Automated Identification of Potential Conflict-of-Interest in Biomedical Articles Using Hybrid Deep Neural Network

Incheol Kim(✉) and George R. Thoma

Lister Hill National Center for Biomedical Communications,
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
{ickim, gthoma}@mail.nih.gov

Abstract. Conflicts-of-interest (COI) in biomedical research may cause ethical risks, including pro-industry conclusions, restrictions on the behavior of investigators, and the use of biased study designs. To ensure the impartiality and objectivity in research, many journal publishers require authors to provide a COI statement within the body text of their articles at the time of peer-review and publication. However, author's self-reported COI disclosure often does not explicitly appear in their article, and may not be very accurate or reliable. In this study, we present a two-stage machine learning scheme using a hybrid deep learning neural network (HDNN) that combines a multi-channel convolutional neural network (CNN) and a feed-forward neural network (FNN), to automatically identify a potential COI in online biomedical articles. HDNN is designed to simultaneously learn a syntactic and semantic representation of text, relationships between neighboring words in a sentence, and handcrafted input features, and achieves a better performance overall (accuracy exceeding 96.8%) than other classifiers such as support vector machine (SVM), single/multi-channel CNNs, Long Short-term Memory (LSTM), and an Ensemble model in a series of classification experiments.

Keywords: Conflict-of-interest · Two-stage machine learning
Hybrid deep neural network · MEDLINE®

1 Introduction

Conflict of interest (COI) is defined as a situation where a primary interest will be compromised or unduly influenced by a secondary interest. From the biomedical field point of view, primary interests represent health of patients, integrity of research, or duties of public office. A secondary interest generally includes a financial gain for the author (or author's spouse or dependents) received from, or personal relationship with, individuals or "for-profit" organizations such as pharmaceutical companies. Financial conflict-of-interest (FCOI) in biomedical research may cause a number of potential ethical risks, including an increased possibility of pro-industry conclusions, restrictions on the behavior of the investigators, and the use of biased study designs.

MEDLINE®, the U.S. National Library of Medicine (NLM)'s premier online bibliographic database containing more than 25 million citations and abstracts from

P. Perner (Ed.): MLDM 2018, LNAI 10934, pp. 99–112, 2018.
https://doi.org/10.1007/978-3-319-96136-1_9

over 5,600 biomedical journals published in the United States and in other countries, recently announced that it will add COI information to article abstracts available through PubMed [1] when COI declaration statements are supplied by the publishers, to allow users to judge the credibility of findings in published articles. Many biomedical journal publishers also require authors to provide a COI statement within the body text of their articles at the time of peer-review and publication, thereby letting reviewers and readers easily know the integrity of research. However, author's self-reported COI disclosure often does not explicitly appear in their article, and may not be very accurate or reliable due to the lack of author's understanding of relatedness between a certain financial gain they received and their current research. Moreover, there have been no means or systems to verify the accuracy of such authors' COI disclosure.

In this paper, we present an automated method for identifying a potential COI from online biomedical articles using a deep learning-based text classification technique. Our idea is to identify a sentence called COI sentence that contains information of funding support from "for-profit" organizations from the body text of a given biomedical article. This task is quite challenging due to the wide range of linguistic expressions and writing styles, and especially similar expressions for "non-profit" funding sources and a personal acknowledgment.

In order to tackle such challenges, we designed and developed a two-stage machine learning scheme. In stage 1, we distinguish all "support" sentences containing information on any financial support authors received for their research from the body text of an article. In stage 2, these "support" sentences are then classified into two classes according to their funding sources: "for-profit" and "non-profit". Our two-stage machine learning scheme is implemented using a hybrid deep neural network (HDNN) built on combining a multi-channel convolutional neural network (CNN) and a feed-forward neural network (FNN). The CNN component in the proposed HDNN is responsible for learning a syntactic and semantic representation of a text, and contextual relationships between neighboring words in a sentence, while the FNN section takes care of handcrafted input features.

We evaluated the proposed HDNN by comparing its classification performance with that of other types of classifiers such as support vector machine (SVM) with a radial basis kernel function (RBF), single/multi-channel CNNs, Long Short-term Memory (LSTM), voting scheme, and Ensemble model. Three types of word vectors (embeddings): two dense and distributed representations known as Word2Vec [2] and GloVe [3] and a dictionary-based sparse and discrete representation of words, are employed to convert an input sentence into two-dimensional input vector representation and to build an embedding layer for the CNNs. In addition, a bag of words (BOW) based on unigram word statistics representing how differently a word is distributed in "support" and other sentence classes is also used as an input feature for the SVM and the FNN section in HDNN.

2 Related Works

Identifying a sentence that suggests “for-profit” or “non-profit” funding support to determine a potential COI belongs to a text classification or categorization task, a popular topic in the field of natural language processing (NLP). Automated text classification is the process of automatically assigning one or more of a set of predefined categories to a given text or document based on its content, and has been addressed by various methods based on statistical theories and machine learning techniques such as Naïve Bayes [4], decision tree [5], and SVMs [6]. In recent years, deep learning techniques [7] have set a new breakthrough trend in machine learning due to the remarkable success in tackling complex learning problems, and ease of access to high performance computing resources and state-of-the-art open source libraries.

CNN has emerged as the most commonly and widely used architecture in deep learning. It was originally developed for computer vision-related tasks but has been also shown to be very effective for various text classification and understanding tasks such as sentiment analysis and question-answering [8–10]. CNN can learn syntactic and semantic representations of a text, and capture relationships between neighboring words in a sentence automatically through convolution and pooling operations for a sequence of 1-dimensional word or character embeddings. A simple CNN architecture achieves very strong results [11], and can therefore serve as a drop-in replacement for the abovementioned conventional machine learning methods.

Recurrent neural network (RNN) is another popular deep learning architecture especially for NLP tasks. Unlike CNN where the architecture is hierarchical and zero padding or rip out is required to make a fixed-sized word (or character) sequence, RNN is sequential and able to naturally handle word sequences of any length. Other variants such as LSTM [12, 13] were also designed to avoid the problem of exploding or vanishing gradients in the standard RNN and to better capture long-term dependencies.

More recently, an ensemble approach [14] which combines different types of multiple pre-trained classifiers has been proposed to achieve better performance by compensating for errors from individual classifiers. Our proposed HDNN also combines multiple neural network (NN) architectures: a multi-channel CNN section designed to employ different types of word embeddings, and a conventional FNN section for high-performing handcrafted input features. However, our approach totally differs in that each individual NN section in the HDNN is not pre-trained and tightly combined during a learning phase, through the full connection between hidden layers and output layer. We demonstrate the effectiveness of our method by comparing it with the other abovementioned deep neural models, as well as other conventional machine learning techniques such as SVM and voting scheme, with respect to their accuracy in identifying potential COI information from biomedical articles.

3 Conflict-of-Interest: Issues and Challenges

There is an increased concern about the impact of financial relationships between biomedical researchers or their institutions and the pharmaceutical industry on the integrity of biomedical research. Thus, the National Institutes of Health (NIH), as the nation's biomedical research agency, has strict regulations regarding FCOI to ensure impartiality and objectivity in the research it funds [15]. According to NIH regulations, FCOI may exist if investigators or their spouse and dependents received financial support such as "consulting fee", "honoraria", "travel cost", and "royalty" from the third-party private companies. Financial support from "non-profit" organizations such as a local or federal government agency is not considered as a potential COI. Typical examples of author's self-declared COI statements are shown in Table 1.

Table 1. Examples of author's self-declared COI statements.

Supports	Author's self-declared COI statements
Consulting fee	Dr. Hodi reports receiving consulting fees from Bristol-Myers Squibb-Medarex, Novartis, and Genentech; Dr. O'Day, receiving consulting fees, grants, honoraria, and fees for participation in speakers' bureaus from Bristol-Myers Squibb.
Patent	Dr. NL Saccone is the spouse of Dr. SF Saccone, who is also listed as an inventor on the above patent.
Stock	Diane Warden, Ph.D., M.B.A. has owned stock in Pfizer, Inc. within the last five years.
Royalty & travel cost	JAS receives licensing royalties from Genzyme/Sanofi for eliglustat tartrate (Cerdelga) and related compounds. BES has received travel support from Shire HGT and Genzyme.

Table 2. Examples of author's COI disclosure (bold and italicized) in the acknowledgment section in biomedical articles.

We thank the staff at the Metabolic Research Unit at the Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, and at the Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine for their work on the study. LC was supported by grant DK007651. ***This research was supported by the Unilever Corporate Research, Bedfordshire, UK***.
The study was supported by NIMH R01MH070437 and K24MH075867. Dr. Druss had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis***. Dr. Druss has received funding in the past from Pfizer for a research study.***
This project was funded by the American Academy of Child and Adolescent Psychiatry Physician Scientist Program in Substance Abuse K12 Award (DA 000357-06AK12) and National Institute on Drug Abuse grants U10 DA013732, DA012845 and 5R01DA022284. ***Medication and matching placebo were supplied by Eli Lilly.***
The authors thank the Sansum Diabetes Research Center and the University of California, Santa Barbara, which provided the Artificial Pancreas Software developed by Dr. Eyal Dassau and colleagues***. Product support from both Insulet, for the OmniPod Insulin Management System, and DexCom, for the STS-Seven System and sensors, is acknowledged.***

Currently, biomedical journal publishers rely on an author's self-reported disclosure in determining the existence of potential COI. However, such a disclosure may not be very accurate or reliable due to the lack of author's understanding of relationship between a certain financial gain they received and their current research. Authors also often do not provide a separate section or paragraph for the explicit COI disclosure in their article. Instead, COI statements or sentences appear implicitly and in a subtle manner at the end of the body text or within the acknowledgment, footnote, or appendix section along with other similar sentences that acknowledge a personal support or a funding support received from the "non-profit" organizations as shown in Table 2. Moreover, the wide variety of author's linguistic expressions and writing styles makes the problem of identifying COI information even more challenging. Furthermore, identifying COI manually is time-consuming and labor intensive. To our knowledge, there are no automated systems to verify the accuracy of such authors' COI disclosure.

4 Proposed Method

Our strategy is to identify COI sentences containing a "for-profit" funding support by adopting a two-stage machine learning scheme as shown in Fig. 1. First, an input text which is usually the body text of a given article is divided into sentences in the preprocessing step. Next, all "support" sentences containing information of any financial supports authors received for their research are determined in the first classification stage. Finally, these "support" sentences are classified into two classes according to their funding sources: "for-profit" or "non-profit" in the second stage.

Actually, this task could have been considered a three-class classification problem as sentences belonging to (1) "for-profit", (2) "non-profit", or (3) Others. However, support sentences, especially "for-profit" support sentences, are much rarer compared to "other" sentences within the body text of biomedical articles, thereby heavily skewing the distribution of sentences in each class. Thus, we choose to employ a machine learning scheme having two separate and sequential classification stages. Each stage of classification is performed using a HDNN model that combines a multichannel CNN and a conventional FNN.

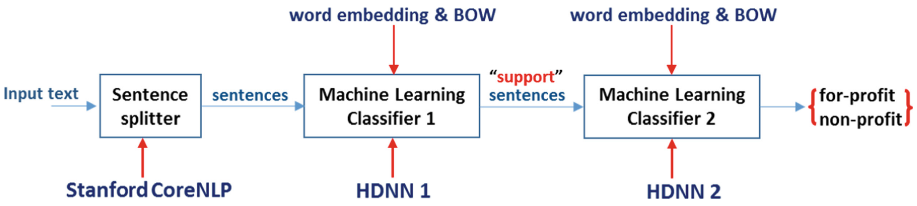

Fig. 1. The proposed two-stage machine learning scheme to identify potential COI.

4.1 Preprocessing: Sentence Splitting

Splitting the text in the body of a biomedical article into individual sentences is an important preprocessing step for the next classification step. Generally, the text in Acknowledgment, Footnote, and Note sections where a “support” sentence is most frequently located, are found to have a more complicated structure than those in the body text. This can be seen from the examples in Table 3, where acronyms or abbreviations for author names, organizations, and degree titles are commonly found in the text. Moreover, the text in these sections often consists of irregular or incomplete sentences. Therefore, it would be not easy to extract sentences using some simple rules based on delimiters such as punctuation marks. In order to deal with this problem, we employed Stanford CoreNLP [16], which is a widely used integrated NLP toolkit including Part-of-speech (POS) tagger, Named entity recognizer (NER), Dependency parser, etc. Its built-in tokenizer has the ability to efficiently and rapidly split sentences.

Table 3. Examples of text containing acronyms or abbreviations for author names, organizations, and degree titles.

Disclosure Summary: J.J.G., K.C., K.A.S., S.B.S.K., E.V.R., and J.M.Z. have nothing to declare. S.M.W.R. consults for Vanda Pharmaceuticals, Inc., through Monash University. C.A.C. has received consulting fees from or served as a paid member of scientific advisory boards for Cephalon, Inc.; Eli Lilly and Co.; Johnson & Johnson; Koninklijke Philips Electronics, N.V./Philips Respironics, Inc.; Sanofi-Aventis Groupe; Sepracor, Inc.; Somnus Therapeutics, Inc.; and Zeo, Inc.
The authors thank the study subjects and their parents for participating in these studies, and the following contributors: Study 1 (M06-888), Edward A. Cherlin, M.D. of Valley Clinical Research, Inc., Andrea Corsino, R.N., M.S.N. of Consultants in Neurology, Ltd., Judith C. Fallon, M.D. of NeuroScience, Inc., David G. Krefetz, D.O., M.D. of CRI Worldwide, Alan J. Levine, M.D. of Alpine Clinical Research Center. Statistical experts were Weining Z. Robieson, Ph.D. (M06-888) and Coleen M. Hall, M.S. (M10-345), of Abbott.

4.2 Input Vector Representation

Our proposed HDNN accepts two different types of input vector representations for a given input sentence: (1) a sequence of word embeddings for the multi-channel CNN section and (2) bag of words (BOW) for the FNN section. These input vector representations are also employed in other types of machine learning models implemented and tested in our study for comparison.

Word Embeddings. In order to perform a text classification task or natural language processing at large using CNN (or other deep neural models such as LSTM), we first need to convert an input sentence or a document to an $n \times k$ matrix. Each row in the matrix is k-dimensional vector representation called word embedding for each individual word in the sentence of length (number of words) n. In our study, we define n as the maximum number of words in a sentence we can find from our training dataset and pad a sentence of length m (<n) with $n - m$ of zeros to make the same size of input matrix for CNN.

We employ three word embedding methods: (1) dictionary-based (look-up table), (2) Word2Vec [2], and (3) Glove [3]. Dictionary-based embedding is a sparse and discrete vector representation of word. A dictionary is created using vocabulary words collected from the training dataset. A word vector represents the index of its corresponding word in the dictionary. Word2Vec is a "prediction-based" unsupervised (more precisely, self-supervised) neural network language model. Unlike a dictionary-based word vector, it generates a dense and distributed numerical vector representation of a word. The basic idea of using Word2Vec is to map semantically similar words or words having similar context to nearby points in a lower dimensional vector space. We actually use the publicly available Word2Vec model pre-trained on 100 billion words from Google News. Lastly, we also use another pre-trained model, GloVe, a new global log-bilinear regression model trained on 840 billion tokens of word data for unsupervised learning of global word-word co-occurrence statistics.

All of the three embedding methods generate a 300-dimensional vector representation for each word in a given sentence. Words not present in the dictionary or pre-trained model are represented by an all-zero (in dictionary-based) or a randomly and uniformly initialized (in Word2Vec and GloVe) vector. Although other approaches exploiting character-level embeddings [17] have been also reported, we mainly focus on these word-level embedding methods in this study.

Bag of Words (BOW). We adopt a bag of words (BOW) based on word statistics representing how differently a word is distributed in "support" and "others" sentence classes, to build an input feature vector for the FNN section in HDNN. Word selection to build a BOW is accomplished by sorting words according to their importance measured by simplified $\chi^2(s\chi^2)$ statistics [18].

In our task, $s\chi^2$ of word t_k for sentences in the "support" class (class c_0) and those in the "others" class (class c_1) can be defined as follows:

$$s\chi^2(t_k, c_i) = P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i) \quad i = 0, 1 \tag{1}$$

where $P(t_k, c_i)$ denotes the probability that, for a random sentence x, word t_k occurs in x, x belongs to class c_i, and is estimated by counting its occurrences in the training set. The importance of word t_k is finally measured as follows:

$$s\chi^2_{max}(t_k) = max_i s\chi^2(t_k, c_i) \quad i = 0, 1 \tag{2}$$

Accordingly, the more differently a word is distributed in "support" and "others" classes, the higher its $s\chi^2_{max}(t_k)$. Words are sorted according to their $s\chi^2_{max}$, and a BOW is then created by selecting words having highest $s\chi^2_{max}$ scores. Through a series of experiments to investigate the influence of word reduction, we discovered that this BOW feature shows the best performance when its word dictionary size is 500. Finally, the BOW is converted to a binary vector: the vector dimension corresponds to the number of words (=500) in the dictionary, and each vector component is assigned 1 if the corresponding word in the dictionary is found in a given sentence or 0 otherwise. Another BOW for “for-profit” and “non-profit” sentence classes is also obtained through the same procedure above.

4.3 HDNN Architecture

Our proposed HDNN consists of two neural network sections: multi-channel CNN and FNN sections, as illustrated in Fig. 2. Each section takes different types of input vector representations for a given input sentence and then proceeds with the learning process separately until they are combined through a full connection between the hidden layers and output layer. As a result, a hybrid structure is created.

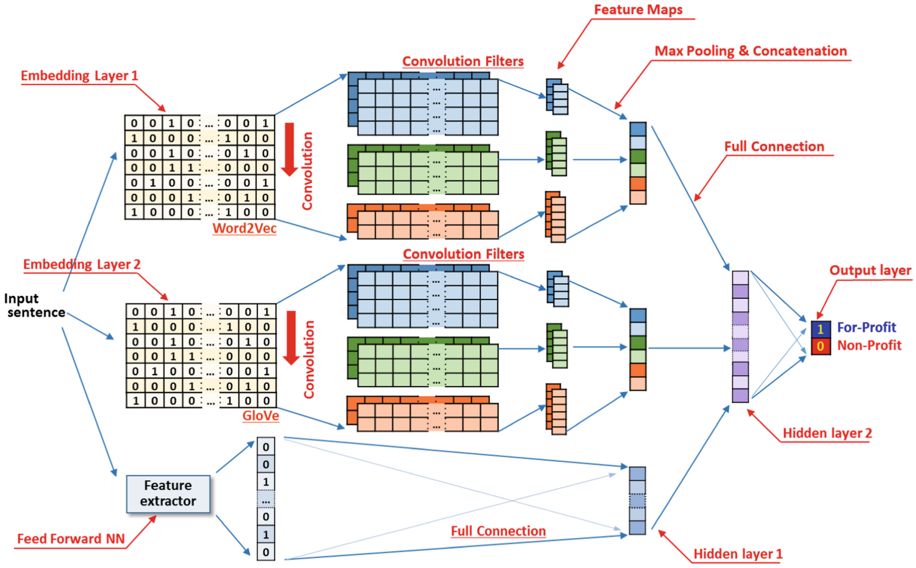

Fig. 2. The structure of proposed HDNN.

Multi-channel CNN Section. The multi-channel CNN section in our HDNN is built on combining two single-channel CNNs. These single-channel CNNs both have the same structures; the size of the matrix word embeddings, the number and size of convolution filters, and other hyperparameters are all identical. However, they employs different types of word embeddings as an input: Word2Vec and GloVe, respectively.

In each CNN, we first apply a convolution filter, $w \in \mathbb{R}^{d \times k}$ and a nonlinear function f with a bias term $b \in \mathbb{R}$ on a window of d rows in the matrix word embeddings, $X \in \mathbb{R}^{n \times k}$, where i^{th} row is the k-dimensional word embedding x_i representing i^{th} word in a given input sentence of length n, as follows;

$$c = f(w \cdot X_d + b) \tag{3}$$

By repeatedly performing such convolution operation over the entire word embedding matrix with stride 1, we can then obtain a feature map $\boldsymbol{c} = [c_1, c_2, \ldots, c_{n-d+1}]$ with $\boldsymbol{c} \in \mathbb{R}^{n-d+1}$. Since we employ the multiple convolution filters (=*h*) with different width or windows sizes (=*l*) for our CNN, we finally have a set of feature maps $\boldsymbol{C} = \left[\boldsymbol{c^1}, \boldsymbol{c^2}, \ldots, \boldsymbol{c^{h \times l}}\right]$.

Next, max-pooling operation is performed to obtain the maximum value from each feature map;

$$y^j = max_i c_i^j \tag{4}$$

where $j = 1, 2, \ldots, h \times l$ and $i = 1, 2, \ldots, n - d + 1$. Such max-pooling operation not only reduces the output dimensionality while keeping the most salient information but also induces a fixed-length of feature vector from the different size of feature maps resulting from applying a different width of convolution filters. These maximum values called features generated from each CNN are concatenated together to form a multi-channel structure, and along with the outputs from the hidden layer in the FNN section, fed to the fully connected next hidden layer. All CNN models implemented in our experiments have 128 convolution filters with three different window sizes (*l* = 3, 4, and 5) generating a total of 384 feature maps, dropout rate of 0.5, mini-batch size of 64, and "Adam" optimizer. All these hyperparameters were obtained through a grid search method. In addition, rectified linear unit (ReLU) and softmax nonlinear activation functions are applied to the hidden layers and the final output layer, respectively.

On the other hand, a similar concept of two-channel approach has been recently suggested by Kim [11]. Unlike our multi-channel approach where each channel of CNN accepts a different type of word embedding and performs a separate convolution operation, his method employs Word2Vec only as an input representation for both channels of CNN; Word2Vec in one channel is kept unchanged (static) and the other is fine-tuned (non-static) during a learning phase.

FNN Section. We introduce the FNN section into HDNN to take advantage of a handcrafted input feature experimentally found to be effective. As mentioned previously, we employ a 500-dimension binary vector representing a BOW as an input feature vector for the FNN section. Words in this BOW are selected and sorted according to their corresponding $s\chi^2_{max}$ scores that reflect the difference of their statistical distributions between the "support" and "others", or "for-profit" and "non-profit" classes.

5 Classification Experiments

5.1 Ground-Truth Dataset and Tools

In order to build a ground-truth dataset for our experiments, we first downloaded 2,800 HTML-formatted online biomedical articles having citation information of grant support or ClinicalTrials.gov from NLM's PubMed Central (PMC) [19]. These articles were published in 938 different journals and indexed in MEDLINE between 2010 and 2015. We then collected a total of 21,822 sentences from these articles and divided them into two classes: "support" and "others" according to whether they contain information of a funding support or not. Sentences in the "support" class were further divided into two sub-classes: "for-profit" and "non-profit".

Among these, 16,753 sentences consisting of 4,528 in the "support" class and 12,225 in the "others" class were randomly selected to train the classifiers for the stage 1 classification experiment—distinguishing "support" sentences from the body text of biomedical articles. The remaining 5,069 sentences (1,509 from the "support" class + 3,560 from the "others" class) were used as a test set to evaluate the performance of our classifiers. Next, for the stage 2 classification experiment—classifying a "support" sentence into "for-profit" or "non-profit" class, we reemployed the "support" sentences already used for the stage 1 classification experiment. Accordingly, each training and testing set has 4,528 (1,937 from the "for-profit" + 2,591 from the "non-profit") and 1,509 (645 from the "for-profit" + 864 from the "non-profit") "support" sentences, respectively.

All DNN models employed in our study including the proposed HDNN, single/multi-channel CNNs, and LSTM were implemented based on Tensorflow [20], very well-known open source library developed by the Google Brain team, Keras [21], a simple and high-level model definition interface, and Nvidia's CUDA toolkit and CuDnn for GPU-acceleration. In addition, SVM with RBF kernel function, another classifier widely used in text classification and other machine learning tasks, was implemented using LibSVM [22], a free software package.

5.2 Experimental Results

As mentioned earlier, our proposed method of identifying potential COI from an online biomedical article adopts a two-stage machine learning scheme. In experiments, we implemented and evaluated a total of 9 classifiers for both stage 1 and 2 classification tasks: SVM with a RBF, 3 single-channel CNNs with different word embeddings, LSTM, multi (two)-channel CNN, voting scheme, Ensemble model, and our proposed HDNN.

First, it can be clearly seen from Tables 4 and 5 that all DNN models consistently outperform SVM. In the case of LSTM, a better performance was achieved than that of any of single-channel CNNs in the stage 1 classification experiments. However, a reversal in performance is observed in the stage 2 classification. Note that the size of the training dataset used in the stage 2 experiments is significantly smaller (about 25%) than that in the stage 1 experiments. Thus, LSTM is analyzed to be more susceptible to the size of the training dataset than other classifiers, thereby resulting in a degradation of the classification performance in stage 2.

We can also see that our multi-channel CNN and especially HDNN both accepting multiple input representations: "word2Vec + GloVe" and "Word2Vec + GloVe + BOW", respectively, yield the best performance overall in both stage 1 and 2 classification experiments. The ensemble model which also employs three types of input representations the same as those used in the proposed HDNN, through the combination

Table 4. Stage 1 classification results.

Models	Accuracy	Precision	Recall	F_1
SVM	96.21	98.59	95.98	97.27
Dic CNN	96.61	96.67	98.57	97.61
W2v CNN	97.04	97.28	98.54	97.91
Glv CNN	97.18	98.03	97.95	97.99
W2v+Glv CNN	97.93	98.40	98.65	98.53
LSTM	97.57	98.02	98.54	98.28
HDNN	**98.11**	**98.52**	**98.79**	**98.65**
Ensemble	97.40	97.69	98.62	98.15
voting	97.73	97.97	98.82	98.39

Table 5. Stage 2 classification results.

Models	Accuracy	Precision	Recall	F_1
SVM	95.16	94.89	96.76	95.82
Dic CNN	96.02	95.79	97.34	96.56
W2v CNN	96.16	95.90	97.45	96.67
Glv CNN	96.22	96.11	97.34	96.72
W2v+Glv CNN	96.62	96.67	97.45	97.06
LSTM	95.89	95.67	97.22	96.44
HDNN	**96.89**	**97.44**	**97.11**	**97.28**
Ensemble	96.49	96.56	97.34	96.95
voting	96.49	96.23	97.69	96.96

of pre-trained 2 CNNs and FNN, is found to provide a slight improvement over the individual classifiers having a single input vector representation such as SVM, single channel CNNs, and LSTM, but to be not as good as HDNN. Rather, a simple majority voting scheme for the outputs of 5 pre-trained individual classifiers (SVM + 3 CNNs + LSTM) performs better. Therefore, we conclude that our HDNN is a more effective architecture for combining multiple learning models and taking advantage of different types of input representations to boost the overall classification performance further.

Finally, in Table 6 we show some examples of false-negative (FN) and false-positive (FP) errors in stage 2 classification. Here, FN means that "non-profit" support sentence is misclassified into "for-profit" class. FP is the reverse of the above. The first "support" sentence in the FN error examples contains two NIH grant numbers "GM103429" and "GM103450". However, this sentence is very short, and there is no contextual description associating these grant numbers with an NIH financial support. The second sentence of FN errors is analyzed to be misclassified due to several pairs of words such as "senior advisor", "pharmaceutical company", and "international market" which are also frequently found in "for-profit" sentences, even though it has the word, "nonprofit". In contrast, FP errors shown in Table 6 result from the existence of "non-profit" organizations names ("Cleveland Clinic" and "NIH") along with a description of "for-profit" funding support within the sentence.

Table 6. Error examples showing FN and FP errors in the stage 2 classification.

Error types	Support sentences
False Negative	JS received financial support from GM103429 and GM103450. He was senior clinical advisor of a nonprofit (501c3) pharmaceutical company studying a lower-cost IUD for the U.S. and international markets (Medicines 360).
False Positive	Mass spectrometry studies were performed in the Cleveland Clinic Mass Spectrometry core facility, which is partially supported by a Center of Innovation Award from AB SCIEX. The vitamin E softgels and matching placebo were provided by Pharmavite through a Clinical Trial Agreement with the NIH.

6 Conclusions

Conflicts of interest have a major negative impact on the integrity of biomedical research. Many biomedical journal publishers thus require authors to provide a COI disclosure statement in their article at the time of peer-review and publication. However, authors often declares a COI implicitly and in a subtle manner. In addition, there are also rising concerns about the accuracy and reliability of author's self-declared COIs.

In this paper, we have presented a sequential two-stage machine learning-based text classification scheme to automatically ascertain potential COI from the body text of online biomedical articles. The first stage of classification is for distinguishing "support" sentences from other sentences in the body text of a given biomedical article, and the second stage is for categorizing those "support" sentences into "for-profit" and "non-profit" classes according to their funding sources. Each stage of classification is implemented using a deep learning model that has a hybrid architecture to combine a multi-channel convolutional neural network (CNN) and a feed-forward neural network

(FNN). This hybrid deep neural network (HDNN) aims at simultaneously learning syntactic and semantic representations and word relationships in a sentence, and handcrafted input features.

Experiments on a total of 21,822 sentences from 2,800 HTML-formatted online biomedical articles published in 938 different journals show that our proposed HDNN yields a consistently higher performance in both stage 1 and 2 classification tasks, compare to other classifiers having a single type of input vector representation such as SVM, single-channel CNNs, and LSTM. Our HDNN is also found to be a superior architecture for combining multiple input representations than an ensemble model or voting scheme both based on pre-trained learning models.

Acknowledgment. This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

1. http://www.ncbi.nlm.nih.gov/pubmed
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2013), pp. 3111–3119, Lake Tahoe (2013)
3. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1532–1543, Doha, Qatar (2014)
4. Ting, S.L., Ip, W.H., Tsang, A.H.C.: Is Naïve Bayes a good classifier for document classification? Int. J. Softw. Eng. Appl. **5**(3), 37–46 (2011)
5. Mercer, R.E., Di Marco, C.: A design methodology for a biomedical literature indexing tool using the rhetoric of science. In: Proceedings of the HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, pp. 77–84, Boston (2004)
6. Athar, A.: Sentiment analysis of citations using sentence-structure-based features. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), pp. 81–87, Portland (2011)
7. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nature **521**(7553), 436–444 (2015)
8. Kalchbrenne, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 655–665, Baltimore (2014)
9. dos Santos, C.N., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers, pp. 69–78, Dublin, Ireland (2014)
10. Dong, L., Wei, F., Zhou, M., Xu, K.: Question answering over freebase with multi-column convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 260–269, Beijing, China (2015)
11. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746–1751, Doha, Qatar (2014)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

13. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2014), pp. 3104–3112, Montreal, Canada (2014)
14. Ghosal, D., Bhatnagar, S., Akhtar, M.S., Ekbal, A., Bhattacharyya, P.: IITP at SemEval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluations, pp. 899–903, Vancouver, Canada (2017)
15. https://grants.nih.gov/grants/policy/coi/tutorial2011/fcoi.htm
16. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, Baltimore (2014)
17. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2015), pp. 649–657, Montreal, Canada (2015)
18. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Borbinha, J., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45268-0_6
19. http://www.ncbi.nlm.nih.gov/pmc/
20. Abadi, M., Agarwal, A. et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. Software (2015). tensorflow.org
21. Chollet, F., et al.: Keras. GitHub (2015). https://github.com/fchollet/keras
22. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). http://www.csie.ntu.edu.tw/∼cjlin/libsvm