

Neural Network Architecture for Credibility Assessment of Textual Claims

Nurendra Choudhary*, Rajat Singh*, Ishita Bindlish and Manish Shrivastava

Language Technologies Research Centre (LTRC)
Kohli Center on Intelligent Systems (KCIS)
International Institute of Information Technology, Hyderabad, India
{nurendra.choudhary, rajat.singh}@research.iiit.ac.in
ishita.bindlish@students.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract. Text articles with false claims, especially news, have recently become aggravating for the Internet users. These articles are in wide circulation and readers face difficulty discerning fact from fiction. Previous work on credibility assessment has focused on factual analysis and linguistic features. The task’s main challenge is the distinction between the features of true and false articles. In this paper, we propose a novel approach called Credibility Outcome (CREDO) which aims at scoring the credibility of an article in an open domain setting.

CREDO consists of different modules for capturing various features responsible for the credibility of an article. These features includes credibility of the article’s source and author, semantic similarity between the article and related credible articles retrieved from a knowledge base, and sentiments conveyed by the article. A neural network architecture learns the contribution of each of these modules to the overall credibility of an article. Experiments on Snopes dataset reveals that CREDO outperforms the state-of-the-art approaches based on linguistic features.

1 Introduction

“Fake news is a type of hoax or deliberate spread of misinformation, be it via the traditional news media or social media, with the intent to mislead, in order to gain financially or politically.” [14]

The fake articles do not base their information on facts but use convenient, seemingly true, logical inferences to make readers trust the news. In recent times, the difference between real and fake news articles has become bleak. Verification of the information needs checking for reliable sources. Most readers find the task cumbersome and hence don’t perform it. Therefore, automated systems that detect such articles are a major requirement.

Correction of misinformation is difficult. Fabricated news persists because of people casually inferring based on available information about a given event. As a result, false information continues to influence opinions, beliefs and attitudes

* These authors have contributed equally to this work.

even after being debunked, unless replaced by an alternate factual explanation. The situation deteriorates when the fake articles have a financial or political agenda. The manner in which these articles shape public opinion affects the society as a whole, making it a very serious problem.

The modules of CREDO's (Credibility Outcome) include keyword extraction, document retrieval, author credibility scores, website trust scores, semantic similarity and sentiment analysis. Keyword extraction module chooses significant words in the given input article which, in the document retrieval phase, fetches documents from a knowledge base, primarily, news pages and Wikipedia. Querying the entire document gives us negative results and is not efficient, considering the long queries. Subsequent phases learn trust scores of the source websites of the retrieved articles and credibility of their authors. To ensure that our retrieved document and the given article are not dissimilar, we also compute semantic similarity. An article is a sequence of words. Semantic similarity modules need to include a method that captures this sequence of words and also learn a distance metric between them. Long Short-Term Memory (LSTM) based recurrent neural networks have proven helpful in capturing sequences [19]. Also, siamese networks have shown promising results in distance-based learning methods [5]. Hence, we use a combination of these by utilizing a bidirectional LSTM to map articles to a semantic space in conjunction with a siamese network to learn the similarity metric between them.

A factual article has more probability of being neutral [18]. Hence, a system to capture this feature is essential. To tackle the problem, we use a sentiment analysis [20] tool to evaluate the neutrality of a given article.

The rest of the paper is organized as follows. We discuss previous approaches in the field, motivating us for the task in section 2. In section 3, we discuss the pipeline and individually explore each module of CREDO. Section 4 describes the datasets and baselines considered for the task. Section 5 explains our experiments and their evaluation. Finally, section 6 concludes the paper.

2 Related Work

The work connects a significant number of research areas, including but not limited to automatic fact checking, rumour detection, sentiment analysis, semantic similarity and credibility analysis.

[29] discusses the belief system of computers and credibility analysis' necessity. [28] discusses fact checking, its definition and motivations, posing it as a classification task. [7] proposes credibility analysis in a social media context, thereby using features that describe users' posting behaviour. Joint model based on CRF [22] employs linguistic features like assertive verbs, discourse markers among others to establish a common style of writing across fake articles and news. This approach proves effective and achieves more than 80% accuracy on snopes dataset.

CoTruth [15] uses a similar approach in employing linguistic features as a joint model based on CRF [22] but does not perform a credibility check on the author and website.

The above approaches model the writing style. However, they do not consider the available knowledge bases. The writing style also depends on the content’s authors and the website’s type. A unified model to capture the style without taking into account the information of authors and website will undoubtedly be fragile.

In semantic similarity, we use siamese networks, to compute the similarity between two sentences. [5] first introduced siamese networks in 1994 to solve the problem of signature verification. Since then, there have been attempts to understand their relevance in the context of sentence similarity. In SCQA model [9], siamese networks solve the task of community question answering and in DSSM model [12], they handle the task of website ranking. The above methods use the siamese network based on the task. Here, the task is the semantic similarity. Hence, we use LSTM models [19] to project our articles in the semantic space to learn a similarity metric between them.

Article	Keywords
In May 1946, Einstein made a rare public appearance outside of Princeton, New Jersey, when he traveled to the campus of Pennsylvania’s Lincoln University, the United States’ first degree-granting black university, to take part in a ceremony conferring upon him the honorary degree of doctor of laws.	[('degree-granting black university', 8.5), ('lincoln university', 4.5), ('ceremony conferring', 4.0)]
There are two fatal problems with the JATO story. First, anybody who understood the extreme forces involved well enough to attach a JATO unit to a car so that it would keep the car going in a straight line would probably know better than to do it in the first place.	[('extreme forces involved', 9.0), ('fatal problems', 4.0), ('straight line', 4.0)]

Table 1: Result of Keyword Extraction

3 CREDO: Methodology and Architecture

CREDO or Credibility Outcome is a neural network architecture combining information retrieval, semantic similarity, and sentiment analysis. The following subsections explain the architecture’s modules.

3.1 Keyword extraction and Document retrieval

Keywords provide us with most of the significant information contained in the article. Hence, keyword extraction becomes an essential initial step for document retrieval. The module creates a query using keywords of input text to retrieve relevant documents. We use Rapid Automatic Keyword Extraction (RAKE) described in [25]. RAKE involves three sub-modules -

1. **Candidate selection** extracts all potential keywords (words, phrases and concepts).
2. **Properties calculation** measures the candidate’s indication of being a keyword. For example, a likely keyword is a candidate in the title of a book.
3. **Scoring and selecting keywords** scores the candidates by either combining the properties into a formula or utilizes a machine learning technique to determine the probability of a candidate being a keyword. The final set are the keywords with a score above an empirical threshold. This assigns the keywords to an article, with their respective scores representing their importance in the article. Some examples of the extracted keywords are given in table 1.

Keyword extraction and document retrieval module uses the keywords to query over the knowledge bases (proprietary set of indexed documents, such as Wikipedia, or a Web search engine like Google search and Bing search) using an information-retrieval system. The result of this document retrieval stage is a set of relevant documents.

Although the set of documents is relevance-ranked, the top entities are probably not credible enough. Keyword relevance is not an appropriate ranking metric for a credibility scoring system. The existence of a highly relevant and large document full of fake news is easily possible. Therefore, we rely on other modules to check document credibility.

3.2 Website and Author Trust Scores

In this module, given an article’s source, the Web of Trust API¹ initializes the credibility score of the website. Web of Trust, a website reputation and review service, provides information about the trustworthiness of a website. The information’s basis is a combination of crowd-sourced reviews and identification of networks involved in malware distribution.

With the scores from Web of Trust, we perform a logarithmic gradient descent based on the tag of the input in the dataset. i.e.

$$WTS(w_{i+1}) = \begin{cases} (i + \log(1 - t(w_i))) \times WTS(w_i), t(w_i) \leq 0.5 \\ (i + \log(1 + t(w_i))) \times WTS(w_i), t(w_i) > 0.5 \end{cases} \quad (1)$$

where $WTS(w_i)$ is the web trust score of website w for its i^{th} instance in the sample and $t(w_i)$ represents the tag of the i^{th} article.

Similarly, initial articles of an author decide his score for the present one. An initial score of 0.5 is set for each author. If the author is anonymous or irretrievable from the dataset, the probability of an author being credible and not credible is assumed to be equal and hence, the ACS score is set to 0.5. The metric for author credibility score is:

$$ACS(a_{i+1}) = \begin{cases} (i + \log(1 - t(a_i))) \times ACS(a_i), t(a_i) \leq 0.5 \\ (i + \log(1 + t(a_i))) \times ACS(a_i), t(a_i) > 0.5 \end{cases} \quad (2)$$

¹ www.mywot.com

where $ACS(a_i)$ is the author credibility score of author a at his i^{th} article in the sample and $t(a_i)$ denotes the tag of the i^{th} article.

New training instances dynamically update these calculated scores according to their contribution to the final credibility Score of the source website and the article’s author.

3.3 Document Summarization

Only specific sections of the document demonstrate the significant relevant information in a text article. Hence, based on the keywords and TextRank algorithm [17] with BM25 ranking function [2] (implementation by Gensim²), we summarize the document by extracting the top k sentences based on the algorithm’s ranking. We choose k such that the summary size is approximately the same as that of the input article. Summarization of large input articles emphasize the most significant information, while at the same time, simplifying keyword extraction.

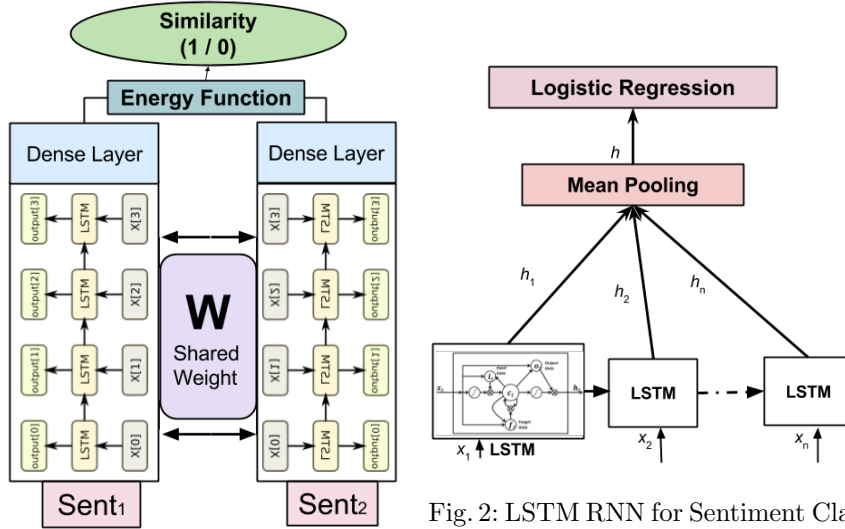


Fig. 1: Siamese Architecture for Semantic Similarity

Fig. 2: LSTM RNN for Sentiment Classification

3.4 Semantic Similarity

Semantic similarity is of paramount importance in the scoring task. Given the summarized versions of the original input article and the documents retrieved in the previous steps, we calculate the semantic similarity score between them. This is necessary because only keywords influence the documents retrieval from the knowledge base. So, it is possible that the keywords match the negative meaning

² <https://radimrehurek.com/gensim/summarization/summariser.html>

or the document retrieved and the original input are unrelated, apart from just some words.

For example, a web search on the query “Donald Trump is the president of India” results in titles - “How US President Donald Trump’s world-view affects India”, “US President Donald Trump to speak to PM Modi as India hopes to build on US ties” etc, which are considered good evidences by the system but are semantically different.

In this model, siamese architecture with LSTMs calculates the semantic similarity, similar to how the DSSM model [12] computes semantic similarity with primarily two differences:

- A fully connected neural network is the DSSM model’s basis, whereas, we employ an LSTM here instead. LSTMs capture sequential information data types like texts, in the sense that they capture order and history.
- The basic unit in their model constitutes character n-grams, whereas we use tokens or words. We do not need character based vectors because news articles are not susceptible to spelling errors and out of vocabulary words.

As illustrated in figure 1, the architecture consists of twin LSTM networks with a similarity-based energy function on the top. The LSTM networks map the article to a vector in the semantic space and the energy function learns the similarity between them. The similarity metric is learned using contrastive learning. The sentences with similar meaning are labeled +1 (positive samples) and dissimilar meaning sentences are labeled -1 (negative samples). The results of semantic similarity are explained later in Section 5.2.

3.5 Sentiment Analysis

Sensationalized or opinionated news articles are sentiment heavy whereas factual articles are more objective and neutral. To establish a relation between credibility and sentiment of an article, we employ a sentiment analysis tool based on LSTM (Long Short-Term Memory) networks. LSTM networks show remarkable results in learning sequential information like texts [27]; [11]. As depicted in Figure 2, we use an LSTM model with single layer. It transforms the input sequence $\langle x_i \rangle$ to a representation sequence $\langle h_i \rangle$. Then a pooling layer averages representations h_i s over all time steps, resulting in a single representation. Finally, we train a logistic regression model on the representations whose target is the sentiment label (+1 or -1) corresponding to the input sequence. Our motivation to use an LSTM based sentiment classification model without using any manual linguistic features is to avoid any linguistic encoding throughout CREDO and rely essentially on machine intelligence. Section 5.2 describes the influence of this module over the system.

3.6 Ensemble of the Modules

Taking Naive Bayes assumption (or conditional independence assumption), we consider all the above features independent of each other. They together contribute to the article’s overall credibility. So, credibility contribution of each

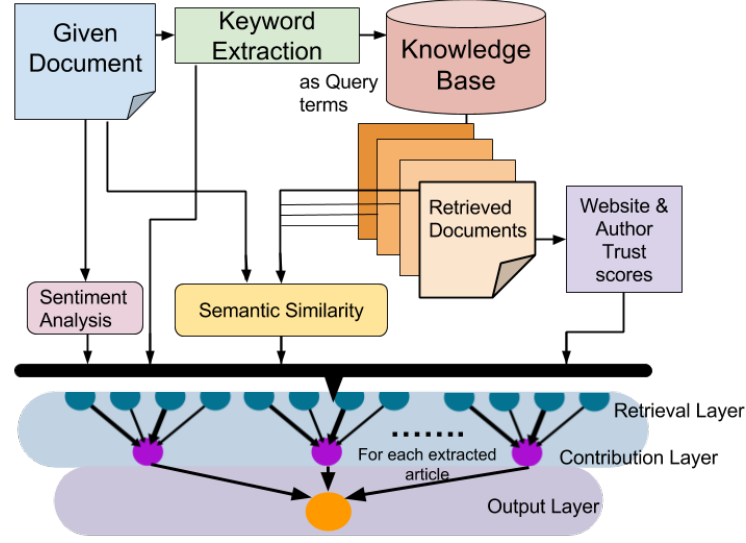


Fig. 3: Architecture of CREDO

article is given by:

$$CC(a_r, a_i) = (w_1 \times (\sum_{i \in k} kscore(i))) + (w_2 \times ACS + w_3 \times WTS) \\ + (w_4 \times NS) + ((w_5) \times SS(a_i, a_r))$$

$$\text{given that, } w_1 + w_2 + w_3 + w_4 + w_5 = 1 \ \& \ w_1, w_2, w_3, w_4, w_5 > 0 \quad (3)$$

where a_r, a_i are the retrieved article and given input article respectively, $CC(a_r, a_i)$ represents the credibility contribution of an article a_r with respect to the given article a_i , k represents the keywords in the article a_r $kscore(i)$ represents the keyword scores of the i^{th} keyword, ACS represents the author's credibility score derived from equation 2, WTS represents the trust score of the website derived from equation 1, $SS(a_i, a_r)$ represents the similarity score between the input and retrieved article and NS represents neutral sentiment's value of a_i , scaled in 0-1. w_1, w_2, w_3, w_4, w_5 are the respective weight parameters given to these measures in the overall credibility.

Since we retrieve multiple articles, we repeat the same process for every one. We assume that the credibility of all the retrieved articles, weighted with a function of their respective retrieval ranks, will contribute to the final credibility score for the given input article.

$$Credo(a_i) = \frac{\sum_{r \in R} (e^{1 - \frac{rank(a_r)}{n}} \times CC(a_r))}{\sum_{r \in R} e^{1 - \frac{rank(a_r)}{n}}} \quad (4)$$

where $Credo(a_i)$ is the input article's overall credibility a_i , R denotes the set of retrieved documents, n denotes the number of documents retrieved, $rank(a_r)$

Experiment	Overall Accuracy	True Claims Accuracy	False Claims Accuracy	Macro-averaged Accuracy	AUC	Fake Claims Precision	Fake Claims Recall	Fake Claims F1-score
CRF	81.39	83.21	80.78	82.00	0.88	0.93	0.81	0.87
LG + SR	71.96	75.43	70.77	73.10	0.80	0.89	0.71	0.79
RBF-SVM	82.8	85.7	81.5	83.6	0.87	0.94	0.85	0.89
MLP-NN	83.3	86.2	81.2	83.7	0.83	0.92	0.84	0.88
Multi-Class	56.4	59.8	55.2	57.5	0.63	0.63	0.53	0.57

Table 2: Comparison of CREDO with baselines (LG+SR and CRF).

represents the retrieval rank of article a_r and $CC(a_r)$ denotes the credibility contribution of a_r given by equation 3. Equation 4 gives us the input article’s credibility in the range $[0, 1]$. Figure 3 details the overall architecture of CREDO.

The weights in the equations 3 and 4 define the parameters for the module’s linear combination scores. Various types of classifiers learn these parameters. Section 5 explains the different trials. The weights learned assigns the credibility score to any new input text article.

4 Dataset & Baselines

This section explains the choice of Dataset for the task and also describes the construction of baselines from the previous approaches to compare with CREDO.

4.1 Dataset

The following datasets are considered for training and evaluation of CREDO.

Snopes Dataset: The dataset used for comparative analysis of the Credibility Scoring Algorithm is Snopes Dataset. Previous approaches to the problem [22], [21] use the same dataset.

Each article on Snopes verifies a single claim. The Snopes’ editors assign a manual credibility verdict to each such claim: True or False. Few of the claims have labels *Mostly True*, *Mostly False*, *Partially True* or *Partially False*. A description of the editors’ arrival to the claim accompanies the credibility verdict (e.g., collected from a Facebook post or received by email etc.) - an *Origins* section describes the claim’s origin, and a *Description* section justifies the verdict. We consider *True* and *Mostly True* as *True* claims. Similarly, we consider *False* and *Mostly False* as *False*. The dataset has 4856 total claims with 1277 *True* claims and 3579 *False* claims.

SemEval 2016 Dataset: For evaluating the Semantic Similarity module, we adopt the standard SemEval-2016 Task1 - English Semantic Textual Similarity (STS) Dataset ³. The dataset consists of sentence pairs with their similarity scores in binary.

³ <http://alt.qcri.org/semeval2016/task1/index.php?id=data-and-tools>

The dataset has five different sentences' types drawn from various sources. The dataset of *Answer-Answer* and *Question-Question* is taken from Stack Exchange Q&A Forums⁴, *Headlines* data is taken from Europe Media Monitor (EMM) [3], data of *Plagiarism* is taken from corpus of plagiarized short answers [8] and data of *Postediting* is taken from WMT quality estimation shared task [6].

4.2 Baselines

We test CREDO system against the previous approaches in the problem [22] and [21] and evaluate it against the same metrics.

- **LG + SR:** This approach uses language stylistic features and source reliability to determine the credibility tag of an article using a distant supervision model. [21]
- **CRF:** This approach improves upon the *LG+SR* and attempts to solve the problem using Conditional Random Fields (CRF) dependent on web-sources, articles, claims and claim credibility labels. [22]

5 Experiments

We conducted an array of experiments to understand the effect of different modules on the system's overall performance. For this, we first excluded the modules and then observed the effect of this exclusion on the system's evaluation metrics. We also conducted experiments to understand the effect of different classifiers on the system. We calculated the evaluation metrics for different classifiers and attempted to understand the reasons for the difference in results. Semantic similarity, being a major module, was evaluated independently to improve results.

5.1 CREDO System

We train and test the CREDO system on the Snopes dataset. The experiment is conducted for binary credibility classification. The *mostly true* and *true* are considered positive labels and *mostly false* and *false* are considered negative labels. However, for our second experiment, we conduct multi-class credibility classification. For this, *true,mostly true,mostly false* and *false* are considered separate labels in a decreasing order.

Accuracy, Precision, Recall and F1-score are the evaluation metrics of this experiment. For training the weights, different classifiers were used - Support Vector Methods(SVM) with RBF kernel [1] and Neural Networks (Multi-Layer Perceptron) [24] had the best results. Evaluation metrics for multi-class SVM use methods from [10].

K-fold cross validation (K=5) was used for training and testing on Snopes dataset. The evaluation metrics obtained are averaged across all the evaluation sets. The results of experiments with different classifiers are given in table 2.

⁴ <https://archive.org/details/stackexchange>

Excluding Modules	Overall	True	False	Macro-		Fake	Fake	Fake
	Accuracy	Claims Accuracy	Claims Accuracy	averaged Accuracy	AUC	Claims Precision	Claims Recall	Claims F1-score
CREDO <i>w/o</i> ACS	80.5	82.3	81.3	81.8	0.83	0.90	0.82	0.86
CREDO <i>w/o</i> WTS	81.8	84.7	80.4	82.55	0.85	0.92	0.83	0.87
CREDO <i>w/o</i> SS	50.6	53.4	50.4	51.9	0.53	0.58	0.56	0.57
CREDO <i>w/o</i> SA	76.2	78.4	75.4	76.9	0.76	0.77	0.76	0.76
CREDO	83.3	86.2	81.2	83.7	0.83	0.92	0.84	0.88

Table 3: Comparison between the full CREDO system and CREDO system without some modules. SS here is Semantic Similarity and SA is Sentiment Analysis

Dataset	Ans. -Ans.	Headlines	Plagiarism	Postediting	Ques.-Ques.	Mean
Baseline STS	0.41	0.54	0.70	0.83	0.04	0.51
CREDO Similarity	0.54	0.71	0.73	0.82	0.61	0.68
Improvement (%)	31.7	31.5	4.3	-1.2	1425	33.3

Table 4: Correlation Score in STS 2016 Task-1 of CREDO semantic similarity module

5.2 Dependency of CREDO on its modules

CREDO is a system based on 5 modules, but not all of them contribute equally. So, this experiment studies each modules' contribution to the overall system. For this, the effect on the performance of CREDO due to the exclusion of modules is analyzed in contrast to the original system.

Semantic similarity has to be tested independently to observe its effect on the overall performance of CREDO. The performance of semantic similarity module is tested on standard SemEval-STS 2016 English task.

5.3 Dependency of CREDO on the classifier

The overall problem is of classification and hence the choice of classifier has to be appropriate for the data points. Quadratic Discriminant Analysis (QDA) [26], Gaussian Naive Bayes [13], Decision Trees [23], Random Forests (ensemble of Decision Trees) [4], AdaBoost Classifier [16], Support Vector Methods with Radial Basis Function kernel (SVM-RBF) [1] and Multi Layer Perceptron Neural Network (MLP-NN) [24] were chosen for the experiments based on the variance of their application. All the evaluation metrics, tested for the original system, were tested for the classifiers as well.

Classifier	Overall	True	False	Macro-		Fake	Fake	Fake
	Accuracy	Claims	Claims	averaged	AUC	Precision	Recall	F1-score
Quadratic DA	61.6	63.5	59.5	64.5	0.67	0.63	0.58	0.60
Naive Bayes	65.8	67.1	64.3	66.7	0.68	0.71	0.65	0.69
Decision Trees	66.3	67.4	64.5	65.4	0.69	0.70	0.67	0.68
AdaBoost	70.6	72.3	69.4	71.1	0.74	0.75	0.72	0.73
Random Forest	78.2	80.1	79.8	80.8	0.81	0.82	0.80	0.81
RBF-SVM	82.8	85.7	81.5	83.6	0.87	0.94	0.85	0.89
MLP-NN	83.3	86.2	81.2	83.7	0.83	0.92	0.84	0.88

Table 5: Comparison between different classifiers.

5.4 Evaluation of the Experiments

Table 2 shows the comparative results of CREDO on Snopes dataset. CREDO outperforms the state-of-the-art approaches across all the metrics. It is also observed that CREDO shows improvements over other approaches, even without the help of the Web of Trust module. Hence, the combination of Web of Trust and author credibility scoring modules helps in boosting the performance.

The results for the experiment on dependency of CREDO on modules (given in Section 5.2) is shown in table 3. The results show that ACS and WTS scores help the model but the improvements are insignificant compared to the enhancement in the metrics brought by the semantic similarity module. Including sentiment analysis module increments the accuracy, but it is minor compared to the improvement given by semantic similarity. At the same time, also not as insignificant as ACS and WTS scores. The results for the semantic similarity module are given in table 4. The results show the improvements the model had over the baseline system of the organizers in terms of Pearson Correlation score between the true similarity in the dataset and the scores of the given model.

The comparative results of the experiment with different classifiers (given in Section 5.3) are given in table 5. As observed, the RBF-SVM model and MLP-NN model are better than the other architectures and have similar scores. Hence, these models were chosen for the main architecture.

6 Conclusion

We have proposed a neural network based approach for credibility analysis of unstructured text articles in an open-domain setting. We use a combination of relevant document retrieval techniques with semantic similarity, sentiment analysis and source reliability of articles then reporting the credibility score of the given input. Experiments on Snopes data demonstrate the effectiveness of our approach compared to the previous approaches to the problem.

Experiment on contribution of modules (given in Section 5.2) shows that the website and author reliability are helpful, but only to a limited extent. The accuracy scores had a difference of about 1% . Semantic similarity and document

retrieval are the major contributors to the system. The exclusion of the semantic similarity module resulted in a slash of 32.7%. Sentiment Analysis contributed 7.1% accuracy. Hence it can be concluded that the most important module is semantic similarity, followed by sentiment analysis, and then author, website scores.

The experiments also show that the choice of the classifier plays a major role in the output. Support Vector Methods with RBF kernel and Neural Network architecture (Multilayer Perceptron) give the best results for the problem. This is because the data points are not trivially classifiable and require non-linear classification. The neural network solves it by using multiple lines to classify the dataset whereas SVM utilizes the non-linear RBF kernel to address the problem.

As part of future work, we would like to apply this system to domains like social media. Also, we would like to enhance CREDO with more handcrafted features like the writing style.

References

1. Amari, S.i., Wu, S.: Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* 12(6), 783–789 (1999)
2. Barrios, F., López, F., Argerich, L., Wachenchauser, R.: Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606* (2016)
3. Best, C., van der Goot, E., Blackler, K., Garcia, T., Horby, D.: Europe media monitor. Technical Report EUR221 73 EN, European Commission (2005)
4. Breiman, L.: Random forests. *Machine learning* 45(1), 5–32 (2001)
5. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a” siamese” time delay neural network. In: *Advances in Neural Information Processing Systems*. pp. 737–744 (1994)
6. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2012 workshop on statistical machine translation. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. pp. 10–51. WMT ’12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), <http://dl.acm.org/citation.cfm?id=2393015.2393018>
7. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: *Proceedings of the 20th international conference on World wide web*. pp. 675–684. ACM (2011)
8. Clough, P., Stevenson, M.: Developing a corpus of plagiarised short answers. *Language Resources and Evaluation* 45(1), 5–24 (2011)
9. Das, A., Yenala, H., Chinnakotla, M., Shrivastava, M.: Together we stand: Siamese networks for similar question retrieval. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. vol. 1, pp. 378–387 (2016)
10. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. pp. 22–30. Springer (2004)
11. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems* (2017)

12. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 2333–2338. ACM (2013)
13. John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. pp. 338–345. Morgan Kaufmann Publishers Inc. (1995)
14. Leonhardt, D., Thompson, S.A.: Trump’s lies. *New York Times* 21 (2017)
15. Liu, J., Sun, Y., Zhang, Q., Jiang, H.: Truth discovery: Cotruth (2017)
16. Mathanker, S., Weckler, P., Bowser, T., Wang, N., Maness, N.: Adaboost classifiers for pecan defect classification. *computers and electronics in Agriculture* 77(1), 60–68 (2011)
17. Mihalcea, R., Tarau, P.: Texttrank: Bringing order into texts. *Association for Computational Linguistics* (2004)
18. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd international conference on Knowledge capture. pp. 70–77. ACM (2003)
19. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R.: Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24(4), 694–707 (2016)
20. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2), 1–135 (2008)
21. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 2173–2178. ACM (2016)
22. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In: 26th International Conference on World Wide Web. ACM (2017)
23. Quinlan, J.R.: Induction of decision trees. *Machine learning* 1(1), 81–106 (1986)
24. Riedmiller, M., Lernen, A.M.: Multi layer perceptron. *Machine Learning Lab Special Lecture, University of Freiburg* (2014)
25. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. *Text Mining* pp. 1–20 (2010)
26. Srivastava, S., Gupta, M.R., Frigiyik, B.A.: Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research* 8(Jun), 1277–1305 (2007)
27. Sundermeyer, M., Schlüter, R., Ney, H.: Lstm neural networks for language modeling. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
28. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. *ACL 2014* p. 18 (2014)
29. Weikum, G.: What computers should know, shouldn’t know, and shouldn’t believe. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1559–1560 (2017)