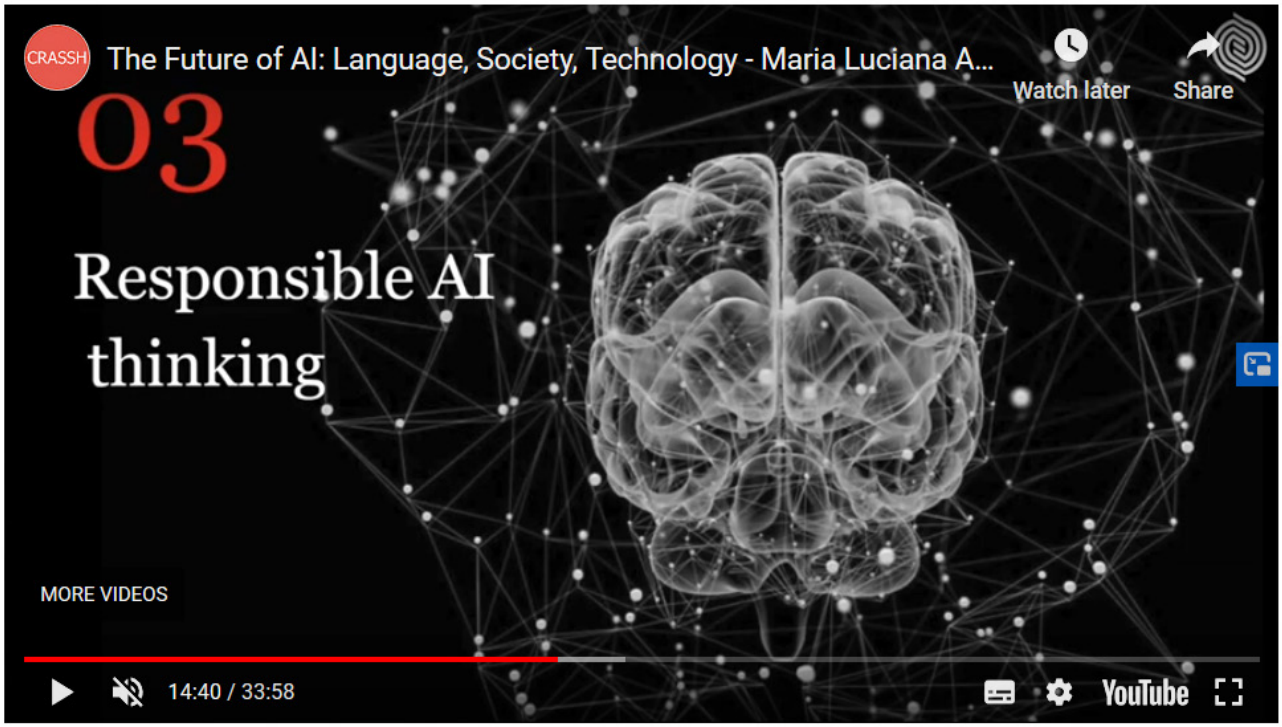
https://www.festivalofideas.cam.ac.uk/speaker-spotlight-marcus-tomalin-senior-research-associate-machine-intelligence-laboratory-and

The audio of the first talk is available here:

http://www.crassh.cam.ac.uk/gallery/audio/artificial-intelligence-and-social-change

Both sessions were packed, and there was an opportunity for Q&A with the audience during each. The audience was extremely diverse, consisting of members of the general public as well as academics and students.

The main event that has taken place so far in 2020 is the Fact-Checking Hackathon (10-12 January). This event brought together people with different kinds of expertise to develop new approaches for tackling the problems posed by fake news, misinformation, and disinformation. Taking an existing automated fact-checking system as a baseline, the main hackathon task was to find ways of improving its performance. The experimental framework used was that developed for the FEVER: Fact Extraction and VERification challenge. This genuinely interdisciplinary event was a great



Slide from one of the presentations at the third GVDD workshop (available to view online)

success, and the attendees worked tirelessly to develop various parts of the baseline system. The core task took the following form:

> *given a factual statement (S1)*
>
> S1: “Richard Nixon was reelected.”
>
> and a corpus of Wikipedia pages, then the system should determine whether or not S1 can confirmed by the corpus.
>
> For instance, if one Wikipedia page (https://en.wikipedia.org/wiki/Richard_Nixon) contains the statement (S2)
>
> S2: “He was reelected in one of the largest electoral landslides in US history in 1972 when he defeated George McGovern.”
>
> then it is possible to conclude that S1 can be verified by the corpus (specifically by S2). However, establishing the connection between S1 and S2 automatically is not trivial, since the antecedent of the pronoun (‘He’) has to be determined.

The first day of the Hackathon started with a series of short talks by experts in the domain of fact-checking – for example, Mevan Babakar the Head of Automated Fact-checking at Full Fact, and Jonty Page, a fourth-year Engineering student supervised by Dr Marcus Tomalin (see Section 6) who is developing a state-of-the-art automated fact-checking system. These provided some context for the topics explored during the event. The attendees were then divided into teams and each team focused on different aspects of the main task. There were research periods interspersed with some sessions that introduced participants to the programming language Python, and other sessions that showed how Python could be used for natural language processing. The advert for the event is available here: http://www.crassh.cam.ac.uk/events/28814

A blog summary of the event is available below (including short interviews with some of the attendees):

http://www.crassh.cam.ac.uk/blog/post/fact-checking-hackathon-a-write-up?fbclid=IwAR16e1_8_uYPgulFiuVZwRswVtrYQb1TMFVGNlH_DlVv76VjKibQcelFHRc

Attendees at the Hackathon

In addition to these main events, the members of the GVDD team have given numerous talks and lectures, including:

- 10 September 2019: paper on 'The Social Impact of Automatic Hate Speech Detection' at the Technology and Society conference, Katholieke Universiteit Leuven
- 22 October 2019: talk on 'The Social Impact of Automatic Hate Speech Detection', Darwin College, Cambridge
- 15 November 2019: talk on 'AI and Gender' at the Department of Computer Science and Technology, Cambridge
- 17 January 2020: paper on 'Ethical Tech: Bias, Algorithms and Social Justice' at the Algorithms for Her? conference, King's College London
- 12 February 2020: Participation in an event at the Oxford Internet Institute
- 17 February 2020: meeting with the Cambridgeshire County Council to advise them on their use of AI systems
- 18-20 February 2020: paper at the conference on 'Data in Discourse Analysis', Technical University of Darmstadt
- 5 March 2020: talk on 'Natural Language and Artificial Intelligence' at Trinity College, Cambridge
- 25-30 April 2020: paper on 'Public Engagement with Open Data' at the ACM CHI Conference on Human Factors in Computing Systems, Honolulu

4. New Postgraduate Courses in Ethics and AI

One strand of the GVDD project that we have been determined to implement since the very beginning involves the teaching of ethics to students of information engineering and computer science. Currently, students who study Medical Sciences at university have to study Medical Ethics; it is an obligatory requirement. However, those who study AI-related technologies are not currently obliged to consider the ethical implications of the systems they are designing and developing. We feel that this situation is untenable, and therefore, under the auspices of GVDD, we were determined to introduce some ethics-related modules to the core technology-focused teaching offered at Cambridge.

Accordingly, Dr Marcus Tomalin delivered two courses in the Department of Computer Science and Technology during the Lent Term 2020 (January to March). The first of these was a module, 'Bias in Datasets', that could be taken by current MPhil students. There was considerable interest in this course, and the various sessions were extremely well attended. After an initial lecture, four of the MPhil students presented work on state-of-the-art approaches to handling data bias in language-based AI systems, and there was an emphasis on word embeddings and machine translation. The second course, 'Building AI Systems Ethically', was aimed at current PhD students and Research Associates (RAs). It explored a wide range of ethical issues that software engineers and AI system designers need to consider, with a focus on algorithmic decision making and the problems caused by bias in datasets. We are hopeful that these courses will become an established part of the teaching offered to the MPhil and PhD students, and RAs, at Cambridge.

5. MPhil Projects

In March 2019, the first MPhil student began working on a project proposed by GVDD. This project required the student to develop an automated fact-checking system that can verify claims in relation to a corpus of textual sources. For instance, given both the assertion 'Trump is the most popular president in US history' and a corpus of Wikipedia pages, the task would be (i) to determine automatically whether the assertion is verifiable or not given the corpus, and (ii) if it is verifiable, to determine automatically whether it is supported or refuted by the corpus (for example, https://en.wikipedia.org/wiki/Historical_rankings_of_presidents_of_the_United_States indicates that George Washington is the most popular president). The project was successfully completed and submitted for examination in July 2019.

Two new projects have been proposed for the 2020 academic year, and the topics of these are:

- Multimodal Hate Speech Detection
- The Automated Detection of Cyberbullying

Detailed overviews of both projects are given in Appendix 2 below.

6. Fourth-year Projects

In the summer of 2019, the first two fourth-year Engineering students began work on their GVDD projects. The two projects were:

- Offensive Language Detection and Classification
- Fact Checking Fake News

Both were supervised by Dr Marcus Tomalin.

The 'Offensive Language Detection and Classification' project required the student to build a system that automatically detects offensive language in a dataset of natural language media posts, and to categorise all instances of hate speech into more specific subtypes. The problem of offensive language and hate speech on social media platforms has attracted considerable attention in machine learning (ML) and natural language processing (NLP) in recent years. The ability to detect offensive or hateful posts automatically would provide the possibility of improving and increasing measures to better protect potential victims. Therefore, it has become a crucial research area for companies such as Facebook, Microsoft, Twitter, and YouTube. The student has now developed a state-of-the-art hate speech detection system that outperforms existing systems. He has also developed an app that demonstrates how the system can be used to protect users from hate speech.

The other project, 'Fact Checking Fake News', involved developing the system that had been put in place by the MPhil student mentioned in Section 5 above. Both fourth-year projects will be completed and submitted for examination in June 2020.

Two new projects have been proposed for the 2020-2021 academical year, and the topics of these are:

- Multimodal Hate Speech Detection
- Automated Suicide Risk Detection from Social Media Posts

Detailed overviews of both projects are given in Appendix 1 below.

7. Future Plans

GVDD's future plans will undoubtedly be severely disrupted by the Covid-19 pandemic. For instance, we were planning to welcome our first visiting lecturer, Dr Jeffrey Watumull, in May 2020. Jeffrey was a Masters student at Cambridge in 2010, when he was supervised by Professor Ian Roberts, and he subsequently obtained a PhD in Linguistics from MIT in 2015. He is currently Director of Artificial Intelligence at Oceanit in Honolulu, and he has specific research interests in language-related AI systems. We are hopeful that he will still be able to join us as planned, but that may prove to be infeasible.

As mentioned in Section 2 above, our article 'Rethinking Bias Reduction in AI Systems: The Practical Ethics of Adaptation' has been submitted to *Ethics and Information Technology*, and will hopefully be published later this year. The research summarised there highlights the naivety of asserting that all sets of training data should be debiased before they are used to train AI systems. As the article demonstrates, for certain kinds of systems and certain kinds of data, that is simply impossible. Nonetheless, the article shows how fully trained highly biased systems can be rendered far less biased by using a technique known as domain-specific adaptation. In other words, the biased trained system is largely debiased as a result of fine-tuning it using a very small set of adaptation data. This innovative approach to the well-known data bias problem is likely to be extremely influential since it provides a way of reducing bias while also retaining high general system performance.

The next project workshop is scheduled to take place on 28 September 2020. We are currently in the process of finalising the thematic focus of that event and contacting potential speakers. In addition, we were planning to participate in various academic conferences during the next few months, but those plans are being revised and will be subject to current UK guidelines on social distancing.

Appendix 1: Fourth-year project proposals

Project Task: Multimodal Hate Speech Detection

The Problem

In recent years, the problem of online hate speech has attracted considerable attention from the Machine Learning (ML) and Natural Language Processing (NLP) research communities. The ability automatically to detect hateful social media posts (for instance) would enable potential victims to be protected more effectively by means of quarantining (Ullmann and Tomalin 2019). However, exclusively text-based approaches to this problem are increasingly limited, since online hate speech frequently involves both texts and images (eg, offensive memes). Consequently, the goal of this project is to build a multimodal system that automatically detects hate speech using textual and image-based sources.

Some Background

With the rapid proliferation of computer-mediated communication, online hate speech on social networking platforms continues to increase. In the worst cases, it has led to the public shaming of victims, or even to their death.[1] In the UK from 2006 onwards, different laws have been implemented which prohibit racial hatred, religious hatred, and hatred on the grounds of sexual orientation.[2] However, the anonymity of online environments presents particular complications.

In 2016, a "Code of Conduct on countering illegal hate speech online" was set up between the European Commission and leading IT companies. According to this Code, hate speech is understood as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin".[3] Although there has been an intense research focus on developing automated hate speech detection systems in the last few years (eg, Davidson et al 2017, Fortuna and Nunes 2018), such systems cannot deal with all forms of online hate speech. This is because offensive messages are not always communicated only via text-based means. With increasing frequency, images are used in conjunction with words in order to form hateful messages, and these multimodal communications present non-trivial challenges for automated detection systems (eg, Gomez et al 2019, Sàbat 2019, Sàbat et al 2019, Yang et al 2019). However, an automatic multimodal hate speech detection system could enable offensive social media posts to be quarantined in the manner proposed for offensive text-based posts in Ullmann and Tomalin 2019.

The Data

The dataset provided for the task, MMHS150K, was originally compiled by Gomez et al 2019.[4] It is a manually annotated multimodal hate speech dataset formed by 150,000 tweets gathered from September 2018 to February 2019, each one of them containing a text and an associated image. The tweets were annotated by crowdsourced workers using Amazon Mechanical Turk. The workers classified each text–image pair into one of six categories: No attacks to any community, racist, sexist, homophobic, religion-based attacks, or attacks to other communities. Each tweet was labelled by three different workers.

---

1. Cocking & van den Hoven 2018
2. See *Public Order Act 1986, Racial and Religious Hatred Act 2006, and Criminal Justice and Immigration Act 2008*.
3. European Commission 2016
4. https://gombru.github.io/2019/10/09/MMHS

The Task and the System

The central focus of this research is a classification task. The trained system developed for the project should be able to determine whether a given combination of text and image constitutes an instance of hate speech or not. Initially, a baseline system that achieves state-of-the-art performance (currently around 68% accuracy) will be constructed, and it will then be improved and refined in various ways to improve its performance. The main components of the system will be constructed using the ML libraries available in TensorFlow or PyTorch.

Bibliography

Cocking, D. and J. van den Hoven. 2018. *Evil Online*. Hoboken, N.J.: Wiley-Blackwell

Davidson, T, D. Warmsley, M. Macy, and I. Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. arXiv:1703.04009

European Commission. 2016. *Code of conduct on countering illegal hate speech online*. https://ec.europa.eu/info/files/code-conduct-countering-illegal-hate-speech-online_en, last accessed on 16/11/2019

Fortuna, P. and S. Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* 51(4) Article 85:1–30. https://doi.org/10.1145/3232676

Gomez, Raul, Jaume Gibert, Lluis Gomez, Dimonsthenis Karatzas. 2019. "Exploring Hate Speech in Multimodal Publications". https://arxiv.org/pdf/1910.03814.pdf

Sàbat, Benet Oriol. 2019. *Multimodal Hate Speech Detection in Memes*. BA Thesis: https://upcommons.upc.edu/handle/2117/165996

Sàbat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro-i-Niet. 2019. "Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation". *AI for Social Good workshop at NeurIPS (2019)*: https://arxiv.org/pdf/1910.02334.pdf

Yang, Fan, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification". *Proceedings of the Third Workshop on Abusive Language Online (2019)*, 11–18: https://pdfs.semanticscholar.org/ee8e/a1a02a7cea51ae9191e0c0dad7ab080741d8.pdf

Ullmann, Stefanie and Marcus Tomalin. 2019. "Quarantining online hate speech: technical and ethical perspectives". *Ethics and Information Technology* https://link.springer.com/article/10.1007/s10676-019-09516-z

Project Task: Automated Suicide Risk Detection from Social Media Posts

The Problem

In recent years, there has been a growing interest in the task of identifying warning signs of suicidal behaviour by automatically analysing text-based social media activity. The underlying assumption is that patterns of linguistic behaviour can reveal the extent to which certain users are potentially at risk. The automated suicide risk detection task has attracted the attention of the ML and NLP research communities, primarily because the ability automatically to detect potentially vulnerable users from their social media posts would enable such individuals to be protected more effectively. Therefore, the goal of this project is to build an automated system that classifies social media posts as either indicating suicidal thoughts, or not indicating suicidal thoughts.

Some Background

With the rapid proliferation of computer-mediated communication, social networking platforms provide ever more detailed information about users' views and states of mind. It is well known that suicide rates have been increasing in many countries, and the World Health Organization's figures currently indicate that about 800,000 people die as a result of suicide each year.[5] In the UK alone, there were 11.2 deaths by suicide for every 100,000 people in 2018. This was a 19-year high, and

the largest increases were in the age range 10 to 24.[6] Part of the problem is that young people are increasingly likely to reach out for help on social media rather than book an appointment with a counsellor or call a mental health hotline.[7] Consequently, if it were possible to identify social media posts that indicated a tendency towards suicidal thoughts, then that would potentially provide a way of offering mental health advice to especially vulnerable users.

The Data

The dataset provided for the task – the Reddit Suicidality Dataset, Version 2 – was originally compiled by the University of Maryland for the Computational Linguistics and Clinical Psychology Workshop 2019.[8] The corpus contains data from 11,129 users who posted on SuicideWatch, and another for 11,129 users who did not. For each user, there is full longitudinal data from the 2015 Full Reddit Submission Corpus, including, for each post, the post ID, anonymised user ID, timestamp, subreddit, de-identified post title, and de-identified post body. In addition, there are two sets of human risk-level annotations for subsets of the users, obtained via crowdsourced annotation (621 users who posted on SuicideWatch and 621 who did not) and expert annotations (245 users who posted on SuicideWatch, paired with 245 control users who did not).

The Task and the System

The central focus of this research will be on a classification task. The trained system developed for the project should be able to determine whether or not a given social media post potentially indicates suicidal thoughts. Initially, a baseline system that achieves state-of-the-art performance will be constructed, and then it will be improved and refined in various ways. Current suicide risk detection systems use different modelling approaches (eg, support vector machines, Neural Nets), and different corpora of training and test data. Therefore, one useful contribution of the project will be to produce a series of comparisons for various systems, using the same training and test data, so that their respective performances can be compared accurately in a meaningful way. The main components of the system will be constructed using the ML libraries available in TensorFlow or PyTorch.

Bibliography

Coppersmith, Glen, et al (2018), 'Natural Language Processing of Social Media as Screening for Suicide Risk', *Biomedical Informatics Insights* 10, 1-11: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111391/pdf/10.1177_1178222618792860.pdf

Ji, Shaoxiong, et al. (2019), 'Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications': https://arxiv.org/pdf/1910.12611.pdf

Kshirsagar, Rohan et al (2019), 'Detecting and Explaining Crisis': https://arxiv.org/pdf/1705.09585.pdf

O'dea, Bridianne, et al. (2015), 'Detecting Suicidality on Twitter', *Internet Interventions* 2:2, 183-188: https://www.sciencedirect.com/science/article/pii/S2214782915000160

Reardon, Sara (2017), 'AI Algorithms to Prevent Suicide Gain Traction' *Nature*: https://www.nature.com/articles/d41586-017-08307-0

Shing, Han-Chin, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daume III , and Philip Resnik (2018), 'Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online', Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic: https://www.aclweb.org/anthology/W18-0603.pdf

5. https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/
6. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2018registrations
7. https://www.nature.com/articles/d41586-017-08307-0
8. http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html

Appendix 2: MPhil Project Proposals

Project Task: The Automated Detection of Cyberbullying

The Problem

In recent years, the problem of cyberbullying on social media platforms has started to attract considerable attention from the ML and NLP research communities. While the definitions used sometimes differ slightly, cyberbullying is essentially an aggressive intentional act carried out by a group or individual, using electronic forms of communication, against victims who cannot easily defend themselves. It is often targeted at especially vulnerable groups (eg, people with disabilities), and therefore it can have an extremely harmful psychological and emotional impact. Since the ability to detect cyberbullying automatically offers a way of protecting potential victims, this topic is becoming an important research area for social media platforms such as Facebook, Instagram, Twitter, and YouTube.

The goal of this project is to build a system that automatically detects cyberbullying in a dataset of natural language social media posts. This involves developing an automated system trained on text-based sources that can determine whether or not cyberbullying is occurring. Therefore, the research will combine approaches from ML and NLP, and it will use existing corpora of training and test data.

Some Background

In recent years, the increase of cyberbullying across social media platforms has become a cause of widespread concern. In the UK Government's recent white paper, *Online Harms* (2019), it was estimated that 1 in 5 young people aged between 11 and 19 are the victims of it (p 19), while other estimates suggest that 54% of young people have witnessed some form of cyberbullying online (Cox 2014). The likelihood of being a victim is significantly greater if the individuals concerned are female, religious minorities, members of the LGBT+ community, and/or disabled in some way. Given this situation, it would be helpful if cyberbullying could be identified automatically since this would potentially offer a greater level of protection for vulnerable social media users. For instance, if bullying messages could be identified automatically in real time, with sufficient accuracy, then they could be handled via the kinds of quarantining methods that have recently been proposed for online hate speech (Ullmann and Tomalin 2019).

The Data

The dataset provided for the task was originally compiled for Van Hee et al 2018. This corpus consists of data collected from the social networking site ASK.fm, which enables users to create profiles and ask or answer questions, with the option of doing so anonymously. ASK.fm data consist of question–answer pairs published on a user's profile. The data were retrieved by crawling a number of seed profiles using GNU Wget software in April and October 2013. After language filtering (ie non-English content was removed), the experimental corpus comprised 113,698 posts. The data were hand-annotated to indicate several types of textual cyberbullying and verbal aggression, their severity, and the author participant roles. For instance,

Post	Classification
I'm going to find out who you are & I swear you're going to regret it	THREAT
No one likes you	INSULT

The Task and the System

The central focus of this research is a classification task. The trained system developed for the project should be able to determine whether a given text constitutes an instance of cyberbullying or not, and it could also specify the subtype of cyberbullying involved. Initially, a baseline system that achieves state-of-the-art performance will be constructed, and that system will then be improved and refined in various ways using techniques from the recent ML and NLP literature. The main components of the system will be constructed using the ML libraries available in TensorFlow or PyTorch.

Bibliography

Chatzakou, Despoina, et al (2019), 'Detecting Cyberbullying and Cyberaggression in Social Media': https://arxiv.org/pdf/1907.08873.pdf

Cheng, Lu et al. (2019), 'Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network': https://www.researchgate.net/publication/331314806_Hierarchical_Attention_Networks_for_Cyberbullying_Detection_on_the_Instagram_Social_Network

Emmery, Chris et al (2019), 'Current Limitations in Cyberbullying Detection: on Evaluation Criteria, Reproducibility, and Data Scarcity': https://arxiv.org/pdf/1910.11922.pdf

Rosa, Hugo, et al (2019), 'Automatic Cyberbullying Detection: A Systematic Review' Computers in Human Behaviour 93, 333-345: https://www.sciencedirect.com/science/article/pii/S0747563218306071

Ullman, Stefanie and Marcus Tomalin (2019): 'Quarantining Online Hate Speech: Ethical and Technical Perspectives', Ethics and Information Technology, 1-12: https://link.springer.com/article/10.1007/s10676-019-09516-z

Van Hee, et al (2018), 'Automatic Detection of Cyberbullying in Social Media Text': https://arxiv.org/pdf/1801.05617.pdf

The Futures Company (2014), '2014 Teen Internet Safety Survey' https://www.cox.com/content/dam/cox/aboutus/documents/tween-internet-safety-survey.pdf

Project Task: Multimodal Hate Speech Detection

The Problem

In recent years, the problem of online hate speech has attracted considerable attention from the Machine Learning (ML) and Natural Language Processing (NLP) research communities. The ability automatically to detect hateful social media posts (for instance) would enable potential victims to be protected more effectively by means of quarantining (Ullmann and Tomalin 2019). However, exclusively text-based approaches to this problem are increasingly limited, since online hate speech frequently involves both texts and images (e.g., offensive memes). Consequently, the goal of this project is to build a multimodal system that automatically detects hate speech using textual and image-based sources.

Some Background

With the rapid proliferation of computer-mediated communication, online hate speech on social networking platforms continues to increase. In the worst cases, it has led to the public shaming of victims, or even to their death. In the UK from 2006 onwards, different laws have been implemented which prohibit racial hatred, religious hatred, and hatred on the ground of sexual orientation. 9 However, the anonymity of online environments presents particular complications.

In 2016, a "Code of Conduct on countering illegal hate speech online" was set up between the European Commission and leading IT companies. According to this Code, hate speech is understood as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or

9. Cocking & van den Hoven 2018

10. See *Public Order Act 1986, Racial and Religious Hatred Act 2006*, and *Criminal Justice and Immigration Act 2008*

ethnic origin". Although there has been an intense research focus on developing automated hate speech detection systems in the last few years (e.g., Davidson et al 2017, Fortuna and Nunes 2018), such systems cannot deal with all forms of online hate speech. This is because offensive messages are not always communicated only via text-based means. With increasing frequency, images are used in conjunction with words in order to form hateful messages, and these multimodal communications present non-trivial challenges for automated detection systems (e.g., Gomez et al 2019, Sàbat 2019, Sàbat et al 2019, Yang et al 2019). However, an automatic multimodal hate speech detection system could enable offensive social media posts to be quarantined in the manner proposed for offensive text-based posts in Ullmann and Tomalin 2019.

The Data

The dataset provided for the task, MMHS150K, was originally compiled by Gomez et al 2019. It is a manually annotated multimodal hate speech dataset formed by 150,000 tweets gathered from September 2018 to February 2019, each one of them containing a text and an associated image. The tweets were annotated by crowdsourced workers using Amazon Mechanical Turk. The workers classified each text-image pair into one of 6 categories: No attacks to any community, racist, sexist, homophobic, religion-based attacks, or attacks to other communities. Each tweet was labelled by 3 different workers.

The Task and the System

The central focus of this research is a classification task. The trained system developed for the project should be able to determine whether a given combination of text and image constitutes an instance of hate speech or not. Initially, a baseline system that achieves state-of-the-art performance (currently around 68% Accuracy) will be constructed, and it will then be improved and refined in various ways to improve its performance. The main components of the system will be constructed using the ML libraries available in Tensorflow or PyTorch.

Bibliography

Cocking, D. and J. van den Hoven. 2018. Evil Online. Hoboken, N.J.: Wiley-Blackwells

Davidson, T, D. Warmsley, M. Macy, and I. Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017). arXiv:1703.04009

European Commission. 2016. Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/files/code-conduct-countering-illegal-hate-speech-online_en, last accessed on 16/11/2019

Fortuna, P. and S. Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." ACM Computing Surveys 51(4) Article 85:1–30. https://doi.org/10.1145/3232676

Gomez, Raul, Jaume Gibert, Lluis Gomez, Dimonsthenis Karatzas. 2019. "Exploring Hate Speech in Multimodal Publications". https://arxiv.org/pdf/1910.03814.pdf

Sàbat, Benet Oriol. 2019. "Multimodal Hate Speech Detection in Memes". BA Thesis: https://upcommons.upc.edu/handle/2117/165996

Sàbat, Benet Oriol, Cristian Canton Ferrer, and Xavier Giro-i-Niet. 2019. "Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation". AI for Social Good workshop at NeurIPS (2019): https://arxiv.org/pdf/1910.02334.pdf

Yang, Fan, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification". Proceedings of the Third Workshop on Abusive Language Online (2019), 11–18: https://pdfs.semanticscholar.org/ee8e/a1a02a7cea51ae9191e0c0dad7ab080741d8.pdf

Ullmann, Stefanie and Marcus Tomalin. 2019. "Quarantining online hate speech: technical and ethical perspectives". Ethics and Information Technology https://link.springer.com/article/10.1007/s10676-019-09516-z

11. European Commission 2016

12. https://gombru.github.io/2019/10/09/MMHS

Contact:

Professor Steven Connor
Director - CRASSH

Centre for Research in the Arts, Social Sciences and Humanities
Alison Richard Building
7 West Road
Cambridge
CB3 9DT

Tel: + 44 (0)1223 765275
Email: skc45@cam.ac.uk
http://www.crassh.cam.ac.uk/

Aaron Westfall
Director of Development

University of Cambridge Development and Alumni Relations
1 Quayside, Bridge Street
Cambridge
CB5 8AB

Tel: +44 (0)1223 766116
Email: aaron.westfall@admin.cam.ac.uk
www.cam.ac.uk/YoursCambridge

Dear World...
Yours, Cambridge

The campaign for the University and Colleges of Cambridge