The Case for Latent Variable vs Deep Learning Methods in Misinformation Detection: An Application to COVID-19

Caitlin Moroney[1], Evan Crothers[2], Sudip Mittal[3], Anupam Joshi[4], Tülay Adalı[4], Christine Mallinson[4], Nathalie Japkowicz[1], and Zois Boukouvalas[1]

[1] American University, Washington, D.C. 20016, USA
cm0246b@student.american.edu
{boukouva,japkowic}@american.edu
[2] University of Ottawa, Ottawa, ON, Canada
ecrot027@uottawa.ca
[3] Mississippi State University, Mississippi State, MS 39762, USA
mittal@cse.msstate.edu
[4] University of Maryland, Baltimore County, Baltimore, MD 21250, USA
{joshi,adali,mallinson}@umbc.edu

Abstract. The detection and removal of misinformation from social media during high impact events, e.g., COVID-19 pandemic, is a sensitive application since the agency in charge of this process must ensure that no unwarranted actions are taken. This suggests that any automated system used for this process must display both high prediction accuracy as well as high explainability. Although Deep Learning methods have shown remarkable prediction accuracy, accessing the contextual information that Deep Learning-based representations carry is a significant challenge. In this paper, we propose a data-driven solution that is based on a popular latent variable model called Independent Component Analysis (ICA), where a slight loss in accuracy with respect to a BERT model is compensated by interpretable contextual representations. Our proposed solution provides direct interpretability without affecting the computational complexity of the model and without designing a separate system. We carry this study on a novel labeled COVID-19 Twitter dataset that is based on socio-linguistic criteria and show that our model's explanations highly correlate with humans' reasoning.

Keywords: Misinformation Detection · Knowledge Discovery · Independent Component Analysis · Explainability.

1 Introduction

With the evolution of social media, there has been a fundamental change in how misinformation is propagated especially during high impact events. A recent example of a high impact event is the COVID-19 disease where misinformation is dangerously spreading and includes conspiracy theories, harmful health advises, misinformation related to and racism, among many others.

Recent machine learning advances have shown significant promise for the detection of misinformation. Examples include approaches based on hand-crafted features and approaches based on deep learning. The idea behind hand-crafted approaches is to define features from textual information in order to capture certain characteristics of misinformation and then train a separating hyper-plane by using a selected classifier [23,12,33]. Although these approaches provide at a certain level interpretable results, in most cases the selection of the features is tied to the particular application affecting the generalization ability of the model. On the other hand, approaches based on deep learning effectively learn the latent representations and have shown great promise in terms of prediction performance [24,25,29,26,9]. However, the connections between high level features and low representation space are usually accessed by using or designing a separate system resulting in high computational or construction overhead [13]. In this study, we present a computationally efficient data-driven solution that is based on a latent variable model called independent component analysis (ICA) such that detection of misinformation and knowledge discovery can be achieved jointly. Our method achieves a prediction performance close to that of deep learning while at the same time offering the kind of interpretability that deep learning, even with the help of a separate explainability system, cannot achieve.

This work makes several contributions. First, it proposes a new method for misinformation detection based on ICA. Second, it demonstrates how to highlight the connections between the low dimensional representation space and the high level features. This enables researchers to ensure that their methods are truly learning to represent relevant information from the data and enables them to understand in an efficient manner the causes of the decisions made by a classification algorithm. Finally, it makes available a new labeled and annotated COVID-19 Twitter dataset[5] as well as a set of rules for label generation based on socio-linguistic criteria.

2 Development of Labeled Twitter COVID-19 Dataset

In constructing our labeled Twitter dataset we initially randomly collected a sample of 282,201 Twitter users from Canada by using the Conditional Independence Coupling (CIC) method [30]. CIC matches the prior distribution of the population, in this case the Canadian general population, ensuring that the sample is balanced for gender, race and age. All tweets posted by these 282,201 people from January 1, 2020 to March 13, 2020 were collected and a random subset of 1,600 tweets was further analyzed to create a manageable and balanced dataset of both real tweets and tweets that contain misinformation. Note here that we follow current literature that defines misinformation as *an umbrella term to include all false or inaccurate information that is spread in social media.* This is a useful heuristic because, on a social media platform where any user can publish anything, it is otherwise difficult to determine whether a piece of misinformation is deliberately created or not. In addition, more specific categories

[5] Dataset is available at `https://zoisboukouvalas.github.io/Code.html`

Table 1. 17 linguistic characteristics identified on the 560 Twitter dataset

Linguistic Attribute	**Example from Dataset**
Hyperbolic, intensified, superlative, or emphatic language [2,20]	e.g., 'blame', 'accuse', 'refuse', 'catastrophe', 'chaos', 'evil'
Greater use of punctuation and/or special characters [2,19]	e.g., 'YA THINK!!?!!?!', 'Can we PLEASE stop spreading the lie that Coronavirus is super super super contagious? It's not. It has a contagious rating of TWO'
Strongly emotional or subjective language [2,28,31,20,18]	e.g., 'fight', 'danger', 'hysteria', 'panic', 'paranoia', 'laugh', 'stupidity' or other words indicating fear, surprise, alarm, anger, and so forth
Greater use of verbs of perception and/or opinion [19]	e.g., 'hear', 'see', 'feel', 'suppose', 'perceive', 'look', 'appear', 'suggest', 'believe', 'pretend'
Language related to death and/or war [11]	e.g., 'martial law', 'kill', 'die', 'weapon', 'weaponizing'
Greater use of proper nouns [14]	e.g., 'USSR lied about Chernobyl. Japan lied about Fukushima. China has lied about Coronavirus. Countries lie. Ego, global'
Shorter and/or simpler, language [14]	e.g., '#Iran just killed 57 of our citizens. The #coronavirus is spreading for Canadians Our economy needs support.'
Hate speech [11] and/or use of racist or stereotypical language	e.g., 'foreigners', 'Wuhan virus', reference to Chinese people eating cats and dogs
First and second person pronouns [20,19]	e.g., 'I', 'me', 'my', 'mine', 'you', 'your', 'we', 'our'
Direct falsity claim and/or a truth claim [2]	e.g., 'propaganda', 'fake news', 'conspiracy', 'claim', 'misleading', 'hoax'
Direct health claim	e.g., 'cure', 'breakthrough', posting infection statistics
Repetitive words or phrases [14]	e.g., 'Communist China is lying about true extent of Coronavirus outbreak - If Communist China doesn't come clean'
Mild or strong expletives, curses, slurs, or other offensive terms	e.g., 'bitch', 'WTF', 'dogbreath', 'Zombie homeless junkies', 'hell', 'screwed'
Language related to religion	e.g., 'secular', 'Bible'
Politically biased terms	e.g., 'MAGA', 'MAGAt', 'Chinese regime', 'deep state', 'Communist China'
Language related to financial or economic impact	e.g., 'THE STOCK MARKET ISN'T REAL THE ECONOMY ISN'T REAL THE CORONAVIRUS ISN'T REAL FAKE NEWS REEEEEEEEEEEEEEEEE'
Language related to the Trump presidential election, campaign, impeachment, base, and rallies	e.g., 'What you are watching with the CoronaVirus has been planned and orchestrated. '

(e.g. fake news, rumor, disinformation) often overlap and are not exclusive [32]. Two subject matter experts from our group independently reviewed each these tweets.

Tweets were labeled as misinformation if 1) they include content that promotes *political bias*, *conspiracy*, *propaganda*, *anger*, or *racism* and thus could affect decision making and create social and political unrest during COVID-19 and 2) they were labeled as misinformation tweets by both experts. This resulted in 280 misinformation tweets. To create a balanced dataset, the true class was randomly down-sampled to 280 and each tweet was checked for consistency and validity with respect to reliability.

The set of tweets that contains misinformation was further analyzed for the presence of linguistic attributes that might indicate unreliability and provide a

set of linguistic rules of potential use to label further data sets and to assess the interpretation ability of our model. This was done by reviewing each tweet for, first, the presence of linguistic characteristics previously identified in the literature as being indicative of or associated with misinformation, bias, and/or less reliable sources in news media; and second, for the presence of any additional distinguishing linguistic characteristics that appeared to be indicative of misinformation in this dataset. A list of 17 linguistic characteristics was developed and is presented in Table 1 along with instances of each characteristics drawn from the dataset.

3 Tweet Representations Generation

3.1 Transformer Language Models

For comparison to state-of-the-art deep learning methods, we compare our results against a suite of Transformer language models. Specifically, we evaluate "base" and "large" variants of Bidirectional Encoder Representations from Transformers (BERT) [10], Robustly Optimized BERT Pretraining Approach (RoBERTA) [17], and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [6].

The bidirectional aspect of BERT comes from using a masked language model (MLM) pre-training objective, which allows the model to incorporate information from both the left and right contexts. Upon release, BERT was shown to achieve state-of-the-art results across eleven natural language processing tasks, including sentence-level and token-level tasks [10]. RoBERTA and ELECTRA represent subsequent improvements of these state-of-the-art results through further investigation of improved pre-training methods [6][17]. To facilitate reproducibility, we use the pre-trained HuggingFace PyTorch implementation of each of these models. The base version of each model contains 12 encoder layers, 768 hidden units, and 12 attention heads (for a total of over 110M parameters), while the large version of each model contains 24 encoder layers, 1024 hidden units, and 16 attention heads (for a total of over 330M parameters).

The classification model takes as input the Transformer output encoding; for each tweet, we obtain a vector of length equal to the dimension of the final hidden layer of the model (i.e., 768 for base models and 1024 for large models). Aggregate representations for the input sequence (in our experiments, a tweet) can be generated by taking the output vector of a `[CLS]` token prepended to the input sequence. However, we found that taking the mean across the vectors for all of a segment's tokens provided a better classification performance, mirroring results from other research into BERT-based sequence representations [21]. As such, our sequence representations are generated through taking the mean across all of a segment's token vectors.

3.2 Independent Component Analysis

We formulate the problem of tweet representation generation for detection of misinformation posts in the following way. Let $\mathbf{X} \in \mathbb{R}^{d \times V}$ be the observation

matrix which denotes the word-word co-occurrence matrix and incorporates contextual information from the raw text data. The model is given by $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A} \in \mathbb{R}^{d \times N}$ is the mixing matrix and $\mathbf{S} \in \mathbb{R}^{N \times V}$ are the latent variables. Since in our study $d > N$, we reduce to the case where $d = N$ using principal component analysis (PCA).

The most popular way to estimate the latent variables is by using ICA [1,7,15]. The assumption of independence of the sources not only enables a unique decomposition under minimal model assumptions but also results in interpretable contextual representations through their linear mixing coefficients. The estimated columns of the mixing matrix $\mathbf{A}$ denote the weight feature vectors that will be used for the construction of the tweet representations. An estimate of $\mathbf{A}$ is computed as $\hat{\mathbf{A}} = (\mathbf{F})^{\dagger}(\mathbf{W})^{-1}$, where $(\mathbf{F})^{\dagger}$ denotes the pseudo-inverse of the matrix that is formed by the eigen-vectors with the first N highest eigenvalues of $\mathbf{X}$ and $\mathbf{W}$ is the estimated demixing matrix resulted from ICA. Though there are many ICA algorithms, in this work, we used the entropy bound minimization (ICA-EBM) algorithm [16], due to the fact that it has shown superior performance in a wide range of applications. To construct the individual tweet representations, we average over the estimated rows of $\mathbf{A}$ for the words in each tweet to obtain a single N-dimensional vector representation for each tweet. For our study we have selected $N = 250$.

4 Results and Discussion

4.1 Prediction Performance

We evaluate the same classification algorithm, Support Vector Machines (SVM) [8] using a linear kernel[6]. For completeness, in addition to the BERT versions and ICA tweet representations, we consider three other popular latent variable methods: Non-negative matrix factorization (NMF) [3], Dictionary Learning (DL) [27], and Latent Dirichlet allocation (LDA) [4]. To construct tweet representations using NMF, DL, and LDA we followed a similar procedure as we did with ICA. To measure performance, we employed the standard suite of evaluation metrics, i.e., accuracy, F1 score, precision, and recall. We report the macro-averaged versions of these scores. For all of the experiments, hyper-parameter optimization and model training and testing is done using a nested five fold cross validation scheme.

From Table 4.1, we see that prediction accuracy using RoBERTa-Large word representations performs the best in terms of accuracy and F1 score. However, the ICA method is able to achieve very similar performance to that of RoBERTa-Large and in some cases better than other BERT versions such as ELECTRA-Base and ELECTRA-Large. Performance using tweet representations derived from NMF, DL, and LDA was significantly lower than that of ICA and BERT across all metrics.

[6] It is worth mentioning that for all methods similar results were obtained with the sigmoid and and the rbf kernel.

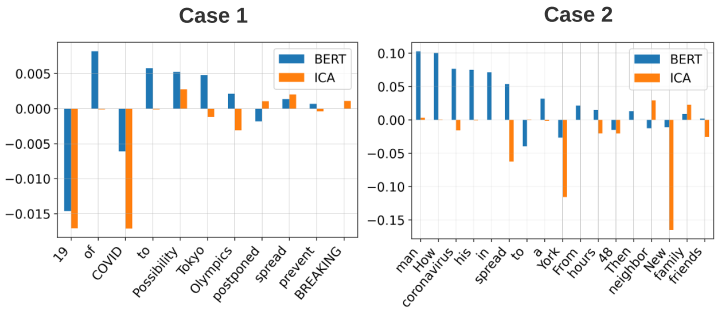
Method	Accuracy	Recall	Precision	F1
LDA	0.754	0.868	0.712	0.777
ICA	**0.862**	**0.931**	**0.820**	**0.871**
DL	0.779	0.740	0.799	0.767
NMF	0.832	0.872	0.810	0.838
BERT-Base-Uncased	0.868	0.883	0.858	0.869
BERT-Base-Cased	0.866	0.889	0.855	0.870
BERT-Large-Uncased	0.880	0.888	0.877	0.881
BERT-Large-Cased	0.875	**0.895**	0.863	0.876
ELECTRA-Base	0.848	0.837	0.861	0.847
ELECTRA-Large	0.832	0.838	0.828	0.832
RoBERTa-Base	0.873	0.862	**0.888**	0.873
RoBERTa-Large	**0.886**	0.891	0.883	**0.886**

4.2 Explainability

ICA has the advantage over Deep Learning techniques of being able to provide contextual interpretations through the estimated mixing matrix $\hat{\mathbf{A}}$. It does so by first ordering the ICA features by magnitude for a given tweet vector representation. For each tweet, the most important features, which may be considered as topics, are then extracted. From the matrix $\hat{\mathbf{A}}$, we then select the columns corresponding to the most important topics for the chosen tweet, and for each column we sort the rows, corresponding to vocabulary words, by magnitude. This allows us to obtain the most important words in each topic for the most important topics in each tweet.

The results of this extraction are shown at the bottom of Figures 1 and 2. Feature 1 and Feature 2 represent the dominant words belonging to the highest two features extracted by ICA on 2 real and 2 misinformation tweets that were all classified correctly by the ICA-based method. From Figure 1 and Figure 2 we see that the words listed in the main features of the two real cases (Cases 1 and 2) do not match the rules extracted in Table 1, except for one: hyperbolic language ("apocalyptic"). This suggests that the two main features extracted in each of these two real cases support the classifier's decision, since the words most strongly associated with the features that caused that decision do not trigger many rules believed to represent language used in misinformation. On the other hand, in the two misinformation cases (Cases 3 and 4), many rules are triggered including hate speech, hyperbolic language, and strongly emotional language. This suggests that the two main features extracted in each of these misinformation cases support the classifier's decision, since the words most strongly associated with the features that caused that decision do trigger many rules believed to represent language used in misinformation. Furthermore, the fact that the two closely related misinformation tweets of Cases 3 and 4 triggered the same two features shows the consistency of our approach. Figure 3, (Case 5), shows the explainability results for the case where the ICA-based method did not classify the tweet correctly. In this case, we see that the second main feature is the same as the one that was picked by the two unreliable tweets en-

abling a user to understand why the ICA-based method predicted this tweet as misinformation.



Case 1

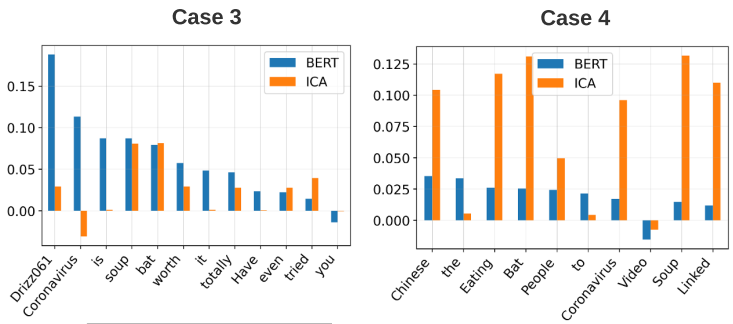
Feature 1	Feature 2
podcast	downsize
road	ceremony
10th	torch
5	olympics
24thminute	classifies
kevlaramee	apocalyptic
ring	consequence
possibility	kuwait
cancellation	vega
tokyo	juno

Case 2

Feature 1	Feature 2
seriously	delay
etc	c.d.c
friend	n.y.
distance	cuomo
coronavirusoutbreak	overnight
excite	double
age	york
social	jersey
im	76
globalhealth	attack

Fig. 1. Top: Local explanations by LIME for both BERT and ICA. Orange bars correspond to the ICA-based approach and the blue bars correspond to the BERT-based approach. Furthermore, 0 is neutral whereas positive values are misinformation and negative values real; Bottom: Feature 1 and Feature 2 represent the dominant words belonging to the highest two features extracted by ICA on two real tweets that ICA-based method correctly classified; Tweets: **Case 1 (predicted as real by ICA and by BERT)**: BREAKING: Possibility Tokyo Olympics postponed to prevent spread of COVID-19; **Case 2 (predicted as real by ICA and as misinformation by BERT)**: From a man to his family. Then to a neighbor. Then to friends. How coronavirus spread in New York in 48 hours.

While, as just discussed, the context in which a decision is made can easily be extracted from the ICA-based method, in recent years, efforts have been made to extract information from opaque classifiers. One such effort is the popular local interpretable model-agnostic explanations (LIME) system [22] which produces local explanations for classifier decisions. This technique, however, comes at a cost since, for example, LIME took, on average, 6,400.1 seconds to process a single tweet explanation for the BERT and SVM pipeline and 70.3 seconds for the ICA and SVM pipeline whereas the extraction of ICA's main features was instantaneous. In addition to the cost, we argue that LIME does not consistently



Case 3

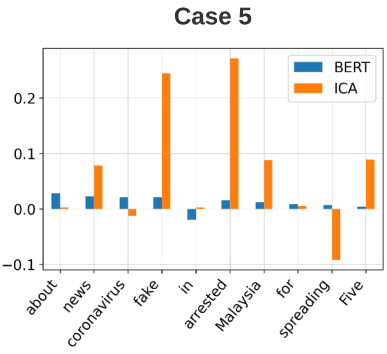
Feature 1	Feature 2
ting	chinavirus
tho	animal
99	barbaric
soup	wuhansars
bat	sicken
worth	wan
totally	eat
drizz061	na
found	chinese
troll	soup

Case 4

Feature 1	Feature 2
ting	chinavirus
tho	animal
99	barbaric
soup	wuhansars
bat	sicken
worth	wan
totally	eat
drizz061	na
found	chinese
troll	soup

Fig. 2. Top: Local explanations by LIME for both BERT and ICA. Orange bars correspond to the ICA-based approach and the blue bars correspond to the BERT-based approach. Furthermore, 0 is neutral whereas positive values are misinformation and negative values real; Bottom: Feature 1 and Feature 2 represent the dominant words belonging to the highest two features extracted by ICA on two misinformation tweets that ICA-based method correctly classified; Tweets: **Case 3 (predicted as misinformation by ICA and by BERT)**: @Drizz061 Have you even tried bat soup? Coronavirus is totally worth it; **Case 4 (predicted as misinformation by ICA and as real by BERT)**: Chinese People Eating Bat Soup Linked to the Coronavirus-Video

outfit the BERT-based method (or the ICA-based method for that matter) with a satisfying explainability. In particular, looking at the top graphs in Figure 1 and Figure 2, we notice inconsistencies in LIME's explanations. In these graphs, the orange bars correspond to the ICA-based approach and the blue bars correspond to the BERT-based approach. Furthermore, 0 is neutral whereas positive values are misinformation and negative values real. In Case 1, BERT issued the correct classification. However, LIME's explanation for this classification is that Covid and 19 were reliable words, whereas irrelevant stop words such as "Of" and "to" gave the system indication that it was misinformation. In case 2 that BERT wrongly classified as misinformation, that unreliability is given by the irrelevant words "man", "How", "Coronavirus", "his", "in", "spread", "a" etc. This does not inspire confidence in LIME the way ICA's features did for the ICA-based explainability method, since whether BERT issues the correct classification or



Case 5	
Feature 1	**Feature 2**
topublichealth	ting
malaysia	tho
five	99
publishing	soup
abc	bat
rvh	worth
ignore	totally
fall	drizz061
contact	found
bird	troll

Fig. 3. Left: Local explanations by LIME for both BERT and ICA. Orange bars correspond to the ICA-based approach and the blue bars correspond to the BERT-based approach. Furthermore, 0 is neutral whereas positive values are misinformation and negative values real; Right: Feature 1 and Feature 2 represent the dominant words belonging to the highest two features extracted by ICA on one real tweet that ICA classified incorrectly; Tweet: **Case 5 (predicted as misinformation by ICA and as real by BERT)**: Five arrested in Malaysia for spreading fake news about coronavirus

not, the words LIME underlines as reliable or indicative of misinformation are not convincing in any situation (real or misinformation). Furthermore, unlike the features extracted by ICA, LIME does not associate the words of a tweet with words associated in other tweets (local explanations). This makes knowledge discovery and explainability a real challenge, since there is no direct way to associate the linguistic attributes of Table 1 with the words contained in a given misinformation tweet. That is because in misinformation detection, the context in which tweets are written matters a lot and while BERT is capable of using that context to classify tweets, LIME is unable to retrieve the context that was used in its explanations. On the other hand, ICA can simultaneously use the context for classification and make it explicit in its explanations. To illustrate this idea, let‘s take the example of the expression “bat soup”. If the expression was used only in the context of sentences such as “Did you know that in some countries bat soup is a delicacy?” in the corpus, then the context, would yield a classification of “real” as well as ICA features that do not trigger any misinformation rules from Table1, whereas in the context of this corpus, the expression “bat soup” triggers many of these rules as mentioned when discussing the ICA results of Figure2. Furthermore, in the two misinformation cases (Cases 3 and 4), the lack of confidence in BERT is, in fact, supported by the fact that the two closely related tweets which elicited the same main features and classification by ICA received opposite classifications by BERT (correct misinformation for Case 3 and incorrect real for Case 4) and seemed to have decided on misinformation in Case 3 based on the unknown reference: “@Drizz061”, which, in passing, ICA did not give much credence to.

Case 5 in Figure 3, which ICA wrongly classified as unreliable was correctly classified by BERT, but no good reason emerges from the LIME graphs, except for the fact that the values it associated with the words hover over 0, whereas in the case of ICA, LIME picked up on the words "Fake" and "arrested" which is different from the explanation given by the feature and, in some way, does make sense according to Table 1 rules: Direct falsity claim ("Fake") and Language related to Death and/or War ("arrested"), which suggests that ICA features together with LIME give a relatively full picture of ICA's mechanism, but that LIME, which is the only tool (or the only category of tools) that can be used in Deep learning settings does not provide the user with a clear window into its decision making approach. Conversely, we believe that the mistake made by ICA in Case 5 can easily be caught by a human operator who can notice that Feature 2 is only very tenuously related to the Tweet (Fake in the Tweet and troll in the feature, perhaps), and that this association is probably erroneous and should be discarded.

5 Conclusion

Although Deep Learning recently became the approach of choice for practical NLP tasks such as the detection and removal of misinformation from social media, this study argues that latent variable decomposition methods can be quite competitive and come with added advantages: *simplicity*, *efficiency*, and most importantly, *built-in explainability*. After presenting a new Covid-19 misinformation data set, we demonstrate that an ICA- based classification approach is almost as accurate as a BERT-based approaches while efficiently extracting features bearing more resemblance to the socio-linguistic rules used to build the data set than the information extracted by LIME, a state-of-the-art explainability tool.

The success of the proposed method raises several interesting questions that can be explored in future work. Although the assumption that the latent representation space discovered by ICA carries meaningful characteristics and would yield reliable performance in terms of prediction accuracy and explainability, quantitatively evaluating such an assumption is a non-trivial task since ICA is an unsupervised process. Thus, as a future task we propose to create formal settings where humans can evaluate whether a set of extracted feature embeddings have human-identifiable semantic coherence and where humans can evaluate whether the associations between a pre-labeled particular malicious post and a set of feature embeddings make sense. These quantitative methods have been similarly used for measuring semantic meaning in inferred topics [5]. By developing human-based evaluation metrics, we will not only assess the ICA representation space, but more importantly, we will be able to identify potential biases related to certain characteristics of the collected social media posts enabling us to correct our model before it is deployed at scale. Finally, in terms of a computational socio-linguistic perspective, the development of validation techniques for the extracted features and how they can be used to answer questions

that record the behavior and interactions of individuals in virtual worlds is a significant research direction and deserves further investigation.

Acknowledgement

Computing resources used for this work were provided by the American University Zorro High Performance Computing System.
We thank Dr. Kenton White, Chief Scientist at Advanced Symbolics Inc, for providing the initial Twitter dataset.

References

1. Adalı, T., Anderson, M., Fu, G.S.: Diversity in Independent Component and Vector Analyses: Identifiability, algorithms, and applications in medical imaging. IEEE Signal Processing Magazine **31**(3), 18–33 (May 2014). https://doi.org/10.1109/MSP.2014.2300511
2. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., Nakov, P.: Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:1810.01765 (2018)
3. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Computational statistics & data analysis **52**(1), 155–173 (2007)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
5. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems. pp. 288–296 (2009)
6. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
7. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent component analysis and applications. Academic press (2010)
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
9. Cui, L., Wang, S., Lee, D.: Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 41–48 (2019)
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
11. Fatemeh Torabi Asr: The language gives it away: How an algorithm can help us detect fake news. https://theconversation.com/the-language-gives-it-away-how\-an-algorithm-can-help-us-detect\-fake-news-120199 (2019), online
12. Gupta, A., Kumaraguru, P.: Credibility ranking of tweets during high impact events. In: Proceedings of the 1st workshop on privacy and security in online social media. p. 2. ACM (2012)

13. Hansen, L.K., Rieger, L.: Interpretability in intelligent systems–a new concept? In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 41–49. Springer (2019)
14. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Eleventh International AAAI Conference on Web and Social Media (2017)
15. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis, vol. 46. John Wiley & Sons (2004)
16. Li, X., Adali, T.: Independent component analysis by entropy bound minimization. IEEE Transactions on Signal Processing **58**(10), 5151–5164 (Oct 2010). https://doi.org/10.1109/TSP.2010.2055859
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
19. Perez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. arXiv preprint arXiv:1708.07104 (2017)
20. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937 (2017)
21. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier `http://arxiv.org/abs/1602.04938`
23. Shu, K., Bernard, H.R., Liu, H.: Studying fake news via network analysis: Detection and mitigation. CoRR **abs/1804.10233** (2018), `http://arxiv.org/abs/1804.10233`
24. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter **19**(1), 22–36 (2017)
25. Shu, K., Wang, S., Liu, H.: Beyond news contents: The role of social context for fake news detection. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 312–320. ACM (2019)
26. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). pp. 39–47. IEEE (2019)
27. Tošić, I., Frossard, P.: Dictionary learning. IEEE Signal Processing Magazine **28**(2), 27–38 (2011)
28. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
29. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. CoRR **abs/1705.00648** (2017), `http://arxiv.org/abs/1705.00648`
30. White, K., Li, G., Japkowicz, N.: Sampling online social networks using coupling from the past. In: 2012 IEEE 12th International Conference on Data Mining Workshops. pp. 266–272. IEEE (2012)
31. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing. pp. 347–354 (2005)

32. Wu, L., Morstatter, F., Carley, K.M., Liu, H.: Misinformation in social media: definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter **21**(2), 80–90 (2019)
33. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1395–1405. International World Wide Web Conferences Steering Committee (2015)