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 Do not cite references in the abstract.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* —** Natural Language Processing, Machine Learning, Deep Learning

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

Extracting meaningful data from large corpora is undoubtedly a complex task, and in order to do this successfully, we must first construct word embeddings that capture the semantic and syntactic relationships among words. In this section, we examine the process of encoding textual data in numeric vector form - known as vectorization, and so we are motivated to answer the question: how can we best vectorize text so as to encompass the important features of language?

Perhaps the simplest vectorization approach is *Bag of Words (BOW)*, which encodes text as a single vector containing a count of the number of occurrences of each word in the snippet. *Term Frequency-Inverse document frequency (TF-IDF)* [3] extends the BOW model by providing each word with a score that reflects its level of importance based upon its frequency within the document. Predictably, vectors produced in this way do not preserve the context within which words are found, complicating the task of discerning their meaning. Hence, these approaches are unlikely to be suitable in the application domain of detecting sarcasm - a highly contextual phenomenon. The *Bag of N-grams* approach is a generalisation of Bag of Words, whereby instead of counting the number of occurrences of each unigram, we count the number of occurrences of each N-gram. This allows us to capture a small window of context around each word, however this results in a much larger and sparser feature set.

First introduced in Mikolov et al. [1], *Word2Vec* describes a group of related models that can be used to produce high-quality vector representations of words - two such models are *Continuous Bag-of-Words* and *Continuous Skip-Gram*. It consists of a shallow (two-layer) neural network that takes a large corpus and produces a high-dimensional vector space. Unlike previous techniques, words that share a similar context tend to have collinear vectors, that is to say they are clustered together in the feature space. Consequently, Word2Vec is able to preserve the semantic relationships between words, even constructing analogies by composing vectors e.g. king - man + woman $\approx$ queen. Likewise, it captures syntactic regularities such as the singular to plural relationship e.g. cars - car $\approx$ apples - apple. In Word2Vec, intrinsic statistical properties of the corpus, which were key to earlier techniques, are neglected, therefore global patterns may be overlooked.To mitigate against this, the *GloVe* [2] approach generates word embeddings by constructing an explicit word co-occurrence matrix for the entire corpus, such that the dot product of two word embeddings is equal to log of the number of times these words co-occur (within a defined window). Despite the ability to preserve semantic relationships between words, Word2Vec and GloVe do not accommodate polysemy, which describes the co-existence of alternative meanings for the same word. In addition, they cannot generalise to words that were not specifically included in the training set. Hence, we may need to use a more robust approach.

B Anomaly Detection Techniques

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

C References

[3] [2] [1] 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 (**?**), we . . . ". Where an authors has more than one publication in a year, add 'a', 'b' etc. after the year.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.