Zeerak Butt
Registration number: 160260775
Computer Science
Programme: Computer Science (PhD/Computer Sci E FT) - COMR33

Dear Zeerak

PROJECT TITLE: Automatic Annotation of Abusive Language
APPLICATION: Reference Number 017306

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 10/04/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 017306 (dated 27/12/2017).

The following optional amendments were suggested:

Please, address the following: Some of the jargon in the "Aims & objectives" is too specific and difficult to grasp for non experts. More specifically, please, amend the research objectives and fix some of the minor typos present in the application.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely

Alice Tucker
Ethics Administrator
Computer Science

Application 017306

Section A: Applicant details

Date application started:
Tue 12 December 2017 at 16:34

First name:
Zeerak

Last name:
Butt

Email:
z.w.butt@sheffield.ac.uk

Programme name:
Computer Science (PhD/Computer Sci E FT) - COMR33

Module name:
COMR33
Last updated:
10/04/2018

Department:
Computer Science

Applying as:
Postgraduate research

Research project title:
Automatic Annotation of Abusive Language

Similar applications:
Domain Adaptation for Abusive Language, Multi-task Learning for Hate Speech Detection

Section B: Basic information

Supervisor

Name	Email
Kalina Bontcheva	k.bontcheva@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Sat 7 April 2018

Anticipated end date (of project)
Wed 12 February 2020

3: URMS number (where applicable)

URMS number
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
No

Involves human tissue?
No

Clinical trial?
No

Social care research?
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
- not entered -

Vulnerabilities

Involves potentially vulnerable participants?
No

Involves potentially highly sensitive topics?
Yes

Section C: Summary of research

1. Aims & Objectives

Content moderation, and in particular abusive language and hate speech on social media platforms has recently received a growth in media attention [Flynn, 2017], political attention [Furness, 2017, Home Office, 2016], and research attention in NLP [Waseem, 2016, Waseem and Hovy, 2016, Davidson et al., 2017, Wulczyn et al., 2017]. To combat inappropriate content, social media platform employ either manual moderation [Chen, 2012] or a mixture of manual and automated moderation [Bogle, 2016]. Given the volume of data published on a daily basis, with over 300M tweets per day [Twitter Inc, 2012], it is necessary to find methods for dealing with content moderation at scale.

Considering an international scope, it has been made clear by the European Union as well as individual member states that online khate speech must be removed and addressed. Specifically, in 2016 the European Commission and a number of technology companies agreed to a code of conduct for the treatment of illegal hate speech online [European Commission, 2016] and Germany imposed e50M fines on social media companies for systematically failing to remove illegal hate speech online [The Guardian, 2017].

Some of the main issues with abusive language research are related to the question of data set construction and usability to closely related tasks. In this project will seek to improve upon methods for abusive language detection. Specifically, we will be considering methods for abusive language research that can allow for computationally gathering data sets with a minimised need for human annotation, thus limiting adverse mental health risks associated with annotating large corpora of abusive language.

This project aims to expand upon existing methods and data sets to incorpo- rate an intersectional feminist methodology [McIntosh, 1988, Crenshaw, 1989] to the area of abusive language research to show that considering societal privilege and oppression can aid in the collection of data sets can have a profound analysis and prediction of abuse and allow for a richer analysis and understanding of the abuse which can allow for policy teams in industry and government to make decisions on a more informed platform.

Research Questions
The research questions we will seek to address in this project pertain the interaction of privilege and different forms of oppression.
1. Design and develop methodology for applying intersectional feminist methodology to automatic gathering of abusive language data.
2. Explore the application of natural language processing and machine learning methods to collect data sets of abusive comments.

2. Methodology

Initially, we will apply the previous released data sets [Waseem, 2016, Waseem and Hovy, 2016, Davidson et al., 2017, Wulczyn et al., 2017] as well as public data sets from Reddit. In addition, we will collect tweets around hashtags that see a great deal of abusive comments such as #whitepride, #blacklivesmatter, #religionofpeace, #feminazi, #feminism, outright slurs such as “nigger”, “kike”, and “paki”, and “raghead”, and finally code words employed by the extreme right, i.e. “googles”, “skypes”, and “skittles” referring to black, jewish, and muslim people respectively. Using such seed words to collect tweets, we will gather all tweets by all of the users. Henceforth, user tweets will refer to all tweets from a given user. Using the previously previously published and public data sets, we will use the positive classes to measure distances to each tweet written by each user. Initially focusing on high recall, if a single tweet by a user has a small distance to the labelled positive classes, all tweets are labeled as abusive. From this point, a triage system using multiple natural langauge processing tools to increase precision, selecting only the tweets that are hate speech. This will, amongst other methods, include using the same distance metric as previously used, sentiment analysis, sentiment towards individuals or demographic targets (cross-referenced with legally protected classes), examine for co-occurrences of terms in our positive classes and each tweet.

3. Personal Safety

Raises personal safety issues? No

- not entered -

Section D: About the participants

1. Potential Participants

Participants will be identified using keywords that generate a great deal of hate speech (e.g. #whitepride, #blacklivesmatter, #religionofpeace, #feminazi, #feminism), outright slurs (e.g. “nigger”, “kike”, and “paki”, and “raghead”), and finally code words employed by the extreme right, (e.g. “googles”, “skypes”, and “skittles” referring to black, jewish, and muslim people respectively).

2. Recruiting Potential Participants

Participants will be recruited using Twitter's API to collect tweets of users that use the aforementioned keywords.

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CiCS? Yes

We will be recruiting using the university mailing lists to ensure that volunteers are physically located in Sheffield. We choose to do this as being exposed to abusive comments can have strains on mental health and by ensuring physical proximity, we can ensure that our volunteers can be given mental health counselling should they require it.

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) No

Informed consent will not be sought as collecting data from social media does not require informed consent. In addition, our aim is to observe behaviour and contacting users may alter their communication patterns.
Further, no tweets from protected accounts will be collected, as these explicitly re- strict access and communicate that they do not give consent for their use.

4. Payment

Will financial/in kind payments be offered to participants? No

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

There are no harms to the participants.

How will this be managed to ensure appropriate protection and well-being of the participants?

To ensure that there will be no harm to participants, we will ensure that all data is stored on encrypted devices in efforts to further

minimise any risk to the participants.

Section E: About the data

1. Data Confidentiality Measures

We will comply to the Data Protection Act (DPA) to further ensure safety and privacy of participants. Furthermore, data access will be restricted to the researchers (Zeerak Waseem) and the supervisors (Kalina Bontcheva, Andreas Vlachos). Further, all data will be stored on encrypted and password protected devices.

2. Data Storage

Only the researchers involved with the study will have access to the data, which will be stored in password protected encrypted folders. The data will be analysed by the research team all of whom are associated with the university of Sheffield. The project will take place at the university of Sheffield.
The data generated by the project will not be stored after the end of the project and will not be made available for future research projects. The data will be deleted at the end of the project.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
No

Consent forms relevant to project?
No

Additional Documentation

External Documentation

References

[Bogle, 2016] Bogle, A. (2016). Instagram is rolling out its tool to filger offensive comments to all users. Last accessed, Dec. 12.
[Chen, 2012] Chen, A. (2012). Inside facebook's outsourced anti-porn and gore brigade, where 'camel toes' are more offensive than 'crushed heads'.
[Crenshaw, 1989] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. University of Chicago Legal Forum, 1989(1).
[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive lan- guage. In Proceedings of ICWSM.
[European Commission, 2016] European Commission (2016). Code of conduct on countering illegal hate speech online. Technical report.
[Flynn, 2017] Flynn, K. (2017). Why celebrities leave twitter. Last accessed, Dec. 12.
[Furness, 2017] Furness, H. (2017). Prince william launches anti-bullying plan to combat 'banter escalation scenarios'. Last accessed Nov. 20, 2017.
[Home Office, 2016] Home Office (2016). Action against hate the uk govern- ment's plan for tackling hate crime. Technical report.
[McIntosh, 1988] McIntosh, P. (1988). White privilege and male privilege: A personal account of coming to see correpondences through work in women's studies.
[The Guardian, 2017] The Guardian (2017). Germany approves plans to fine social media firms up to e50m.
[Twitter Inc, 2012] Twitter Inc (2012). Twitter turns six. Last Accessed Dec. 12.
[Waseem, 2016] Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 138–142, Austin, Texas. Association for Computational Linguistics.
[Waseem and Hovy, 2016] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, Califor- nia. Association for Computational Linguistics.
[Wulczyn et al., 2017] Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th Inter- national Conference on World Wide Web, WWW '17, pages 1391–1399, Re- public and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Section G: Declaration

Signed by:
Zeerak Waseem
Date signed:
Wed 27 December 2017 at 22:44

Offical notes

- not entered -