# TI-CNN: Convolutional Neural Networks for Fake News Detection

Yang Yang
Beihang University
Beijing, China
Email: yangyangfuture@gmail.com

Lei Zheng
University of Illinois at Chicago
Chicago, United States
Email: lzheng21@uic.edu

Jiawei Zhang
Florida State University
Florida, United States
Email: jzhang@cs.fsu.edu

Qingcai Cui
Beihang University
Beijing, China
Email: cqc@cuiqingcai.com

Xiaoming Zhang
Beihang University
Beijing, China
Email: yolixs@163.com

Zhoujun Li
Beihang University
Beijing, China
Email: lizj@buaa.edu.cn

Philip S. Yu
University of Illinois at Chicago
Chicago, United States
Email: psyu@uic.edu

*Abstract*—With the development of social networks, fake news for various commercial and political purposes has been appearing in large numbers and gotten widespread in the online world. With deceptive words, people can get infected by the fake news very easily and will share them without any fact-checking. For instance, during the 2016 US president election, various kinds of fake news about the candidates widely spread through both official news media and the online social networks. These fake news is usually released to either smear the opponents or support the candidate on their side. The erroneous information in the fake news is usually written to motivate the voters' irrational emotion and enthusiasm. Such kinds of fake news sometimes can bring about devastating effects, and an important goal in improving the credibility of online social networks is to identify the fake news timely. In this paper, we propose to study the "fake news detection" problem. Automatic fake news identification is extremely hard, since pure model based fact-checking for news is still an open problem, and few existing models can be applied to solve the problem. With a thorough investigation of a fake news data, lots of useful explicit features are identified from both the text words and images used in the fake news. Besides the explicit features, there also exist some hidden patterns in the words and images used in fake news, which can be captured with a set of latent features extracted via the multiple convolutional layers in our model. A model named as TI-CNN (Text and Image information based Convolutinal Neural Network) is proposed in this paper. By projecting the explicit and latent features into a unified feature space, TI-CNN is trained with both the text and image information simultaneously. Extensive experiments carried on the real-world fake news datasets have demonstrate the effectiveness of TI-CNN in solving the fake new detection problem.

## I. INTRODUCTION

Fake news is written in an intentional and unverifiable language to mislead readers. It has a long history since the 19th century. In 1835, New York Sun published a series of articles about "the discovery of life on the moon". Soon the fake stories were printed in newspapers in Europe. Similarly, fake news widely exists in our daily life and is becoming more widespread following the Internet's development. Exposed to the fast-food culture, people nowadays can easily believe something without even checking whether the information is correct or not, such as the "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide". These fake news frequently appear during the United States presidential election campaign in 2016. This phenomenon has aroused the attention of people, and it has a significant impact on the election.

Fake news dissemination is very common in social networks [1]. Due to the extensive social connections among users, fake news on certain topics, e.g., politics, celebrities and product promotions, can propagate and lead to a large number of nodes reporting the same (incorrect) observations rapidly in online social networks. According to the statistical results reported by the researchers in Stanford University, 72.3% of the fake news actually originates from the official news media and online social networks [1], [8]. The potential reasons are provided as follows. Firstly, the emergence of social media greatly lower down the barriers to enter in the media industry. Various online blogs, "we media", and virtual communities are becoming more and more popular in recent years, in which everyone can post news articles online. Secondly, the large number of social media users provide a breeding ground for fake news. Fake news involving conspiracy and pitfalls can always attract our attention. People like to share this kind of information to their friends. Thirdly, the 'trust and confidence' in the mass media greatly dropped these years. More and more people tend to trust the fake news by browsing the headlines only without reading the content at all.

Fake news identification from online social media is extremely challenging due to various reasons. Firstly, it's difficult to collect the fake news data, and it is also hard to label fake news manually [43]. News that appears on Facebook and Twitter news feeds belongs to private data. To this context so far, few large-scale fake news detection public dataset really exists. Some news datasets available online involve a small number of the instances only, which are not sufficient to train a generalized model for application. Secondly, fake news is written by human. Most liars tend to use their language strategically to avoid being caught. In spite of the attempt to control

what they are saying, language leakage occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage [10]. Thirdly, the limited data representation of texts is a bottleneck of fake news identification. In the bag-of-words approach, individual words or "n-grams" (multiword) frequencies are aggregated and analyzed to reveal cues of deception. Further tagging of words into respective lexical cues for example, parts of speech or "shallow syntax" [28], affective dimensions [42], and location-based words [32] can all provide frequency sets to reveal linguistic cues of deception [31], [14]. The simplicity of this representation also leads to its biggest shortcoming. In addition to relying exclusively on language, the method relies on isolated n-grams, often divorced from useful context information. Word embedding techniques provide a useful way to represent the meaning of the word. In some circumstances, sentences of different lengths can be represented as a tensor with different dimensions. Traditional models cannot handle the sparse and high order features very well.



(a) Cartoon in fake news.
(b) Altered low-resolution image.

(c) Irrelevant image in fake news.
(d) Low-resolution image.

Fig. 1. The images in fake news: (a) 'FBI Finds Previously Unseen Hillary Clinton Emails On Weiner's Laptop', (b)'BREAKING: Leaked Picture Of Obama Being Dragged Before A Judge In Handcuffs For Wiretapping Trump', (c) 'The Amish Brotherhood have endorsed Donald Trump for president', (d) 'Wikileaks Gives Hillary An Ultimatum: QUIT, Or We Dump Something Life-Destroying'. The news texts of images (c) and (d) are represented in Section VI-A

Though the deceivers make great efforts in polishing fake news to avoid being found, there are some leakages according to our analysis from the text and image aspect respectively. For instance, the lexical diversity and cognition of the deceivers are totally different from the truth teller. Beyond the text information, images in fake news are also different from that in real news. As shown in Fig. I, cartoons, irrelevant images (mismatch of text and image, no face in political news) and altered low-resolution images are frequently observed in fake news. In this paper, we propose a TI-CNN model to consider both text and image information in fake news detection. Beyond the explicit features extracted from the data, as the development of the representative learning, convolutional neural networks are employed to learn the latent features which cannot be captured by the explicit features. Finally, we utilize TI-CNN to combine the explicit and latent features of text and image information into a unified feature space, and then use the learned features to identify the fake news. Hence, the contributions of this paper are summarized as follows:

- We collect a high quality dataset and take in-depth analysis on the text from multiple perspectives.
- Image information is proved to be effective features in identifying the fake news.
- A unified model is proposed to analyze the text and image information using the covolutoinal neural networks.
- The model proposed in this paper is an effective way to recognize fake news from lots of online information.

In the rest of the paper, we first define the problem of fake news identification. Then we introduce the analysis on the fake news data. A unified model is proposed to illustrate how to model the explicit and latent features of text and image information. The details of experiment setup is demonstrated in the experiment part. At last, we compare our model with several popular methods to show the effectiveness of our model.

## II. RELATED WORK

Deception detection is a hot topic in the past few years. Deception information includes scientific fraud, fake news, false tweets etc. Fake news detection is a subtopic in this area. Researchers solve the deception detection problem from two aspects: 1) linguistic approach. 2) network approach.

### A. Linguistic approaches

Mihalcea and Strapparvva 2009 [29] started to use natural language processing techniques to solve this problem. Bing Liu et.al. [19] analyzed fake reviews on Amazon these years based on the sentiment analysis, lexical, content similarity, style similarity and semantic inconsistency to identify the fake reviews. Hai et al. [13] proposed semi-supervised learning method to detect deceptive text on crowdsourced datasets in 2016.

The methods based on word analysis is not enough to identify deception. Many researchers focus on some deeper language structures, such as the syntax tree. In this case, the sentences are represented as a parse tree to describe syntax structure, for example noun and verb phrases, which are in turn rewritten by their syntactic constituent parts [9].

### B. Network-based approaches

Another way to identify the deception is to analyze the network structure and behaviors, which are important complementary features. As the development of knowledge graph, it will be very helpful to check fact based on the relationship among entities. Ciampaglia et al. [6] proposed a new concept 'network effect' variables to derive the probabilities of news. The methods based on the knowledge graph analysis can achieve 61% to 95% accuracy. Another promising research direction is exploiting the social network behavior to identify the deception.

## C. Neural Network based approaches

Deep learning models are widely used in both academic community and industry. In computer vision [25] and speech recognition [12], the state-of-art methods are almost all deep neural networks. In the natural language processing (NLP) area, deep learning models are used to train a model that can represent words as vectors. Then researchers propose many deep learning models based on the word vectors for QA [3] and summarization[21], etc. Convolutional neural networks (CNN) utilize filters to capture the local structures of the image, which performs very well on computer vision tasks. Researchers also find that CNN is effective on many NLP tasks. For instance, semantic parsing [45], sentence modeling [22], and other traditional NLP tasks [7].
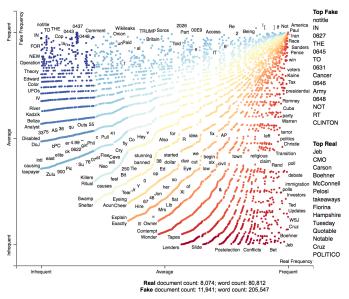


Fig. 2. Word frequency in titles of real and fake news. If the news has no title, we set the title as 'notitle'. The words on the top-left are frequently used in fake news, while the words on the bottom-right are frequently used in real news. The 'Top Fake' words are capital characters and some meaningless numbers that represent special characters, while the 'Top Real' words contain many names and motion verbs, i.e., 'who' and 'what' — the two important factors in the five elements of news: when, where, what, why and who.

## III. PROBLEM DEFINITION

Given a set of $m$ news articles containing the text and image information, we can represent the data as a set of text-image tuples $\mathcal{A} = \{(A_i^T, A_i^I)\}_i^m$. In the fake news detection problem, we want to predict whether the news articles in $\mathcal{A}$ are fake news or not. We can represent the label set as $\mathcal{Y} = \{[1, 0], [0, 1]\}$, where $[1, 0]$ denotes real news while $[0, 1]$ represents the fake news. Meanwhile, based on the news articles, e.g., $(A_i^T, A_i^I) \in \mathcal{A}$, a set of features (including both explicit and latent features to be introduced later in Model Section) can be extracted from both the text and image information available in the article, which can be represented as $\mathbf{X}_i^T$ and $\mathbf{X}_i^I$ respectively. The objective of the *fake news detection* problem is to build a model $f : \{\mathbf{X}_i^T, \mathbf{X}_i^I\}_i^m \in \mathbb{X} \to \mathcal{Y}$ to infer the potential labels of the news articles in $\mathcal{A}$.

## IV. DATA ANALYSIS

To examine the finding from the raw data, a thorough investigation has been carried out to study the text and image information in news articles. There are some differences between real and fake news on American presidential election in 2016. We investigate the text and image information from various perspectives, such as the computational linguistic, sentiment analysis, psychological analysis and other image related features. We show the quantitative information of the data in this section, which are important clues for us to identify fake news from a large amount of data.

### A. Dataset

The dataset in this paper contains 20,015 news, i.e., 11,941 fake news and 8,074 real news. It is available online[1]. For fake news, it contains text and metadata scraped from more than 240 websites by the Megan Risdal on Kaggle[2]. The real news is crawled from the well known authoritative news websites, i.e., the New York Times, Washington Post, etc. The dataset contains multiple information, such as the title, text, image, author and website. To reveal the intrinsic differences between real and fake news, we solely use the title, text and image information.

### B. Text Analysis

Let's take the word frequency [23] in the titles as an example to demonstrate the differences between real and fake news in Fig. 2. If the news has no title, we set the title as 'notitle'. The frequently observed words in the title of fake news are *notitle, IN, THE, CLINTON* and many meaningless numbers that represent special characters. We can have some interesting findings from the figure. Firstly, much fake news have no titles. These fake news are widely spread as the tweet with a few keywords and hyperlink of the news on social networks. Secondly, there are more capital characters in fake news. The purpose is to draw the readers' attention, while the real news contains less capital letters, which is written in a standard format. Thirdly, the real news contain more detailed descriptions. For example, names (*Jeb Bush, Mitch McConnell*, etc.), and motion verbs (*left, claim, debate and poll*, etc.).

#### 1) Computational Linguistic:

*a) Number of words and sentences:* Although liars have some control over the content of their stories, their underlying state of mind may leak out through the style of language used to tell the story. The same is true for the people who write the fake news. The data presented in the following paragraph provides some insight into the linguistic manifestations of this state of mind [14].

As shown in Fig. 3(a), fake news has fewer words than real news on average. There are 4,360 words on average for real news, while the number is 3,943 for fake news. Besides, the number of words in fake news distributes over a wide range,

---

(a) The number of words in news.

(b) The average number of words in a sentence.

(c) Question mark in news.

(d) Exclamation mark in news.

(e) The exclusive words in news.

(f) The negations in news.

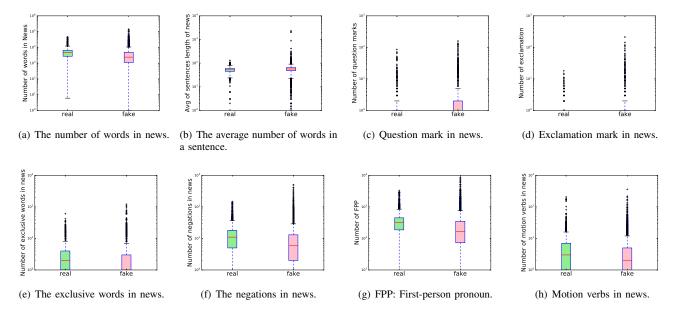(g) FPP: First-person pronoun.

(h) Motion verbs in news.

Fig. 3. Analysis on the news text.

which indicates that some fake news have very few words and some have plenty of words. The number of words is just a simple view to analyze the fake news. Besides, real news has more sentences than fake news on average. Real news has 84 sentences, while fake news has 69 sentences. Based on the above analysis, we can get the average number of words in a sentence for real and fake news, respectively. As shown in Fig. 3(b), the sentence of real news is shorter than that of fake news. Real news has 51.9 words on average in a sentence. However, the number is 57.1 for fake news. According to the box plot, the variance of the real news is much smaller than that of fake news. And this phenomenon appears in almost all the box plots. The reason is that the editor of real news must write the article under certain rules of the press. These rules include the length, word selection, no grammatical errors, etc. It indicates that most of the real news are written in a more standard and consistent way. However, most of the people who write fake news don't have to follow these rules.

*b) Question mark, exclamation and capital letters:* According to the statistics on the news text, real news has fewer question marks than fake news, as shown in Fig. 3(c). The reasons may lie in that there are many rhetorical questions in fake news. These rhetorical questions are always used to emphasize the ideas consciously and intensify the sentiment.

According to the analysis on the data, we find that both real and fake news have very few exclamations. However, the inner fence of fake news box plot is much larger than that of real news, as shown in Fig. 3(d). Exclamation can turn a simple indicative or declarative sentence into a strong command or reflect an emotional outburst. Hence, fake news is inclined to use the words with exclamations to fan specific emotions among the readers.

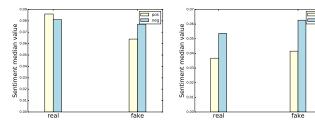Capital letters are also analyzed in the real and fake news.

The reason for the capitalization in news is to draw readers attention or emphasize the idea expressed by the writers. According to the statistic data, fake news have much more capital letters than real news. It indicates that fake news deceivers are good at using the capital letters to attract the attention of readers, draw them to read it and believe it.

*c) Cognitive perspective:* From the cognitive perspective, we investigate the exclusive words (e.g., 'but', 'without', 'however') and negations (e.g.,, 'no', 'not' ) used in the news. Truth tellers use negations more frequently, as shown in Fig. 3(e) and 3(f). The exclusive words in news have the similar phenomenon with the negations. The median of negations in fake news is much smaller than that of real news. The deceiver must be more specific and precise when they use exclusive words and negations, to lower the likelihood that being caught in a contradiction. Hence, they use fewer exclusive words and negations in writing. For the truth teller, they can exactly discuss what happened and what didn't happen in that real news writer witnessed the event and knew all the details of the event. Specifically, individuals who use a higher number of exclusive words are generally healthier than those who do not use these words [35].

*2) Psychology Perspective:* From the psychology perspective, we also investigate the use of first-person pronouns (e.g., I, we, my) in the real and fake news. Deceptive people often use language that minimizes references to themselves. A person who's lying tends not to use "we and "I", and tend not to use person pronouns. Instead of saying "I didnt take your book," a liar might say "That's not the kind of thing that anyone with integrity would do" [31]. Similarly, as shown in Fig. 3(g), the result is the same with the point of view from the psychology perspective. On average, fake news has fewer first-person pronouns. The second-person pronouns

(e.g., you, yours) and third-person pronouns (e.g., he, she, it) are also tallied up. We find that deceptive information can be characterized by the use of fewer first-person, fewer second-person and more third-person pronouns. Given space limitationswe just show the first-person pronouns figure. In addition, the deceivers avoid discussing the details of the news event. Hence, they use few motion verbs, as shown in Fig. 3(h).

*3) Lexical Diversity:* Lexical diversity is a measure of how many different words that are used in a text, while lexical density provides a measure of the proportion of lexical items (i.e. nouns, verbs, adjectives and some adverbs) in the text. The rich news has more diversity. According to the experimental results, the lexical diversity of real news is $2.2e$-06, which is larger than $1.76e$-06 for fake news.

*4) Sentiment Analysis:* The sentiment [26] in the real and fake news is totally different. For real news, they are more positive than negative ones. The reason is that deceivers may feel guilty or they are not confident to the topic. Under the tension and guilt, the deceivers may have more negative emotion [28], [35]. The experimental results agree with the above analysis in Fig. 4. The standard deviation of fake news on negative sentiment is also larger than that of real news, which indicates that some of the fake news have very strong negative sentiment.



(a) The median sentiment values: positive and negative.

(b) The standard deviation sentiment values: positive and negative.

Fig. 4. Sentiment analysis on real and fake news.

### C. Image Analysis

We also analyze the properties of images in the political news. According to some observations on the images in the fake news, we find that there are more faces in the real news. Some fake news have irrelevant images, such as animals and scenes. The experiment result is consistent with the above analysis. There are 0.366 faces on average in real news, while the number is 0.299 in fake news. In addition, real news has a better resolution image than fake news. The real news has $457 \times 277$ pixels on average, while the fake news has a resolution of $355 \times 228$.

## V. MODEL – THE ARCHITECTURE

In this section, we introduce the architecture of TI-CNN model in detail. Besides the explicit features, we innovatively utilize two parallel CNNs to extract latent features from both textual and visual information. And then explicit and latent features are projected into the same feature space to form new

representations of texts and images. At last, we propose to fuse textual and visual representations together for fake news detection.

As shown in Fig. 5, the overall model contains two major branches, i.e., text branch and image branch. For each branch, taking textual or visual data as inputs, explicit and latent features are extracted for final predictions. To demonstrate the theory of constructing the TI-CNN, we introduce the model by answering the following questions: 1) How to extract the latent features from text? 2) How to combine the explicit and latent features? 3) How to deal with the text and image features together? 4) How to design the model with fewer parameters? 5) How to train and accelerate the training process?

TABLE I
SYMBOLS IN THIS PAPER.

| Parameter | Parameter Name | Dimension |
|---|---|---|
| $\mathbf{X}_{i,j}^{Tl}$ | latent word vector $j$ in sample $i$ | $\mathbb{R}^k$ |
| $\mathbf{X}_{i,1:n}^{Tl}$ | sentence for sample $i$ | $\mathbb{R}^{n \times k}$ |
| $\mathbf{X}_i^{Te}$ | explicit text feature for sample $i$ | $\mathbb{R}^k$ |
| $\mathbf{X}_i^{Ie}$ | explicit image feature for sample $i$ | $\mathbb{R}^k$ |
| $\mathbf{X}_i^{Il}$ | latent image feature for sample $i$ | $\mathbb{R}^k$ |
| $\theta$ | weight for the word | $\mathbb{R}^{h \times k}$ |
| $\mathbb{Y}$ | label of news | $\mathbb{R}^{n \times 2}$ |
| $w$ | filter for texts | $\mathbb{R}^{h \times k}$ |
| $b$ | bias | $\mathbb{R}$ |
| $\mathbf{c}$ | feature map | $\mathbb{R}^{n-h+1}$ |
| $\hat{c}$ | the maximum value in feature map | $\mathbb{R}$ |
| $M$ | number of maps | $\mathbb{R}$ |
| $M_i$ | the i-th filter for images | $\mathbb{R}^{K_\alpha \times K_\beta}$ |
| $\tau$ | the scores in tags of label | $\mathbb{R}$ |
| $T$ | the number of tags in label | $\mathbb{R}$ |
| $s_w(\mathbb{X})_\tau$ | the predicted probability | $\mathbb{R} \in [0,1]$ |

### A. Text Branch

For the text branch, we utilize two types of features: textual explicit features $\mathbf{X}^{Te}$ and textual latent features $\mathbf{X}^{Tl}$. The textual explicit features are derived from the statistics of the news text as we mentioned in the data analysis part, such as the length of the news, the number of sentences, question marks, exclamations and capital letters, etc. The statistics of a single news can be organized as a vector with fixed size. Then the vector is transformed by a fully connected layer to form a textual explicit features.

The textual latent features in the model are based on a variant of CNN. Although CNNs are mainly used in Computer Vision tasks, such as image classification [25] or object recognition [38], CNN also show notable performances in many Natural Language Processing (NLP) tasks [24], [46]. With the convolutional approach, the neural network can produce local features around each word of the adjacent word and then combines them using a max operation to create a fixed-sized word-level embedding, as shown in Fig. 5. Therefore, we employ CNN to model textual latent features for fake news detection. Let the $j$-th word in the news $i$ denote as $\mathbf{x}_{i,j} \in \mathbb{R}^k$, which is a k-dimensional word embedding vector. Suppose the
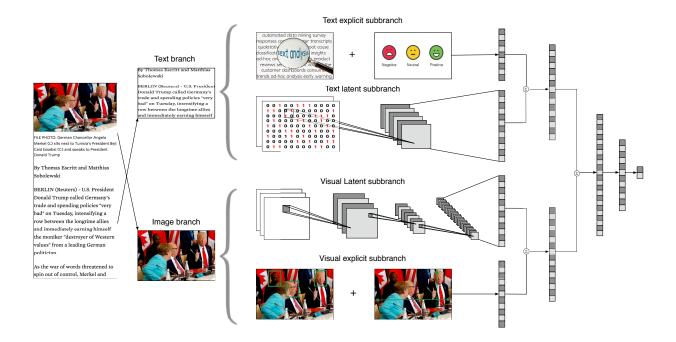
Fig. 5. The architecture of the model. The rectangles in the last 5 layers represent the hidden dense layers. The dropout, batch normalization and flatten layers are not drawn for brevity. The details of the structure are shown in Table III.

maximum length of the news is $n$, s.t., the news have less than $n$ words can be padded as a sequence with length $n$. Hence, the overall news can be written as

$$\mathbf{X}_{i,1:n}^{Tl} = \mathbf{x}_{i,1} \oplus \mathbf{x}_{i,1} \oplus \mathbf{x}_{i,2} \oplus ... \oplus \mathbf{x}_{i,n}. \quad (1)$$

It means that the news $\mathbf{X}_{i,1:n}^{Tl}$ is concatenated by each word. In this case, each news can be represented as a matrix. Then we use convolutional filters $w \in \mathbb{R}^{h \times k}$ to construct the new features. For instance, a window of words $\mathbf{X}_{i,j:j+h-1}^{Tl}$ can produce a feature $c_i$ as follows:

$$c_i = f(\mathbf{w} \cdot \mathbf{X}_{i,j:j+h-1}^{Tl} + b), \quad (2)$$

where the $b \in R$ is the bias, and $\cdot$ is the convolutional operation. $f$ is the non-linear transformation, such as the sigmoid and tagent function. A feature map is generated from the filter by going through all the possible window of words in the news.

$$\mathbf{c} = [c_1, c_2, ..., c_{n-h+1}], \quad (3)$$

where $\mathbf{c} \in \mathbb{R}^{n-h+1}$. A max-pooling layer [30] is applied to take the maximum in the feature map $\mathbf{c}$. The maximum value is denoted as $\hat{c} = max\{\mathbf{c}\}$. The max-pooling layer can greatly improve the robustness of the model by reserving the most important convolutional results for fake news detection. The pooling results are fed into a fully connected layer to obtain our final textual latent features for predicting news labels.

### B. Image Branch

Similar to the text branch, we use two types of features: visual explicit features $\mathbf{X}^{Ie}$ and visual latent features $\mathbf{X}^{Il}$. As shown in Fig. 5, in order to obtain the visual explicit features,

we firstly extract the resolution of an image and the number of faces in the image to form a feature vector. And then, we transform the vector into our visual explicit feature with a fully connected layer.

Although visual explicit features can convey information of images contained in the news, it is hand-crafted features and not data-driven. To directly learn from the raw images contained in the news to derive more powerful features, we employ another CNN to learn from images in the news.

*1) Convolutional layer:* In the convolutional layer, filters are replicated across entire visual field and share the same parameterisation forming a feature map. In this case, the network have a nice property of translation-invariant. Suppose the convolutional layer has $M$ maps of size $(M_\alpha, M_\beta)$. A filter $(K_\alpha, K_\beta)$ is shifted over all the regions of the images. Hence the size of the output map is as follows:

$$M_\alpha^n = M_\alpha^{n-1} - K_\alpha^n + 1 \quad (4)$$

$$M_\beta^n = M_\beta^{n-1} - K_\beta^n + 1 \quad (5)$$

*2) Max-pooling layer:* A max-pooling layer [30] is connected to the convolutional layer. Then we apply maximum activation over the rectangular filters $(K_\alpha, K_\beta)$ to the output of max-pooling layer. Max-pooling enables position invariance over larger local regions and downsamples the input image by a factor of $K_\alpha$ and $K_\beta$ along each direction, which can make the model select invariant features, converge faster and improve the generalization significantly. A theoretical analysis of feature pooling in general and max-pooling in particular is given by [2].

| Title | News text | Type |
|---|---|---|
| The Amish Brotherhood have endorsed Donald Trump for president. | The Amish, who are direct descendants of the protestant reformation sect known as the Anabaptists, have typically stayed out of politics in the past. As a general rule, they don't vote, serve in the military, or engage in any other displays of patriotism. This year, however, the AAB has said that it is imperative that they get involved in the democratic process. | Fake |
| Wikileaks Gives Hillary An Ultimatum: QUIT, Or We Dump Something Life-Destroying | On Sunday, Wikileaks gave Hillary Clinton less than a 24-hour window to drop out of the race or they will dump something that will destroy her completely.Recently, Julian Assange confirmed that WikiLeaks was not working with the Russian government, but in their pursuit of justice, they are obligated to release anything that they can to bring light to a corrupt system and who could possibly be more corrupt than Crooked Hillary? | Fake |

## C. Rectified Linear Neuron

The sigmoid and tanh activation functions may cause the gradient explode or vanishing problem [34] in convolutional neural networks. Hence, we add the ReLU activation to the image branch to remede the problem of gradient vanishing.

$$y = max(0, \sum_{i=1}^{k} x_i \theta_i + b) \tag{6}$$

ReLUs can also improve neural networks by speeding up training. The gradient computation is very simple (either 0 or 1 depending on the sign of $x$). Also, the computational step of a ReLU is easy: any negative elements are set to 0.0 – no exponentials, no multiplication or division operations.

Logistic and hyperbolic tangent networks suffer from the vanishing gradient problem, where the gradient essentially becomes 0 after a certain amount of training (because of the two horizontal asymptotes) and stops all learning in that section of the network. ReLU units are only 0 gradient on one side, which is empirically superior.

## D. Regularization

As shown in Table III, we empoly dropout [40] as well as $l_2$-norms to prevent overfitting. Dropout is to set some of the elements in weight vectors as zero with a probability $p$ of the hidden units during the forward and backward propagation. For instance, we have a dense layer $z = [z_1, ..., z_m]$, and $r$ is a vector where all the elements are zero. When we start to train the model, the dropout is to set some of the elements of $r$ as 1 with probability as $p$. Suppose the output of dense layer is $y$. Then the dropout operation can be formulated as

$$y = \theta \cdot (z \circ r) + b, \tag{7}$$

where $\theta$ is the weight vector. $\circ$ is the element-wise multiplication operator. When we start to test the performance on the test dataset, the deleted neurons are back. The deleted weight are scaled by $p$ such that $\hat{\theta} = p\theta$. The $\hat{\theta}$ is used to predict the test samples. The above procedure is implemented iteratively, which greatly improve the generalization ability of the model. We also use early stopping [36] to avoid overfitting. It can also be considered a type of regularization method (like L1/L2 weight decay and dropout).

## E. Network Training

We train our neural network by minimizing the negative likelihood on the training dataset $D$. To identify the label of a news $\mathbb{X}$, the network with parameter $\theta$ computes a value $s_w(x)_\tau$. Then a sigmoid function is used over all the scores of tags $\tau \in T$ to transform the value into the conditional probability distribution of labels:

$$p(\tau|\mathbb{X}, \theta) = \frac{e^{s_\theta(\mathbb{X})_\tau}}{\sum_{\forall i \in T} e^{s_\theta(\mathbb{X})_i}} \tag{8}$$

The negative log likelihood of Equation 8 is

$$E(W) = -lnp(\tau|\mathbb{X}, \theta) = s_\theta(\mathbb{X})_\tau - log\left(\sum_{\forall i \in T} e^{s_\theta(\mathbb{X})_\tau}\right) \tag{9}$$

We use the RMSprop [16] to minimize the loss function with respect to parameter $\theta$:

$$\theta - > \sum_{(\mathbb{X},\mathbb{Y}) \in D} -log\ p(\mathbb{Y}|\mathbb{X}, \theta) \tag{10}$$

where $\mathbb{X}$ is the input data, and $\mathbb{Y}$ is the label of the news. We naturally choose back-propagation algorithm [15] to compute the gradients of the network structure. With the fine-tuned parameters, the loss converges to a good local minimum in a few epochs.

## VI. EXPERIMENTS

### A. Case study

A case study of the fake news is given in this section. The two fake news in Table II correspond to the Fig. 1(c) and 1(d). The first fake news is an article reporting that 'the American Amish Brotherhood endorsed Donald Trump for President'. However, the website is a fake CNN page. The image in the fake news can be easily searched online, and it is not very relevant with the news texts[3]. For the second fake news – 'Wikileaks gave Hillary Clinton less than a 24-hour window to drop out of the race', it is actually not from Wikileaks. Besides, the composite image [4] in the news is low quality.

---

[3]http://cnn.com.de/news/amish-commit-vote-donald-trump-now-lock-presidency/

[4]http://thelastlineofdefense.org/wikileaks-gives-hillary-an-ultimatum-quit-or-we-dump-something-life-destroying/

| Text Branch | | Image Branch | |
|---|---|---|---|
| Textual Explicit | Textual Latent | Visual Latent | Visual Explicit |
| Input 31×1 | Emb 1000×100 | Input 50×50×3 | Input 4×1 |
| | Dropout $D_\alpha$ | (2×2) Conv(32) | |
| | | ReLU | |
| Dense 128 | Emb 1000×100 | Dropout $D_\beta$ | Dense 128 |
| | (3,3) Conv1D(10) | (2×2) Maxpo | |
| | 2 MaxPo1D | (2×2) Conv(32) | |
| | Flatten | ReLU | |
| BN | Dense 128 | Dropout $D_\beta$ | BN |
| | | (2×2) Maxpo | |
| | BN | (2×2) Conv(32) | |
| | | ReLU | |
| | ReLU | Dropout $D_\beta$ | |
| | | (2×2) Maxpo | |
| | | Flatten | |
| ReLU | Dropout $D_\beta$ | Dense 128 | ReLU |
| | | BN | |
| | | RelU | |
| Merge | | Merge | |
| Merge | | | |
| ReLU | | | |
| Dense 128 | | | |
| BN | | | |
| Sigmoid | | | |

## B. Experimental Setup

We use 80% of the data for training, 10% of the data for validation and 10% of the data for testing. All the experiments are run at least 10 times separately. The textual explicit subbranch and visual explicit subbranch are connected with a dense layer. The parameters in these subbranches can be learned easily by the back-propagation algorithm. Thus, most of the parameters, which need to be tuned, exist in the textual latent subbranch and visual latent subbranch. The parameters are set as follows.

*1) Text branch:* For the textual latent subbranch, the embedding dimension of the word2vec is set to 100. The details of how to select the parameters are demonstrated in the sensitivity analysis section. The context of the word2vec is set to 10 words. The filter size in the convolutional neural network is $(3, 3)$. There are 10 filters in all. Two dropouts are adopted to improve the model's generalization ability. For the textual explicit subbranch, we add a dense layer with 100 neurons first, and then add a batch normalization layer to normalize the activations of the previous layer at each batch, i.e. applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1. The outputs of textual explicit subbranch and textual latent feature subbranch are combined by summing the outputs up.

*2) Image branch:* For the visual latent subbranch, all the images are reshaped as size $(50 \times 50)$. Three convolutional layers are added to the network hierarchically. The filters size is set to $(3, 3)$, and there are 32 filters for each convolutional layer followed by a ReLU activation layer. A maxpooling layer with pool size $(2, 2)$ is connected to each convolutional layer to reduce the probability to be over-fitting. Finally, a flatten, batch normalization and activation layer is added to the model to extract the latent features from the images. For the explicit image feature subbranch, the input of the explicit features is connected to the dense layer with 100 neurons. And then a batch normalization and activation layer are added. The outputs of image convolutional neural network and explicit image feature subbranch are combined by summing the outputs up. We concatenate the outputs of text and image branch. An activation layer and dense layer are transforming the output into two dimensions. The labels of the news are given by the last sigmoid activation layer. In Table III, we show the parameter settings in the TI-CNN model. The total number of parameters is 7,509,980, and the number of trainable parameters is 7,509,176.

## C. Experimental Results

We compare our model with several competitive baseline methods in Table IV. With image information only, the model cannot identify the fake news well. It indicates that image information is insufficient to identify the fake news. With text information, traditional machine learning method — logistic regression [18] is employed to detect the fake news. However, logistic regression fails to identify the fake news using the text information. The reason is that the hyperplane is linear, while the raw data is linearly inseparable. GRU [5] and Long short-term memory [17] with text information are inefficient with very long sequences, and the model with 1000 input length performs worse. Hence, we take the input length 400 as the baseline method. With text and image information, TI-CNN outperforms all the baseline methods significantly.

| Method | Precision | Recall | F1-measure |
|---|---|---|---|
| **CNN-image** | 0.5387 | 0.4215 | 0.4729 |
| **LR-text-1000** | 0.5703 | 0.4114 | 0.4780 |
| **CNN-text-1000** | 0.8722 | 0.9079 | 0.8897 |
| **LSTM-text-400** | 0.9146 | 0.8704 | 0.8920 |
| **GRU-text-400** | 0.8875 | 0.8643 | 0.8758 |
| **TI-CNN-1000** | **0.9220** | **0.9277** | **0.9210** |

## D. Sensitivity Analysis

In this section, we study the effectiveness of several parameters in the proposed model: the word embedding dimensions, batch size, the hidden layer dimensions, the dropout probability and filter size.
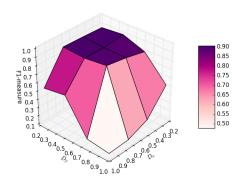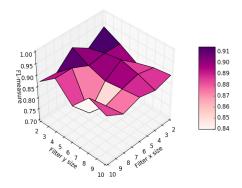
(a) Word embedding dimension and F1-measure.

(b) Batch size and F1-measure.

(c) Hidden layer dimension and F1-measure.

Fig. 6. Word embedding dimension, batch size and the performance of the model.

*a) word embedding dimensions:* In the text branch, we exploit a 3 layer neural network to learn the word embedding. The learned word vector can be defined as a vector with different dimensions, i.e., from 50 to 350. In Fig. 6(a), we plot the relation between the word embedding dimensions and the performance of the model. As shown in figure 6(a), we find that the precision, recall and f1-measure increase as the word embedding dimension goes up from 50 to 100. However, the precision and recall decrease from 100 to 350. The recall of the model is growing all the time with the increase of the word embedding dimension. We select 100 as the word embedding dimension in that the precision, recall and f1-measure are balanced. For fake news detection in real world applications, the model with high recall is also a good choice. The reason is that publishers can use high recall model to collect all the suspected fake news at the beginning, and then the fake news can be identified by manual inspection.

*b) batch size:* Batch size defines the number of samples that going to be propagated through the network. The higher the batch size, the more memory space the program will need. The lower the batch size, the less time the training process will take. The relation between batch size and the performance of the model is shown in Fig. 6(b). The best choice for batch size is 32 and 64. The F1 measure goes up from batch size 8 to 32 first, and then drops when the batch size increases from 32 to 128. For batch size 8, it takes 32 seconds to train the data on each epoch. For batch size 128, it costs more than 10 minutes to train the model on each epoch.

*c) hidden layer dimension:* As shown in Fig. 5, there are many hidden dense layers in the model. Deciding the number of neurons in the hidden layers is a very important part of deciding the overall neural network architecture. Though these layers do not directly interact with the external environment, they have a tremendous influence on the final output. Using too few neurons in the hidden layers will result in underfitting. Using too many neurons in the hidden layers can also result in several problems. Some compromise must be reached between too many and too few neurons in the hidden layers. As shown in Fig. 6(c), we find that 128 is the best choice for hidden layer dimension. The performance firstly goes up with the increase of the hidden layer dimension from 8 to 128. However, the



(a) Dropout probabilities ($D_\alpha, D_\beta$) and the performance of the model.



(b) Filter size and the performance of the model.

Fig. 7. Dropout probabilities ($D_\alpha, D_\beta$), filter size and the performance of the model.

dimension of the hidden layer reaches 256, the performance of the model drops due to overfitting.

*d) Dropout probability and filter size:* We analyze the dropout probabilities, as shown in Table III. $D_\alpha$ in Fig. 7(a) is the dropout layer connected to the text embedding layer, while $D_\beta$ is used in both text and image branches. We use the grid search to choose the dropout probabilities. The model performs

well when the $D_\alpha$ in the range [0.1,0.5] and the $D_\beta$ in range [0.1,0.8]. In this paper, we set the dropout probabilities as (0.5,0.8), which can improve the model's generalization ability and accelerate the training process.

The filter size of a 1-dimension convolutional neural network layer in the textual latent subbranch is also a key factor in identifying the performance of the model. According to the paper [24], the model prefers small filter size for text information. It is consistent with the experimental results in Fig. 7(b). When the filter size is set to (3,3), the F1-measure of the model is 0.92-0.93.

## VII. Conclusions and Future Work

The spread of fake news has raised concerns all over the world recently. These fake political news may have severe consequences. The identification of the fake news grows in importance. In this paper, we propose a unified model, i.e., TI-CNN, which can combine the text and image information with the corresponding explicit and latent features. The proposed model has strong expandability, which can easily absorb other features of news. Besides, the convolutional neural network makes the model to see the entire input at once, and it can be trained much faster than LSTM and many other RNN models. We do experiments on the dataset collected before the presidential election. The experimental results show that the TI-CNN can successfully identify the fake news based on the explicit features and the latent features learned from the convolutional neurons.

The dataset in this paper focuses on the news about American presidential election. We will crawl more data about the France national elections to further investigate the differences between real and fake news in other languages. It's also a promising direction to identify the fake news with much social network information, such as the social network structures and the users' behaviors. In addition, the relevance between headline and news texts is a very interesting research topic, which is useful to identify the fake news. As the development of Generative Adversarial Networks (GAN) [11], [37], the image can generate captions. It provides a novel way to evaluate the relevance between image and news text.

## References

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[2] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[3] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[4] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[7] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[8] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[9] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

[10] Vanessa Wei Feng and Graeme Hirst. Detecting deceptive opinions with profile compatibility. In *IJCNLP*, pages 338–346, 2013.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[13] Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, Xiao-Li Li, Guangxia Li, and Ant Financial. Deceptive review spam detection via exploiting task relatedness and unlabeled data. EMNLP, 2016.

[14] Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

[15] Robert Hecht-Nielsen et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988.

[16] Geoffrey Hinton, NiRsh Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini–batch gradient descent. 2012.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[19] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[21] Khosrow Kaikhah. Automatic text summarization with neural networks. In *Intelligent Systems, 2004. Proceedings. 2004 2nd International IEEE Conference*, volume 1, pages 40–44. IEEE, 2004.

[22] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[23] Jason S. Kessler. Scattertext: a browser-based tool for visualizing how corpora differ. 2017.

[24] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[27] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[28] David M Markowitz and Jeffrey T Hancock. Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4):435–445, 2016.

[29] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics, 2009.

[30] Jawad Nagi, Frederick Ducatelle, Gianni A Di Caro, Dan Cireşan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks

for vision-based hand gesture recognition. In *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, pages 342–347. IEEE, 2011.

[31] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[32] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *HLT-NAACL*, pages 497–501, 2013.

[33] Zizi Papacharissi and Maria de Fatima Oliveira. Affective news and networked publics: The rhythms of news storytelling on# egypt. *Journal of Communication*, 62(2):266–282, 2012.

[34] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318, 2013.

[35] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[36] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[39] Victoria L Rubin and Tatiana Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917, 2015.

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[41] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[42] Aldert Vrij, Samantha Mann, and Ronald P Fisher. An empirical test of the behaviour analysis interview. *Law and human behavior*, 30(3):329–345, 2006.

[43] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[44] Guangyu Wu, Derek Greene, Barry Smyth, and Pádraig Cunningham. Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*, pages 10–13. ACM, 2010.

[45] Scott Wen-tau Yih, Xiaodong He, and Chris Meek. Semantic parsing for single-relation question answering. 2014.

[46] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.