DETECTING HEALTH MISINFORMATION IN WEB PAGE TEXT USING DEEP LEARNING METHODS

Dione Morales

Bachelor of Engineering
Computer Engineering Stream

School of Engineering
Macquarie University

November XX, 2018

Supervisor: Associate Professor Adam Dunn

ACKNOWLEDGMENTS

I would like to acknowledge ...

STATEMENT OF CANDIDATE

I, (insert name here), declare that this report, submitted as part of the requirement for the award of Bachelor of Engineering in the School of Engineering, Macquarie University, is entirely my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualification or assessment an any academic institution.

Student's Name:

Student's Signature:

Date:

ABSTRACT

This is where you write your abstract ...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the key components required to minimize the propagation of misinformation online is to have the ability of automatically evaluating and quantifying the credibility of articles. However, traditional automated methods - such as shallow learning-based techniques, still require the domain knowledge of experts to be able to develop the features required by the model. Thus, this project aims to investigate the performance of Deep Learning-based (DL) techniques in evaluating the credibility of information within domain-specific articles via the classification of set criteria that have deemed to be highly correlated with articles that have low credibility. Specifically, this project will focus on evaluating the credibility of online health articles related to vaccination due to the commonly misinformed and controversial views associated with its effects [8].

1.1 Project Overview

This section details the scope of the project and its associated outcomes outlining the various tasks that must be accomplished to successfully complete the project.

1.1.1 Project Scope

With the primary objective of this project being the evaluation on the effectiveness of deep learning models in determining the credibility of online health-related articles. Due to the complexity of this project, a set of activities - divided into main goals and stretch goals, have been defined to ensure that the completion of this project remains feasible in the given time frame. The completion of all activities categorized as main goals will signal the realization of the primary objective and the completion of the project. Stretch goals are activities of interest that have been identified as non-essential to the completion of the primary objective but (talk about the overarching goal that all stretch goals have in common e.g. understand the model, utilize the model etc.) and will be worked on after the completion of the project.

Main Goals

- Develop a set of criteria that will be used to determine the credibility of online articles.
- Evaluate the performance of common ML-based methods on the classification of online vaccine-related articles, based on the criteria developed.
- Evaluate the performance of the proposed DL model on the classification of online vaccine-related articles, based on the criteria developed.
- Evaluate the effect of transfer learning methods in the performance of the proposed DL method (assuming that the chosen method doesn't rely on transfer learning)
- Evaluate the effectiveness of various transfer learning methods for the classification task

Stretch Goals

- Utilize attention mechanisms to understand how the aforementioned DL model classifies the criteria for credibility.

Chapter 2

Background and Related Work

A literature review has been conducted to develop an understanding on the research that has been done in the assessment of the credibility of information, specifically in the context of information related to health and the limitations and capabilities of shallow learning techniques and how it differs from deep learning-based methods for the task of text classification.

2.1 Assessing the Credibility of Information

To have the capability of automating the process of evaluating the quality or credibility of online information, a definition that outlines of what is required by an article to be considered as a credible source of information must be developed. While there has been a significant amount of research that has been done on the development of tools and frameworks that aim to assess the credibility of online health information, there is currently no standardized method or benchmark that is universally used. Tools and frameworks that have been identified to be applicable in the context of this project are: DISCERN [1], HealthNewsReview [2] and Quality Index for health-related Media Reports (QIMR) [16].

2.1.1 DISCERN

DISCERN is a questionnaire designed to assess the reliability of a publication, it consists of 16 questions each with a Likert scale, ranging from 1 (no) to 5 (yes) and is divided into 3 sections. The first section (questions 1 - 8) investigate the reliability of the information and is comprised of questions such as "Are the aims clear?", "Is it balanced and unbiased?" and "Does it refer to areas of uncertainty?". The second section (questions 9 - 15) assesses the quality of information provided by the publication for treatment choices and is composed of questions such as "Does it describe the risks/benefits of each treatment?" and "Does it provide support for shared decision-making?". The final section (question 16) assesses the overall rating of the publication ("Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices").

Note: Need to talk about advantages/disadvantages, limitations, relevance to this project etc.

2.1.2 HealthNewsReview

HealthNewsReview provides a set of 10 criteria designed to act as a framework for evaluating the credibility of health-related media. The criteria is based on the various elements that all health-related media should consist of, the criteria is composed of criterion such as "Does the story compare the new approach with existing alternatives?", "Does the story use independent sources and identify conflicts of interest?" and "Does the story appear to rely solely or largely on a news release?".

Note: Need to talk about advantages/disadvantages, limitations, relevance to this project etc.

2.1.3 QIMR

QIMR is a tool developed to monitor the quality of health research reports presented in the media. The tool considers 5 main factors that have been deemed to be correlated with the quality of research reports based on interviews with health journalists and researchers. The 5 main factors are: background information provided, sources of information used, manner in which results were analysed, context of the research and the validity of their methodology.

2.2 Prior Approaches

The task of text classification is a heavily researched topic due to its wide number of applications in various domains. Formally, text classification is defined to be the task of finding a classifier $f : D \to L$ where $D = \{d_1, d_2, \ldots, d_n\}$ is a collection of documents and $L = \{l_1, l_2, \ldots, l_k\}$ is the set of possible labels that a document d can be classified as.

$$f : D \to L \text{ where } f(d) = l \tag{2.1}$$

Common applications of text classification algorithms are: organization and filtering of news articles, document retrieval, opinion mining, email classification and spam filtering [4]. For the purpose of this literature review, a shallow-based method refers to one in which it is unable to capture a high-level, general representation of information.

2.2.1 Shallow Learning Models

Traditional shallow learning-based approaches for text classification, such as Naive Bayes, Support Vector Machines (SVM) and Random Forests, require the manual extraction of features [4] [6] [11] [13] as they are incapable of performing feature learning [7]. These features are typically either hand-crafted and domain specific, requiring expert-level knowledge for the task domain [13] or are simple and general but prone to the loss of information (e.g. being unable to account for context) since these features, such as term frequency-inverse document frequency (TF-IDF) scores or bag of words (BoW) models, are unable to represent and account for the sequential nature of text.

Naive Bayes Classifier

Naive Bayes is a probabilistic classifier that functions on an assumption in how the data (i.e. the words in a document) is generated. The assumption that these Bayesian classifiers are based on is that the distribution of different words within a corpus are independent from each other. Despite this assumption clearly being wrong when considering the distribution of words within a document (due to the sequential nature of text), Naive Bayes classifiers are still able to perform well in text classification tasks such as the filtering of spam emails [15].

Naive Bayes classifiers utilize Bayes' theorem, which attempts to find the probability of an event B occurring given some prior condition A i.e. $P(B|A)$. In the context of text classification, Naive Bayes classifiers classifies a document d by calculating the probabilities that the document belongs for all labels $l_i \in L$ and then selecting the highest probability [6]:

$$P(L = l_i|d) = \frac{P(l_i)P(d|l_i)}{P(d)} \tag{2.2}$$

According to Aggarwal et al. [4], there are two main types of Naive Bayes classifiers that are commonly used, they are the multivariate Bernoulli model and the multinomial model. The main discriminating factor between these two models is that the Bernoulli model does not take into account the frequency of words as it represents a document using a vector of binary features which signify the presence or absence of words, based on some vocabulary, for a given document. This is in contrast with the multinomial model, which accounts for the frequency of words as the document is represented using a BoW model. Deciding on which model to use for text classification largely depends on the size of the vocabulary as, according to McCallum et al. [12], the multinomial models almost always outperforms the Bernoulli models if the size of the vocabulary is large (> 1000) or even if the size has been optimally chosen for both models.

Support Vector Machines

SVMs are a type of supervised machine learning-based binary classifiers that are extensively used for text classification due to their capability of handling the high dimensional and sparse nature of the common techniques used to represent text documents [10]. SVMs are able to classify a document by using a hyperplane to separate the different classes into separate regions. The hyperplane used is determined by choosing the one that maximises the margin of separation (i.e. The Euclidean distance between the hyperplane and all data points in the representation space) between the two classes [4].

Consider the illustration shown in figure 2.1 that presents three two-dimensional hyperplanes, A, B and C, which separates the classes 'x' and 'o'. Visually, we can determine that A is the hyperplane that maximises the margin of separation, thus, A will be used as a decision rule to classify any new document based on its location with respect to A. Mathematically, determining which hyperplane to use as the decision rule is an optimisation problem that attempts to maximise the margin represented as shown in Equation 2.3:

$$\text{maximise: } \frac{1}{2}||\frac{1}{w}||^2 \tag{2.3}$$

This optimisation problem is often re-framed to the minimisation of Equation 2.4:

$$\text{minimise: } \frac{1}{2}||w||^2 \tag{2.4}$$

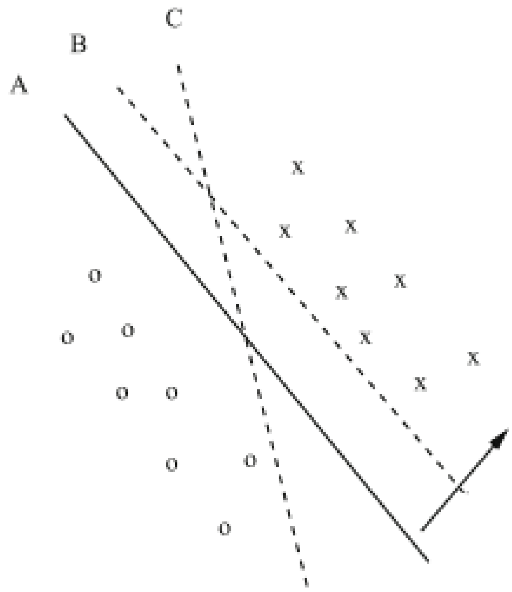

Figure 2.1: What is the Best Separating Hyperplane? [4]

This technique of determining the hyperplane requires that the two classes are linearly separable. In situations where this isn't true, the kernel trick [5] is applied which is essentially a function that maps data in a particular representation space to another representation space of a different dimension. This change in dimensionality can allow the classes to become linearly separable and thus a hyperplane can then be constructed that separates each class as illustrated in Figure 2.2.

2.2.2 Shallow Textual Representation Techniques

Due to the unstructured nature of natural text, it is common practice for natural language processing (NLP) tasks to transform the text into a structured representation that minimises the loss of the information required by the model. However, despite common

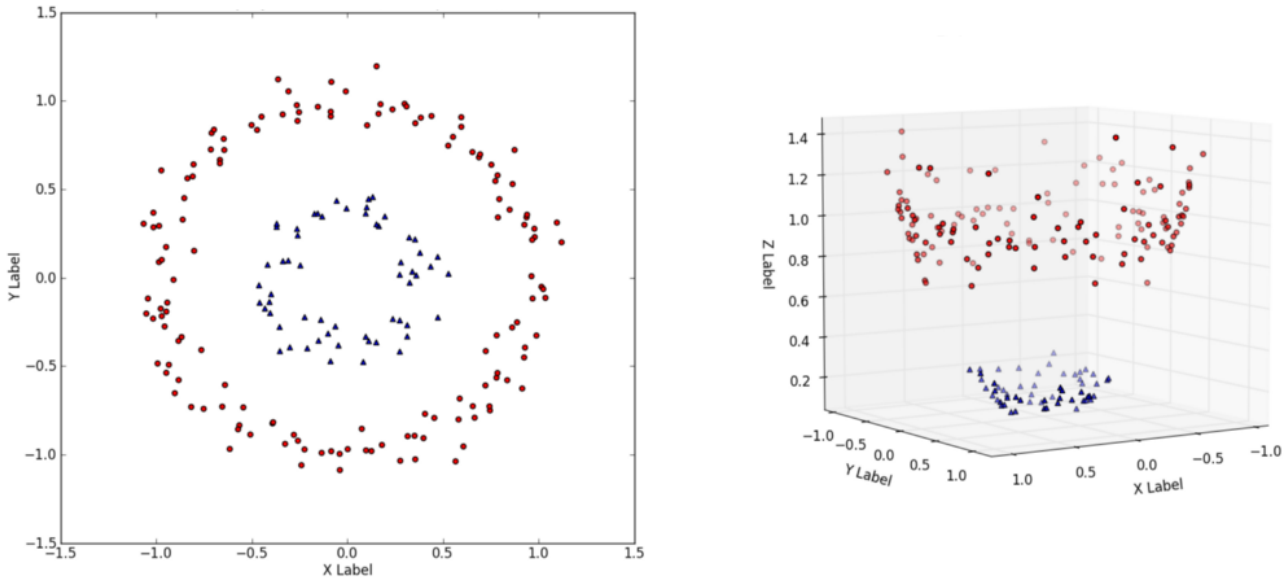

Figure 2.2: (Left) Non-linearly distributed data of two classes in a 2D representation space.
(Right) Linearly separable data of the same classes in a 3D representation space after the application of a kernel [3].

shallow representation techniques of texts such as BoW, TF-IDF or N-Grams showing respectable performances in the classification of texts [17] [18], these representations are prone to being unable to generalise across multiple domains [14] as the texts are typically represented using simple mechanisms that are unable to represent high-level information such as context or the semantic relationship of words..

TF-IDF Score

REEEEEEEEEEEEE The TF-IDF score of a word represents its importance in a specific document relative to all other documents within a corpus. Based on the BoW model, the TF-IDF calculates both the frequency of a word within a document (term frequency) and the

N-Grams

Word2Vec

Word Embeddings

- ngrams
- neural language models e.g. word2vec

2.3 Deep Learning Approaches

Deep learning models are a class of machine learning models that have the capability of automatically learning a hierarchical representation of data [7]. These hierarchical representations are constructed through the use of artificial neural networks, the main underlying mechanism of deep learning models. Typically, large amounts of training data is required to train a model in learning the language model required to attain state of the art results, in the task of text classification for instance, the size of commonly used non-domain specific datasets range from hundreds of thousands of training examples to millions [9] [17] *(note: look into the datasets used by state of the art approaches)*. Due to these constraints, it is not feasible to procure a dataset for the domain specific task of this project due to the aforementioned knowledge expertise and time requirements to manually label the articles required. Hence, *(Talk about transfer learning/N-shot learning/domain adaptation here)* will be used to overcome this issue.

Introduce the state-of-the-art DL based approaches for text classification and try to compare it performance with state-of-the-art ML approaches
For each model, talk about the following:
NOTE: REMEMBER WHEN WRITING THIS SECTION TO ALWAYS CONSIDER HOW IT DIFFERS TO ML TECHNIQUES

- *How it works and the mechanisms involved*
- *Advantages*
- *Limitations*

Introduce the typical architectures used for text classification e.g. RNNs, LSTMs, CNNs, GRUs?

2.3.1 Deep Learning Models

Recurrent Neural Networks

Gated Recurrent Unit Networks

Long Short-Term Memory Networks

Convolutional Neural Networks

2.3.2 Deep Textual Representation Techniques

2.3.3 GET PROPER NAME FOR THIS SECTION

Transfer Learning

Talk about transfer learning and how it works and how it is applicable to this project.

N-Shot Learning

Talk about zero/few/etc-shot learning and how it works and how it is applicable to this project.

2.4 Conclusion

Summarize lit review and describe why DL-based approaches should be preferred over ML-based for this type of problem. Also talk about Transfer/N-Shot learning and describe which one will be feasible given the project's time constraints

Chapter 3

Proposed Approach

3.1 Rationale

Introduce and discuss the factors that led to me choosing the proposed approach

3.2 Credibility Criteria

Introduce and discuss the 7 criteria that will be classified and describe how the criteria was determined

3.3 Study Data

Talk about the data I'll be using, how we got it, its characteristics etc.

3.4 System Model

Describe the architecture of the model

3.5 Experiments

Describe the experiments that I'm planning to do (in such a way that they are easily reproducible)

3.6 Outcome Measures

Talk about the type of analyses that I'll be doing to determine the performance of my proposed model

Chapter 4

Conclusions and Future Work

4.1 Conclusions

The end

4.2 Future Work

Chapter 5

Abbreviations

AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BS	Base Station
CSI	Channel State Information
CSIR	Channel State Information at Receiver
CSIT	Channel State Information at Transmitter
dB	Decibels
DPC	Dirty Paper Coding
GS	Gram-Schmidt
RVQ	Random Vector Quantisation
SISO	Single Input Single Output
SNR	Signal to Noise Ratio
SINR	Signal to Interference plus Noise Ratio
MISO	Multiple Input Single Output
SIMO	Single Input Multiple Output
MIMO	Multiple Input Multiple Output
MMSE	Minimum Mean Square Error
MRC	Maximum Ratio Combining
QoS	Quality of Service
TDD	Time Division Duplex
FDD	Frequency Division Duplex
ZF	Zero-Forcing
ZFBF	Zero-Forcing Beamforming
ZMCSCG	Zero Mean Circularly Symmetric Complex Gaussian

Appendix A

name of appendix A

A.1 Overview

here is the Overview of appendix A ...

A.2 Name of this section

here is the content of this section ...

Appendix B

name of appendix B

B.1 Overview

here is the Overview of appendix B ...

B.2 Name of this section

here is the content of this section ...

Bibliography

[1] "DISCERN - The DISCERN Instrument." [Online]. Available: http://www.discern.org.uk/discern{_}instrument.php

[2] "Our Review Criteria - HealthNewsReview.org." [Online]. Available: https://www.healthnewsreview.org/about-us/review-criteria/

[3] "Understanding the kernel trick. – Towards Data Science." [Online]. Available: https://towardsdatascience.com/understanding-the-kernel-trick-e0bc6112ef78

[4] C. C. Aggarwal and C. Zhai, "A SURVEY OF TEXT CLASSIFICATION ALGORITHMS," 2012. [Online]. Available: http://alias-i.com/lingpipe/

[5] M. A. Aizerman, E. A. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, June 1964.

[6] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," Tech. Rep., 2017. [Online]. Available: http://en.wikipedia.org/wiki/Statistics

[7] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," Tech. Rep., 2000. [Online]. Available: www.elsevier.com/locate/jmicmeth

[8] D. C. Burgess, M. A. Burgess, and J. Leask, "The MMR vaccination and autism controversy in United Kingdom 1998-2005: Inevitable community outrage or a failure of risk communication?" *Vaccine*, vol. 24, pp. 3921–3928, 2006. [Online]. Available: https://ac.els-cdn.com/S0264410X06002076/1-s2.0-S0264410X06002076-main.pdf?{_}tid=46d1dda6-f576-4f5e-ad53-d550f1cd9990{&}acdnat=1534726962{_}1b237371d8bb916694f34f0f951c84bc

[9] A. Conneau, H. Schwenk, Y. Le Cun, and L. Loıc Barrault, "Very Deep Convolutional Networks for Text Classification," 2017. [Online]. Available: https://arxiv.org/pdf/1606.01781.pdf

[10] F. Informatik, "UNIVERSIT AT D ORTMUND Text Categorization with Support Vector Machines: Learning with Many Relevant F eatures Thorsten Joachims," Tech. Rep., 1997. [Online]. Available: https://www.cs.cornell.edu/people/tj/publications/joachims{_}97b.pdf

[11] V. Korde and C. N. Mahender, "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY," *International Journal of Artificial Intelligence & Applications (IJAIA)*, vol. 3, no. 2, 2012. [Online]. Available: http://aircconline.com/ijaia/V3N2/3212ijaia08.pdf

[12] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," Tech. Rep., 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324{&}rep=rep1{&}type=pdf

[13] K. Pasupa, "A Comparison between Shallow and Deep Architecture Classifiers on Small Dataset," 2016.

[14] R. Rosenfeld, "TWO DECADES OF STATISTICAL LANGUAGE MODELING: WHERE DO WE GO FROM HERE?" Tech. Rep., 2000. [Online]. Available: https://www.cs.cmu.edu/{~}roni/papers/survey-slm-IEEE-PROC-0004.pdf

[15] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," Tech. Rep. Cohen, 1998. [Online]. Available: http://research.microsoft.com/en-us/um/people/horvitz/junkfilter.htm

[16] D. Zeraatkar, M. Obeda, J. S. Ginsberg, and J. Hirsh, "The development and validation of an instrument to measure the quality of health research reports in the lay media," *BMC Public Health*, vol. 17, no. 1, p. 343, dec 2017. [Online]. Available: http://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-017-4259-y

[17] X. Zhang, J. Zhao, and Y. Lecun, "Character-level Convolutional Networks for Text Classification," 2015. [Online]. Available: https://arxiv.org/pdf/1509.01626.pdf

[18] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," 2015. [Online]. Available: https://arxiv.org/pdf/1511.08630.pdf