

# Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News

LUÍS BORGES, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

BRUNO MARTINS, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

PÁVEL CALADO, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Fake news are nowadays an issue of pressing concern, given their recent rise as a potential threat to high-quality journalism and well-informed public discourse. The Fake News Challenge (FNC-1) was organized in early 2017 to encourage the development of machine learning-based classification systems for stance detection (i.e., for identifying whether a particular news article agrees, disagrees, discusses, or is unrelated to a particular news headline), thus helping in the detection and analysis of possible instances of fake news. This article presents a novel approach to tackle this stance detection problem, based on the combination of string similarity features with a deep neural network architecture that leverages ideas previously advanced in the context of learning efficient text representations, document classification, and natural language inference. Specifically, we use bi-directional Recurrent Neural Networks (RNNs), together with max-pooling over the temporal/sequential dimension and neural attention, for representing (i) the headline, (ii) the first two sentences of the news article, and (iii) the entire news article. These representations are then combined/compared, complemented with similarity features inspired on other FNC-1 approaches, and passed to a final layer that predicts the stance of the article towards the headline. We also explore the use of external sources of information, specifically large datasets of sentence pairs originally proposed for training and evaluating natural language inference methods, in order to pre-train specific components of the neural network architecture (e.g., the RNNs used for encoding sentences). The obtained results attest to the effectiveness of the proposed ideas and show that our model, particularly when considering pre-training and the combination of neural representations together with similarity features, slightly outperforms the previous state-of-the-art.

CCS Concepts: • **Computing methodologies** → **Neural networks; Natural language processing; Supervised learning by classification;**

Additional Key Words and Phrases: Fake News, Fact Checking, Stance Detection, Deep Learning, Natural Language Processing, Recurrent Neural Networks

## ACM Reference Format:

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News. *ACM J. Data Inform. Quality* 9, 4, Article 39 (April 2019), 24 pages. <https://doi.org/0000001.0000001>

## 1 INTRODUCTION

Fake news (i.e., made up stories with the intention of deceiving, and that most of the times are used to achieve secondary gains) are undoubtedly one of the most serious challenges that journalists and the news industry are facing today. Given the ease in obtaining and spreading information through social networking platforms, it

---

Authors' addresses: Luís Borges, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol, 9, Lisbon, 1000-029, Portugal, [luis.borges@tecnico.ulisboa.pt](mailto:luis.borges@tecnico.ulisboa.pt); Bruno Martins, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol, 9, Lisbon, 1000-029, Portugal, [bruno.g.martins@tecnico.ulisboa.pt](mailto:bruno.g.martins@tecnico.ulisboa.pt); Pável Calado, INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol, 9, Lisbon, 1000-029, Portugal, [pavel.calado@tecnico.ulisboa.pt](mailto:pavel.calado@tecnico.ulisboa.pt).

---

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

© 2019 Association for Computing Machinery.

1936-1955/2019/4-ART39 \$15.00

<https://doi.org/0000001.0000001>

is increasingly harder to know for sure what to trust, with the absorption of fake news by the masses having increasingly harmful consequences [48]. Automatically dealing with fake news has drawn considerable attention in several research communities [24, 26, 34, 36, 40, 41, 45]. However, the task of evaluating the veracity of news articles is still very demanding and complex, even for trained specialists and much more for automated systems.

A useful first step towards identifying fake news articles relates to understanding what other news agencies, in a given moment, are reporting about the same topic. This sub-task is often referred to as stance detection, and automating this process might be useful in developing automated assistants to help in fact checking. In particular, an automatic approach to stance detection would allow, for example, someone to insert an allegation or a news title, and recover related articles that either agree, disagree, or discuss that title. Then, the human checker would use her own judgment to assess the situation.

Based on the aforementioned general ideas, a Fake News Challenge (FNC-1) was organized in early 2017 by a group of academics and contributors from the journalism industry to foster the development of systems applying artificial intelligence and machine learning for evaluating what a news source is saying about a particular issue. Specifically, FNC-1 involved developing models that, given the body of a news article and a news headline, estimate the stance of the article towards the headline (i.e., the article can either agree, disagree, discuss, or be unrelated to the headline). More information on the FNC-1 task, its rules, the dataset, and the evaluation metrics, can be found on the official website<sup>1</sup>. A total of 50 teams actively participated on the challenge, and the training/testing data splits have been released afterwards, in order to encourage further developments.

The winning entries in FNC-1 relied on ensemble models combining similarity features computed between the headline and the body, with representations built from the words occurring in the texts (e.g., by processing word embeddings with convolutional neural networks). Despite the good results achieved by the participants, with scores of 82.02% and 81.98% for the two winning teams, in terms of the accuracy metric considered for the challenge, we believe there are several opportunities for further improvements. For instance, although deep neural networks are known to outperform other approaches (e.g., linear models based on careful feature engineering) in several natural language processing problems related to stance detection (e.g., in tasks related to measuring semantic similarity between sentences, and/or related to performing natural language inference), the results from FNC-1 showed that standard approaches based on convolutional and/or recurrent neural networks are not well-suited to this task, as they fail to model semantic relations with respect to large pieces of text (i.e., the bodies of news articles, composed of multiple sentences). A FNC-1 baseline that was introduced by the task organizers, which leverages a combination of feature engineering heuristics (i.e., word and  $n$ -gram overlap features, as well as indicators for refutation and polarity), achieves a competitive performance with the best systems, and it even outperforms several widely used deep learning architectures.

In this article, we address the stance detection problem proposed in the context of FNC-1, using a novel approach based on the combination of similarity features with deep neural networks for generating effective representations. The main ideas and contributions of our work can be summarized as follows:

- Taking inspiration on previous studies addressing natural language inference [4, 49], we propose a deep neural network architecture for stance detection, as defined in the FNC-1 task. Our architecture leverages pre-trained word embeddings, uses shared parameters for some of the components (e.g., the sentence encoders), and attempts to match multiple representations learned from the inputs. Specifically, we combine/match representations inferred from three different inputs (i.e., the headline, the first two sentences of the news article, and the entire document) through a series of operations known to perform well on natural language inference [12], namely the vector concatenation, difference, and element-wise product. A final layer processes this result, and predicts the stance of the news article towards the headline.

<sup>1</sup><http://fakenewschallenge.org>

- The proposed neural network architecture leverages a hierarchical approach for modeling the body of news articles, taking inspiration on previous studies addressing the classification of long documents [14, 52]. In this approach, a Recurrent Neural Network (RNN) is used for modeling the sequence of sentences, which in turn are individually modeled by a separate RNN encoding sequences of words. Specifically, we evaluated the use of bi-directional RNNs (i.e., either Gated Recurrent Units or Long Short-Term Memory units, optionally also in a stacked arrangement with shortcut connections), together with max-pooling and/or a neural attention mechanism that weights the individual word representations, for encoding sentences in our model (i.e., the headline, and the sentences in the news article being matched). Separate layers of bi-directional RNNs, also combined with max-pooling and/or neural attention, are used for encoding the sentences that constitute the document. This approach can directly incorporate knowledge of the document structure in the model architecture, at the same time also exploring the intuition that not all sentences/words will be equally relevant for predicting the stance;
- We used the SNLI [4] and MultiNLI [49] datasets of sentence pairs, previously proposed for evaluating natural language inference models (i.e., models for checking if a given hypothesis sentence entails, contradicts, or is neutral towards a given premise sentence), to pre-train some of the components involved in our neural network architecture. Previous studies have already attested to the benefits of similar pre-training procedures in other natural language processing and text classification tasks [12]. We specifically used SNLI/MultiNLI sentence pairs to pre-train the components involved in encoding sentences, and also the part of the network that matches the headline with the first two sentences of the document (although this second component is not used in the neural architecture that achieved the best experimental results);
- Learning effective representations for news article bodies can be quite challenging for neural methods, although this is essential for achieving good performance on the FNC-1 task [2, 30]. We therefore propose to combine the representations learned through neural network layers, with external similarity features. We specifically leverage features from previous FNC-1 methods (e.g., from the baseline method introduced by the organizers of the challenge), together with other methods proposed for similar text matching problems (e.g., BLEU and ROUGE scores [27, 33], the soft-cosine similarity metric [6], or the CIDEr score [47]).
- We report on the results of an extensive set of experiments, evaluating the contribution of each of the aforementioned components. The results confirm that model pre-training can indeed improve the overall accuracy. Moreover, despite the use of the hierarchical attention method for encoding the news articles, which has been shown to perform well on tasks related to the classification of long documents [52], our results also show that hand-crafted similarity features are highly beneficial. The complete method establishes a new state-of-the-art result for the FNC-1 dataset, slightly outperforming the previous approach described by Bhatt et al. [2]. The source code supporting the experiments reported on this article has also been made available online through GitHub<sup>2</sup>.

The remainder of this article is organized as follows: Section 2 presents fundamental concepts related to the use of deep neural networks in tasks such as natural language inference or stance detection, together with an overview of previous work related to FNC-1. Section 3 presents the neural network architecture proposed for handling the FNC-1 stance detection task, detailing (i) the use of bi-directional RNNs together with max-pooling and with an attention mechanism for encoding sentences, (ii) the hierarchical attention model for encoding larger pieces of text (i.e., the body of a news article), (iii) the method used for combining the representations generated for the headline and the body, (iv) the integration of similarity features inspired on other approaches for the FNC-1 task, and (v) the use of the SNLI and MultiNLI datasets for model pre-training. Section 4 describes the experimental evaluation of the proposed method, specifically detailing the evaluation methodology and

---

<sup>2</sup><http://github.com/LuisPB7/fnc-msc>

presenting ablation tests that validate the contribution of the different model components. Finally, Section 5 presents the main conclusions and highlights possible directions for future work.

## 2 CONCEPTS AND RELATED WORK

This section starts by reviewing neural network methods for modeling textual information, which are the fundamental building blocks of most modern approaches for tasks such as natural language inference or stance detection. It then overviews previous work developed in the context of the Fake News Challenge (FNC-1).

### 2.1 Deep Neural Networks for Natural Language Processing

Supervised machine learning is extensively used for Natural Language Processing (NLP). In general, supervised learning concerns with inferring the parameters of models that take as input vector representations  $\mathbf{x}$  and return as output another vector, where each dimension reflects the probability of the input belonging to a certain class. In NLP tasks such as document classification,  $\mathbf{x}$  typically encodes features like words or characters occurring in the text. Bag-of-words approaches, and extensions considering  $n$ -grams, are arguably the most commonly used representations, treating words and/or phrases as unique discrete symbols, and weighting their contributions through heuristics such as the Term Frequency multiplied by the Inverse Document Frequency (TF-IDF). More recently, noting that bags-of-words often fail in capturing similarities between words, at the same time suffering from sparsity and high dimensionality, methods using neural networks to learn distributed vector representations of words (i.e., word embeddings) have gained popularity. These word embeddings can be (pre-)trained in an unsupervised manner over large corpora (e.g., by learning to predict target words using their neighboring words), through methods like *word2vec* [29] or GloVe [20]. One can then average the embedding vectors to generate representations of larger pieces of text, loosing word order as in bag-of-words approaches, or instead use sequences with the word embeddings themselves as the inputs to be processed by the learning models.

Different machine learning methods have been used in NLP applications, and deep neural networks are nowadays a popular choice. In general, neural networks can be seen as as nested composite functions that transform vector representations, and whose parameters can be trained directly to minimize a given loss function computed over the outputs and the expected results. This training procedure involves an algorithm known as back-propagation, in combination with some variation of gradient descent optimization [15]. An optimization procedure that has been frequently used to train deep neural networks is the Adaptive Moment Estimation (Adam) algorithm [22]. Adam computes parameter updates leveraging an exponentially decaying average of past gradients, together with adaptive learning rates for each parameter. In practice, it performs larger updates for infrequent parameters, and smaller updates for frequent parameters.

The Multi-Layer Perceptron (MLP) is a simple neural network architecture, which consists of a set of nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes. The input signal propagates through the network layer-by-layer in a feed-forward manner, until it reaches the output node(s). Considering a single hidden layer, the corresponding computations can be written as shown in Equation 1.

$$\mathbf{y} = \sigma(\sigma'(\mathbf{x} \cdot \mathbf{A} + \mathbf{a}) \cdot \mathbf{B} + \mathbf{b}) \quad (1)$$

In the previous equation,  $\mathbf{x}$  is a vector of inputs and  $\mathbf{y}$  a vector of outputs. The matrix  $\mathbf{A}$  represents the weights of the first layer and  $\mathbf{a}$  is a bias vector for the first layer, while  $\mathbf{B}$  and  $\mathbf{b}$  are, respectively, the weight matrix and the bias vector of the second layer. The functions  $\sigma'(\cdot)$  and  $\sigma(\cdot)$  both denote an element-wise non-linearity, i.e. the activation functions respectively associated to nodes in the hidden layer, and in the output layer of the network. The softmax function (i.e., a normalized exponential function that produces as output a probability distribution) is often used as the activation in the final layer of a MLP classifier, training the network to minimize a cross-entropy loss defined over the predictions and the ground-truth labels.

While MLPs have been extensively used in NLP applications, the problems in this area often involve capturing regularities over structured data of arbitrary sizes (e.g., sequences of word embeddings). In many cases, this means encoding the structure as a fixed width vector, which we can then be used for further processing. Other network architectures besides MLPs are thus commonly used in NLP, in order to transform a sequence of word embeddings  $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^d$  into a vector (e.g., a sentence representation)  $\mathbf{s} \in \mathbb{R}^h$ . These include Convolutional Neural Networks (CNNs) and different forms of Recurrent Neural Networks (RNNs).

CNNs involve the application of  $h$  filters, sliding them over the input sequence. Each filter performs a local convolution (i.e., an element-wise matrix multiplication followed by a summation) on the sub-sequences of the input, to obtain a set of feature maps. Then, a global average- or max-pooling over time is performed to obtain a scalar, and the scalars from the  $h$  filters are finally concatenated into the sequence representation vector  $\mathbf{s} \in \mathbb{R}^h$ .

Assuming an input sequence  $\mathbf{x}_1, \dots, \mathbf{x}_T$  (e.g., a sequence of embeddings for a document with  $T$  words), a convolution layer of width  $k$  works by moving a sliding window of size  $k$  over the sequence, creating several instances of windows  $\mathbf{w}_i = [\mathbf{x}_i; \mathbf{x}_{i+1}; \dots; \mathbf{x}_{i+k-1}]$ . A filter, i.e. a linear transformation followed by an activation function, is then applied to each window, resulting in  $m$  vectors  $\mathbf{p}_1, \dots, \mathbf{p}_m$  where each is defined as follows:

$$\mathbf{p}_i = \sigma(\mathbf{w}_i \cdot \mathbf{A} + \mathbf{a}) \quad (2)$$

In the previous equation,  $\sigma(\cdot)$  is an activation function that is applied element-wise, while  $\mathbf{A}$  and  $\mathbf{a}$  are parameters of the network. The  $m$  vectors are then passed through a max-pooling layer and a final representation vector  $\mathbf{r}$  is obtained. Each element  $j$  of  $\mathbf{r}$  is obtained as follows, where  $\mathbf{p}_i[j]$  denotes the  $j$ -th component of  $\mathbf{p}_i$ :

$$r[j] = \max_{1 < i \leq m} \mathbf{p}_i[j]. \quad (3)$$

RNNs can, in turn, be seen as time-dependent neural networks, which at time step  $t$  (i.e., at position  $t$  for a given input sequence) compute a hidden state vector  $\mathbf{h}_t$ , which is obtained by a non-linear transformation with two inputs, i.e. the previous hidden state  $\mathbf{h}_{t-1}$  and the current word input  $\mathbf{x}_t$ . The most elementary RNN is often called the Elman RNN, corresponding to the following equation:

$$\mathbf{h}_t = \sigma(\mathbf{W}_1 \cdot \mathbf{h}_{t-1} + \mathbf{W}_2 \cdot \mathbf{x}_t) \quad (4)$$

However, previous research has noted that the Elman RNN has difficulties in modeling long sequences. Extensions have been proposed to handle this problem, and two well-known examples are Long Short-Term Memory (LSTM) units [17] and Gated Recurrent Units (GRUs) [11]. Both these approaches involve different components, i.e. gating mechanisms, which interact in a particular way in order to combine previous states with the current inputs. For instance, GRUs correspond to the following equations:

$$\mathbf{z}_t = \varphi_g(\mathbf{W}_z \cdot \mathbf{x}_t + \mathbf{U}_z \cdot \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (5)$$

$$\mathbf{r}_t = \varphi_g(\mathbf{W}_r \cdot \mathbf{x}_t + \mathbf{U}_r \cdot \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (6)$$

$$\tilde{\mathbf{h}}_t = \varphi_h(\mathbf{W}_h \cdot \mathbf{x}_t + \mathbf{U}_h \cdot (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (7)$$

$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t \quad (8)$$

In the previous equations, the operator  $\odot$  denotes the Hadamard product (i.e., the entry-wise product of two matrices), while  $\mathbf{x}_t$  denotes the input vector at time step  $t$ , and  $\mathbf{h}_t$  denotes the hidden state at time step  $t$ . The activations  $\varphi_g(\cdot)$  and  $\varphi_h(\cdot)$  are typically chosen to be the sigmoid and the hyperbolic tangent functions, respectively, while the multiple parameters  $\mathbf{W}$ ,  $\mathbf{U}$  and  $\mathbf{b}$  denote the different weight matrices and bias vectors, adjusted when training the neural network.

Notice that a GRU involves two gates, namely a reset gate  $\mathbf{r}$ , that determines how to combine the new input with the previous memory, and an update gate  $\mathbf{z}$ , that defines how much of the previous memory to keep around. If we set the reset gate to all ones, and the update gate to all zeros, we again arrive at the Elman RNN model

that was discussed previously. The gating mechanism allows GRUs to better handle long-term dependencies. By learning the parameters for its gates, the network learns how its internal memory should behave, given that the gates define how much of the input and previous state vectors should be considered.

LSTMs are an alternative type of RNN, with more parameters than GRUs (e.g., they have an extra gate) although also reported to outperform them when more training data is available, and in tasks requiring modeling longer-distance relations [53]. LSTMs correspond to the following equations, where  $\odot$ ,  $\varphi_g(\cdot)$ ,  $\varphi_h(\cdot)$ ,  $\mathbf{x}_t$ ,  $\mathbf{h}_t$ ,  $\mathbf{W}$ ,  $\mathbf{U}$  and  $\mathbf{b}$  share the same meaning as in the GRU equations.

$$\mathbf{i}_t = \varphi_g(\mathbf{W}_i \cdot \mathbf{x}_t + \mathbf{U}_i \cdot \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (9)$$

$$\mathbf{f}_t = \varphi_g(\mathbf{W}_f \cdot \mathbf{x}_t + \mathbf{U}_f \cdot \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (10)$$

$$\mathbf{o}_t = \varphi_g(\mathbf{W}_o \cdot \mathbf{x}_t + \mathbf{U}_o \cdot \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (11)$$

$$\mathbf{g}_t = \varphi_h(\mathbf{W}_g \cdot \mathbf{x}_t + \mathbf{U}_g \cdot \mathbf{h}_{t-1} + \mathbf{b}_g) \quad (12)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{g}_t \odot \mathbf{i}_t \quad (13)$$

$$\mathbf{h}_t = \varphi_h(\mathbf{c}_t) \odot \mathbf{o}_t \quad (14)$$

Notice that LSTMs apply different gating mechanisms, namely through the use of (i) a forget gate that controls how much of the previous memory will be kept, (ii) an input gate that controls how much of the proposed update  $\mathbf{g}_t$  should be kept, and (iii) an output gate that controls the output at time  $t$ .

Both LSTMs and GRUs have been shown to achieve remarkable performance on many sequential learning problems [53], particularly when processing short texts. In addition, hierarchical arrangements of LSTMs or GRUs can be used for modeling long documents, accounting not only with word order but also with sentence structure [14, 52]. For instance Yang et al. [52] proposed a hierarchical attention network for document classification, considering documents to be composed of sentences, and modeling sentences as sequences of words. They used a bi-directional GRU (i.e., they concatenate the states produced by two GRUs to generate the result for each time step, with one GRU processing the input in the forward direction, and the other processing the input in reverse) to encode each word in a sentence, followed by an attention mechanism to weight relevant words in the aggregated representation of each sentence. The attention mechanism corresponds to the following equations, effectively weighting the contribution of each word in the sentence, when building the representation:

$$\mathbf{u}_{it} = \tanh(\mathbf{W}_h \cdot \mathbf{h}_{it} + \mathbf{b}_w) \quad (15)$$

$$\alpha_{it} = \frac{\exp(\mathbf{u}_{it}^\top \mathbf{u}_w)}{\sum_t \exp(\mathbf{u}_{it}^\top \mathbf{u}_w)} \quad (16)$$

$$\mathbf{s}_i = \sum_t \alpha_{it} \times \mathbf{h}_{it} \quad (17)$$

The vector  $\mathbf{h}_{it}$  corresponds to the word representation produced by the bi-GRU at time  $t$ , when processing sentence  $i$ , while  $\mathbf{W}_h$  and  $\mathbf{b}_w$  are parameters to be learned. The vector  $\mathbf{u}_{it}$  is a hidden representation of  $\mathbf{h}_{it}$ ,  $\mathbf{u}_w$  is a word-level context vector to be learned,  $\alpha_{it}$  is an importance weight, and  $\mathbf{s}_i$  is the sentence vector, calculated as the weighted sum of the word representations. To create a document vector, the same methodology can be applied. A bi-directional GRU encodes the sentence vectors, and an attention mechanism is used to determine the importance of each sentence, yielding a document vector. The equations for the document-level attention mechanism are analogous to those from Equations 15 to 17. To generate the final prediction, the document vector is processed through feed-forward layer(s) with a final a softmax activation.

Besides document classification, another NLP task that is strongly related to the subject of this article, and where deep learning has been extensively employed, is Natural Language Inference (NLI). In brief, NLI concerns with determining if a given hypothesis sentence  $h$  can be inferred from a premise sentence  $p$ . Generalized versions

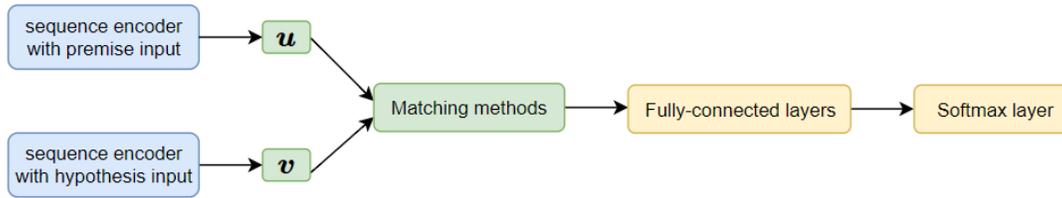


Fig. 1. Overview of a generic NLI approach, adapted from the description by Conneau et al. [12].

of the task have also considered multiple possible relations between the hypothesis and the premise (e.g., relations like entailment, contradiction, or neutrality), and large datasets such as the Stanford Natural Language Inference (SNLI) [4] or the Multi-Genre Natural Language Inference (MultiNLI) [49] corpora have enabled significant progress in terms of deep learning methods for building effective semantic representations of natural language information (e.g., models trained for NLI tasks can provide rich domain-general semantic representations).

For instance, Conneau et al. [12] demonstrated that the supervised training of sentence embeddings, based on the aforementioned NLI datasets, can consistently outperform other state-of-the-art approaches for representing sentences in different NLP tasks (e.g., taking the average of *word2vec* [29] or GloVe [20] embeddings, using unsupervised methods such as skip-thought sentence embeddings [23], or using other supervised models such as the *paragram-phrase* approach described by Wieting et al. [21]). The model parameters obtained when solving the NLI problem can thus be used for initializing other NLP models. Conneau et al. have also advanced a generic architecture for addressing the NLI task, illustrated in Figure 1. In brief, the premise and the hypothesis can both be encoded by a CNN or an RNN, creating a vector representation for each sentence. These representations are then matched in some way (e.g., through a concatenation of the vectors, through the vector difference, and/or through an element-wise product), fed into a set of fully-connected layers, and finally processed through a feed-forward layer with a softmax activation, that generates a final prediction.

As a sentence encoder (i.e., to generate sentence representations  $\mathbf{u}$  and  $\mathbf{v}$ , respectively from the premise and the hypothesis), Conneau et al. used a bi-directional LSTM, followed by a max-pooling operation over the sequence of output states. Then, the authors combined both these representations through the concatenation of  $\mathbf{u}$  and  $\mathbf{v}$ , the element-wise product, and the difference between the vectors (i.e., they used a combination of all three operations). Finally, the result was fed to multiple fully-connected layers, with a final softmax activation.

Besides the aforementioned general model, several other NLI approaches have been proposed in the literature [7, 16, 31], and evaluated over the SNLI and MultiNLI corpora (e.g., the authors of SNLI maintain a leader-board on the website<sup>3</sup> describing the corpus). For instance, Nie and Bansal [31] developed a model that scores an accuracy of 86.1% on the SNLI test set, an accuracy of 74.6% on the MultiNLI matched testing set, and 73.6% on the MultiNLI mismatched testing set. Following the same general approach from Figure 1, these authors used a sentence encoder that takes as input pre-trained GloVe embeddings [20] for the words in the sentence, and processes/refines the embeddings through three layers of bi-directional LSTMs with shortcut connections (i.e., the input sequences for the  $i$ -th bi-LSTM layer are the concatenated outputs of all the previous layers, plus the original word embedding sequence). The final representations of the premise and hypothesis are generated through a max-pooling operation between all the hidden states of the last bi-directional LSTM. A similar approach is used in the present study to encode sentences in the stance detection problem from the Fake News Challenge, although also using inner-attention as in the study by Yang et al. [52], instead of just using max-pooling.

<sup>3</sup><http://nlp.stanford.edu/projects/snli/>

## 2.2 Stance Detection in the Context of the Fake News Challenge

The Fake News Challenge (FNC-1) consisted of a competition for evaluating NLP methods designed to solve a stance detection problem between the headline of a news article, and the main body of text for a news article. The leader-board of the competition points to several interesting methods, some of them reviewed in this section.

The organizers of the competition also provided a baseline consisting of a gradient boosting classifier leveraging hand-crafted features, such as multiple similarity scores between the headline and the body (e.g., based on  $n$ -gram overlap). Using 10-fold cross validation, the baseline achieved a weighted accuracy of 75.20%, following the evaluation metric detailed in Section 4.1 (i.e., an accuracy metric that gives extra weight to some of the classes).

The first place in the competition was obtained by the team *SOLAT in the SWEN*<sup>4</sup>, which used an ensemble of two sub-models, each outputting predictions associated with a confidence score. The aggregated model made its decisions with basis on the weighted average of both sub-models, achieving a weighted accuracy of 82.02%. One of the sub-models was a gradient boosting classifier similar to that of the baseline, leveraging features like the number of overlapping words between the headline and the body, or similarity scores computed from word  $n$ -gram representations. The second sub-model was based on Convolutional Neural Networks (CNNs) for encoding the headline and the body, using pre-trained *word2vec* embeddings [29]. The outputs for the headline CNN and body CNN were concatenated and put through feed-forward layers.

In second place came *Team Athene*<sup>5</sup>, with a weighted accuracy of 81.97% and using an ensemble of five identical sub-models with randomly initialized parameters. These sub-models were Multi-layer Perceptrons (MLPs) with seven hidden layers and a softmax activation at the end. Seven distinct feature types were considered, computed from either the headline, the body, or a combination of both. Examples of features include vectors of unigram occurrences, or the cosine distance between headline and body representations computed from a factorization of the unigram occurrence matrix (i.e., from representations produced through non-negative matrix factorization). The final prediction came from hard voting between the five sub-models in the ensemble.

In third place, team *UCL Machine Reading* [37] achieved a weighted accuracy of 81.72%. This model was also fairly simple, consisting of a MLP with one hidden layer and a final layer with a softmax activation. The input feature vector resulted from a concatenation of Term-Frequency (TF) vectors built from the headline and the body, with the cosine similarity between TF-IDF vectors for the headline and the body.

When considering the top three teams, it is important to notice that all of them leveraged hand-crafted features, together with other neural approaches. In a recent publication, Tosik et al. [45] reported on tests evaluating the individual contribution of hand-crafted features, using them to feed a single gradient boosting classifier (when directly considering the four class classification problem), or an ensemble of gradient boosting classifiers, where the instances are first classified as either *unrelated* or *related*, and then the *related* instances are assigned to one of the remaining three classes (i.e., *agrees*, *disagrees*, or *discusses*).

Specifically, Tosik et al. leveraged text similarity features such as  $n$ -gram or word overlap, or the cosine similarity between TF-IDF vectors representing the headline and the body, as well as other miscellaneous features such as the presence of several refuting words in the headline/body, or the length of the headline/body. After achieving a weighted accuracy score of 78.63% on the FNC-1 testing dataset, the authors conducted ablation tests and concluded that the most helpful features were the overlap features between  $n$ -grams and words of the headline and body, the refuting features based on a lexicon, and distance measures such as the cosine similarity between TF-IDF vectors of the headline and the body, and the Word Mover's Distance [25] between the headline and body text. On the other hand, sentiment features and other distance scores, such as the Hamming distance, did not contribute to the obtained result.

<sup>4</sup><http://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html>

<sup>5</sup><http://medium.com/@andre134679/team-athene-on-the-fake-news-challenge-28a5cf5e017b>

Pfohl and Legros [35] reported on other interesting submissions to FNC-1. These authors experimented with four different approaches, namely (i) a bag-of-words method, (ii) a basic LSTM approach, (iii) a method leveraging an LSTM together with an inner-attention mechanism, and (iv) a more sophisticated approach which the authors named conditional encoding LSTM with attention (CEA-LSTM).

The bag-of-words model was built by averaging embeddings for words occurring in the headline, for words occurring in the body, and then concatenating these vectors. This result is processed by a feed-forward network with a softmax output layer. Both the basic LSTM model, and the model combining an LSTM with an inner-attention mechanism, processed a concatenation of the headline and the article body to classify the stance. The inner-attention was computed over a window with the first 15 tokens of the concatenated text. Finally, the CEA-LSTM processed the headline and the body with two separate LSTMs, using the final hidden state of the headline LSTM as the first hidden state of the body LSTM. An attention mechanism operated over a window with the last 15 output states from the headline LSTM, together with the final hidden state of the body LSTM. The best results reported by Pfohl and Legros were achieved with the CEA-LSTM method, corresponding to a weighted accuracy of 80.8% (i.e., the method based on LSTMs failed to outperform other simpler approaches).

More recently, after the announcement of the winners for FNC-1, Bhatt et al. [2] described an approach based on a MLP for combining neural representations, statistical summaries of the data, and feature engineering heuristics. The neural representations leveraged skip-thought vectors [23] (i.e., sentence representations learned in an unsupervised manner, by encoding a sentence to predict the sentences around it in a given text corpora) to encode the headline and the body. Given the encodings for the headline and the body, the component-wise product and the absolute difference between the vectors were computed and used as features. The statistical features correspond to unigram occurrences within the headline and the body, weighted according to TD-IDF. Finally, the external heuristic features included the number of similar words in the headline and the body, the cosine similarity between vector encodings for the headline and the body, the number of matching  $n$ -grams, the sentiment difference between the headline and the body, etc. Each of the three main sets of features is processed through feed-forward layers before being combined and processed by a final layer with a softmax activation. This approach achieved a weighted accuracy of 83.08%, scoring higher than the winning team in FNC-1. To the best of our knowledge, this result corresponds to the current state-of-the-art in the FNC-1 task.

Chaudhry et al. [1] described six different models for stance detection, which were also tested on FNC-1 data. The first model was a baseline leveraging the Jaccard similarity between unigrams from the headline and from the sentences in the body. The second model was a MLP leveraging representations for the headline and the body, e.g. obtained by averaging pre-trained GloVe embeddings [20]. The third model explored the use of two independent LSTMs, one encoding the headline and another encoding the body. The two final state vectors are passed to a softmax layer that generates the final prediction. The fourth model also used an LSTM to encode the headline, but now its final state vector was used to initialize another LSTM that encodes the body. The fifth approach augmented the fourth model by considering bi-directional LSTMs. Finally, the sixth approach extended the fifth model with a self-attention mechanism on top of the encoder LSTMs. All six models were tested on a custom training/testing data split (i.e., the official test set was not available at the time the paper was written, and thus the results cannot be directly compared against those reported for the FNC-1 competition). The best results were achieved by the fifth model (i.e., the bi-directional encoder), corresponding to a weighted accuracy of 95.3%.

Similarly to Tosik et al. [45], Bourgonje et al. [3] also addressed the separate tasks of (a) determining whether a headline-body pair is *unrelated* or *related* and, in the latter case, (b) the task of determining whether the body *agrees*, *disagrees*, or *discusses* the subject of the headline. Again, experiments leveraged data from FNC-1. In the first task, the authors start by gathering two sets of  $n$ -grams, one for the headline and another for the body. Then, the number of matching  $n$ -grams is multiplied by the length and IDF value of the matching  $n$ -grams, and divided by the total number of  $n$ -grams. If the resulting score is above a threshold, the pair is considered to be *related*. For the second task (i.e., the three class classification problem), Bourgonje et al. used a logistic regression classifier

trained on features extracted from the headlines of the FNC-1 dataset. If the distance between the best and the second best scoring classes is below a given threshold, the authors use a separate binary classifier, trained on features from both the headlines and the bodies of FNC-1 news articles. Three binary classifiers were trained for this second-level model, namely one for discriminating between *agrees* and *disagrees*, one for *agrees* versus *discusses*, and another one for *discusses* versus *disagrees*. Bourgonje et al. evaluated the proposed approach by leveraging 50 different tests with random 90-10 splits of the FNC-1 dataset, scoring a weighted accuracy of 89.59%. However, these results are again not directly comparable to those from the FNC-1 participants.

Zeng et al. [54] tested six different encoders in a neural method for the FNC-1 task, based on concatenating representations for the headline and the body, afterwards generating the final classification through a softmax layer. The first encoder consisted of separate bi-directional GRUs, one for the headline and another for the body, with representations obtained from the final GRU states. In the second approach, the headline is fed to a bi-directional GRU, whose final hidden state is used to initialize another bi-directional GRU that processes the body. The third encoder uses a bi-directional GRU to process the concatenation of the headline and the body text.

The fourth and fifth models both leverage attention mechanisms. Specifically, the fourth model compares every state of the bi-directional GRU processing the body with the final hidden state of the bi-directional GRU that processes the headline. The attention weights are generated as follows:

$$\alpha_i = \text{softmax}(\mathbf{q}^\top \cdot \mathbf{W}_s \cdot \tilde{\mathbf{p}}_i). \quad (18)$$

In the previous equation,  $\mathbf{q}$  is the final state of the headline bi-directional GRU,  $\tilde{\mathbf{p}}_i$  is the  $i$ -th hidden state of the body bi-directional GRU,  $\mathbf{W}_s$  are trainable weights, and  $\alpha_i$  is the attention weight of the  $i$ -th body token. The attention weights are used to compute a weighted sum of the hidden states from the body bi-directional GRU, and this result is then concatenated with the final hidden state of the bi-directional GRU encoding the headline.

The fifth model is very similar to the fourth, in this case computing the attention weights by comparing the hidden states of the body GRU with all the hidden states of the headline GRU, instead of only the final state. The final attention weight for each body hidden state is given by the maximum value over all the weights calculated for every hidden state of the headline GRU.

Finally, the sixth approach was named the Bilateral Multiple Perspective Matching model. First, the headline and the body are encoded using separate bi-directional GRUs. Then, a cosine similarity is calculated between every hidden state from the headline GRU, and every hidden state from the body GRU:

$$\mathbf{m}_k = \text{cosine}(\mathbf{w}_k \odot \mathbf{v}_1, \mathbf{w}_k \odot \mathbf{v}_2). \quad (19)$$

In the previous equation,  $\mathbf{w}_k$  are trainable weights, and  $\mathbf{v}_1$  and  $\mathbf{v}_2$  represent the hidden states to be compared. For every headline hidden state, several representations were created (i.e., as many  $\mathbf{m}_k$  vectors as the length of the body). A max-pooling is then applied, in order to obtain a single vector for every headline state. For the body hidden states, the process is analogous. After this first layer of representations, a second layer again uses bi-directional GRUs to further process the headline and the body representations. The two resulting hidden states are finally concatenated, and fed to feed-forward layers that perform the classification.

The aforementioned six models were evaluated on a custom split of the FNC-1 dataset, and the fifth approach described by the authors (i.e., the attention model comparing the body hidden states with every headline hidden state) achieved the best results, with a weighted accuracy of 85.2%. Again, this result cannot be directly compared against those from the FNC-1 participants.

More recently, Mohtarami et al. [30] introduced a memory network for stance detection, evaluating it on the FNC-1 testing dataset. The proposed memory network can be seen as 6-tuple  $\{M, I, F, G, O, R\}$ , where the memory  $M$  is a sequence of representations,  $I$  is a mapping from inputs to their representations,  $F$  is an inference component that identifies the relevant parts of the input,  $G$  is a generalization component that updates the

memory according to  $F$ ,  $O$  is an output generated for each new input given the current memory state, and finally  $R$  is conversion from  $O$  into a desired response format.

Given a news article and a headline, the input component  $I$  first converts the news article into a 3D tensor  $\mathbf{d} = (X, W, E)$ , where  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  is a set of paragraphs that constitute the document,  $W = \{\mathbf{w}_1, \dots, \mathbf{w}_v\}$  are the words that represent each paragraph  $\mathbf{x}_j$ , and  $E = \{\mathbf{e}_1, \dots, \mathbf{e}_v\}$  are the word embeddings. Every  $\mathbf{x}_j$  is separately processed through an LSTM and a CNN, generating representations  $\mathbf{m}_j$  and  $\mathbf{c}_j$ . A similar procedure is considered for the headline, generating representations  $\mathbf{s}_{lstm}$  and  $\mathbf{s}_{cnn}$ .

The inference component  $F$  takes the previously computed representations and generates two similarities:

$$\mathbf{p}_{lstm}^j = \mathbf{s}_{lstm}^\top \cdot \mathbf{M} \cdot \mathbf{m}_j \quad (20)$$

$$\mathbf{p}_{cnn}^j = \mathbf{s}_{cnn}^\top \cdot \mathbf{M}' \cdot \mathbf{c}_j \quad (21)$$

In the previous equations,  $\mathbf{M}$  and  $\mathbf{M}'$  are trainable similarity matrices. Additionally, another similarity vector  $\mathbf{p}_{tfidf}^j$  is computed by applying the cosine similarity metric to TF-IDF representations for the news headline and for each sentence of the news article body.

The memory component  $M$  and the generalization component  $G$  update the  $\mathbf{m}_j$  vectors in the following way:

$$\tilde{\mathbf{m}}_j = \mathbf{m}_j \odot \mathbf{p}_{tfidf}^j \quad (22)$$

Then, the updated  $\tilde{\mathbf{m}}_j$  and  $\mathbf{s}_{lstm}$  are used by the inference component  $F$  to compute separate  $\tilde{\mathbf{p}}_{lstm}^j$  vectors, using the same procedure detailed above. The new  $\tilde{\mathbf{p}}_{lstm}^j$  vectors are then used to update the  $\mathbf{c}_j$  vectors:

$$\tilde{\mathbf{c}}_j = \mathbf{c}_j \odot \tilde{\mathbf{p}}_{lstm}^j \quad (23)$$

Finally, the updated  $\tilde{\mathbf{c}}_j$  representations in conjunction with  $\mathbf{s}_{cnn}$  are leveraged to compute new  $\tilde{\mathbf{p}}_{cnn}^j$  vectors, using the same procedure as explained above. All the aforementioned vectors are used in the output component  $O$  to compute the following vector:

$$\mathbf{o} = [\text{mean}(\{\mathbf{c}_j\}); \text{max}(\{\mathbf{p}_{cnn}^j\}); \text{mean}(\{\tilde{\mathbf{p}}_{cnn}^j\}); \text{max}(\{\mathbf{p}_{lstm}^j\}); \text{mean}(\{\tilde{\mathbf{p}}_{lstm}^j\}); \text{max}(\{\mathbf{p}_{tfidf}^j\}); \text{mean}(\{\tilde{\mathbf{p}}_{tfidf}^j\})] \quad (24)$$

In the response component  $R$ , the concatenation  $[\mathbf{o}; \mathbf{s}_{lstm}; \mathbf{s}_{cnn}]$  is fed to feed-forward layers, and a final softmax layer gives the prediction for the stance. In their tests, Mohtarami et al. achieved a weighted accuracy of 81.23%, again attesting to the difficulties associated to the application of modern neural approaches to the FNC-1 task.

### 3 PROPOSED NEURAL NETWORK ARCHITECTURE

Our neural network architecture takes inspiration on the models by Yang et al. [52] and by Nie and Bansal [31], described on the previous section. Figure 2 provides an high-level overview of the proposed approach.

The headline is processed through the sentence encoder described in Section 3.1, which outputs the corresponding representation. In turn, the body of the news article is processed through the document encoder described in Section 3.2, which leverages a two-layer hierarchical model combining the sentence encoder with an encoder for the sequence of sentences. A third optional branch compares the headline with the first two sentences of the body, leveraging the sentence encoder to build the involved representations. This matcher, as well as the common sentence encoder, can be pre-trained with the SNLI and MultiNLI datasets, as explained in Section 3.3.

The representations for the headline, the entire body, and the first two sentences from the body, are matched through vector operations such as the element-wise product, the vector difference, or the concatenation. These results are then combined with external features computed from the headline and/or the body – see Section 3.4 – and the result is processed by two fully-connected layers, with the final one producing the stance classification. The following sections detail the components of this architecture.

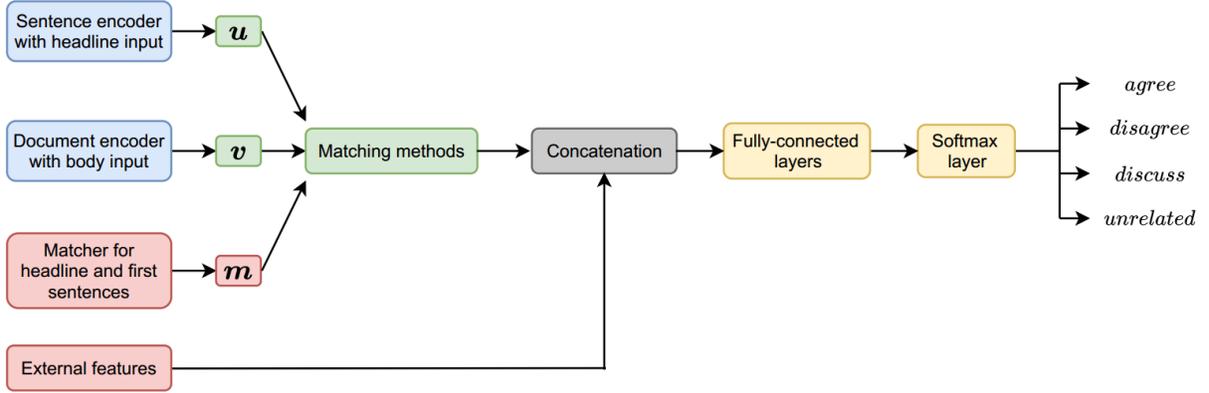


Fig. 2. Overview of the proposed approach.

### 3.1 Encoding Sentences with Bi-Directional RNNs and Summarization Mechanisms

The proposed neural architecture takes inspiration from Conneau et al. [12], leveraging a sentence encoder responsible for building representations for input sentences, e.g., for the news headlines. This encoder takes as input a sequence of  $w_l$  words (i.e., a sentence), with  $l \in [0, L]$  and where  $L$  is the length of the sequence, replacing each word by a pre-trained GloVe embedding [20]. The resulting matrix of embedded words  $E$  is processed through a bi-directional RNN, which generates a hidden state matrix  $H_1$  as follows:

$$h_{tf} = \text{RNN}_{\text{forward}}(E_t), t \in [1, \dots, L] \quad (25)$$

$$h_{tb} = \text{RNN}_{\text{backward}}(E_t), t \in [1, \dots, L] \quad (26)$$

$$h_t = [h_{tf}; h_{tb}], t \in [1, \dots, L] \quad (27)$$

$$H = [h_1, \dots, h_L]. \quad (28)$$

In the previous equations, the RNN function can be instantiated either with the LSTM or with the GRU equations, both shown in Section 2.1 and hence not repeated here. Finally, the RNN states  $h_t$  from matrix  $H$  are processed through a summarization mechanism, which outputs a single vector. For this work, we consider the summarization mechanism to be either a max-pooling operation, and/or an inner-attention mechanism defined as shown in the next equations:

$$u_t = \tanh(W \cdot h_t), t \in [1, \dots, L] \quad (29)$$

$$\alpha_t = \frac{\exp(u_t^T)}{\sum_t \exp(u_t^T)}, t \in [1, \dots, L] \quad (30)$$

$$s = \sum_t \alpha_t \times h_t, t \in [1, \dots, L] \quad (31)$$

In the equations above, the matrix  $W$  corresponds to trainable weights,  $\alpha_t$  is an importance weight assigned to each hidden state, and  $s$  is the final representation of the input sentence.

In order to further augment the sentence encoder described above, we also experimented with stacking two layers of bi-directional RNNs, feeding the second bi-directional RNN layer with the concatenation of original embeddings  $E$  and the hidden states  $H$  from the first bi-RNN (i.e., considering shortcut connections between the two layers). In this case, a second hidden state matrix  $H_2$  is produced through the same procedure described through Equations 25-28. Finally, the RNN states  $h_t$  from matrix  $H_2$  are processed through a summarization mechanism, similar to what was detailed in the previous paragraph.

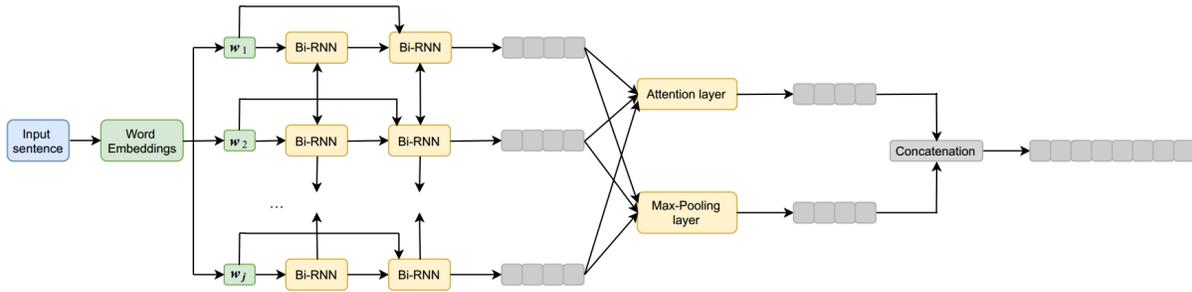


Fig. 3. An extension of the previously detailed sentence encoder.

Figure 3 illustrates the sentence encoder combining all the aforementioned mechanisms, thus using two stacked layers of bi-RNNs (i.e., GRUs or LSTMs), shortcut connections, and a combination of max-pooling and neural attention obtained through a concatenation operation.

### 3.2 Hierarchical Approach for Encoding Documents

The neural encoder for the news article body leverages the sentence encoder from the previous section, combining it with the hierarchical approach described by Yang et al. [52]. Each sentence in the news article is thus encoded through the procedure described in Section 3.1, and the resulting sequence of sentence vectors is then processed through a similar encoder (i.e., using a bi-directional RNN or a stacked model with two bi-RNNs and shortcut connections, followed by max-pooling and/or an inner-attention mechanism).

### 3.3 Model Pre-Training with the SNLI and MultiNLI Datasets

Besides matching the headline against the entire body of the news article, we also experimented with a neural architecture that additionally attempts to match the headline against the first two sentences of the body, through a separate branch. This approach leverages the intuition that the opening sentences of news articles often contain a summary for the entire document. The sentence encoder described in Section 3.1 is used to represent the headline and the first sentences of news article, and these representations are then combined through vector operations (i.e., the element-wise multiplication, the difference, and the concatenation).

Moreover, instead of randomly initializing all the parameters of the proposed neural network architecture, we used the SNLI [4] and MultiNLI [49] datasets to pre-train the parts of the network that do (i) the encoding of sentences, and (ii) the matching of the headline against the first two sentences of the body, as described in the previous paragraph. Previous studies have already attested to the benefits of similar pre-training procedures in other natural language processing and text classification tasks [12].

The headline is seen as the hypothesis sentence from the NLI task, and the first two sentences from the body are seen as the premise. Our complete NLI model thus involves the same sentence encoder and the aforementioned matching strategy (i.e., combining the representations for the hypothesis and the premise through the element-wise multiplication, vector difference, and vector concatenation), combined with a final softmax layer that returns the NLI class. After pre-training this model, we simply ignore the final softmax layer and re-use the remaining components on our complete FNC-1 neural architecture.

### 3.4 Combining the Matched Representations with External Features

The previous section described the vector operations used for matching the headline against the first two sentences of the body. A similar approach is also used for matching the representation for the headline against the representation for the entire body of the news article. Thus, the vector resulting from the pre-trained NLI network, which matches the headline against the first two sentences, is concatenated with the vector encoding the entire body, and with the results from the element-wise product and the difference between the vectors representing the headline and the body.

Moreover, taking inspiration on the baseline model proposed by the organizers of FNC-1, we also combine the aforementioned representations built through the neural network with external features computed from the headline and/or the body. The following list enumerates the considered features:

- (1) The number of words in common between (a) the headline and the body of the news article, and (b) the headline and the first two sentences of the body;
- (2) Refutation features, based on the presence of refuting words, listed in a given dictionary, in the headline (e.g., words like *deny*, *doubt*, *fraud* or *debunk*);
- (3) Polarity features, based on the presence of words with high emotion/sentiment polarity in (a) the headline, (b) the entire body, and (c) the first two sentences;
- (4) The number of word tokens that are common to (a) the headline and the body of the news article, and (b) the headline and the first two sentences of the body;
- (5) The number of word  $n$ -grams that are common to (a) the headline and the body of the news article, and (b) the headline and the first two sentences of the body;
- (6) The soft cosine similarity metric [6], computed between representations leveraging word occurrences for (a) the headline and the entire body, or (b) the headline and the first two sentences of the body;
- (7) The BLEU score [33] computed between (a) the headline and the set of sentences from the body, and (b) the headline and the first two sentences of the body;
- (8) ROUGE scores [27] computed between (a) the headline and the set of sentences from the body, and (b) the headline and the first two sentences of the body;
- (9) The CIDEr similarity score [47], computed between (a) the headline and the set of sentences of the article, and (b) the headline and the first two sentences of the body;
- (10) The cosine similarity metric, computed between TF-IDF vector representations for the words occurring in the headline, and in the body of the news article.
- (11) A vector representation of the headline, with 50 dimensions, resulting from a Singular Value Decomposition (SVD) of a matrix with TF-IDF representations for the texts;
- (12) A vector representation for the body of the article, with 50 dimensions, resulting from a Singular Value Decomposition (SVD) of a matrix with TF-IDF representations for the texts;
- (13) The cosine similarity metric, computed between the SVD vectors for the headline and the body;
- (14) A vector representation for the headline, with 300 dimensions, produced by averaging the *word2vec* embeddings for the words occurring in the headline;
- (15) A vector representation for the body, with 300 dimensions, produced by averaging the *word2vec* embeddings for all the words occurring in the body of the article;
- (16) The cosine similarity metric, computed between averaged *word2vec* embeddings for the headline and body;
- (17) Sentiment polarity scores for the headline and the body, computed with basis on a word polarity lexicon.

The first 5 features from the previous enumeration were taken from the official FNC-1 baseline system, provided by the organizers. Features 10 to 17 were taken from the system<sup>6</sup> that won the FNC-1 challenge, developed by the team *SOLAT in the SWEN*. Finally, in addition to external features based on previous work within the context

<sup>6</sup><http://github.com/Cisco-Talos/fnc-1>

of FNC-1, we also considered features from previous work concerned with assessing text similarity (e.g., the soft cosine similarity metric), or concerned with the evaluation of NLP methods for automated translation (i.e., the BLEU score), text summarization (i.e., ROUGE scores), or caption generation (i.e., the CIDEr score). The three aforementioned metrics all attempt to assess the similarity between a given input text and a set of reference texts. When computing them, we considered the input text (i.e., the candidate) to be the headline, and the set of reference texts to be formed by each sentence that composes the corresponding news article. When considering only the first two sentences from an article, we instead calculated the scores between the headline (i.e., the candidate) and the concatenation of the referred first two sentences (i.e., a reference set with a single instance).

BLEU [33] works by counting the number of matching unigrams between the candidate text and the references. For every unigram in the candidate, BLEU saves the maximum amount of times it appears in a reference. BLEU then takes the minimum values from the aforementioned counts, and the number of times that the corresponding unigram appears in the candidate, hence generating a value  $m$  for every unigram in the candidate. Finally, the BLEU score between the candidate and the references is computed by summing all the  $m$  values for every unigram, and then dividing the result by the number of unigrams in the candidate text.

ROUGE [27] is similarly based on matches between the candidate and the references. Different variations of the metric can be computed, and we specifically used ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 is based on the average of the number of overlapping unigrams between the candidate and the references. ROUGE-2 is similar, considering instead the number of overlapping bi-grams. ROUGE-L averages the number of longest common sub-sequences between the candidate and the references.

CIDEr [47] begins by representing each sentence (i.e., the candidate and the references) as a TF-IDF vector, with basis on the  $n$ -grams that compose it. The CIDEr score between a candidate sentence and a set of reference sentences is then computed as follows:

$$\text{CIDEr}(c, S) = \sum_{n=1}^N w_n \times \left( \frac{1}{m} \times \sum_j \frac{\mathbf{g}_c^n \cdot \mathbf{g}_{s_j}^n}{\|\mathbf{g}_c^n\| \times \|\mathbf{g}_{s_j}^n\|} \right) \quad (32)$$

In the previous equation,  $c$  is the candidate sentence,  $S$  is the set of references,  $\mathbf{g}_c^n$  is the TF-IDF vector representation for the  $n$ -grams occurring in the candidate, and  $\mathbf{g}_{s_j}^n$  is the TF-IDF vector for the  $j$ -th reference. The number of references is represented by  $m$ , while  $w_n$  is a weight, and  $N$  is the maximum length of  $n$ -grams to consider. In our case, we defined  $w_n = 1/N$  and  $N = 4$ .

Finally, the Soft Cosine Similarity (SCS) metric computes the similarity between two texts leveraging bag-of-words representations, being computed as follows:

$$\text{SCS}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \cdot \mathbf{M} \cdot \mathbf{b}}{\sqrt{\mathbf{a}^\top \cdot \mathbf{M} \cdot \mathbf{a}} \times \sqrt{\mathbf{b}^\top \cdot \mathbf{M} \cdot \mathbf{b}}} \quad (33)$$

In the above equation,  $\mathbf{a}$  and  $\mathbf{b}$  are the representations for the input texts, and  $\mathbf{M}$  is a relation matrix whose element  $m_{ij}$  expresses some relation between word  $i$  and word  $j$ , hence guaranteeing that two texts without any word in common have a score above 0 as soon as they share related words. For our work,  $\mathbf{M}$  is a sparse term similarity matrix computed from 50-dimensional GloVe word embeddings [20].

#### 4 EXPERIMENTAL EVALUATION

We evaluated the stance detection model described on the previous section by leveraging the problem definition and the evaluation methodology (i.e., the dataset and the metrics) from the Fake News Challenge (FNC-1). This challenge aimed to evaluate models for estimating the stance of a news article body towards a given news

Table 1. Characterization statistics for the FNC-1 dataset.

Property	Training Split	Testing Split
Number of instances	49,972	25,413
Number of different news headlines	1,648	893
Number of different news article bodies	1,683	899
Headline average length (tokens)	13	12
Body average length (tokens)	428	396
Percentage of <i>unrelated</i> pairs	73.131%	72.203%
Percentage of <i>discuss</i> pairs	17.828%	17.566%
Percentage of <i>agree</i> pairs	7.360%	7.488%
Percentage of <i>disagree</i> pairs	1.681%	2.742%

headline. One such model should take as input a headline and a body text, from the same article or from different articles, and return as output a classification for the stance, considering the following four categories:

- (1) **Agrees:** The body agrees with the headline;
- (2) **Disagrees:** The body disagrees with the headline;
- (3) **Discusses:** The body is related with the headline, but it does not take a position regarding its subject;
- (4) **Unrelated:** The body discusses a different topic from that of the headline.

#### 4.1 The FNC-1 Dataset and Experimental Methodology

To support the training and testing of models participating in the FNC-1, the organizers of the challenge released training and testing datasets. The training dataset contains 49,972 instances (i.e., pairs of headline and body texts) classified with a stance. The labeled testing dataset was, in turn, released after the end of the competition, and it contains 25,419 instances. Table 1 presents elementary characterization statistics for the training/testing datasets provided by the FNC-1 organizers.

Due to the imbalance in the class distribution, and also due to the fact that the distinction between *agree*, *disagree* and *discusses* is much more relevant to fake news detection, the FNC-1 organizers suggested a weighted scoring system. If a test instance is *unrelated* and the model labels it correctly, the score will be incremented by 0.25. If the test instance is related, i.e., the correct label is either *agree*, *disagree* or *discusses*, then the score will be incremented by 0.25 if the model labels the pair with one of the previously mentioned labels. In case the model chooses the correct label of a related test instance, the score will be incremented by an additional 0.75. Summing up, the equation for the proposed weighted accuracy metric is as follows:

$$Acc_{FNC} = 0.25 \times Acc_{Related, Unrelated} + 0.75 \times Acc_{Agree, Disagree, Discuss} \quad (34)$$

#### 4.2 Parameters Involved in the Proposed Approach

We evaluated the complete model outlined in Section 3 against previous alternatives for the FNC-1 task, specifically focusing on previous studies that used the same training/testing splits. We also performed an ablation study, removing some of the components from the full model and seeing how that affects the performance.

Our deep neural network relied on representations for the word tokens based on pre-trained GloVe embeddings with 300 dimensions [20]. Out-of-vocabulary words in the testing dataset were represented by the GloVe embedding of their most similar word, as given by the Jaro-Winkler similarity metric [51].

Table 2. Accuracy on the natural language inference task, over the SNLI and MultiNLI testing datasets.

Method	SNLI	MultiNLI	
		Matched	Mismatched
Most Frequent Class	34.3	36.5	35.6
Continuous BOW (Averaging Word Embeddings)	75.3	65.2	64.6
BiLSTM	83.3	67.5	67.1
Enhanced Sequential Inference Model [7]	<b>88.0</b>	72.4	71.9
Nie and Bansal [31]	86.1	74.6	73.6
Chen et al. [8]	85.5	74.9	74.9
Conneau et al. [12]	85.0	–	–
Densely Interactive Inference Network [16]	<b>88.0</b>	<b>79.2</b>	<b>79.1</b>
Directional Self-Attention Encoders [38]	85.6	71.0	71.4
Compare-Propagate Alignment-Factorized Encoders [42]	85.9	78.7	77.9
Gumbel TreeLSTM Encoders [10]	86.0	–	–
Reinforced Self-Attention Network [39]	86.3	–	–
Distance-Based Self-Attention Network [18]	86.3	74.1	72.9
Bi-GRU sentence encoder + attention	83.2	69.4	69.5
Bi-LSTM sentence encoder + attention	83.5	69.9	70.4
Bi-GRU sentence encoder + max-pooling	81.6	66.5	67.3
Bi-LSTM sentence encoder + max-pooling	84.5	70.7	71.1
Bi-GRU sentence encoder + max-pooling + attention	82.8	67.2	68.2
Bi-LSTM sentence encoder + max-pooling + attention	84.1	70.8	70.8
Stacked Bi-GRUs + shortcut connections + max-pooling	84.0	68.5	68.9
Stacked Bi-LSTMs + shortcut connections + max-pooling	84.8	71.4	72.2
Stacked Bi-GRUs + shortcut connections + max-pooling + attention	83.2	68.9	68.4
Stacked Bi-LSTMs + shortcut connections + max-pooling + attention	84.4	70.7	70.5

The model was implemented with the keras<sup>7</sup> deep learning framework, and the corresponding source code is available on GitHub<sup>8</sup>. Given the implementations for RNNs available within keras, every sentence (i.e., the headlines, the premise and hypothesis sentences from the NLI datasets, and the sentences within the body of news articles) was either zero-padded or truncated to have 50 tokens. Every news article body was also zero-padded or truncated to have to 30 sentences. We used GRUs and LSTMs with hidden states of 300 dimensions (i.e., the bi-directional RNN representations have 600 dimensions), and the feed-forward layers before the final softmax layer are composed of 600 and 300 neurons, respectively.

We trained the neural models leveraging the Adam [22] optimizer with default parameters, for a maximum of 100 epochs. An early stopping was activated in case the training loss does not decrease for 2 consecutive epochs.

### 4.3 The Obtained Results

The complete model described in Section 3 involved pre-training the branch that matches the headline against the first two sentences of the body (and also the common sentence encoder used in the remaining branches) with data from the SNLI and MultiNLI corpora. We therefore evaluated the performance of this sub-component against previous alternatives for Natural Language Inference (NLI). The results are given in Table 2, presenting

<sup>7</sup><http://keras.io/>

<sup>8</sup><http://github.com/LuisPB7/fnc-msc>

Table 3. Results obtained over the testing split of the FNC-1 dataset.

Method	Weighted Accuracy	Per-Class Accuracy			
		Unrelated	Discuss	Agree	Disagree
FNC-1 baseline <sup>9</sup>	75.20	97.97	79.65	9.09	1.00
Baseline based on word2vec + hand-crafted features [2]	72.78	96.05	53.38	50.70	<b>9.61</b>
Baseline based on skip-thought embeddings [2]	76.18	91.18	81.20	31.80	0.00
Baseline based on TF-IDF vectors [2]	81.72	97.90	81.38	44.04	6.60
Best set of features from Tosik et al. [45]	78.63	97.98	<b>90.95</b>	1.42	0.00
Neural baseline based on bi-directional LSTMs [2]	63.11	78.27	58.13	38.04	4.59
Neural method from Mohtarami et al. [30]	78.97	–	–	–	–
Neural method from Mohtarami et al. + TF-IDF [30]	81.23	–	–	–	–
3 <sup>rd</sup> place at FNC-1 – Team UCL Machine Reading [37]	81.72	97.90	81.38	44.04	6.60
2 <sup>nd</sup> place at FNC-1 – Team Athene <sup>10</sup>	81.97	<b>99.25</b>	80.89	44.72	9.47
1 <sup>st</sup> place at FNC-1 – Team SOLAT in the SWEN <sup>11</sup>	82.02	98.70	76.18	58.50	1.86
Previous state-of-the-art – Bhatt et al. [2]	<b>83.08</b>	98.04	85.68	43.82	6.31
MLP with the considered external features	81.95	97.86	77.93	52.55	2.87
Bi-LSTM + max-pooling	81.29	96.99	80.76	45.35	5.16
Bi-GRU + max-pooling + attention	80.76	97.36	76.08	52.08	4.16
Bi-LSTM + max-pooling + attention ( <i>best encoder</i> )	82.23	96.74	81.52	51.34	10.33
Stacked bi-LSTMs + shortcuts + max-pooling	82.16	96.13	79.39	56.59	11.91
Stacked bi-GRUs + shortcuts + max-pooling + attention	81.95	96.22	77.71	52.86	<b>19.08</b>
Stacked bi-LSTMs + shortcuts + max-pooling + attention	81.16	95.25	73.32	<b>66.47</b>	6.74

accuracy scores over the SNLI testing dataset, and over the *matched* (i.e., including news subjects that also appear in the training dataset) and *mismatched* portions of the MultiNLI testing dataset.

Besides the complete sentence encoder described in Section 3.1, considering a stacked arrangement of bi-GRUs or bi-LSTMs with shortcut connections, we also evaluated variants leveraging (a) a single bi-directional RNN layer, or (b) using only max-pooling or neural attention as the summarization method, instead of concatenating the results of max-pooling and attention. The best results on the SNLI dataset were obtained with a single bi-directional LSTM combined with max-pooling, whereas for the MultiNLI dataset the best results were achieved with a single bi-LSTM together with a combination of max-pooling and neural attention. Bi-directional GRUs achieved consistently worse results, and the combination of max-pooling with neural attention achieved almost similar results to the usage of max-pooling alone (i.e., better results when leveraging GRUs, and in some cases neural attention lead to worse results when leveraging LSTMs).

Although we outperform simpler NLI baselines, our results are still far from the current state-of-the-art. However, it should be noted that our results are not directly comparable to those of previous systems, given that we trained the NLI model on a combination of the training data from the SNLI and MultiNLI corpora. Moreover, we did not optimize results for the NLI task, instead being more concerned with pre-training the network for building effective representations of textual inputs. It is perhaps the case that a more sophisticated matching method would improve the results for the NLI task [16], but in our pre-training scenario it is more important to have a good performance on the part of the network that encodes the semantics of the input sentences [12].

Table 3 presents results obtained over the FNC-1 testing dataset, including results from (a) previous baseline methods leveraging feature engineering, (b) baseline methods using simpler neural networks, (c) the best

Table 4. Experimental results with different variations of the proposed method, over the testing split of the FNC-1 dataset.

Method	Weighted Accuracy	Per-Class Accuracy			
		Unrelated	Discuss	Agree	Disagree
Best encoder with the complete architecture	82.23	96.74	<b>81.52</b>	51.34	10.33
Best encoder - pre-training	81.85	<b>98.26</b>	73.61	63.16	10.19
Best encoder - matching first sentences	<b>83.38</b>	97.27	80.60	59.64	13.06
Best encoder - external features	75.31	92.66	72.20	43.72	4.73
Best encoder - pre-training - first sentences	82.06	97.38	67.20	<b>74.62</b>	<b>22.53</b>
Best encoder - pre-training - first sentences - features	76.08	86.23	72.65	61.90	15.49

submissions to the FNC-1 challenge, (d) the previous state-of-the-art on the FNC-1 testing dataset, (e) a Multi-Layer Perceptron (MLP) leveraging the complete set of external features outlined in Section 3.4, and (f) the proposed neural architecture, considering all the components and model pre-training with the NLI datasets. Each group in Table 3 corresponds to one of the previous six items, with the last group presenting the results obtained with different sentence encoders, i.e. considering only one bi-RNN or a stack of two bi-RNNs with shortcut connections, and a combination of max-pooling and neural attention. The best weighted accuracy was achieved with a sentence encoder leveraging a single bi-LSTM, similarly to what was obtained in the case of the NLI datasets. The complete neural model, leveraging the best sentence encoder, outperformed the baseline method based on a MLP with the considered set of external features (which also achieved competitive results, outperforming the third best team at the FNC-1 competition and performing similarly to the more complex neural network architecture from Mohtarami et al. [30], in terms of the weighted accuracy metric).

We also evaluated variations of the complete method introduced in Section 3, using the best sentence encoder but without considering (a) model pre-training with the NLI datasets, (b) external features complementing the neural representations, (c) the matching between the headline and the first two sentences of the article, and (d) combinations of the three components above. Table 4 displays the results for this set of ablation tests.

Our full model without the branch matching the headline against the first two sentences of the body slightly outperforms the previous state-of-the-art in the FNC-1 test split, obtaining a weighted accuracy score of 83.38%. The different ablation tests in which this branch has been removed seem to indicate that the matching against the first two sentences does not provide useful/additional information to the FNC-1 task. The results also show that both model pre-training and the external features are essential to achieving good results with the proposed neural network architecture, confirming previous studies that have shown that the FNC-1 task is challenging for methods based exclusively on representations built through deep neural networks.

When the weights of the network are not initialized with the weights from the NLI task (i.e., when there is no model pre-training), the weighted accuracy drops to 81.85% (or 82.02% if the matching between the headline and the first two sentences is also removed). If the external features are instead removed, the weighted accuracy drops to 75.31%, which is better than a baseline method leveraging a bi-directional LSTM [2], but only marginally better than the official baseline released by the organizers of FNC-1, and much worse than the winning entries at the challenge, or the previous state-of-the-art by Bhatt et al. [2].

Besides evaluating model quality in terms of predictive accuracy, we also attempted to see if the neural attention weights associated to the sentence encoder, and to the hierarchical encoder for the news article body, could provide clues as to the rationale behind particular classification decisions. Figure 4 illustrates the attention weights computed by the neural network for two example instances from the FNC-1 testing dataset, with darker colors corresponding to the words/sentences to which the model assigned a higher weight. The two example instances shown in Figure 4 were correctly classified by our model (i.e., the best approach from Table 4), with

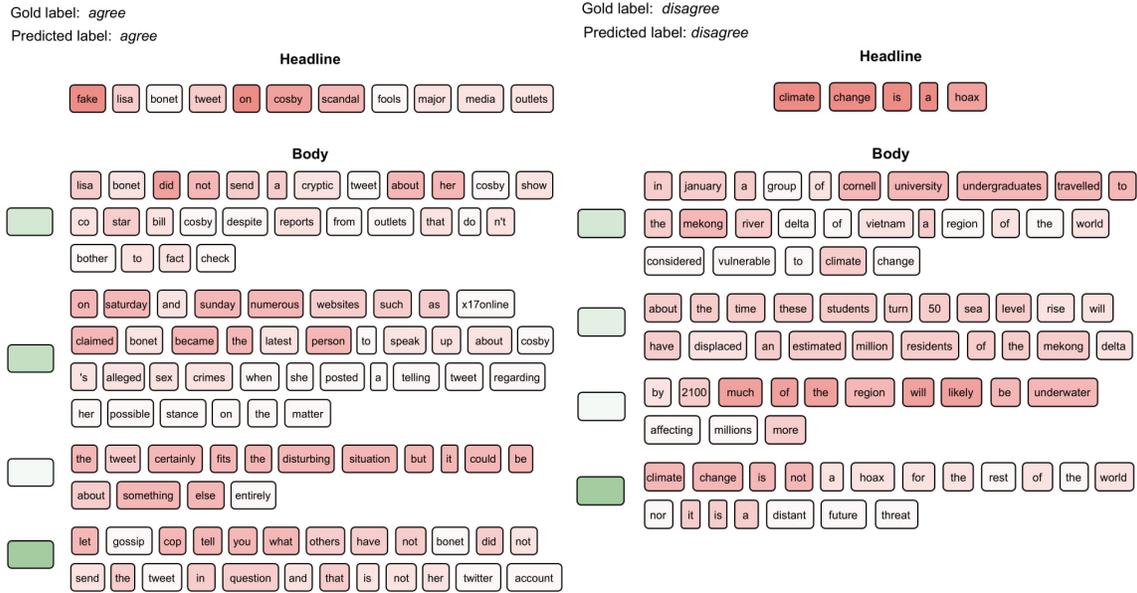


Fig. 4. Attention weights computed with the complete model, for two example instances from the FNC-1 test split.

one of them belonging to the *agree* class and the other to the *disagree* class. In both examples, the headline and the body of the news article contain words that can be highly indicative of the stance (e.g., words such as *fake*, *fraud* or *hoax*, present in the lexicons used to compute the refutation and polarity features).

The headline for the *agree* example refers to a fake tweet, with the word *fake* having a high weight. Some of the sentences in the body confirmed that the supposed author did not send the tweet, with words/expressions such as *did not send* or *claimed* also having a high weight. On what regards the *disagree* example, its headline

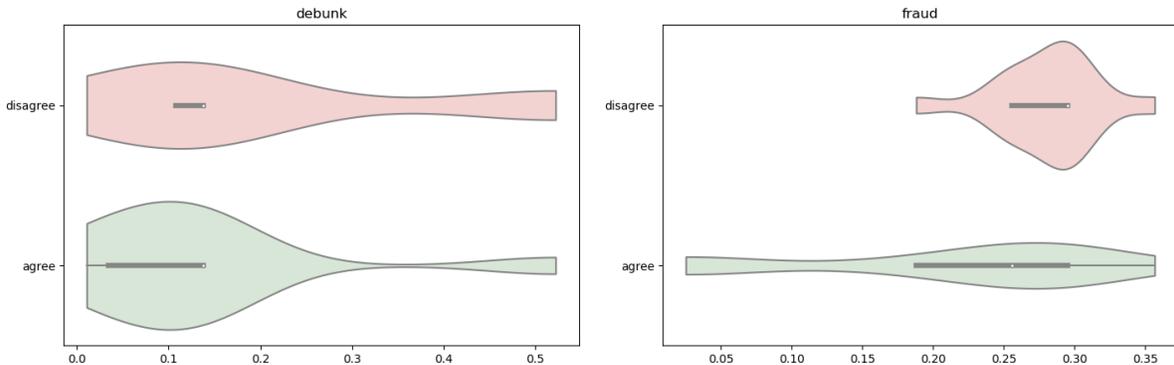


Fig. 5. Distribution for the neural attention weights associated to two distinct refutation words, within FCN-1 testing instances belonging to the *disagrees* versus the *agrees* class.

simply states *climate change is a hoax*, while the last sentence of the corresponding news article (i.e., the sentence with the highest weight) denies the claim from the headline, containing the words *climate change is not a hoax*.

Figure 5 further illustrates the weights computed for the neural attention layers in association to specific words, specifically by presenting 2 violin plots with the distribution for the attention weights associated to the words *fraud* and *debunk* (i.e., words that belong to the FNC-1 refutation dictionary and that also appear frequently in the test instances). The plots contrast normalized values for the attention weights (i.e., we normalized the attention scores associated to words in the headline and in the body text, using the min-max normalization scheme together with the minimum and maximum attention weights that are seen in these two parts of the input) that are estimated by the classifier when analyzing test instances belonging to the *disagrees* versus the *agrees* category. As expected, the weights associated to aforementioned words tend to be higher in the case of documents belonging to the *disagrees* class, thus confirming that the attention layers are indeed providing interesting information that can support model interpretability and the design of user interfaces for manually inspecting the classification results.

In brief, the obtained results with the FNC-1 dataset confirm the importance of using external similarity features for tasks involving the modeling/matching of long pieces of text, and also that model pre-training can be a simple, yet effective, way of improving the representations build through deep neural networks.

## 5 CONCLUSIONS AND FUTURE WORK

This article presented a deep learning method for addressing the stance detection problem from the Fake News Challenge (FNC-1), leveraging bi-directional RNNs together with max-pooling and neural attention mechanisms for building representations from headlines and from the body of news articles, and combining these representations with external similarity features. We also explored the use of external sources of information (e.g., large datasets proposed for the natural language inference task) in order to pre-train specific components of the neural network architecture (e.g., the RNNs used for encoding sentences).

The obtained results show that our model, particularly when considering pre-training and the combination of neural representations together with external similarity features, slightly outperforms the previous state-of-the-art. Most of the previous methods proposed for the FNC-1 task leveraged careful feature engineering, instead of representation learning through deep neural networks. Our results also confirm the challenges in addressing the FNC-1 task with modern neural approaches, given that the external features were essential to the good performance of our model (i.e., they increased the weighted accuracy score by 6.92 percentage points).

Our experiments also confirmed the effectiveness of model pre-training leveraging Natural Language Inference data, confirming previous results by Conneau et al. [12]. When pre-training the sentence encoder with the SNLI and MultiNLI datasets, our model reaches a new state-of-the-art result on the FNC-1 testing dataset.

Despite the interesting results, there are also many possible ideas for future work. For instance, within the context of NLI, previous studies have proposed more advanced methods for modeling sentences [9, 43, 44, 50], and/or for matching representations for the premise and hypothesis sentences [16, 43, 50]. Examples include methods that leverage discourse markers (i.e., words used to link two clauses in a sentence) through transfer-learning approaches [32], methods for modeling the interactions between the sentence representations as a tensor, afterwards applying a CNN over this tensor [16], or methods that create representations for each sentence that depend on the other sentence, e.g. by initializing the RNN responsible for encoding a sentence with the last hidden state of the RNN responsible for encoding the other sentence [43, 44, 50]. For future work, we plan to experiment with these and other similar ideas for extending the deep neural network used in the FNC-1 task, leveraging more recent developments within the general context of natural language understanding (e.g., methods that, instead of using RNNs for encoding text, consider only feed-forward computations and attention approaches, similar to those from the transformer architecture [5, 13, 46]).

It is also our belief that extensions of the method presented in this article can later find application in the context of tools for fake news detection, going beyond predicting the stance [24, 34, 36]. Moreover, the integration of text analysis methods into practical tools for fake news detection can benefit from mechanisms to help interpreting the predictions, and neural attention mechanisms can be explored in this direction [52]. For future work, we would like to extend the experiments related to the analysis of the attention weights given to specific parts of the headline and/or the body text, for instance considering sparsemax [28] as an alternative to the softmax normalization in the attention weights, in the hope of obtaining a more selective and compact attention focus, facilitating interpretability. The previous study by Mohtarami et al. [30], exploring neural mechanisms for FNC-1, also provided some ideas regarding model interpretability, interesting to explore as future work (i.e., the authors extracted snippets of evidence for the predictions, although not with basis on neural attention mechanisms).

Besides leveraging textual information and the stance of headlines towards longer pieces of text, fake news detection can also benefit from the combination of text mining methods, such as the one advanced in this article, with other types of approaches. For example, some previous studies have noted that fake news stories are often relatively simple (i.e., fake stories are usually *flatter* than ordinary news stories), using exaggeration in a way that makes it easy to detect and follow them [19], or mentioning facts that can easily be disproven through inconsistencies against external sources of information<sup>12</sup>. Fake news stories also have particular geographical diffusion patterns [19], with rapid reproduction and elevated mutation rates (i.e., fake news rise and fall in weeks and often in days). Future work in the area can perhaps consider the combination/extension of text mining methods with other types of external features, capturing different characteristics of news articles.

## ACKNOWLEDGEMENTS

This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the project grant with reference CMUPERI/TIC/0046/2014 (GoLocal), as well as through the INESC-ID multi-annual funding from the PIDDAC programme, which has the reference UID/CEC/50021/2013. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used in the experiments.

## REFERENCES

- [1] Darren Baker, Ali K. Chaudhry, and Philipp Thun-Hohenstein. 2017. Stance Detection for the Fake News Challenge: Identifying Textual Relationships with Deep Neural Nets. *CS224n: Natural Language Processing with Deep Learning* (2017).
- [2] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2017. On the Benefit of Combining Neural, Statistical and External Features for Fake News Identification. *arXiv preprint arXiv:1712.03935* (2017).
- [3] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).
- [6] Delphine Charlet and Geraldine Damnati. 2017. SimBow at SemEval-2017 Task 3: Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the International Workshop on Semantic Evaluation*.
- [7] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [8] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Recurrent Neural Network-Based Sentence Encoder with Gated Attention for Natural Language Inference. In *Proceedings of the Workshop on Evaluating Vector Space Representations for NLP*.
- [9] Jihun Choi, Taek Kim, and Sang goo Lee. 2018. Cell-aware Stacked LSTMs for Modeling Sentences. *arXiv preprint arXiv:1809.02279* (2018).
- [10] J. Choi, K. M. Yoo, and S.-g. Lee. 2017. Learning to Compose Task-Specific Tree Structures. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*.

<sup>12</sup><http://fakenews.publicdatalab.org/>

- [11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the NIPS Workshop on Deep Learning*.
- [12] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [13] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and ÁÅukasz Kaiser. 2018. Universal Transformers. *arXiv preprint arXiv:1807.03819* (2018).
- [14] Francisco Duarte, Bruno Martins, Cátia Sousa Pinto, and Mário J. Silva. 2018. A Deep Learning Method for ICD-10 Coding of Free-Text Death Certificates. *Journal of Biomedical Informatics* 30 (2018).
- [15] Yoav Goldberg. 2016. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research* 57 (2016).
- [16] Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348* (2017).
- [17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997).
- [18] Jinbae Im and Sungzoon Cho. 2017. Distance-based Self-Attention Network for Natural Language Inference. *arXiv preprint arXiv:1712.02047* (2017).
- [19] Krzysztof Janowicz and Grant McKenzie. 2017. How "Alternative" are Alternative Facts? Measuring Statement Coherence via Spatial Analysis. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.
- [20] Richard Socher Jeffrey Pennington and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [21] Kevin Gimpel John Wieting, Mohit Bansal and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In *Proceedings of the International Conference on Learning Representations*.
- [22] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- [23] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the Neural Information Processing Systems Conference*.
- [24] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. In *Proceedings of the EMNLP Workshop on Fact Extraction and Verification*.
- [25] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings to Document Distances. In *Proceedings of the International Conference on Machine Learning*.
- [26] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018).
- [27] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out*.
- [28] Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the International Conference on Machine Learning*.
- [29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*.
- [30] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluis Marquez, and Alessandro Moschitti. 2018. Automatic Stance Detection Using End-to-End Memory Networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.
- [31] Yixin Nie and Mohit Bansal. 2017. Shortcut-Stacked Sentence Encoders for Multi-Domain Inference. In *Proceedings of the Workshop on Evaluating Vector Space Representations for NLP*.
- [32] Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. 2018. Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*.
- [34] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the International Conference on Computational Linguistics*.
- [35] Oskar Triebe Pfohl and Ferdinand Legros. 2017. Stance Detection for the Fake News Challenge with Attention and Conditional Encoding. *CS224n: Natural Language Processing with Deep Learning* (2017).
- [36] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

- [37] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv preprint arXiv:1707.03264* (2017).
- [38] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*.
- [39] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced Self-Attention Network: a Hybrid of Hard and Soft Attention for Sequence Modeling. *arXiv preprint arXiv:1801.10296* (2018).
- [40] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286* (2018).
- [41] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting Tri-Relationship for Fake News Detection. *arXiv preprint arXiv:1712.07709* (2017).
- [42] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. A Compare-Propagate Architecture with Alignment Factorization for Natural Language Inference. *arXiv preprint arXiv:1801.00102* (2017).
- [43] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Co-Stack Residual Affinity Networks with Multi-level Attention Refinement for Matching Text Sequences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [44] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. A Compare-Propagate Architecture with Alignment Factorization for Natural Language Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [45] M. Tosik, A. Mallia, and K. Gangopadhyay. 2018. Debunking Fake News One Feature at a Time. *arXiv preprint arXiv:1808.02831* (2018).
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Neural Information Processing Systems Conference*.
- [47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [48] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018).
- [49] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv preprint arXiv:1704.05426* (2017).
- [50] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2018. DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.
- [51] William E Winkler. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (1990).
- [52] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- [53] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923* (2017).
- [54] Qi Zeng, Quan Zhou, and Shanshan Xu. 2017. Neural Stance Detectors for Fake News Challenge. *CS224n: Natural Language Processing with Deep Learning* (2017).

Received March 2018; revised March 2018; accepted July 2018