Universität Koblenz - Landau
Department für Informatik,
Institute for Web Science and Technologies
Prof. Dr. Steffen Staab
Dr. Chandan Kumar
Lukas Schmelzeisen

6. November 2018

Master Thesis Kandhasamy Rajasekaran

Fake news classification through Wikipedia using recurrent neural networks

1 Introduction

According to Lazer et al. [1], fake news is a false information, constructed intentionally (disinformation) or unintentionally (misinformation) where the publishers do not have rigorous news media editorial norms for making sure of its accuracy and credibility. The focus of this master thesis is to classify a claim as true or fake news by using Wikipedia as ground truth. In this thesis, different approaches will be used to model the fake news classification process using Wikipedia.

Fake news is prevalent, as can be seen by a research study from Allcot and Gentzkow [2], which states that "the average American adult saw on the order of one or perhaps several fake news stories in the months around the election, with just over half of those who recalled seeing them believing them". A large-scale empirical study with twitter dataset by Vosoughi et al. [3], reveals that fake news spread longer, faster, deeper and broader than legitimate news. It also reveals that the effects of fake news are more prominent in a political context than news about terrorism, natural disaster, science, and other domains.

Most often fake news is detected by people and organizations through their common sense. There are many facts checking websites such as snopes.com, factcheck.org, politifact.com, etc., which use collaborative effort of domain experts. In these websites, each news article is tagged with a fact meter by domain experts to quantify its authenticity. The main strategy to handle fake news is to check against a reliable source of information and classify its integrity. This process works under the assumption that domain experts have good knowledge on the subject of news. Although this method helps in classifying fake news effectively, since it requires manual effort, it is expensive and unscalable to all domains. citation pending - Also, the news articles that are checked by this manual process is very less in comparison to the rate at which the information is produced because of many social networks, blogging sites, etc.

Wikipedia is a free online encyclopedia available in more than 300 languages with a principle that anyone can edit [4]. According to Alexa[1] and SimilarWeb[2], Wikipedia is considered to be the fifth most popular website. The English Wikipedia[3] is the largest among all language editions of Wikipedia and it consists of more than 5 million articles. Also, the frequency of updates in the English Wikipedia is very high and it is approximately equal to 10 updates per second and 600 articles per day. The articles of Wikipedia are grouped into categories such as history, music, sports, etc., based on the

[1] https://www.alexa.com
[2] https://www.similarweb.com
[3] https://en.wikipedia.org/wiki/English_Wikipedia

main theme of its content. According to a research by Halavais et al. [5], it is estimated that categories that constitute at least 5% of total articles in the English Wikipedia are music, history, geography, philosophy, science, technology, and literature. Also, it is found that, for all categories, the average size of an article is at least 20 kilobytes (excluding images) and the average edits of an article are at least 8. Thus, these results show that the breadth and depth of information in the English Wikipedia is substantial. Although it can be edited by anyone, an investigation carried out by Nature[4], reveals that the quality of content is similar to another established encyclopedia such as Britannica[5] [4]. It has been shown that, although there are malicious users in Wikipedia, the community and its culture ensures most of the high impactful errors are rectified very quickly [6]. Thus, we hypothesize that the English Wikipedia can be used as a proxy for a reliable source of information since most of its content is true. It is true that we are dealing only with the english language edition and Wikipedia does not reflect current affairs which is a crucial factor for considering it to cross check fake news since most of the fake news are based on current affairs.

In this master thesis, the English Wikipedia will be considered as a ground reality or as a source of experts opinion and this knowledge will be used to cross-check claims automatically. The presence or absence of a claim in the English Wikipedia will be used as a proxy for classifying that claim as true or fake. Specifically, deep learning [7] will be used to build different models for classifying fake news efficiently. Some of the models will take both claim and context information about claim as inputs. This context information will either be extracted from the English Wikipedia or derived sources like DBpedia or Wikidata. We also hypothesize that a model trained using both input and relevant context information will perform better as a fake news classifier than the model trained only with input. Different representations of context information will be experimented and evaluated based on how well it classifies fake news.

Section 2 will show a brief summary of different research about fake news classification using semantics and platform based approaches. Section 3 will introduce the required deep learning background to build fake news classifiers. Section 4 will explain the architecture of proposed system and the overall approach to be taken in this thesis. Section 5 will discuss different methods for evaluating the proposed fake news classifiers.

2 Related work

There has been much effort to detect fake news in microblogging platforms such as Twitter[6]. Most of these works [8, 9] classify a microblog as truth or fake by using platform/user-specific information such as how popular the post is, the credibility of the user who shared it, diffusion patterns, etc. Zhao, et al. [10] have used cue terms such as “not true”, “unconfirmed” etc., in retweets or comments to detect fake news. The reasoning in their approach is that when people are exposed to fake news they will comment or retweet with such words as a response to the post. Other studies focused on using temporal characteristics of fake news during the spread. Kwon et al. [11] used tweet volume in time series and Ma et al. [9] measured variations of social context features over time.

However, the mentioned works only used comparatively small datasets. Wang [12] curated a dataset consisting of approximately 13,000 short statements by mining politfact.com covering a decade of information. In this approach, six machine learning models have been built ranging from logistic

[4]https://www.nature.com
[5]https://www.britannica.com/
[6]https://twitter.com/

regression to a convolutional neural network. Along with the text data, metadata such as the speaker, subject, speaker history are also included. The convolutional neural network model used to capture surface level linguistics along with metadata performed better than other models.

Jing Ma et al. [13] efforts have been focused on building a recurrent neural network (RNN) to detect rumors from microblogs such as Twitter and Weibo[7] effectively. The training dataset is obtained by using keywords from fake and true news of debunking services such as Snopes[8] and the Sina community management center[9]. These keywords are used in search APIs of microblogging platforms and the found microblogs are labeled as true or fake, respectively. The social context information of a post and all its relevant posts such as comments or retweets are modeled as a variable-length time series. RNNs with different configurations are used to capture textual representations and the temporal patterns associated with those posts. An accuracy of at least 80% was achieved by architectures using RNNs on both datasets. This method avoids handcrafted feature engineering efforts which are biased and time-consuming by only considering the text of posts. It produced better results with datasets from Twitter and Sina Weibo than all of the traditional machine learning methods it was compared against. RNNs with two layers of GRU captured long distance dependencies very well and achieved 88% and 91% accuracy on twitter and weibo dataset respectively. An accuracy of more than 80% was achieved with a training time of 12 hours, whereas the average time taken for a single fake news item by debunking services were more than 50 hours.

However, the mentioned work by Jing Ma et al. is only for microblogging platforms. And also the features specified in one platform will be different from another and it is not guaranteed that this methodology will give the same results for the same news in two different platforms. The semantics of the tweets are not used and if used then it might improve the chance of achieving same results across platforms.

Ciampaglia et al. [14] used DBpedia[10], a knowledge graph, for checking whether a given information is factual or not. DBPedia datasets represent infobox section in Wikipedia. Therefore, it contains only non-controversial and factual information because it stems from human collaboration. Their methodology formulates the problem of checking facts into a network analysis problem which is finding the shortest path between nodes (subject and object of a sentence) in a graph. The aggregated generalities of nodes along a path in a weighted undirected graph are used as a metric for measuring the authenticity of information. The more elements are generic; the weaker the truthfulness is. The genericness of a node is obtained by the degree of that node, that is the number of nodes connected to that node. The truthfulness of the information is improved if there exists at least one path from subject to an object with non-generic nodes. This approach exploits the indirect connections to a great extent with distance constraints in a knowledge graph. The approach was evaluated against four datasets: directors and their movies, US presidents and their spouses, US states and their capital and countries and their capital. The approach assigns better scores to true statements than the false statements with probabilities 0.95, 0.98, 0.61 and 0.95 for director, spouse, US capital and world capital datasets, respectively.

The mentioned work by Ciampaglia et al. is a good step towards an automated fact checker system using only the semantics of data. However, the methodology implemented is primitive and can work only for statements which have a simple relationship between the subject and object. We should also say that the evaluation approach used simple datasets and it is not really fake news. blah blah

[7] https://www.weibo.com
[8] https://www.snopes.com/
[9] https://www.sina.com.cn/
[10] https://wiki.dbpedia.org/

3 Background Study

In this thesis, the proposed system will use many neural network architectures to classify fake news. The required information to understand different components of the proposed system is explained briefly in this section. Every sentence of Wikipedia would be fed and the system will be trained to remember it. Several other techniques will be tried to improve the efficiency in learning. One of it is extracting information related to an input sentence, inputting it along with sentence and training the system. This is achieved by using attention mechanism in neural networks.

3.1 Neural networks and its components

Recently, neural networks have become the most popular approach for building machine learning systems. Neural networks compose many interconnected, fundamental, functional units called neurons. Each neuron in the network takes in multiple scalar inputs and multiplies each input by a weight and then sums them, adds the result with a bias, applies a non-linear function at the end, which gives out a scalar output. There are different architectures of neural networks which vary mostly in how the neurons are connected to each other and how the weights are managed.

Feed-forward neural networks [15] can have multiple layers and each neuron in one layer is connected with every other neuron in the subsequent layer as given in the Figure 1

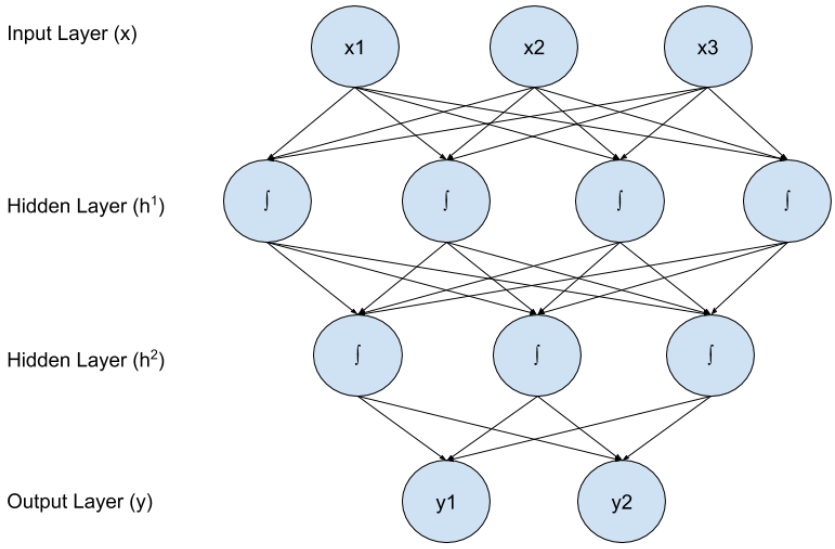

Figure 1: A Feed-Forward neural network.

There are 4 layers in figure 1. Each circle is a neuron with incoming lines as inputs and outgoing lines as outputs to the next layer. Each line carries a weight and the input layer has no weights since it has no incoming lines. The input layer consists of 3 neurons and the extracted features of raw data will be sent through these neurons. The first hidden layer consists of 4 neurons, of which each neuron takes 3 inputs from input layer. Each input to the neuron in first hidden layer is multiplied by a unique weight variable and added together. Finally, the output is added with a bias variable and will be passed to a non-linear activation function as shown in equation 1. The same process is carried out for the second

hidden layer except that it has 3 neurons and each neuron will take 4 inputs from the first hidden layer. The output layer consists of 2 neurons and each will take 3 inputs from second hidden layer as shown in equation 3.

$$h^1 = g^1(xW^1 + b^1) \tag{1}$$

$$h^2 = g^2(h^1W^2 + b^2) \tag{2}$$

$$\mathrm{NN}_{\mathrm{MLP2}}(x) = y = g^3(h^2W^3 + b^3) \tag{3}$$

$$\begin{aligned} x \in \mathbb{R}^{d_{in}}, y \in \mathbb{R}^{d_{out}} \\ W^1 \in \mathbb{R}^{d_{in} \times d_1}, b^1 \in \mathbb{R}^{d_1}, h1 \in \mathbb{R}^{d_1} \\ W^2 \in \mathbb{R}^{d_1 \times d_2}, b^2 \in \mathbb{R}^{d_2}, h2 \in \mathbb{R}^{d_2} \\ W^3 \in \mathbb{R}^{d_2 \times d_{\mathrm{out}}}, b^3 \in \mathbb{R}^{d_{out}} \end{aligned}$$

Here $\mathbf{W^1}$, $\mathbf{W^2}$, $\mathbf{W^3}$, and $\mathbf{b^1}$, $\mathbf{b^2}$ and $\mathbf{b^3}$ are matrices and bias vectors for first, second and third linear transforms, respectively. The functions g^1, g^2 and g^3 are activation functions and they are almost always non-linear. With respect to figure 1, the values of d_{in}, d_1, d_2 and d_{out} are 3, 4, 3, and 2, respectively.

The activation functions help the neural network models to approximate any nonlinear function. Different activation functions pose different advantages. Some popular activation functions are sigmoid, hyperbolic tangent and rectifiers [16]:

1. The sigmoid activation function is a S-shaped function which transforms any value into the range between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

2. The hyperbolic tangent function is also a S-shaped function, but it transforms any value into the range between -1 and 1.

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

3. The rectifier activation function clips values lesser than 0

$$\mathrm{ReLU}(x) = \max(0, x)$$

The sigmoid activation function is not used in internal layers of neural networks since other functions have been giving better results empirically. The rectifier activation function is commonly used since it performs faster and better than sigmoid and hyperbolic tangent functions. Instead of an activation function, the function in output layer g^3 can be a transformation function such as softmax to convert values to represent a discrete probability distribution. Each of the converted values will be between 0 and 1 and sum of all of them will be 1.

$$\begin{aligned} y = \begin{bmatrix} y_1 & y_2 & \dots & y_k \end{bmatrix} \\ s_i = \mathrm{softmax}(\mathrm{y_i}) = \frac{e^{y_i}}{(\sum_{j=1}^{k} e^{y_j})} \end{aligned}$$

3.2 Training a neural network

Training is an essential part of learning and like many supervised algorithms, a loss function is used to compute the error for the predicted output against the actual output. The gradient of the errors is calculated with respect to each weight and bias variable by propagating backward using chain rule of differentiation. The values of the weights and bias are adjusted with respect to the gradient and a learning parameter. Typically a random batch of inputs is selected and a forward pass is carried out which involves multiplying weights, adding bias and applying an activation function to predict outputs. The average loss is computed for that batch and the parameters are adjusted accordingly. This optimization technique is called stochastic gradient descent [17]. A number of extensions exists, such as Nesterov Momentum [18] or AdaGrad [19]. Some loss functions that exist are hinge loss (binary and multiclass), log loss and categorical cross-entropy loss [16].

The categorical cross-entropy loss is used when predicted output refers to a probability distribution. This is typically achieved by using a softmax activation function in the output layer. Let $y = y^1, y^2, \ldots, y^n$ be representing the target multinomial distribution over the labels $1, 2, \ldots, n$ and let $\hat{y} = \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ be the network's output which is transformed by a softmax function. The categorical cross-entropy loss measures the difference between the true label distribution y and the predicted label distribution $\hat{y}$.

$$L_{\text{cross-entropy}}(\hat{y}, y) = -\sum_i y_i \cdot \log(\hat{y_i})$$

For hard classification, y is a one-hot vector representing the true class. Here t is the correct class assignment. Training will attempt to set the correct class t to 1 which inturn will decrease the other class assignment to 0.

$$L_{\text{cross-entropy(hardclassification)}}(\hat{y}, y) = -\log(\hat{y_t})$$

The overfitting in neural networks will cause the trained system to perform well only for trained data but not on the test data. This can be minimized by using regularization techniques such as L_2 regularization and dropout [20]. L_2 regularization works by adding a penalty term equal to sum of the squares of all the parameters in the network to the loss function which is being minimized. Dropout, instead, works by randomly ignoring half of the neurons in every layer and corrects the error only using the parameters of other half of neurons. This helps to prevent the network from relying on only specific weights.

3.3 Neural networks for text data

In case of text data, the input is sequential and of unknown length, where the ordering of words is important. Techniques such as continuous bag of words [21] can be used with feed-forward networks to convert sequential input into fixed length vectors but it will discard the order of words. Convolutional neural networks (CNN) [22] are good at capturing the local characteristics of data irrespective of its position. In this, a nonlinear function is applied to every k-word sliding window and captures the important characteristics of the word in that window. All the important characteristics from each window are combined by either taking maximum or average value from each window. This captures the important characteristics of a sentence irrespective of its location. However, because of the nature of CNNs they fail to recognize patterns that are far apart in the sequence.

Recurrent neural networks (RNN) accept sequential inputs and are often able to extract patterns over long distances [23]. A RNN takes input as an ordered list of input vectors $\mathbf{x_1}, \ldots, \mathbf{x_n}$ with initial state vector $\mathbf{h_0}$ and returns an ordered list of state vectors $\mathbf{h_1}, \ldots, \mathbf{h_n}$ as well as an ordered list of output vectors $\mathbf{o_1}, \ldots, \mathbf{o_n}$. At time step t, a RNN takes as input a state vector $\mathbf{h_{t-1}}$, an input vector $\mathbf{x_t}$ and outputs a new state vector $\mathbf{h_t}$ as shown in figure 2. The outputted state vector is used as input state vector at the next time step. The same weights for input, state, and output vectors are used in each time step.

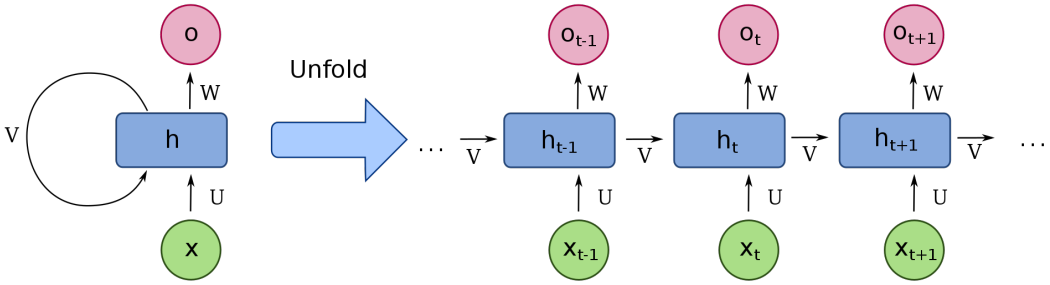

Figure 2: A basic example of RNN architecture [24].

$$\begin{aligned} \text{RNN}(h_0, x_{1:n}) &= h_{1:n}, o_{1:n} \\ h_i &= g^1(h_{i-1}V + x_iU + b^1) \\ o_i &= g^2(h_iW + b^2) \end{aligned}$$

$$\begin{aligned} x_i \in \mathbb{R}^{d_x}, U \in \mathbb{R}^{d_x \times d_h} \\ h_i \in \mathbb{R}^{d_h}, V \in \mathbb{R}^{d_h \times d_h}, b^1 \in \mathbb{R}^{d_h} \\ o_i \in \mathbb{R}^{d_o}, W \in \mathbb{R}^{d_h \times d_o}, b^2 \in \mathbb{R}^{d_o} \end{aligned}$$

Here the functions g^1 and g^2 are non-linear activation functions; $\mathbf{W}$, $\mathbf{V}$, and $\mathbf{U}$ are weight matrices and $\mathbf{b^1}$, $\mathbf{b^2}$ are bias vectors.

To train a RNN, the network is unrolled for a given input sequence and the loss function is used to compute the gradient of error with respect to parameters involved in every time step by propagating backward through time. After that, the parameters are adjusted to reduce the error in prediction [25]. While training RNNs, a common problem that especially occurs with long input sentences is that the error gradients might vanish (become too close to zero) or explode (become too large) which results in numerical instability during the backpropagation step. The gradient explosion can be handled by clipping a given gradient when it goes beyond the threshold. LSTM networks [26] solve the vanishing gradient problem by introducing memory cells which are controlled by gating components. These gating components decide at each time step, what parts of the hidden state should be forgotten and what parts of new input should be included into the memory cells. These memory cells are involved in the computation of hidden states, which in turn are used to compute the output states. This technique has been shown to provide good results in practice, in capturing the dependency between words even though separated by a long distance.

3.4 Neural networks for improving fake news classification

A RNN computes the state of current word $\mathbf{x_i}$ only based on the words in the past, i.e. $\mathbf{x_1}, \ldots, \mathbf{x_{i-1}}$. However, the following words $\mathbf{x_{i+1}}, \ldots, \mathbf{x_n}$ will also be useful in computing the hidden state regarding the current word. The Bidirectional RNN (biRNN) (Schuster and Paliwal, 1997) solves the problem by having two different RNNs. The first RNN is fed with the input sequence $\mathbf{x_1}, \ldots, \mathbf{x_n}$ and the second RNN is fed with input sequence in reverse. The hidden state representation $\mathbf{h_i}$ is then composed of both the forward and backward states. Each state representation consists of the token information along with sentential context from both directions which has shown better results than classical uni-directional RNNs in practice.

The mentioned approaches take only Wikipedia sentence as an input. This can be extended further by feeding information which gives context about the input sentence. This extra information might enable the system to perform efficiently in classifying whether a news item is fake or not. Attention mechanism in neural networks is used to take two inputs and convert it into one common representation. This is done by configuring the level of attention that needs to be given to different parts of each input and combine them to form a single output. Most often the inputs will be matrices and the output will be a vector. Bidirectional RNNs can be used to convert both input and context information into matrices. The fixed output vector is then fed into a deep feed-forward neural network to predict a class. Attention mechanism with RNN gives state of the art results in many NLP applications such as Natural Language Inference [27], Machine Translation [28] and Document classification [29].

4 Approach

The task is to build a fake news classifier using Wikipedia and cross-check claims. The input to this classifier will be a sentence representing a claim and the output from this classifier will be a binary value representing the inputted sentence as true or fake. For example, when the sentence 'Obama is a muslim' is sent as an input to classifier then the output will be a binary value '0' representing the input is fake. Similary, for the input sentence 'Obama is a christian', the classifier will output a binary value '1' representing the input is truthy.

4.1 Research Questions

Through this work, the research questions that would be addressed are:

1. Wikipedia is large and the proposed system may not be able to remember all the sentences. How well can we memorize or recognize sentences from Wikipedia in a neural network? This will give an understanding of the limitations of proposed system.
2. How well Wikipedia be used for fake news detection?. Does Wikipedia contains all relevant information to detect wide spectrum of fake news? This would uncover the limitations of Wikipedia to be used as a fake news classifier.
3. Since it is difficult to remember everything in Wikipedia, depending on the claim to be checked, selected portions of Wikipedia will be retrieved and used as a context. This approach will overcome the limitation of not being able to remember everything in Wikipedia to a certain extent. How good can we recognize fake news information in context data from Wikipedia?

4.2 Model

Model - general architecture Different kinds of context Represent input + context detailed model architecture (type of RNN, type of attention mechanism)

A general architecture of the proposed system is as shown in figure 3

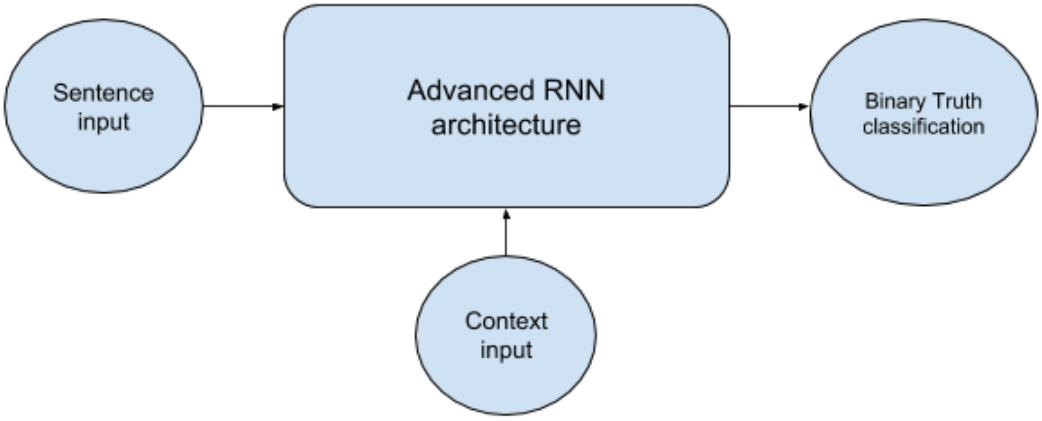

Figure 3: Broad Architecture

The reason behind using context input is that the RNN system would be better equipped with information to make a decision whether it is a true or fake news. And also that, the problem of fake news classification can be seen as a Natural Language Inference [27] problem, which is about finding entailment or contradictory relationship between two sentences such as hypothesis and premise. Here, the context input is the premise and sentence input is the hypothesis. If there is an entailment between sentence input and context then it is true news otherwise false.

There are many ways by which context input can be modeled and inputted to the systems. Extracting the context as relevant to sentence input can be challenging and a suitable information retrieval method will be selected for it.

1. The context can be empty. The system would be trained to understand Wikipedia and use this information alone to make a decision.
2. Top 10 Wikipedia articles relevant to the sentence will be passed as context input. The Wikipedia articles will be long and so will increase the size of context input. The system may face long distance dependency issues and a suitable hyper-parameter selection is critical. Since both the sentence and context input are text from Wikipedia, the system might find it easy to align one with respect to other and find patterns among them.
3. Nearly 100 triples of knowledge graph will be passed as context input. Resource Description Framework (RDF) triples of DBPedia can be used. Extracting such information from DBPedia will be relatively easier. The structure of context input will be simple and straightforward and the length of context input will be shorter than using Wikipedia articles.
4. Information such as parts of speech, named entity and dependency graph of a sentence will be passed as context input. These metadata will bring more clarity about the sentence structure, which will help the system to extract patterns efficiently.

A detailed architecture of an advanced RNN system is shown in the figure 4. The sentence input and context input will be encoded into a rich and convenient representation separately. Bidirectional RNNs can be used which will take a sequential input and convert it into a matrix. The number of rows of the matrix will be equal to number of tokens in a sentence and number of columns can be controlled. Each row in the matrix will represent a token with sentential context. Both the sentence and context representation are then combined to form an efficient common representation. The arbitrariness of both sentence and context representation will be controlled and converted into a representation of fixed length. Attention modules can be used which will figure out how much attention needs to be given to both the representations to produce a common representation. It usually takes two matrices and outputs a fixed vector. But the attention mechanism can be a pure reduction process, where it takes only one matrix and converts into a fixed vector. This can be used when the context input is empty. The predictor takes in a fixed vector as input and predicts which one of the binary class label(true or fake) is most likely the final output. A feed-forward neural network will be a good choice to be used as a predictor.

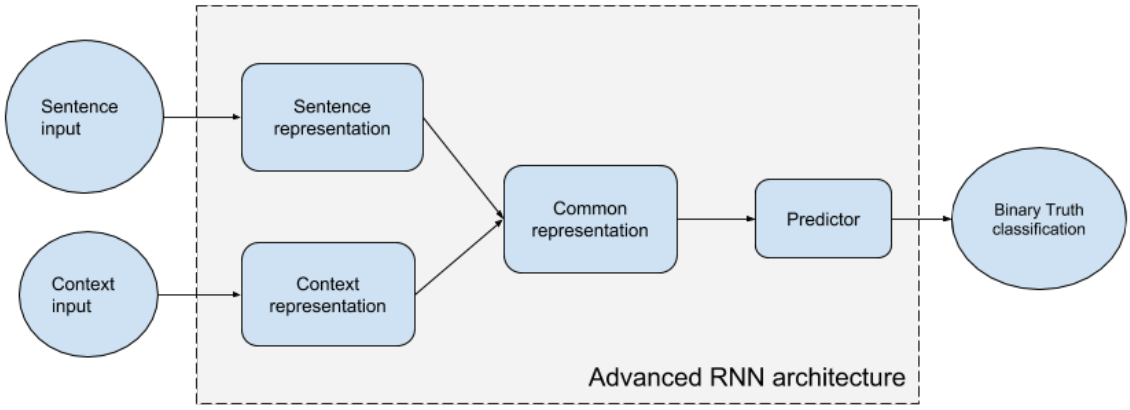

Figure 4: Detailed Architecture

4.3 Experiments

Experiments : 4 experiments Data wikipedia : sentence from wikipedia + negative examples through distortion Data fake news : kaggle - dataset or Ciampaglia dataset

1) Training - 90 percent data wikipedia, Testing - 10 data from wikipedia 2) 100 percent Data wikipedia, Data fake news 3) 90 percent Data fake news, 10 percent data fake news 4) Data from wikipedia and fake news, data fake news

Since we use Wikipedia as a proxy for authentic information, each sentence in the Wikipedia is considered as true news. There is hardly any false news present. For training to be successful to build a good fake news detector, it should be ensured that more or less equal amount of data should be present for both true and fake news. The construction of false news is tricky and different natural language processing (NLP) methods will be used to generate false news. Spacy[11], a python library for NLP tasks will be used for extracting parts of speech, named entity and dependency relations. Some techniques are

[11] https://www.spacy.io

1. Swap the words randomly in each and every sentence extracted from Wikipedia. The RNN models used will remember the sequence information and so a randomly shuffled sentence will be considered as a different sentence by the system. However, the syntax and semantics of the sentence created will be poor. This method is primitive but it will serve as a good baseline.

2. Replace the random words picked in a sentence with a random word from the corpus. This method would retain some of the words as it is and changing only few words will bring a subtle change in sentence. This is similar to random swapping but the syntactic structure of the sentence will be preserved to some extent.

3. Replace all or randomly picked words in a sentence based on their parts of speech. In this method, all the words in the corpus will be tagged with a parts of speech label. This information will be used and randomly a word will be picked from the corpus which has the same parts of speech as the word to be replaced. This method will preserve the syntactic structure of the sentence better than the previous methods.

4. Replace the words in each sentence with their opposites extracted from WordNet[12]. This method will not only preserve the syntax but also likely to have right semantics. The sentence created will more likely to have a meaning which is opposite of the original sentence.

5. Extract the dependency tree for each sentence and select an appropriate one from other sentences to replace it. The dependency relationship will contain information about which words are dependent on which other words. Replacing based on this similarity may create better examples.

Total 80 experiments. 4 ways to create a model. 5 ways to create negative examples. and 4 ways of experiments

4.4 Implementation

Implementation: Tensorflow, spacy

Recurrent neural networks will be used in this master thesis extensively since they perform well with sequential data. TensorFlow[13] is a python library for building intensive numerical computation applications and deploy them in multiple platforms such as CPU and GPU. TensorFlow has built-in libraries to develop different neural network architectures. Wikipedia consists of a lot of articles; each article consists of a lot of sentences. Each sentence will be fed as input to the neural network. ———— An advanced RNN architecture takes in either one or two inputs such as sentence in Wikipedia and context information. The context can be empty. The system processes the input and produces a binary output representing true or fake news. The outputs would contain a probability distribution of how likely it is true or fake news.

The inputs, in general, will be pre-processed to extract tokens, remove common words such as articles, prepositions and perform lemmatization which is converting the words to its root form. Low dimensional and dense word embeddings will be used to represent the inputs instead of one hot vectors. Already available word embeddings such as Word2vec [30] and GloVe[14] will be used to represent the inputs.

[12] https://wordnet.princeton.edu/

[13] https://www.tensorflow.org/

[14] https://nlp.stanford.edu/projects/glove/

The dataset will be split and used for training, validation, and testing of the system. A general rule of thumb is to divide the dataset into 60-20-20 for training, validation, and testing respectively. But if the datasets are in order of millions then it is good enough to have approximately 10000 entries each for validation and testing. The validation dataset is used to try out different hyper-parameter configurations and select the optimal values for each. The hyper-parameters of the advanced neural network system that will be configured are number of layers, learning parameter, batch size, activation functions and epoch number(number of times the training dataset should be iterated).

The proposed methods involve modeling the content of Wikipedia in different ways such that the system can understand it better. These procedures are completely automatic in building the training examples and none of the features will be manually crafted. The different configurations of the system along with unique methods in constructing the training data is expected to understand Wikipedia better and the fake news detection can be carried out in a platform agnostic manner.

5 Evaluation

The system will be evaluated as a fake news classifier. The test data should be different from the training or validation data which is used by the system.

5.1 Datasets

There are two comprehensive open datasets on fake news detection available.

1. One of it is published in Kaggle[15] which consists of structured data of 13,000 rows and 20 columns. Some of the important information in this dataset are title, text and spam score.
2. The LIAR dataset is published by Wang [12] which is mined from politifact.com and it consists of 12,836 short statements. In addition to that, it also contains information about context, labels with 6 classes and justification for the label classification.

The mentioned datasets might be too diverse to compare the performance of the system at the beginning. It might require training from the whole of Wikipedia. Research work from Ciampaglia et al. [14] used many simple datasets to evaluate network analysis techniques on a knowledge graph to build fake news classfier. The datasets are US politicians and their party affiliation, directors and their movies, US president and their spouses, US states and their capitals and world countries and their capitals. It also involves a dataset from Google Relation Extraction Corpus (GREC) which is about education degrees and institutional affiliations of people. These datasets will help to compare the performance of fake news classifier in specific domains of knowledge against their achieved results.

In addition to that, the system can be evaluated based on how well it is able to understand Wikipedia. We use the presence and absence of a information in Wikipedia as a proxy for labeling that information as true or fake news. So, it is good to evaluate our system based on how well it is able to classify whether a given information is present in Wikipedia or not. This is done by inputting either random sentences or the sentences taken from Wikipedia and check whether it is present or not. If the system performs very well in this test and fails to perform well with the mentioned standard datasets for fake news classification then it might help to gain insights on how good the coverage of news in Wikipedia is.

[15] https://www.kaggle.com/mrisdal/fake-news

5.2 Metrics

The baseline system involves only RNN and the improvements that will be added are RNN with LSTM blocks, bidirectional RNNs, and inclusion of context input along with sentence input. The following factors are used to compare the baseline with the improved system

1. How good the system is able to classify the non-Wikipedia news / fake news?
2. How fast the model is build up?

The quality of the system can be measured by metrics such as Precision, Recall, F1 measure, and Accuracy. The F1 measure is preferred since it includes both false positive and false negatives. In this master thesis, the focus is mainly on improving the accuracy of the system and the speed of the system can be improved by using good hardware.

6 Organizational matters

Duration of work:	01-Aug-2018 – 31-Jan-2018
Candidate:	Kandhasamy Rajasekaran
E-Mail:	kandhasamy@uni-koblenz.de
Student number:	216100855
Primary supervisor:	Prof. Dr. Steffen Staab
Supervisor:	Dr. Chandan Kumar
Secondary supervisor:	Lukas Schmelzeisen

7 Time schedule

- Introduction and Literature: 01-May-2018 – 30-June-2018
- Initial phase: 01-Aug-2018 – 15-Oct-2018
  - Prototyping: 01-Aug-2018 – 30-Aug-2018
  - ML pipeline implementation: 01-Sep-2018 – 15-Sep-2018
  - Baseline implementation: 15-Sep-2018 – 30-Sep-2018
  - Testing and refining: 01-Oct-2018 – 15-Oct-2018
- Development phase: 16-Oct-2018 – 30-Dec-2018
  - RNN with different configuration: 16-Oct-2018 – 30-Oct-2018
  - Comprehend benchmark results: 16-Oct-2018 – 30-Oct-2018
  - Analyse and figure out improvements: 22-Oct-2018 – 30-Oct-2018
  - Improvisation using NLP techniques: 01-Nov-2018 – 30-Nov-2018
  - Attention mechanism implementation: 01-Dec-2018 – 30-Dec-2018
  - Comprehend benchmark results: 16-Dec-2018 – 30-Dec-2018
- Final phase: 01-Jan-2019 – 01-Feb-2019
  - Comprehend Benchamark results: 01-Jan-2019 – 07-Jan-2019
  - Revision: 08-Jan-2019 – 22-Jan-2019
  - Thesis report: 01-Jan-2019 – 30-Jan-2019

A meeting with Lukas Schmelzeisen will happen approximately once in two weeks to discuss about the progress made and set the targets and milestones for subsequent weeks.

References

[1] David M. J. Lazer et al. “The science of fake news”. In: *Science* 359.6380 (2018), pp. 1094–1096. ISSN: 0036-8075. DOI: 10.1126/science.aao2998. eprint: http://science.sciencemag.org/content/359/6380/1094.full.pdf. URL: http://science.sciencemag.org/content/359/6380/1094.

[2] Hunt Allcott and Matthew Gentzkow. “Social Media and Fake News in the 2016 Election”. In: *Journal of Economic Perspectives* (2017). ISSN: 0895-3309. DOI: 10.1257/jep.31.2.211. arXiv: 1704.07506.

[3] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online”. In: *Science* 359.6380 (2018), pp. 1146–1151. ISSN: 0036-8075. DOI: 10.1126/science.aap9559. eprint: http://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: http://science.sciencemag.org/content/359/6380/1146.

[4] Jimmy Wales. *Internet encyclopaedias go head to head*. 2005. DOI: 10.1038/438900a.

[5] Alexander Halavais and Derek Lackaff. “An analysis of topical coverage of Wikipedia”. In: *Journal of Computer-Mediated Communication* 13.2 (2008), pp. 429–440. ISSN: 10836101. DOI: 10.1111/j.1083-6101.2008.00403.x.

[6] Reid Priedhorsky et al. “Creating, destroying, and restoring value in wikipedia”. In: *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP ’07*. 2007. ISBN: 9781595938459. DOI: 10.1145/1316624.1316663.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. www.deeplearningbook.org. MIT Press, 2016.

[8] Xiaomo Liu et al. “Real-time Rumor Debunking on Twitter”. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM ’15*. 2015. ISBN: 9781450337946. DOI: 10.1145/2806416.2806651. arXiv: arXiv:1601.00306v1.

[9] Jing Ma et al. “Detect Rumors Using Time Series of Social Context Information on Microblogging Websites”. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM ’15*. 2015. ISBN: 9781450337946. DOI: 10.1145/2806416.2806607.

[10] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. “Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 1395–1405. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741637. URL: https://doi.org/10.1145/2736277.2741637.

[11] Sejeong Kwon et al. “Prominent features of rumor propagation in online social media”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2013. ISBN: 978-0-7695-5108-1. DOI: 10.1109/ICDM.2013.61.

[12] William Yang Wang. “"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection”. In: (2017). DOI: 10.18653/v1/P17-2067. arXiv: 1705.00648. URL: http://arxiv.org/abs/1705.00648.

[13] Jing Ma et al. “Detecting Rumors from Microblogs with Recurrent Neural Networks”. In: ().

[14] Giovanni Luca Ciampaglia et al. “Computational fact checking from knowledge networks”. In: *PLoS ONE* (2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0128193. arXiv: 1501.03471v1.

[15] Daniel Svozil, Vladimir Kvasnieka, and Jie Pospichal. “Chemometrics and intelligent laboratory systems Introduction to multi-layer feed-forward neural networks”. In: *Chemometrics and Intelligent Laboratory Systems* 39 (1997), pp. 43–62.

[16] Yoav Goldberg. “A Primer on Neural Network Models for Natural Language Processing”. In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420.

[17] Léon Bottou. “Stochastic Gradient Descent Tricks”. In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Geneviève B. Montavon Grégoire and Orr and Müller Klaus-Robert. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_25. URL: https://doi.org/10.1007/978-3-642-35289-8{_}25.

[18] Ilya Sutskever et al. “On the importance of initialization and momentum in deep learning”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2010 (2013), pp. 8609–8613. ISSN: 15206149. DOI: 10.1109/ICASSP.2013.6639346. arXiv: arXiv:1301.3605v3.

[19] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2121–2159. ISSN: 15324435. DOI: 10.1109/CDC.2012.6426698. arXiv: arXiv:1103.4296v1. URL: http://jmlr.org/papers/v12/duchi11a.html.

[20] Geoffrey E. Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: (2012). ISSN: 9781467394673. DOI: arXiv:1207.0580. arXiv: 1207.0580.

[21] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *CoRR* abs/1301.3781 (2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781.

[22] Y Bengio. “Convolutional Networks for Images, Speech, and Time-Series Parsing View project Oracle Performance for Visual Captioning View project”. In: (1997). URL: https://www.researchgate.net/publication/2453996.

[23] Jeffrey L Elman. “Finding Structure in Time”. In: *COGNITIVE SCIENCE* 14.1 (), pp. 179–21. DOI: 10.1207/s15516709cog1402_1.

[24] Francois Deloche. *Recurrent Neural Network unfold*. [Online; accessed April 27, 2013; https://commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg; CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0). 2013.

[25] Paul J. Werbos. “Backpropagation Through Time: What It Does and How to Do It”. In: *Proceedings of the IEEE* (1990). ISSN: 15582256. DOI: 10.1109/5.58337.

[26] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[27] Ankur P. Parikh et al. “A Decomposable Attention Model for Natural Language Inference”. In: (2016). ISSN: 0001-0782. DOI: 10.18653/v1/N16-1062. arXiv: 1606.01933. URL: http://arxiv.org/abs/1606.01933.

[28] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: (2014), pp. 1–15. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.26.041002.131047. arXiv: 1409.0473. URL: http://arxiv.org/abs/1409.0473.

[29] Zichao Yang et al. "Hierarchical Attention Networks for Document Classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016), pp. 1480–1489. ISSN: 1606.02393. DOI: 10.18653/v1/N16-1174. arXiv: 1606.02393. URL: http://aclweb.org/anthology/N16-1174.

[30] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality arXiv : 1310 . 4546v1 [cs . CL] 16 Oct 2013". In: (), pp. 1–9. arXiv: arXiv:1310.4546v1.

8 Signatures

Kandhasamy Rajasekaran	Prof. Dr. Steffen Staab
Dr. Chandan Kumar	Lukas Schmelzeisen

9 Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here.

This paper was not previously presented to another examination board and has not been published.

Koblenz, on November 6, 2018

_____

Kandhasamy Rajasekaran