Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings

Gisel Bastidas Guacho
University of California Riverside
gbast001@ucr.edu

Sara Abdali
University of California Riverside
sabda005@ucr.edu

Neil Shah
Snap Inc.
nshah@snap.com

Evangelos E. Papalexakis
University of California Riverside
epapalex@cs.ucr.edu

ABSTRACT

Fake news may be intentionally created to promote economic, political and social interests, and can lead to negative impacts on humans beliefs and decisions. Hence, detection of fake news is an emerging problem that has become extremely prevalent during the last few years. Most existing works on this topic focus on manual feature extraction and supervised classification models leveraging a large number of labeled (fake or real) articles. In contrast, we focus on content-based detection of fake news articles, while assuming that we have a *small* amount of labels, made available by manual fact-checkers or automated sources. We argue this is a more realistic setting in the presence of massive amounts of content, most of which cannot be easily fact-checked. So, we represent collections of news articles as multi-dimensional tensors, leverage tensor decomposition to derive concise article embeddings that capture spatial/contextual information about each news article, and use those embeddings to create an article-by-article graph on which we propagate limited labels. Results on real-world datasets show that our method performs on par or better than existing fully supervised models, in that we achieve better detection accuracy using fewer labels. In particular, our proposed method achieves 75.43% of accuracy using only 30% of labels of a public dataset while an SVM-based classifier achieved 67.43%. Furthermore, our method achieves 70.92% of accuracy in a large dataset using only 2% of labels.

KEYWORDS

Fake news, tensor decomposition, semi-supervised learning, belief propagation.

1 INTRODUCTION

Misinformation on the web is a problem that has been greatly amplified by the use of social media, and the problem of fake news in particular has become ever more prevalent during the last years. Social media is a common platform for consuming and sharing news, due to its ease-of-use in diffusing content and promoting exposure/discussion. In fact, two-thirds of Americans reported getting some of their news from social media in 2017 [1]. Even though social media has become a news source for its advantages, it is especially vulnerable to the propagation of fake news mostly coming from unverified publishers and crowd-based content creators because there is practically no control over the information that is shared. The well-documented spread of misinformation on Twitter during events such as Hurricane Sandy in 2012 [4], the Boston Marathon blasts in 2013 [3] and US Presidential Elections on Facebook in 2016 [22] are all such examples. Since misinformation is intentionally created forjavascript:void(0); malicious purposes such as obtain economic and political benefits or deceiving the public [21], it can clearly lead to negative user experience by either influencing their beliefs and impacting their decisions for the worse. Several approaches in recent literature have been proposed to automatically detect misinformation using supervised classification models. Some works extract manually crafted features from news content such as the number of nouns, length of the article, fraction of positive/negative words, and more in order to discriminate fake news articles [7, 9, 18]. In addition to these works, several others proposed propagation-based models for evaluating news credibility [5, 11, 12]. Nonetheless, they initialized credibility values for the entire network using a supervised classifier. However, the reality is that such labels are often very limited and sparse. Fact-checking websites such as Snopes.com, PolitiFact.com, and FactCheck.org can be used to assess claims, but these websites require domain experts to assign credibility values to claims and are therefore, limited by human capacity. Moreover, fact-checking is a time-consuming process, often requiring surveying multiple articles and sources, evaluating reputation and likelihood of the claims before coming to a decision.

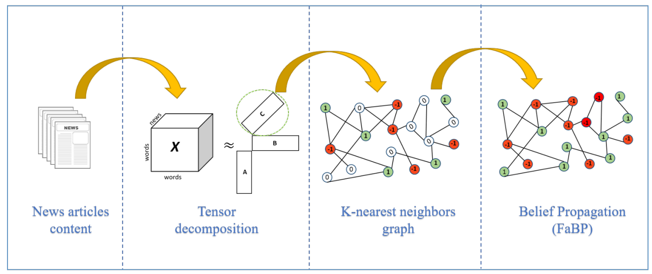

Figure 1: Our proposed method discerns real from misinformative news articles via leveraging tensor representation and semi-supervised learning in graphs.

In this paper, we propose a new *semi-supervised* approach for fake news article detection based on news content, which requires *limited* labels. On a high level, our approach exploits tensor representation and decomposition of news articles, careful construction

[1]http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

of a k-nearest neighbor graph, and propagation of limited labeled article information to conduct inference on a larger set.

Our main contributions are:

- We leverage tensor-based article embeddings, which are shown to produce concise representations of articles with respect to their spatial context, in order to derive a graph representation of news articles.
- We formulate fake news detection as a semi-supervised method that propagates known labels on a graph to determine unknown labels.
- We collect a large dataset of misinformation and real news articles publicly shared on social media.
- We evaluate our method on real datasets. Experiments on two previously used datasets demonstrate that our method outperforms prior works since it requires a fewer number of known labels and achieves comparable performance.

2 PROBLEM DEFINITION

We consider a misinformative, or fake, news article as one that is *"intentionally and verifiably false,"* following the definition used in [21]. With this definition in mind, we aim to discern fake news articles from real ones based on their content. Henceforth, by "content," we refer to the text of the article. We reserve the investigation of other types of content (such as image and video) for future work.

Let $\mathcal{N} = \{n_1, n_2, n_3, ..., n_M\}$ be a collection of news articles of size M where each news article is a set of words and $\mathcal{D} = \{w_1, w_2, w_3, ..., w_I\}$ be a dictionary of words of size I. Note that articles can have varying length. Assuming that labels of some news articles are available. Let $l \in \{-1, 0, 1\}$ denote a vector containing the partially known labels, such that entries of 1 represent real articles, -1 represents fake articles and 0 denotes an unknown status. We address the problem as a binary classification problem; hence, a news article is classified either fake or real.

3 PROPOSED METHOD

Our proposed method consists of the following steps:

Step 1: Tensor Decomposition.

A tensor is a multidimensional array where its dimensions are referred as modes. The most widely used tensor decomposition is Canonical Polyadic (CP) or PARAFAC decomposition [8]. CP/PARAFAC factorizes a tensor into a sum of rank-one tensors. For instance, a three-mode tensor is decomposed into a sum of outer products of three vectors:$\boldsymbol{\mathcal{X}} \approx \Sigma_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$ where $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, $\mathbf{c}_r \in \mathbb{R}^K$ and the outer product is given by $(\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r)(i, j, k) = \mathbf{a}_r(i)\mathbf{b}_r(j)\mathbf{c}_r(k)$ for all i, j, k [16].

We build similar tensor-based article embeddings as proposed in [10]. Specifically, we propose the use of *binary-based tensor* construction method. That is, we build a three-mode tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times I \times M}$ $(words, words, news)$ where for each news article, we create a co-occurrence matrix where all co-occurrence entries are boolean and indicate $(word_1, word_2)$ appeared within a window parameter of w (5-10) words [2] at least once. We then use CP/PARAFAC tensor decomposition [8] to factorize the tensor. As [10] demonstrates, such tensor-based article embeddings captures spatial/contextual nuances of different types of news articles and result in homogeneous article groups. After decomposing the tensor, we obtain the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ whose columns correspond to different latent topics, clustering news articles and words in the latent topic space. More specifically, each row of $\mathbf{C}$ is the representation of the corresponding article in the resulting embedding space.

Step 2: k-NN graph of news articles. The k-nearest-neighbors of a point in n-dimensional space are defined using a "closeness" relation where proximity is often defined in terms of a distance metric [6] such as Euclidean ℓ_2 distance. We use the factor matrix $\mathbf{C}$ in order to construct a k-NN graph G of news articles. As we mentioned before, each column in $\mathbf{C}$ is the representation of the corresponding news article in the latent topic space; thus, by constructing a k-NN graph on $\mathbf{C}$, we can find similar articles in that space. So , we consider each row in $\mathbf{C} \in \mathbb{R}^{M \times R}$ as a point in R-dimensional space. We then compute ℓ_2 distance among news and find the k-closest points for each point in $\mathbf{C}$.

Step 3: Belief Propagation.

Using the graphical representation of the news articles above, and considering that for a small set of those news articles we have ground truth labels, our problem becomes an instance of semi-supervised learning over graphs. We use a belief propagation algorithm which assumes homophily, because news articles that are connected in the k-NN graph are likely to be of the same type due to the construction method of the tensor embeddings; moreover, [10] demonstrates that such embeddings produces fairly homogeneous article groups. More specifically, we use the fast and linearized FaBP variant proposed in [13].

4 EXPERIMENTAL EVALUATION

We implemented our method in MATLAB using Tensor Toolbox [1] and used the MATLAB FaBP implementation [13].

4.1 Dataset description

We use the following three datasets:

Public datasets The two public datasets were used in previous studies. Specifically, *Dataset1* consists of 150 political news articles, balanced to have 75 articles of each class, and was provided by [9]. *Dataset2* contains 68 real and 69 fake news articles, and was provided by [7].Our dataset contains of 31,739 articles of different categories of fake news such as Fake,Conspiracy,Rumor,Satire and Junk Science.

Our dataset In constructing our dataset, we collected news article URLs from Twitter tweets during a 3-month period from June-August 2017. These URLs were filtered based on website domain. We then crawled those URLs to get the news article content. To that end, we used web API boilerpipe[3], Python library Newspaper3k [4], and Diffbot [5]. All real news articles were featured on 367 domains obtained from Alexa [6], and fake news articles belong to 367 domains from the BSDetector browser extension domain list [2].

We measure Accuracy, Precision, Recall and $F1\_score$.

[2]We experimented with small values of that window and results were qualitatively similar.

[3]http://boilerpipe-web.appspot.com/

[4]http://newspaper.readthedocs.io/en/latest/

[5]https://www.diffbot.com/dev/docs/article/

[6]https://www.alexa.com/

4.2 Evaluation

For evaluation, we measure Accuracy, Precision, Recall and $F1\_score$. In order to find the best-performing parameters for our method, we run an iterative process using cross-validation where we evaluated different settings with respect to R (i.e. decomposition rank) and k (i.e. the number of nearest neighbors, controlling the density of the k-NN graph). We considered values of R from 1 to 20, since decomposition rank is often set to be low for time and space reasons in practice [20]. Likewise, we tested k with values from 1 to 100, trading off greater bias for less variance with increasing k. We found that the best accuracy is obtained when both parameters R and k are set to be 10. We find that for values of k and R greater than 10, performance is qualitatively similar as shown in Figure 2, and thus we fix the parameters as such in evaluation. Notice that using a small k value (e.g. 1 or 2), the accuracy is relatively poor; this is because building a k-NN graph with small k results in a highly sparse graph which offers limited propagation capacity. In all experiments, we tested accuracy over the test set of all articles whose labels were "unknown" or unspecified in the propagation step.

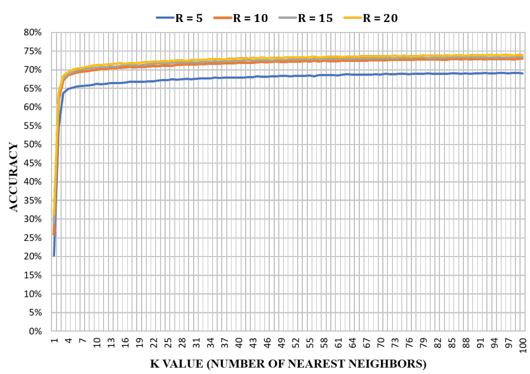

Figure 2: Performance using different parameter settings for decomposition rank (R) and number of nearest neighbors (k).

We evaluated our method with different percentages p of known labels. Table 1 shows the performance of our method using $p \in \{5\%, 10\%, 20\%, 30\%\}$ of labeled news articles from our dataset. Our results demonstrate that we can achieve an accuracy of 70.76% only using 10% of labeled articles. We also evaluated the performance of our approach using extremely sparse known labels. That is, we evaluated our method using $p < 5\%$ and varying the number of nearest neighbors. Figure 3 shows that we can achieve an accuracy of 70.92% using 2% of known labels when the number of nearest neighbors is set to be 200. In fact, the performance of our approach degrades fairly gracefully with even smaller proportions of known labels.

In addition, we evaluated our model using **Dataset1** and **Dataset2**. We compared the accuracy achieved for **Dataset1** to the results reported by [9]. To this extent, we extracted features from news content and use SVM classification as proposed in their work in order to show the performance using different percentages of data to train SVM. Figure 5 shows our approach demonstrates improved accuracy even with fewer labels – specifically, we achieved

Table 1: Performance of the proposed method using our dataset with different percentages of labeled news.

%Labels	Accuracy	Precision	Recall	F1
5%	69.12 ± 0.003	69.09 ± 0.004	69.24 ± 0.009	69.16 ± 0.004
10%	70.76 ± 0.003	70.59 ± 0.003	71.13 ± 0.010	70.85 ± 0.004
20%	72.39 ± 0.001	71.95 ± 0.002	73.32 ± 0.004	72.63 ± 0.002
30%	73.44 ± 0.001	73.13 ± 0.003	74.14 ± 0.003	73.63 ± 0.001

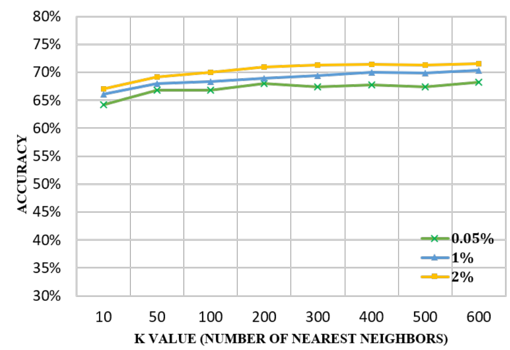

Figure 3: Performance using extremely sparse (<5%) labeled articles and varying number of nearest neighbors.

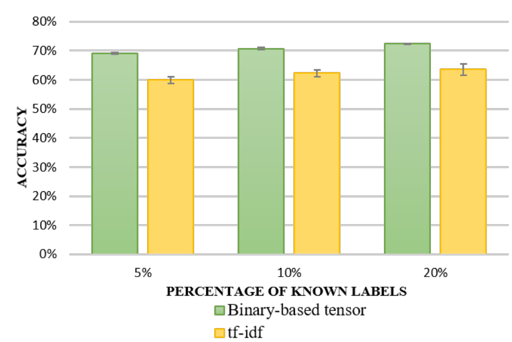

Figure 4: Performance our method using tensor-based article embedding compared to using a graph built form tf-idf matrix.

75.43% accuracy using only the 30% of news labels while in [9], authors reported an accuracy of 71% using 80% of labels in 5-fold cross-validation. In addition, we evaluated our model using *Dataset1* and *Dataset2*. We compare the accuracy achieved by our method to the accuracy achieved by the following approaches: *SVM on content-based features* as proposed in [9]. To this extent, we replicated the feature extraction from news content and used SVM in order to show the performance using different percentages of training data.

Logistic regression on content-based features proposed by [7]. We used their publicly available implementation. In particular, we run their method with linguistic (n-gram) feature extraction using different percentages of training data.

Figure 5 shows the results for *Dataset1*. Our approach demonstrates improved accuracy even with fewer labels – specifically, we achieved 75.43% accuracy using only the 30% of news labels while SVM(30%/70% train/test), SVM(5-fold cross-validation), and logistic regression (30%/70% train/test) attained 67.43%, 71% and 50.09%

of accuracy, respectively. The accuracy achieved by SVM(5-fold cross-validation) was reported by Horne et al. in [9].

For *Dataset2*, we run logistic regression and SVM, using 10%/90% train/test split. These approaches achieved an accuracy of 59.84% and 64.79%, respectively, compared to the 67.38% accuracy achieved by our approach, using the same percentage of labeled articles.

We note that our method is able to achieve this performance only having a small number of labeled news articles due to the quality of the tensor embeddings which define a favorable graph where the node labels are propagated.

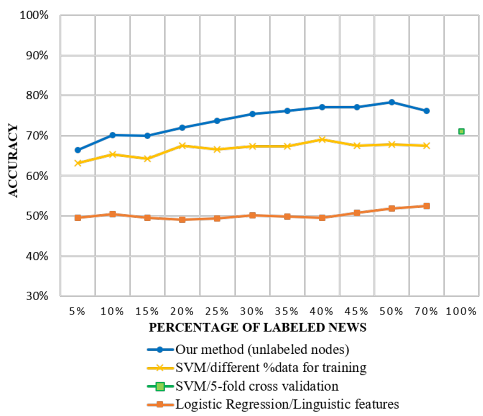

Figure 5: Performance using *Dataset1* provided by Horne et al. [9]

5 RELATED WORK

In [7], the authors proposed a logistic regression classifier using linguistic (*n*-gram), credibility (punctuation, pronoun use, capitalization) and semantic features generated from the news content. [9] used SVM on content-based features that are categorized into stylistic, complexity and psychological features in order to classify real, fake and satirical news. In [17], the authors propose detecting rumors by building naïve-Bayes classifiers on content, network and microblog-specific features. [14] and [19] leverage temporal structure by using recurrent neural network (RNN) based models to represent text and user characteristics. In [15], the authors propose a Dynamic Series-Time Structure (DSTS) model for detecting rumors by capturing the social context of an event from content, user and propagation-based features.

6 CONCLUSIONS

In this paper, we propose a semi-supervised content-based method for detecting misinformative news articles, leveraging ensor-based article embeddings and guilt-by-association. Extensive experiments on over 63K real articles demonstrate that our method distinguish efake from real news only using a small number of labeled articles, with performance on par or better than state-of-the-art.

7 ACKNOWLEDGEMENTS

Research was supported by a gift from Snap Inc and an Adobe Data Science Research Faculty Award Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

REFERENCES

[1] T. G. Kolda B. W. Bader et al. 2015. Matlab tensor toolbox version 2.6. Available online. (February 2015).

[2] B.S. Detector. 2017. http://bsdetector.tech/. (2017).

[3] A. Gupta, H. Lamba, and P. Kumaraguru. 2013. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter. In *2013 APWG eCrime Researchers Summit*. 1–12. DOI: http://dx.doi.org/10.1109/eCRS.2013.6805772

[4] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13 Companion)*. ACM, New York, NY, USA, 729–736. DOI: http://dx.doi.org/10.1145/2487788.2488033

[5] Manish Gupta, Peixiang Zhao, and Jiawei Han. *Evaluating Event Credibility on Twitter.* 153–164. DOI: http://dx.doi.org/10.1137/1.9781611972825.14 arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611972825.14

[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[7] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In Search of Credible News. *Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2016 Lecture Notes in Computer Science* (2016), 172fi?!180. DOI: http://dx.doi.org/10.1007/978-3-319-44748-3_17

[8] R. A. Harshman. 1970. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics* 16, 1 (1970), 84.

[9] Benjamin D. Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *CoRR* abs/1703.09398 (2017). arXiv:1703.09398 http://arxiv.org/abs/1703.09398

[10] Seyedmehdi Hosseinimotlagh and Evangelos E. Papalexakis. 2017. Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles.

[11] Z. Jin, J. Cao, Y. G. Jiang, and Y. Zhang. 2014. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. In *2014 IEEE International Conference on Data Mining*. 230–239. DOI: http://dx.doi.org/10.1109/ICDM.2014.91

[12] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. (2016). https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12128

[13] Danai Koutra, Tai-You Ke, U. Kang, Duen Chau, Hsing-Kuo Pao, and Christos Faloutsos. 2011. Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*. Lecture Notes in Computer Science, Vol. 6912. 245–260.

[14] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 3818–3824. http://dl.acm.org/citation.cfm?id=3061053.3061153

[15] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1751–1754. DOI: http://dx.doi.org/10.1145/2806416.2806607

[16] Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. 2016. Tensors for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms. *ACM Trans. Intell. Syst. Technol.* 8, 2, Article 16 (Oct. 2016), 44 pages. DOI: http://dx.doi.org/10.1145/2915921

[17] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor Has It: Identifying Misinformation in Microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1589–1599. http://dl.acm.org/citation.cfm?id=2145432.2145602

[18] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News.

[19] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News. *CoRR* abs/1703.06959 (2017). arXiv:1703.06959 http://arxiv.org/abs/1703.06959

[20] Neil Shah, Alex Beutel, Brian Gallagher, and Christos Faloutsos. 2014. Spotting suspicious link behavior with fbox: An adversarial perspective. In *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 959–964.

[21] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *CoRR* abs/1708.01967 (2017). arXiv:1708.01967 http://arxiv.org/abs/1708.01967

[22] Craig Silverman. 2016. This Analysis Shows how Fake Election News Stories Outperformed Real News on Facebook. BuzzFeed News.