BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

SPECIALIZATION APPLIED COMPUTATIONAL INTELLIGENCE

DISSERTATION THESIS

Fake news detection

Supervisor

Lect. Dr. Lupșa Dana

Author

**Teodora Ioana NASTE**

UNIVERSITATEA BABEŞ-BOLYAI CLUJ-NAPOCA

FACULTATEA DE MATEMATICǍ ŞI INFORMATICǍ

SPECIALIZAREA INTELIGENȚĂ COMPUTAȚIONALĂ APLICATĂ

LUCRARE DE DIZERTȚIE

Detectarea știrilor false

Conducător ştiinţific

Lect. Dr. Lupșa Dana

Absolvent

**Teodora Ioana NASTE**

Contents

Introduction

1. Natural language processing
   1. Text similarity
   2. Text Classification
   3. Misinformation classification (separate chapter describing the task?)
2. Neural networks
   1. Feed-forward neural networks
   2. Recurrent neural networks
   3. LSTM
3. Presentation of my model
   1. Architecture
   2. Technology and Framework
   3. Experimental Evaluation
4. Future work
5. Conclusion

INTRODUCTION

Here comes some text for the introduction. Will be done last.

CHAPTER 1

Natural Language Processing

* 1. Preprocessing
  2. Text Classification
  3. TF-IDF

CHAPTER 2

Machine learning models

2.2 Artificial neural networks

The idea of a neural networks has first been introduced in 1943 by McCulloch and Pitts who have introduced the concept of a neuron as a conceptual unit capable of simple operations. However, as it is mentioned by Kröse, Ben et al. [6] , only in the 80s the interest towards artificial neural networks has begun to grow.

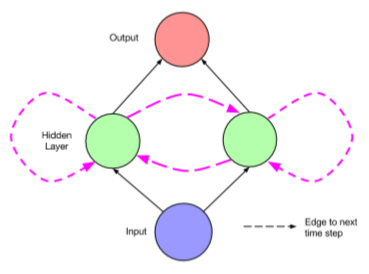
As the name suggests, the model of the artificial neural networks (ANN) is inspired by the structure of the human brain. It’s formed by a big number of artificial neurons, also called units or nodes, united by edges. Each unit *j*  has an associated activation function and each edge has a weight , which represents the cost from unit *j* to the unit *j’*. So, the value of a node, , is computed applying the activation function, like in the formula below:

The activation function is a simple linear function. The most common used ones are the sigmoid function: and the tanh function: . These are especially used for feedforward neural networks, but they have been applied to recurrent neural networks too. Another possible activation function is the rectified linear unit (ReLU) which is mostly used for deep neural networks and for the task of image detection.

The classic neural network is capable of approximating a non-linear function. These networks learn using error propagation. The values calculated by the nodes are sent to the output nodes which compare them with the expected results. Each output node then returns an error. The purpose is to minimize this error. The simplest method is to distribute the error from one output node to all the other nodes connected with it , proportionally with the nodes’ weights, till the input nodes and each edge will update its weight and a new error will be calculated.

2.2.1 Recurrent neural networks

In contrast to traditional neural networks, the recurrent ones can have cycles, introducing the notion of time to the model. Thus, at a time *t* , the nodes with a recurrent edge get information from the current input and from the previous states . A simple recurrent neural network (RNN) is presented in Figure 2.2.1. [7] Moreover, RNNs are able to handle variable-length input, making them suitable candidates for processing sequences like texts.



***Figure 2.2.1. Recurrent neural network***

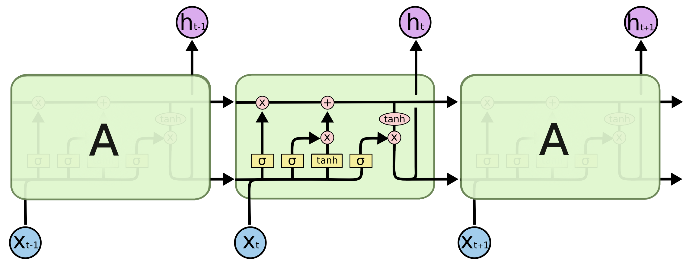
The training for a RNN can be difficult. The main problems that can appear are vanishing and exploding gradients which can occur during error propagation. Which of these two events may occur depends on the weight of the edges and on the activation function. For example, if the activation function is a sigmoid one, the values risk to get too close to 0 and the network won’t learn. In this case, we deal with vanishing gradient. If the activation function is a ReLU max(1,x), then it’s easy to imagine the values growing too high and the exploding gradient might appear.

2.2.2 Long Short-Term Memory Neural Networks

To solve the problems of the vanishing and exploding gradients that appear during training of a standard RNN, the Long Short-Term Memory neural network was proposed in 1997 [4] with an innovative gating system that introduces an intermediate memory space via the memory cell.

The gating system consists of 3 gates: the input, forget and output gates. The input gate, , decides how much of the information processed by the input node, , should be modified. The forget gate, helps to sort out the irrelevant information and the internal state, , is calculated based on the results from the previous 2 gates. Then, the final activation is composed with the help of the output gate, , which decides how much of the information should be passed to the next units. This process can be viewed in the next equations [7]:

Firstly, the hidden state computed by the previous nodes, , is concatenated with the input , then the result is passed through the forget gate, which uses a sigmod activation to ”*forget*” information that’s irrelevant for the current task. Next, the values go throught he input gate which also sues a sigmoid activation. At the same time, the candidate value is composed using a tanh activation. The candidates are then concatenated with results from the input gate. The next step is computing the internal state and the hidden statte that will be forwared to the rest of the nodes . This last step is done by the output gate. The internal state is normalized using a sigmoid function then multiplied with the candidate values . The whole flow can be also seen in Figure 2.2.2, which presents the structure of a LSTM neural network with its gates and activations.



***Figure 2.2.2. Structure of a LSTM neural network [7]***

LSTM neural networks have been successfully applied in many differen domains and are currently the most popular recurrent neural networks, even if they were firstly proposed over 20 years ago. Some of the fields in which they have been used are: text classification, text generation, automated translation, image captioning or predictions.

2.2.3 Gated Recurrent Neural Networks

The gated recurrent unit neural network is a type of recurrent neural networks proposed recently which use a gating system to solve the vanishing gradient problem similar to the LSTM. It was first proposed in 2014 by Kyunghyun Cho et al [2] as a simplified version of the aforementioned LSTM.

The gating mechanism is composed by 2 gates: the reset and the update gates. The reset gate eliminates the past information that is considered to be irrelevant, thus helping the current unit memory at moment *t*  to have only the important information stored. The update gate helps to decide how much of the past information to be passed along to the future. The final activation of the unit is composed using the previous hidden state and the results of the update gate, then passed to the next unit. Mathematically, this process consists of the next 4 equations [2]:

Each element of the input sequence is read by the model and the hidden state is computed based on the element and the previous hidden state. Then, the process is very similar to that of a LSTM in which the result from concatenating the input, however, the GRU doesn’t use an internal state anymore, but is counting on the hidden state to transport information throughout the network. While the update gate has a similar role like the forget and input gate, the reset gate is a bit different, because rather than deciding what to pass to the next unit, it decides what to forget from the past information.

CHAPTER 3

Applications

The raise in popularity of social media has massively increased the amount of user-generated information that can spread uncontrollably throughout the web. This huge amount of data generated in real-time is impossible to be filtered and checked manually. So, there has been a growing interest to research automatic ways of detecting false information. However, considering the complexity of what can be considered fake news and the various forms in which this can be spread in, the task of simply determining the authenticity of the information can be very hard and complex. In this chapter, I will present some interesting and diverse methods that have been applied for this task.

3.1 Stance detection vs. veracity detection

When talking about fake news detection, most people would think of labelling a claim, article, headline or post as either true or false. This is, more or less, what veracity detection aims to do. However, in practice, this is more complex that it sounds. First of all, research in the area has been done heavily only in the past 4-5 years. This leads to data being scarce and unstructured. Stance detection came as a response to these shortages, providing a first step towards fully automated fact checking. The idea behind it is to possibly help the professional fact checkers by offering relevant background information about the text based on the stances of various organizations on the topic, weighted by their credibility.

Both approaches have the same overall goal, but they are different in ways each is approach, the data they need or the inputs and outputs of the models. Also, the solutions from one task can inspire research for the other task and that’s why I think it is relevant to describe some of these methods.

3.1.1 Stance detection

Realising the complexity that fake news detection consists of, as a first step to tackle the task, in 2017 the Fake News Challenge[[1]](#footnote-1) released a contest for stance detection. Given two pieces of text, the purpose of the task was to estimate the relative perspective or stance of a text to the other on a given topic, claim or issue. The dataset for the challenge consisted in pairs of news article headline and its body and the contestants had to estimate the stance of the body towards the headline. The stance was represented by one of the following labels: *agree, disagree, discuss* orunrelated. There were many entries for this challenge, each with different types of models, however the ones that got into top 3 used rather simple methods.

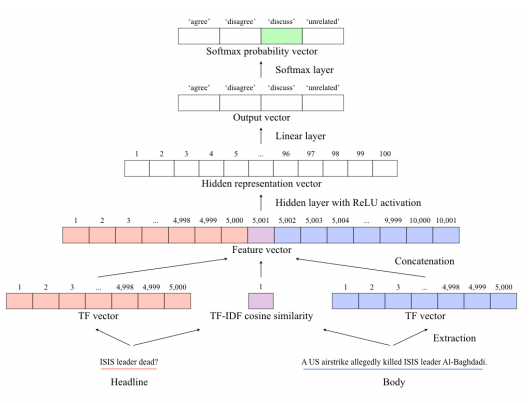
One of the approaches that got into top 3 was proposed by Benjamin Riedel et al. [1]. They probably had the most simple method, yet managed to get a considerable high accuracy.

**Methodology**

The method proposed consists of passing lexical and similarity features through a multi-layer perceptron with one hidden layer. The text was represented by a simple bag-of-words and the features considered were the following:

* Term frequency (TF) vector of the headline
* Term frequency (TF) vector of the body
* Cosine similarity between the TF-IDF vectors of the body and headline

The MLP classifier had 100 units on its single hidden layer. The rectified linear unit (ReLU) was used as the activation function and a softmax function on the output layer. In figure 3.1.1 a schema of the system is presented.

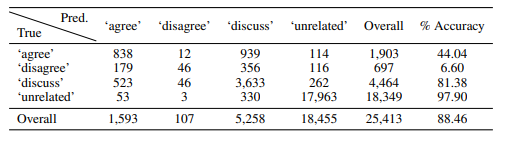


**Figure 3.1.1 Schematic diagram of the stance detection system**

A dropout and the Adam optimiser were also applied to avoid the gradient descent issue. The parameters for the model were learnt using a random search and cross-validation.

**Results**

The proposed method reached an accuracy of 81.72% which took the team on the third position at the end of the challenge. Results on different labels are detailed in Table 3.1.1., where it can be observed that the system performed best on text labelled with “agree”.



**Table 3.1.1 Confusion matrix for the stance detection system**

3.1.2 Veracity detection

When talking about fake news detection, the first concept that pops into anyone’s mind is to check the information (content) from a piece of news and assign if it is true or false, without necessarily taking into account the intention of the creator for example.

Such approach has been taken by Souvick Ghosh an Chirag Shah [8] in their research paper *Toward automatic fake news classification*. Inspired by one of the top 3 teams from Fake News Challenge 2017 they wanted to see if combining information retrieval with stance detection and deep learning, they will be able to obtain good results on labelling text as ‘Fake’, ‘Legit’ or ‘Suspicous’.

**Dataset**

Because of a lack of benchmark datasets , the researchers used multiple datasets which they combined into 2 categories:

* Type 1, formed from LIAR, Kaggle’s fake News dataset and Fake News Challenge dataset, is a set of short texts with the length between 70 and 150 characters, including statements and article headlines.
* Type 2, formed from University of Washington Fake News Dataset, consists of much larger texts with the length between 400 to 700 words.

**Methodology**

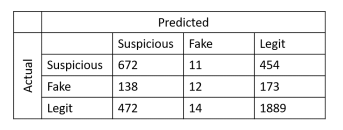
A submodular approach has been used. The first submodule was assigned to detect the veracity of the text using information retrieval, after previously constructing a knowledge base. First, documents referring the claim are retrieved from the knowledge base. For this the TF-IDF method was used, combined with a other advanced algorithms (BM25, Vector Space model, Language model) to improve the performance. After that, for each document retrieved the stance of the claim towards the document is determined. The second step has the task of classifying the claims into one of the 3 categories mentioned above. For that , they used a simple Feed Forward Neural Network that gets as input the bags-of-words vectors for the claim and the document, as well as the cosine similarity between these vectors. For training, the Type 1 dataset was used, as it is more focused on fact-based differences.

The second submodule is concerned with finding patterns in the writing style of the claim. It has been observed that fake news tend to be more aggressive and depict stronger emotions in order to manipulate the audience [9]. For this task, a bidirectional LSTM has been used, as these networks are known for their capabilities to process long sentences. This submodule was trained using the Type 2 dataset, as it has more richness in style compared to the short texts from Type 1.

Finally, results from both submodules are combined based on a voting system with the help of a weighted average.

**Results**

The prediction submodule managed to obtain an accuracy 67.1% for the ternary classification and of 72,12% for the binary classification, while the style based submodule obtained an accuracy of 81.83%. After combining the two submodules, they obtained a total accuracy of 82.4%. The experimental results for the prediction on FNC Dataset can be seen in Table 3.1.2.

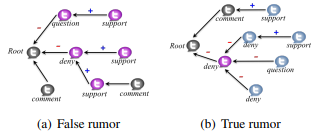


**Table 3.1.2 Confusion matrix for FNC Dataset**

3.2 Fake news detection based on propagation graphs

Social media is the main environment in which fake news are being created and spread at an alarming rate. However, this makes space to research different aspects of fake news, like the propagation pattern and the users’ response via comments or reposts, with the intention to have an automatic method of spotting misinformation in its early stages of spreading. Veracity detection has been used for detecting fake news in social media, but the difference here is manner in which the true/false tags are assigned: instead of checking the information from a post, it looks at the way it propagates.

In [5], have proposed a tree-structured recursive neural network approach that analyses how a rumour is spread and what’s the users’ stance relative to it. They started this researched based on observations they have made that suggest that if a user denies a post with a fake rumour, it tends to trigger positive responses from other users, confirming the denial, while denying a true rumour triggers negative and confusing responses. This pattern can be seen in Figure 3.2.1.



**Figure 3.2.1 Propagation tree containing users’ stance towards the post**

**Methodology**

A rumour (claim) was defined as an original tweet together with all its relevant responsive tweets ordered chronologically. For these claims the task is to be classified into one of the following categories:  *non-rumour, false rumour, true rumour* and *unverified rumour.* To capture the information from different angles, two variants have been adopted:

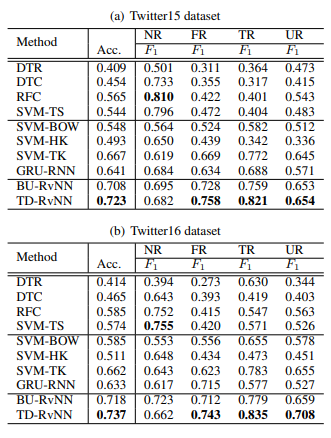
* Bottom-up RNN in which the direction of the edge points to the source, like in Figure 3.3.1
* Top-Down RNN which follows the natural propagation course in which the direction of an edge goes from the source to the user node that responded to that post

The purpose for each tree is to generate a feature vector that can be used as input to an extended version of the GRU neural network that Jing Ma et al. implemented. The difference between the two variants however is that, the bottom-up is going to start from the leaves and create a feature vector for each subtree and merging them together at the root creating the vector for the whole tree. On the top-down version, the creation is starting from the original post , resulting in feature vectors for each propagation line, since the final representation will be in the leaves of the tree. This enables a deeper insight into the complex propagation patterns that rumours can have on social media.

The error function used was the squared error. The parameters were updated using back-propagation through structure and also a gradient-based optimization was applied.

**Results**

The evaluation was done on two real-life Twitter datasets (Twitter15 and Twitter16) which each contain over 1000 propagations trees. To better present the advances of their work, Jing Ma et al. made an elaborate comparison with multiple state-of-the-art methods. The results on both sets can be seen in Table 3.2.1



**Table 3.2.1 Results comparison of rumour detection for each category**

From the table, it can be easily observed how the proposed model surpasses previous ones on most categories. Moreover, after obtaining such amazing results, further experiments were done to see the capabilities to detect rumours in early stages. They concluded that their method needs approximately 8 hours or about 90 tweets to correctly spot a fake rumours which is superior to the best baseline method which needs 36 hours and about 300 posts.

3.3 Satirical fake news detection

Satirical news bring another challenge to fake news detection, because their purpose is for entertainment and thus, it is harder to recognize the intention of the creator. Also, the satire can be subtle and so, users could easily misunderstood it as true news. The style of writing these kind of news it’s different than with other types of articles and it needs special investigation. Therefore, satirical news detection is researched separately from other news detection.

Considering this, Fan Yang et al. [3] have investigated satirical news and proposed a paragraph-level detection model that can deal with such subtle satire, but also highlighting the difference in features between document-level and paragraph-level. As they also pointed, satirical news are still considered fake news if the user believes them and spreads them as true information.

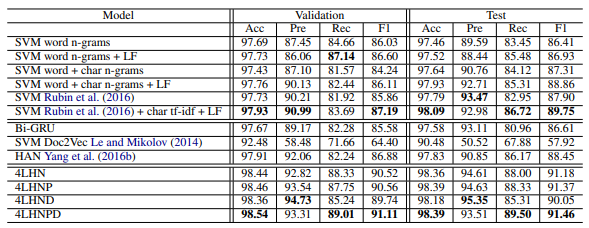
**Methodology**

To achieve their goal, they proposed a 4-level hierarchical neural network, following the hierarchy of character-word-paragraph-document.

Firstly, on the lowest level of the hierarchy, a convolutional neural network (CNN) is used to encode words representations from characters. Then these representations are sent one level up, where based the word embedding takes place. For encoding a Gated recurrent Unit (GRU) was implemented. The third level focuses on encoding at paragraph level, using a bidirectional GRU. Here, the attention mechanism is included because it was observed that not all paragraphs contain satirical cues and some paragraphs can be more relevant than others. For that, a satirical degree is being computed based on the paragraph representations and the results from the hidden states. The final document is the weighted sum of the satirical degree and the paragraph representations. On this level extra linguistics features are added. On the final level the final vector representation is computed via MLP and the prediction is done using a sigmoid activation.

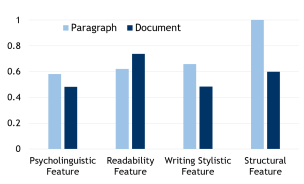
**Results**

For evaluation, over 16.000 satirical news were collected from various sites that focus on such content and over 160.000 true news from major news outlets. Since the data is not annotated, only the binary classification was taken into account. Moreover, different models were implemented to prove the efficiency of the method. These models as well as all the results can be seen in Table 3.3.1.



**Table 3.3.1 Results comparison for satirical news detection**

It can be observed that the model including both paragraph and document level features performs the best in almost all categories. The importance of the linguistic features used was also analysed which gives a very interesting insight into what features are more relevant when trying to detect satire. The results can be seen in Figure 3.3.1.



**Figure 3.3.1 Comparing the importance of the four features sets**

CHAPTER 4

Implemented solution

CHAPTER 5

Future work and improvements

CONCLUSION

**BIBLIOGRAPHY**

[1] Benjamin Riedel et al. , *A simple but tough-to-beat baseline for the Fake News Challenge stance detection task, 2017*

[2] Cho, Kyunghyun et al. , *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation ,* 2014

[3] Fan Yang et al., *Satirical news detection and analysis using Attention Mechanism and Linguistic Features,* 2017

[4] Hochreiter, S., & Schmidhuber, J.. *Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.* doi:10.1162/neco.1997.9.8.1735 , 1997

[5] Jing Ma, Wei Gao and Kam-Fai Wong, *Rumour Detection on Twitter with Tree-structured Recursive neural Networks*, 2018

[6] Kröse, Ben, et al. , *An introduction to neural networks*, 1993

[7] Lipton, Zachary C., John Berkowitz, and Charles Elkan. *A critical review of recurrent neural networks for sequence learning*. arXiv preprint arXiv:1506.00019 , 2015.

[8] Souvick Ghosh and Chirag Shah*, Toward Automatic Fake News Classification, 2017*

[9] Xinyi Zhou and Reza Zafarani, *A Survey of Fake News: Fundamental theories, detection methods and opportunities,* 2020

1. <http://www.fakenewschallenge.org/> [↑](#footnote-ref-1)