Chapter 3

Methods

3.1 Approach

Through a review of the relevant literature, I found that there is value in being able to appraise the credibility of online health information for use in tools that could help limit the spread of misinformation. However, it is a challenge to apply credibility appraisal tools at scale because the process is time-consuming and often requires certain expertise. Tools for automating this task are therefore likely to be of value but to date most research in the area simply aims to label fake news or uses heuristics to identify misinformation available online. In this chapter, I describe the approach I used to address this challenge using machine learning to train classifiers capable of predicting whether an online article meets a set of credibility criteria.

This involved obtaining training and testing data consisting of online vaccine-related articles along with its associated credibility that were manually determined by a team of health professionals using a pre-existing framework (Section 3.2).

Once the dataset was created, a set of machine learning and deep learning models were implemented (Section 3.3). The machine learning models comprised of the implementation of widely used text classifiers, Naive Bayes (NB) and Support Vector Machines (SVMs), in combination with a variety of textual representation methods. The deep learning model involved the construction of a Quasi-Recurrent Neural Network (QRNN) in conjunction with a fine-tuned language model (LM).

The models are then analysed and evaluated on their feasibility for real-world use based on their ability to correctly predict the label for each criteria and other factors such as the training time of the models and their storage requirements (Section 3.4).

3.2 Study Data

3.2.1 Dataset

I obtained a set of vaccine-related online articles via the URLs included in a set of 6.59 million Twitter posts (tweets) from 1.86 million Twitter users collected between January 2017 and March 2018. There were 1.27 million unique URLs included in the set of tweets. A majority of the URLs led to web pages that could not be used as they were either broken or were links to other social media posts and YouTube videos. After removing web pages that could not be used in the analysis, I finally included a set of 3,348 vaccine-related online articles (the corpus). The content of the set of online articles included in the corpus ranged from descriptions of recent research to discussions (Figure 3.1) .

The articles included in the corpus were collected by selectively accessing the extracted URLs embedded within tweets that contained specific keywords related to vaccination on Twitter. This was done to ensure that the articles used represented, to some degree, the articles that are most likely to contribute to the effects of the propagation of misinformation as the articles that have been collected and labelled are the ones being shared and discussed online.

3.2.2 Credibility Criteria and Appraisal

The credibility criteria were developed by a team of three researchers from the Centre for Health Informatics as part of a separate project funded by the National Health & Medical Research Council. The seven criteria describe a set of desirable characteristics for online health information (Table 3.1). The development of the criteria was based on a set of existing tools and checklists (see Section 2.1), which was adapted by the researchers to be applicable to online articles about vaccination.

Prior to creating the dataset for this project, a small pilot experiment was conducted by the team that developed the criteria to ensure that the simplicity and objectivity of each criteria allowed for consistent and reliable labels to be produced. In the pilot test of the credibility criteria, each of the three researchers manually and independently labelled the same set of 30 online articles. The team then measured how well their answers matched, resolving differences by discussion, and updating the definitions of the criteria to improve the consistency.

For the purposes of training and evaluating the performance of the machine learning and deep learning models for this non-standard classification task, the team then manually labelled an additional 470 online articles based on the criteria outlined in Table 3.1. Once labelled, the article's credibility was then quantified using a credibility score which is equivalent to the total number of labels an article has. Due to the infeasibility

Thousands of parents describe their children as fine one day, then their children suddenly develop autism (a neurological regression) after vaccines.

What does the science say? It agrees...

But first lets go through the numbers. The increase in autism cases in the last three decades is truly shocking. Before the 1980s, autism was so rare, it was not even tracked. Remember the eye-opening movie RainMan with Dustin Hoffman? Before that movie came out in the early 1980s, many had no idea what autism was.

At that time, statistics put the rate of autism at 1 in every 8,000 children. By the mid-1990s, it was around 1 in 1000 children, but by the mid-2000s, it had risen to 1 in 250. The epidemic has continued and in 2017, the US National Center for Health Statistics just released the latest rate: 1 in every 36 children now have autism (link to report here). These skyrocketing rates clearly prove there is a true epidemic of autism in the United States.

So what is autism? Autism is brain damage caused by brain inflammation, which can be triggered by the heavy metals used as adjuvants (aka ingredients) in vaccines. Just as thousands of parents saw firsthand with their own children. Read some of these stories here.

Doctors usually try to deny any responsibility and connection to the vaccines they gave by saying Autism is genetic. But autism is even listed as a possible reaction (or side effect) of vaccines on some of the vaccine inserts. Unfortunately, doctors rarely see these inserts, favoring the pharmaceutical marketing sheets over real science.

So lets get to that real science the type that is NOT funded by the pharmaceutical industry that wants to sell vaccines and then the prescription drugs that these children are on to control the symptoms of autism, an issue possibly caused by the vaccines.

Figure 3.1: Excerpt from a low credibility vaccine-related online article [2].

of developing the expert-level knowledge and skill set required to manually label the articles given the time constraints of this project, I did not participate in this labelling process.

When the team sampled articles for expert labelling, they over-sampled from online articles that were linked to published peer-reviewed research literature using Altmetric, an organisation that provides access to information about how often research articles are mentioned in online media. As a consequence, the proportion of online articles that met the first criterion (identifying the source of evidence) was higher than might be expected from a random sample (Figure 3.2). Relatively few articles satisfied criterion 3 or criterion 7, while most articles satisfied criterion 4 and criterion 6. Overall, the resulting data were relatively unbalanced, which has implications in relation to the training of classifiers, as discussed in Chapter 4.

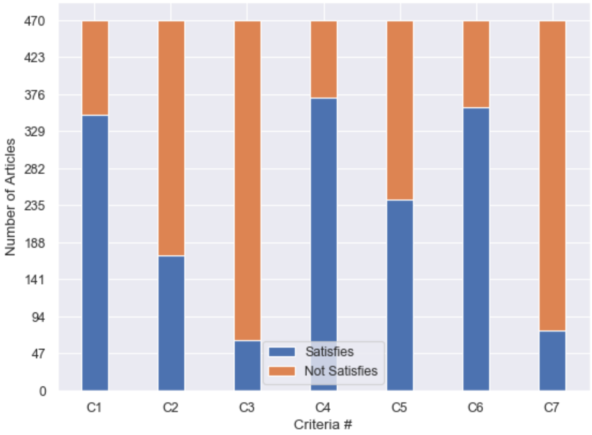


Figure 3.2: Distribution of labels amongst the expert-labelled articles.

I then aggregated the total number of criteria met by an aritcle to produce a credibility score for each of the 470 documents, which produces a score that may take any integer value between 0 and 7. Partly because of the way the online articles were sampled for expert labelling, the credibility scores for the expert-labelled set were unevenly distributed. The largest proportion of the expert-labelled articles received a credibility score of 4; only X and X received a credibility score of 0 and 7, respectively (Figure 3.3).

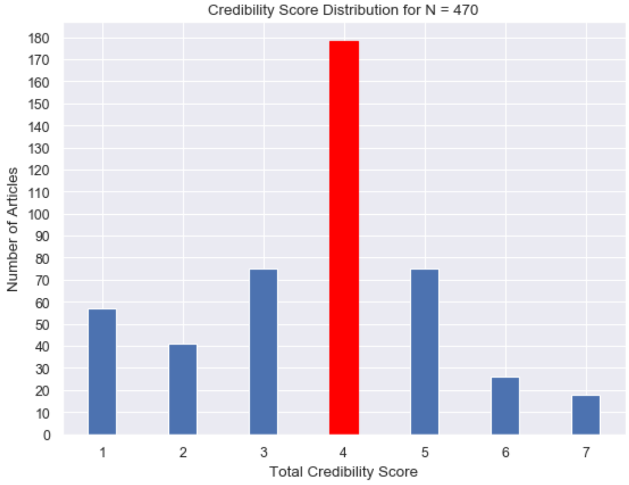


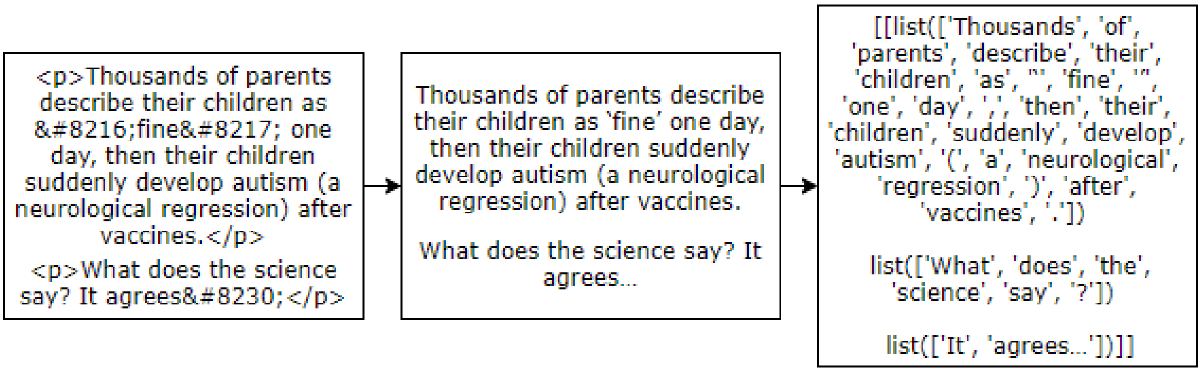
Figure 3.3: Distribution of article scores amongst the expert-labelled articles.

#	Criteria	Description
1	Identifies the information sources supporting a position	Score 1 if the article clearly indicates what sources of information are used to support its main claims or statements.
2	Uses information sources based on objective, scientific research.	Score 1 if the article's main claims are based on objective, scientific research. Score 0 if the article uses anecdotal evidence exclusively to support its main claims.
3	Communicates the strength or weakness of the evidence used to support a position.	Score 1 if the article includes adequate details about the level of evidence offered by the research.
4	Does not exaggerate or overstate a position.	Score 0 if the article: - Uses unjustified sensational language - Presents information in a sensational, emotive or alarmist way - Selectively or incorrectly presents evidence
5	Presents information in a balanced manner.	Score 1 if the article: - Identifies/acknowledges uncertainties and limitations in the research - Acknowledges where an issue is controversial, and includes all reasonable sides in a fair way - Uses a range of information sources - Provides a balanced description of the strengths and weaknesses of the study
6	Uses clear, non-technical language that is easy to understand.	Score 1 if the article: - Is professionally written, with proper grammar, spelling, and composition - Defines any technical jargon, or uses everyday examples to explain the technical terms or concepts
7	Is transparent about sponsorship and funding.	Score 1 if the article: - Clearly distinguishes content intended to promote or sell a product from service from educational and scientific content - Discloses sources of funding for the organisation/website

3.3 Model Implementation

3.3.1 Preprocessing

Preprocessing of the unlabelled and expert-labelled articles of the dataset to obtain its raw article text was performed by removing the superfluous content from the web-scraped pages such as the HTML, CSS and JavaScript elements. They were removed as they were deemed to have no contribution to the credibility of the article's content with regards to the credibility criteria. Once the unstructured article text was obtained, it was transformed into a more functional representation via tokenising each article into sentences that consisted of unigrams (Figure 3.4). The entire process was ensured to be general and simple as the various features that have been implemented and tested would require various kinds of transformations as explained in further detail under their respective section in Section 3.3.2.



```
<p>Thousands of parents describe their children as &#8216;fine&#8217; one day, then their children suddenly develop autism (a neurological regression) after vaccines.</p>
<p>What does the science say? It agrees&#8230;</p>
```

```
Thousands of parents describe their children as 'fine' one day, then their children suddenly develop autism (a neurological regression) after vaccines.

What does the science say? It agrees...
```

```
[[list(['Thousands', 'of', 'parents', 'describe', 'their', 'children', 'as', '‘', 'fine', '’', 'one', 'day', ',', 'then', 'their', 'children', 'suddenly', 'develop', 'autism', '(', 'a', 'neurological', 'regression', ')', 'after', 'vaccines', '.'])

list(['What', 'does', 'the', 'science', 'say', '?'])

list(['It', 'agrees...'])]]
```

Figure 3.4: Example output of preprocessing phase.

3.3.2 Feature Selection

Bag of Words

Two n-gram BoW variants, a unigram and bigram BoW model, was constructed using the expert-labelled articles both of which were limited to a maximum of 20,000 n-grams. Similar BoW models were also constructed for different n-gram amounts, specifically, BoW models for 5,000 and 10,000 n-grams were also tested. Common English stopwords such as 'the', 'is' and 'are' listed in NLTK's English stopwords corpus [12] were removed from the unigram BoW model but remained for the bigram BoW model.

Term Frequency - Inverse Document Frequency

TF-IDF scores of the unigrams extracted from the expert-labelled articles which were limited to the top 10,000 scores were calculated and used for evaluation. The top 10,000 TF-IDF scores consisted of scores for unigrams that appeared in at least two expert-labelled articles with no maximum occurrence limit.

GloVe

A pre-trained GloVe vector [38] was also used for the evaluation of the ML-based models. This was achieved by using the mean value of each word's vector representation. The GloVe vector used was trained on 6 billion tokens, with a vocabulary size of 400,000 and a has dimensionality of 300.

Language Models

A non-domain specific language model (LM) was constructed for evaluating the DL-based model. This was achieved by training a Quasi-Recurrent Neural Network (QRNN) using the same hyperparameter settings and training scheme outlined in [13] on the Wikitext-103 corpus. Wikitext-103 is a corpus consisting of approximately 28,000 Wikipedia articles that meet either the 'Good' or 'Featured' article criteria [31].

The LM was then fine-tuned using the techniques proposed by Howard et al. [23] by using the 2,878 unlabelled articles from the dataset in order to learn the task-specific features for classification. The Wikipedia articles used to create the non-domain specific LM also underwent the same preprocessing process described in Section 3.3.1 sans the removal of webpage elements.

3.3.3 Model Selection

In the context of this section, a model refers to the combination of a feature and classifier. The following sections detail the implementation process of the classifiers and the features used in conjunction to create the various models. For the software used to implement the models, the NB and SVM models were implemented using scikit-learn, an open source machine learning library in Python [37]. The QRNN model which utilised the LM and QRNN classifier was implemented using PyTorch [36] and the open-source software developed by Bradbury et al. [13].

Naive Bayes

A multinomial NB classifier was used for evaluation, based on the reasoning discussed in Section 2.2.2. Along with the four features used in conjunction with the NB classifier (unigram BoW, bigram BoW, TF-IDF and GloVe), a total of 28 NB models (7 criteria × 4 features) were developed. The hyperparameter values were optimised via grid search and evaluated using 10-fold cross-validation.

Support Vector Machine

A linear support vector classifier (linear SVC), an SVM with a linear kernel, was used for evaluation. Similar to the implementation of the NB models, a classifier was trained for each criteria with the same set of features totalling 28 SVM models. The hyperparameter values were also optimised and evaluated using grid search and 10-fold cross-validation.

Quasi-Recurrent Neural Network

In addition to using a QRNN for constructing the fine-tuned LM, another variant was implemented as a classifier. The architecture for the classifier was adapted from the experiments conducted by Bradbury et al. [13] for sentiment classification. From this model, the number of units within each layer was increased to 2500 and the embedding size was changed to 150. The model was trained for up to 20 epochs with a batch size of 100, however the model began to converge after only 5 epochs. In contrast to the NB and SVM models, seven QRNN models (QRNN classifier + LM) were created to predict the correct label for each of the seven criteria via a softmax function applied to the output layer.

3.4 Experiments

Considering the context of this project, other factors such as the size of the model and its training speed must be considered in addition to its capability on correctly producing the correct labels to determine the feasibility of deploying the model for a real-world application.

3.4.1 Accuracy

Two experiments have been conducted to evaluate the performance of a model. The credibility classification experiment is designed to determine the model's ability to produce the correct label on a criteria-level allowing for a more detailed insight on the reasoning for a particular article's credibility score. Whereas the low credibility identification experiment is designed to determine the model's ability to correctly determine whether an article is considered to have low credibility or not for situations where the detailed insight is not required.

Credibility Classification Experiment

The credibility classification experiment aims to determine a model's ability in correctly identifying whether an article either satisfies or doesn't satisfy a certain criteria. For each criteria, the respective NB and SVM model is employed along with the QRNN model and is then evaluated by measuring the micro averaged f1-scores. This experiment attempts to assess the feasibility in providing insight on the factors that attributed to the credibility of an article.

Low Credibility Identification Experiment

The Low Credibility Identification Experiment was designed to evaluate the capability of a model to identify low credibility articles which have been defined to be articles with a credibility score lower than 3. The expert-labelled articles are separated into two separate classes based on this condition and the best performing model, which was determined from the results of the credibility classification experiment, is then applied and evaluated using the resulting micro averaged f1-scores.

3.4.2 Training

The training time for a model is another factor that must be considered when determining the feasibility of deploying and using a model on a real-world application. This is because the model will require to be constantly re-trained when additional article samples are incorporated as the model encounters new and never before seen information. To ensure consistent results, Amazon Web Service's EC2 instances [1] have been used to act as an isolated environment for timing the training of the NB, SVM, QRNN classifiers and also the creation of the LM. The reported training times in Section 4.2 are for the best performing NB and SVM-based models, which were trained within a c5.2xlarge instance utilising all CPU cores. The reported training times for the QRNN classifier and LM were obtained using the P3 instances and utilising only a single K80 GPU.

3.4.3 Storage

To minimise the frequency in which a model will have to be re-trained or re-created, storing the model for later use is required. Since the entire implementation was completed using Python, the Python module *pickle* was used to store the models externally. Pickle serialises any Python object into a byte stream allowing a model to be persistently stored. The NB and SVM-based models were stored using pickle and the size of the resulting byte stream was then measured. The QRNN classifier and LM were stored using PyTorch's inbuilt save function, that acts as a wrapper over the *pickle* module which provides storage optimisations for PyTorch objects. The resulting byte stream is then recorded and reported in Section 4.3. test