Literature Survey: Identifying anomalies within social media textual data

Student Name: Molly Hayward

Supervisor Name: Dr Noura Al-Moubayed

Submitted as part of the degree of BSc Computer Science to the

Board of Examiners in the School of Engineering and Computing Sciences, Durham University

***Abstract* —** These instructions give you guidelines for preparing the design paper. DO NOT change any settings, such as margins and font sizes. Just use this as a template and modify the contents into your design paper. Do not cite references in the abstract.

The abstract must be a Structured Abstract with the headings **Context/Background**, **Aims**, **Method**, and **Proposed Solution**. This section should not be no longer than a page, and having no more than two or three sentences under each heading is advised.

***Keywords* —** Put a few keywords here.

I INTRODUCTION

A Problem Background

Social media can facilitate collaboration, education and the forming of meaningful relationships. However, it can also aid the propagation of fake news, hatred and propaganda. In recent years, social networks have come under mounting pressure to constrain the growth of this problem. Traditionally, users have been able to flag inappropriate comments, some of which are manually removed by the social network. As systems grow in size, traditional reporting techniques do not scale effectively, therefore gathering feedback about user behaviour becomes an expensive task. The aim of this project is to survey anomaly detection techniques in order to develop an effective tool with which to detect anomalies among social media textual data. Such anomalies may include ...

B Terms

Anomaly: inconsistent with or deviating from what is usual, normal, or expected [1]
Unsupervised: whereby training data is not labelled
Text corpus: Large and structured set of texts [2]

II THEMES

A Text Data Vectorization

In order to extract meaningful data from large bodies of text, it must first undergo a pre-processing stage whereby the data is converted from its textual form into numerical data, known as vectorization.

B Anomaly Detection Techniques

In order to extract meaningful data from large bodies of text, it must first undergo a pre-processing stage whereby the data is converted from its textual form into numerical data, known as vectorization.

C References

[**?**] The list of cited references should appear at the end of the report, ordered alphabetically by the surnames of the first authors. The default style for references cited in the main text is the Harvard (author, date) format. When citing a section in a book, please give the relevant page numbers, as in [**?**, p293]. When citing, where there are either one or two authors, use the names, but if there are more than two, give the first one and use "et al." as in , except where this would be ambiguous, in which case use all author names.

You need to give all authors' names in each reference. Do not use "et al." unless there are more than five authors. Papers that have not been published should be cited as "unpublished" [**?**]. Papers that have been submitted or accepted for publication should be cited as "submitted for publication" as in [**?**] You can also cite using just the year when the author's name appears in the text, as in "but according to Futher (2006), we . . . ". Where an authors has more than one publication in a year, add 'a', 'b' etc. after the year.

References

Budgen, D. (2003), *Software Design*, 2nd edn, Addison Wesley.

Euther, K. (2006), Title of paper. unpublished.

Futher, R. (2006), Title of paper 2. submitted for publication.

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.