Downloaded: 06/06/2018
Approved: 30/01/2018

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

Dear Zeerak

PROJECT TITLE: Domain adaptation for abusive language
APPLICATION: Reference Number 017099

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 30/01/2018 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 017099 (dated 03/12/2017).

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 017099

Section A: Applicant details

Date application started:
Wed 29 November 2017 at 11:07

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:
30/01/2018

Department:
Computer Science

Applying as:
Postgraduate research

Research project title:
Domain adaptation for abusive language

Similar applications:
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):
Wed 20 December 2017

Anticipated end date (of project)
Wed 5 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

Reports of hate crime on and offline have increased in the U.K. and U.S. following the Brexit referendum with a 29% percent increase in recorded hate crimes from 2015/2016 to 2016/2017 (O'Neill, 2017). In addition, online abuse has also become a focal point for high profile initiatives such as Prince William's initiative to counter online bullying (Furness, 2017). Finally, online hate crimes are a part of the Government's Hate Crime Action Plan (Home Office, 2016).
Considering an international scope, it has been made clear by the European Union as well as individual member states that online hate 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 50M Euro 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 aiming to consider methods for abusive language research that can allow for the reuse of data sets created for disparate tasks (i.e. cyberbullying, hate speech detection, toxicity detection) to allow for the reuse of data from semantically similar tasks.

This project aims to expand upon existing methods to incorporate an intersectional feminist methodology (McIntosh, 1988, Crenshaw, 1989) to the area of abusive langauge research. We will aim to apply domain transfer methods to overcome annotation gaps in abusive langauge research, as data sets are often annotated either for bullying, hate speech, or toxicity and models to detect hate speech perform poorly even when shifting shifting domain between different forms of hate speech (Waseem, 2016). By considering methods for models that can shift domains, we allow for building models that seek to take advantage of the commonalities of distinct forms of abuse (Waseem et al., 2017).

The main research areas and questions we will seek to explore are:

- How can machine learning methods for domain transfer be applied to various forms of abusive language detection tasks?
- Design and develop methodology for applying intersectional feminist methodology to computational abusive language research.

2. Methodology

The abusive language data will be obtained by using previously released data sets by Waseem (2016), Waseem and Hovy, (2016), Davidson et al. (2017), and Wulczyn et al. (2017). In addition, publicly available data sets from Reddit communities that are known to be toxic will be utilised to train a model to recognise potentially offensive tweets.

We will address the issue of domain adaptation by treating each data set as being distinct and applying machine learning methods such as self-training, co-training, ensemble methods, multi-task learning and joint-learning. We will work with both neural networks and linear models.

3. Personal Safety

Raises personal safety issues? No

There are no personal safety issues as there are no human participants involved in this project.

Section D: About the participants

1. Potential Participants

As we will use previously published data sets, we will not be identifying any new participants.

2. Recruiting Potential Participants

Recruitment will not be necessary

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CiCS? No

- not entered -

3. Consent

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

Given that we will not be recruiting human participants, no informed consent needs to be collected.

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

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:

(Crenshaw, 1989) Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory 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 language. In Proceedings of ICWSM.
(European Commission, 2016) European Commission (2016). Code of conduct on countering illegal hate speech online. Technical report.
(Furness, 2017) Furness, H. (2017). Prince william launches anti-bullying plan to combat 'banter escalation scenarios'. Last accessed Nov. 20, 2017. The Telegraph.
(Home Office, 2016) Home Office (2016). Action against hate the uk government's plan for tackling hate crime. Technical report.
(Levin, 2017) Levin, S. (2017). Moderators who had to view child abuse content sue microsoft, claiming ptsd.The Guardian
(McIntosh, 1988) McIntosh, P. (1988). White privilege and male privilege: A personal account of coming to see correpondences through work in women,Äôs studies.
(O'Neill, 2017) O'Neill, A. (2017). Hate crime, england and wales, 2016/17. Report.
(The Guardian, 2017) The Guardian (2017). Germany approves plans to fine social media firms up to €50m.
(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, California. Association for Computational Linguistics.
(Wulczyn et al., 2017) Wulczyn, E., Thain, N., Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale (to appear in Proceedings of the 26th International Conference on World Wide Web – WWW 2017).

Section G: Declaration

Signed by:
Zeerak Butt
Date signed:
Sun 3 December 2017 at 23:38

Offical notes

- not entered -