Interpretable Fake News Detection with Topic and Deep Variational Models

Marjan Hosseini\textsuperscript{a,\ast}, Alireza Javadian Sabet\textsuperscript{b}, Suining He\textsuperscript{a} and Derek Aguiar\textsuperscript{a}

\textsuperscript{a}Department of Computer Science and Engineering, University of Connecticut, Storrs, 06269, CT, USA.
\textsuperscript{b}Department of Informatics and Networked Systems, University of Pittsburgh, Pittsburgh, 15260, PA, USA.

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Fake News
Misinformation
Social Media
Interpretability
Topic Models
Variational Autoencoder

\textbf{ABSTRACT}

The growing societal dependence on social media and user generated content for news and information has increased the influence of unreliable sources and fake content, which muddles public discourse and lessens trust in the media. Validating the credibility of such information is a difficult task that is susceptible to confirmation bias, leading to the development of algorithmic techniques to distinguish between fake and real news. However, most existing methods are challenging to interpret, making it difficult to establish trust in predictions, and make assumptions that are unrealistic in many real-world scenarios, e.g., the availability of audiovisual features or provenance. In this work, we focus on fake news detection of textual content using interpretable features and methods. In particular, we have developed a deep probabilistic model that integrates a dense representation of textual news using a variational autoencoder and bi-directional Long Short-Term Memory (LSTM) networks with semantic topic-related features inferred from a Bayesian admixture model. Extensive experimental studies with 3 real-world datasets demonstrate that our model achieves comparable performance to state-of-the-art competing models while facilitating model interpretability from the learned topics. Finally, we have conducted model ablation studies to justify the effectiveness and accuracy of integrating neural embeddings and topic features both quantitatively by evaluating performance and qualitatively through separability in lower dimensional embeddings.

1. Introduction

The increased availability and consumption of user-generated content on the internet has provided an ideal environment for propagating fake news, or fabricated information that imitates news media. While fake news has existed since at least the 1\textsuperscript{st} century BC (Posetti and Matthews, 2018), the prevalence of online social media (OSM) and the echo chamber effect of content servicing algorithms have significantly increased the number of fake news domains and their associated web traffic (Cinelli et al., 2021). For instance, there was a significant increase in the number of fake news web domains prior to the 2016 U.S. presidential election (Chalkiadakis et al., 2021) and an estimated 41.8\% of their traffic was driven by OSM (Allcott and Gentzkow, 2017). False information is shared by more people and spreads faster than true information, particularly for political topics (Vosoughi et al., 2018). Platforms like Twitter and Facebook are integral to this increased exposure, allowing users to easily share and promote content without being subject to journalistic norms and ethical standards (Posetti and Matthews, 2018). This is in stark contrast with traditional professional journalism that strives for truthfulness, accuracy, objectivity, fairness, and accountability (Society of Professional Journalists, 2021). OSM are also increasingly exploited as a primary news source; 53% – 62% of U.S. adults often or sometimes rely on OSM as their source of daily news (Carminati et al., 2012).

The rise of fake news and OSM as a preferred news medium has been associated with detrimental effects on society (Hindman and Barash, 2018; Lee, 2019). In the political domain, consuming fake news is associated with decreased trust towards news media and increased political polarization (Guess et al., 2020). Based on the effects of media in general, it has been suggested that fake news can encourage extremism, increase cynicism, and apathy (Lazer et al., 2018), and deepens belief in false claims or conspiracy theories (Guess et al., 2020). One mechanism for combating the sharing of false information is to annotate news articles with the \textit{truthfulness} of the underlying claims. Evidence suggests that fake news headlines are considered less accurate when people are warned about the potentially false or misleading nature of the content (Clayton et al., 2020).

Detecting and classifying fake news is therefore a vital goal to curtail its spread and impact. However, we need to carefully address the following two major challenges to detect and disrupt fake news early in its propagation and to
Interpretable Fake News Detection with Topic and Deep Variational Models

Figure 1: Methods overview. Input text samples are first preprocessed (e.g., tokenized, non-English and stop words removed, and oversampled if needed). (Top) We use the word2vec embeddings as input to the bi-directional LSTM VAE model to extract a latent representation for each news sample. (Bottom) After fitting the LDA model using the preprocessed text, the document specific topic distributions are combined with the latent representation from the VAE to form the input to a classifier.

establish and retain public trust.

Challenge 1 – Model Interpretability while Retaining Accuracy: Facebook, Twitter, and other OSM have begun classifying user generated content using third party fact checking organizations such as Snopes and Politifact and automated AI systems (Babaei et al., 2019). News articles are either classified as a binary (real or fake) or ordinal categorical variable (e.g., Politifact’s Truth-O-Meter) based on a level of “realness”. Snopes and Politifact are highly accurate and explainable, but can be slow due to the human editorial resources that are required to pass judgment (Politifact, 2020; Snopes, 2019). In contrast, automatic classification of fake news using AI is highly efficient and can be accurate (Pennycook and Rand, 2019), but typically rely solely on “black box” models, i.e., deep neural networks (Dosilović et al., 2018). It is often difficult to comprehend how and why these models generated a prediction, i.e. interpret the model, which is a precursor for understanding the model and ultimately establishing user trust (Rudin, 2019; Gilpin et al., 2018).

Challenge 2 – Missing Data Modalities: AI-based fake news classification systems often assume the availability of specific data modalities, including textual content, images, user profiles (sources), network traffic, or audiovisual content (Khattar et al., 2019; Li et al., 2014; Qian et al., 2018; Wang, 2017; Zhang et al., 2020). However, in many real-world scenarios, acquiring many of these modalities is challenging or not possible due to data scarcity, privacy concerns, or technical limitations (Stiegitz et al., 2018). Many news articles do not provide images (One, 2021), and although user profile information might be helpful in some scenarios, it does not necessarily characterize credibility (Moens et al., 2014). Finally, provenance in OSM is difficult to establish since it is common to propagate information without mentioning the original author.

1.1. Contributions
In this work, we consider automated fake news classification considering a data scarcity scenario where only textual content is available. We have developed a novel architecture, LDA\textsubscript{VAE}, that couples Bayesian topic modelling – in particular, latent Dirichlet allocation (LDA) – with a bi-directional long short-term memory (LSTM) based variational autoencoder (VAE) (Fig. 1). Pretraining the bi-directional LSTM and VAE ensures our methods are accurate and can be applied quickly during test time; topic modelling provides a probabilistic mechanism for (a) inferring the topic composition of news articles, which is often assumed to be known, (b) using learned topics as features to improve classification accuracy, and (c) interpreting the model and features. The learned representation is then used to classify news as real or fake.

We have conducted extensive studies to compare LDA\textsubscript{VAE} with state-of-the-art (SOTA) fake news classification models using the Information Security and Object Technology (ISOT) data (Ahmed et al., 2018), COVID-19 (COVID) data (Banik, 2020), and Twitter data (Boididou et al., 2018). First, we have internally evaluated several classification methods with respect to accuracy, F1 score, false positive rates (FPR), and false negative rates (FNR) for six classifiers across each dataset. Second, we have conducted an extensive ablation study to demonstrate the utility of including thematic features by comparing accuracies across each classifier for LDA\textsubscript{VAE}, LDA\textsubscript{VAE} without topic embeddings, and LDA\textsubscript{VAE} without VAE embeddings. Third, we have demonstrated highly competitive performance in terms of above-mentioned four metrics with competing SOTA methods (Khattar et al., 2019; Mikolov et al., 2010; Wang, 2017; Zhang et al., 2020) while retaining interpretability. To
summarize, we make the following four major contributions:

i. designing a novel fake news detection method that combines the strengths of probabilistic and deep generative modelling (interpretability and accuracy as illustrated in Fig. 1) and makes few assumptions about the richness of the input: (a) only requires text and (b) infers topics instead of relying on noisy or discrete topic labels;

ii. providing a procedure for model interpretation that we demonstrate on experimental data;

iii. justifying the choice in classification method and using topic-based features through an extensive ablation study;

iv. demonstrating highly competitive performance on several metrics while providing both model and feature based interpretability.

The source code for LDA-VAE and evaluation scripts are freely available on GitHub.¹

2. Problem Formulation and Background

2.1. Fake News Detection Problem: Definition and Notations

A fake news article is a text-based document that is intentionally false or misleading (Antoun et al., 2020). In this work, fake news detection refers to the binary classification of news articles as fake or real. Let the training set be \( D_{tr} = \{ d_1, \ldots, d_N \} \) where each element \( d_i \) is a news article (sample) indexed by \( i \in \{1, \ldots, N\} \). A news article \( d_i = (x_i, y_i) \) consists of an input and class label pair where \( y_i \in \{0, 1\} \) and \( x_i \) is a multi-dimensional vector that depends on the textual representation of the model. If \( X = (x_1, x_2, \ldots, x_N) \) and \( Y = (y_1, y_2, \ldots, y_N) \) are the input news articles and labels for training data, then the fake news detection problem is formally defined as learning a function \( F \), parameterized by \( \theta \) that uses \( D_{tr} \) to build a classifier for predicting class labels \( \hat{y}_i \), i.e.,

\[
F(x_i; \theta) = \hat{y}_i.
\]

The objective is to minimize an arbitrary error criterion \((\epsilon)\) between actual class labels and predictions (Eq. 1), i.e.,

\[
\theta^* = \arg \min_{\theta} \epsilon(Y, \hat{Y}),
\]

where \( \hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_N\} \).

2.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is an admixture model that represents topics as distributions over a word vocabulary and documents as a collection of words, each of which is sampled from a latent topic (Blei et al., 2003). Documents are then represented as distributions over topics. LDA can be defined graphically (Fig. 2) or through its joint probability distribution (Eq. 2).

\[
p(\phi, \beta, W | \eta, \alpha) = p(\beta | \eta)p(\phi | \alpha)p(Z | \eta)p(W | \beta, Z) \tag{2}
\]

Let \( N \), \( l_i \), and \( K \) be the number of news documents, the number of words in the \( i \)-th news document, and the number of topics respectively. Let matrix \( W \) be the collection of observed words \( (w_{ij}) \) in a news document where \( i = 1, \ldots, N \) and \( j = 1, \ldots, l_i \), and \( \eta \) and \( \alpha \) are model hyper-parameters. During training, we infer the posterior distributions of \( \phi_t \) (distribution of the topics in the \( i \)-th news article), \( \beta_k \) (distribution of the words in topic \( k \)), and \( z_{ij} \) (the mapping from observed word \( w_{ij} \) to a topic).

2.3. Variational Autoencoder (VAE)

A VAE is an unsupervised deep neural architecture that learns to embed and reconstruct input samples (Kingma and Welling, 2013). It first encodes the input to the parameters of a lower dimensional multivariate normal distribution with isotropic covariance, samples from this distribution to construct a \( z \) vector (i.e., the latent space), and then, the decoder reconstructs the input from \( z \). The objective is derived from a variational lower bound on the marginal log-likelihood and consists of maximizing the sum of an expected log-likelihood term \( p(x|z) \) and the negative Kullback–Leibler (KL) divergence between the posterior and prior distributions \( D_{KL}(q(z|x)||p(z)) \).

2.4. Long Short-Term Memory

Long short-term memory (LSTM) is a recurrent neural network (RNN) developed to address the problem of vanishing gradients and to accommodate longer-term dependencies than traditional RNNs (Hochreiter and Schmidhuber, 1997). Long-term dependencies with arbitrary gaps are modelled using a series of LSTM units that consist of input, output, and forget gates. A bidirectional-LSTM combines two independent recurrent layers side-by-side, one of which receives the input sequence while the other receives a reversed copy (Graves and Schmidhuber, 2005). By processing both the forward and backward directions simultaneously, bidirectional-LSTMs exhibit improved performance over traditional LSTMs for sequence classification problems (Liu and Guo, 2019).

2.5. Word2vec

Standard approaches to representing words and documents consider feature vectors in a high dimensional vocabulary \( V \). The word2vec model is a two-layer neural network that learns word embeddings that attempts to maximize

1 https://github.com/Marjam-Hosseini/LDA-VAE
Given an input word \( w_i \), or the log probability of a given word \( w_i \) given an input word \( w_j \) (Mikolov et al., 2013). After training on a corpus, word2vec embeds words into a significantly lower dimensional space \( w \), such that the cosine distance of semantically similar word embeddings is small and semantically dissimilar words is high.

2.6. Dimensionality Reduction

Dimensionality reduction methods are designed to transform high-dimensional data into a lower-dimensional space while attempting to preserve some meaningful properties of the original data (Cunningham and Ghahramani, 2015). Unlike feature selection methods that seek to subselect a set of features without augmenting them, dimensionality reduction methods transform features through this lower dimensional mapping. They are typically applied for low dimensional visualizations and exploration of data and results, data compression or denoising, or as a step prior to unsupervised or supervised learning (Rosipal et al., 2001). In this work, we consider two dimensionality reduction methods: principle component analysis and t-distributed stochastic neighbor embedding.

Principle component analysis (PCA) is a linear dimension reduction method that uses an eigendecomposition to factorize data covariance matrix in terms of its eigenvectors and eigenvalues and find the axes in which data has more variance (Ringnér, 2008). Then it projects the data points to these new orthogonal axes such that the greatest variance by any projection of the data comes to lie on the first coordinate. In other words, it transforms many correlated variables into a smaller number of uncorrelated variables (principal components). The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. However, it is unable to capture non-linear relationship between variables.

T-distributed stochastic neighbor embedding (t-SNE) is a non-linear probabilistic dimensionality reduction method designed for visualizing high dimensional data; it is based on the idea that similar objects in the high dimensional space should be represented proportionally closer to each other than dissimilar objects in lower-dimensional space (Van der Maaten and Hinton, 2008). To accomplish this, t-SNE minimizes the KL divergence between a joint probability distribution, in the high dimensional space and a joint probability distribution in the low dimensional space. The t-SNE cost function is a symmetrized version of the SNE cost function with simpler gradients and a Student t distribution to compute the similarity between pairs of points, which is more appropriate for data with outliers. Unlike PCA, t-SNE captures the non-linear relationships between random variables focusing on local structure, but does not preserve global distances or densities (Narayan et al., 2021).

3. Prior Work in Fake News Detection

Methods for classifying fake news employ a subset of three primary strategies: propagation-based, source-based and content-based. The motivation for propagation-based techniques is that the diffusion and spread of fake news through social media differs systematically from real news in terms of the speed and the patterns of propagation (Vosoughi et al., 2018; Jin et al., 2014). Source-based methods use features derived from the provenance of news articles, including the general behavior of the source in OSM (Baly et al., 2018). When available, detection using source-based methods can be accurate and fast, however it is often the case that information concerning the spread of news or its authorship is not observed for other users; even when observed, source-based methods can suffer from legitimate users unintentionally spreading fake news that was initiated from other sources (Zhou and Zafarani, 2020). We elaborate more thoroughly on content-based approaches since this is the primary focus of LDA-VAE.

Content-based techniques extract features from the content of news such as text, images, or audiovisual content (Khattar et al., 2019; Wang, 2017; Zhang et al., 2020). This approach assumes that features such as the language, topic, and style of the news body are discriminative attributes for validating its authenticity (Afroz et al., 2012; Rashkin et al., 2017; Rubin et al., 2016). Here, we focus on the textual content since this information is typically always available for news articles.

Textual content-based features can be categorized into three main types. Syntactic features include statistical information about the sentences, like sentence complexity and the frequency of different parts of speech or specific patterns. Lexical features concern the usage of specific words or phrases in the texts, such as bi-grams and tri-grams. Semantic features correspond to the meaning surrounding text and are extracted using techniques from natural language processing (NLP) and data mining such as sentiment analysis (Liu et al., 2010) and emotion mining (Yadollahi et al., 2017). Recently, extracting word embeddings (Mikolov et al., 2013) and topics from text (Ito et al., 2015) has been proposed as potentially useful features for supervised learning. Features can be extracted manually through domain knowledge, but this task is tedious and subject to bias; in contrast, deep learning methods learn features automatically from neural embeddings of the textual content, which can be done more efficiently and with less human bias (Ma et al., 2016). Neural embeddings are then typically used as input for machine learning classifiers, such as support vector machines (SVM), random forests (RF), decision trees, or logistic regression (LR), or deep learning methods such as RNNs (Mikolov et al., 2010) or convolutional neural networks (CNN) (Bondielli and Marcelloni, 2019).

Based on these prior studies, various approaches based on deep learning components have been studied. The Multimodal Variational Autoencoder for Fake News Detection (MVAE) uses a shared representation of the news for further classification as fake or real based on multi-modal variational autoencoder (Khattar et al., 2019). FakeDetector uses latent representation of the text, text subject, and information
about the authors profile (e.g., their title, job, and credibility) (Zhang et al., 2020). The Hybrid CNN is a deep learning based method, with parallel CNN and LSTM layers for processing different modalities (Wang, 2017).

Our proposed work differs from most of the prior studies in the following three aspects: (1) Probabilistic model parameters in LDA-VAE are interpretable as topic distributions for each document and word distributions for each topic in a news articles corpus; the posterior means of learned parameters are used as additional features for classification. (2) Our model requires only textual data. (3) We retain excellent efficiency by integrating a deep architecture (VAE) with LDA.

4. Proposed Method

Our method, LDA-VAE, is based on two views of news article text: a deep neural embedding and a probabilistic topic embedding. LDA-VAE uses a VAE for extracting a lower dimensional semantic representation of the news article and an LDA model for extracting topic-based features. The LDA model differs from the VAE in that the parameter values of the latent variables are highly interpretable. Our motivation for combining the two representations is that they capture similar but complementary discriminative features for fake news classification that will ultimately increase model performance, while retaining interpretability of model parameters.

4.1. Variational Autoencoder

The VAE component in our model, composed of an encoder and decoder, is extended to include a classifier and corresponding cross-entropy loss (Khattar et al., 2019). The architecture of the encoder and decoder is composed of bi-directional LSTM (Bi-LSTM; Fig. 3) and fully connected layers. The Bi-LSTM architecture introduces a new layer of LSTM units that processes the sequence tokens in the backward direction. This overcomes the limitations of a traditional scheme by preserving the information from the past in the forward layer and from the future in the backward layer. In the forward unit of our model (Fig. 3 bottom), the information from words positioning before the $j^{th}$ word token $r(j)$ are passed through $s_{j-1}$ and $c_{j-1}$, i.e.,

$$
[f_j, t_j, o_j] = \sigma(Ws_{j-1} + Ut^{f(j)} + b), \quad \tilde{c}_j = \tanh(Ws_{j-1} + Ut^{f(j)} + b),
$$

$$
\quad c_j = f_j \odot \tilde{c}_{j-1} + t_j \odot \tilde{c}_j, \quad s_j = o_j \odot \tanh(c_j).
$$

For the same word $r(j)$ future words are taken into consideration in the backward LSTM unit (upper part in Fig. 3). The information from the future are passed through $s'_{j+1}$ and $c'_{j+1}$, i.e.,

$$
[f'_j, t'_j, o'_j] = \sigma(W's'_{j+1} + U't^{f(j)} + b'), \quad \tilde{c}'_j = \tanh(W's'_{j+1} + U't^{f(j)} + b'),
$$

$$
\quad c'_j = f'_j \odot c'_{j+1} + t'_j \odot \tilde{c}'_j, \quad s'_j = o'_j \odot \tanh(c'_j).
$$

Figure 3: Bi-LSTM layer on a fake news article from the COVID data. The past and future information for the $j^{th}$ word $r^{(j)}$ in the news article is collected by forward (bottom) and backward (top) LSTM units.

$$
c'_j = f'_j \odot c'_{j+1} + t'_j \odot \tilde{c}'_j, \quad s'_j = o'_j \odot \tanh(c'_j),
$$

(4)

where $\sigma$ is the sigmoid function, $f_j, t_j, o_j$ and $c_j$ are forget, input and output gates and memory state at position $j$ in the input text. $W, W', U$ and $U'$ are weight matrices and $b$ and $b'$ are the bias vectors and $\odot$ denotes element-wise matrix multiplication. Then, the output is $r^{(j)} = g(s_j, s'_j)$. Here $g$ concatenates $s_j$ and $s'_j$ (Fig. 3).

To incorporate labels, we couple a classifier along with the encoder and decoder such that during training the parameters are optimized with respect to the unsupervised VAE ($\mathcal{L}_{rec}$) and classifier ($\mathcal{L}_{BC}$) losses. The VAE loss is given by:

$$
\mathcal{L}_{CE} = -\mathbb{E}_{i \sim D} \left[ \sum_{j=1}^{n_L} \sum_{v \in \mathcal{V}} 1_{v = r^{(j)}} \log f^{(j)}_v \right],
$$

$$
\mathcal{L}_{KL} = -\frac{1}{2} \sum_{f=1}^{n_f} (\mu_f^2 + \sigma_f^2 - \log(\sigma_f) - 1),
$$

$$
\mathcal{L}_{rec} = \mathcal{L}_{CE} + \mathcal{L}_{KL},
$$

(5)

where $r^{(j)}$ is word $j$ in sample $i$, $n_f$ is the number of latent features assumed in the encoder, and $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ are the cross entropy and KL divergence loss functions in the news set $D$. The classifier loss is given by:

$$
\mathcal{L}_{BC} = -\mathbb{E}_{i \sim D} \left[ y_i \log(y_i) + (1 - y_i) \log(1 - y_i) \right],
$$

(6)

where $y_i$ and $\hat{y}_i$ denote the label and the probability that news article $i$ is fake computed by the classifier. The optimized parameters $\theta_{VAE}^*$ minimize the total loss function:

$$
\theta_{VAE}^* = \text{argmin}_\theta(\mathcal{L}_{rec} + \mathcal{L}_{BC}),
$$

(7)
The output of the decoder is the generated news articles. The decoder has a similar architecture to the encoder but in the opposite order of layers for the purpose of reconstructing the input from the latent space. In our framework, the desired features to be extracted from the VAE are the extracted latent features which are a \(|D| \times n_f\) matrix.

4.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is the unsupervised component in our model that jointly infers latent topics (distributions over words in a vocabulary) and the distribution of topics for each document. The generative model for LDA is as follows:

\[
\begin{align*}
\beta_k & \sim \text{Dirichlet}_V(\eta), \\
w_{ij} & \sim \text{Multinomial}(\beta_{z_j}), \\
z_{ij} & \sim \text{Multinomial}(\varphi_i), \\
\varphi_i & \sim \text{Dirichlet}_K(\alpha).
\end{align*}
\] (8)

We use stochastic variational inference to approximate the posterior distribution (Fig. 2 and Eq. 2) (Hoffman et al., 2013). In this context, stochastic variation inference uses stochastic optimization on the mean field variational distribution to iteratively optimize the evidence lower bound for the model specified in Equation 8. Here, \(w_{ij}\) is \(j^{th}\) word in the \(i^{th}\) post, and \(z_{ij}\) determines the assignment of \(w_{ij}\) to a topic. The variational posterior for model parameter \(\varphi = (\varphi_i)\) is an \(N \times K\) matrix where \(N\) is the number of news articles, \(K\) is the number of topics, and element \(\varphi_{ik}\) represents the probability of generating a word from topic \(\beta_k\) in news article \(i\). The \(K\)-dimensional \(\varphi_i\) vectors are concatenated to the corresponding \(n_f\)-dimensional latent features from the VAE (Fig. 1).

4.3. Classifier

The classifier receives the concatenated \(\varphi_i\) and VAE embedding feature vector as input and outputs the news article label. We test six classifiers that include discriminative, generative, interpretable, and deep models: MLP, SVM, LR, Naïve Bayes (NB), RF, and KNN. Our motivation for including a diverse set of classifiers is to allow for desired downstream tasks (e.g., prioritizing interpretability of classification model).

5. Results

5.1. Preprocessing

Before the data were input into \textit{LDA-VAE}, the text was tokenized, non-English text and stop words were removed, hyperlinks, parentheses, and characters that are not expected to be in the words, such as ‘-’, ‘@’ and ‘#’ were removed. If the data is imbalanced, we oversample the minority class using a nearest neighbor extension of the Synthetic Minority Oversampling Technique (SMOTE) algorithm for categorical data (SMOTEN) (Chawla et al., 2002). These methods are commonly used for text classification in imbalanced data scenarios (Zhao et al., 2021). Then as a preprocessing step for the VAE component, we transform the words to \(w\)-dimensional vectors by applying the distributed word2vec pre-trained model (Mikolov et al., 2013). As a result of this transformation, semantically similar words are mapped closer to each other in the \(w\) dimensional space than semantically dissimilar words.

5.2. Datasets Studied

We consider three fake news datasets that contain shorter OSM-shared news samples that we denote as news posts.

5.2.1. ISOT

The ISOT fake news dataset contains 44,898 labeled news posts with a maximum length of 36 words (21,417 real, 23,481 fake) (Ahmed et al., 2018). The authentic and fake news are collected from the news agency Reuters and unreliable websites (flagged by PolitiFact) respectively. The majority of the data are political news from 2016 to 2017. Although the data are cleaned, the punctuation and grammatical mistakes in the fake news posts are retained in the original data (Ahmed et al., 2018). After preprocessing, the number of samples was reduced to 44,143 (22,727 fake and 21,416 real). The similarity between real and fake news post word usage in the ISOT dataset suggests the real and fake news posts are not well separated by word frequencies alone (Fig. 4).

Figure 4: ISOT WordCloud of ISOT real (left) and fake (right) data.

5.2.2. COVID

Published in Nov 2020, the COVID dataset is a recently collected and labeled dataset that contains 10,201 news headlines shared on OSM related to COVID-19 (Banik, 2020). After preprocessing, the COVID data consisted of 9,999 news headlines with binary real or fake labels, and a maximum post length of 115 words. Classifying fake news headlines is challenging in this data due to class imbalance. Among the 9,999 headlines, only 463 are labeled as real news, so we synthesized 6,695 additional real news samples for a total sample size of 16,694. Unlike the ISOT data, word usage appears to be qualitatively different between real and fake news posts; e.g., terms associated with China are more prevalent in fake news posts (Fig. 5).

5.2.3. Twitter

The Twitter dataset (Boïdisou et al., 2018) is another benchmark that was originally collected for the MediaEval Workshop in 2016 (Detection and visualization of misleading content on Twitter, 2018). It includes 15,629 labeled
tweets of maximum length 31 words and covering 17 events. The number of fake and real posts are 9,404 and 6,225 respectively. After preprocessing, the number of samples was reduced to 12,242, among which 7,205 are fake and 5,037 are real. The data contains information about the news post such as the source and textual content. Our motivation to use this dataset is three-fold: it is often used to evaluate fake news classifiers (Can and Alatas, 2019; Khattar et al., 2019; Verdoliva, 2020); it contains tweets from different events; and we expect the LDA model will find distinguishable topics after training. However, this dataset is challenging due to the small length of news posts and the presence of noise and many non-meaningful words in the text, which is reflected in the word frequencies (Fig. 6).

5.3. Evaluation criteria

We evaluate the performance of LDA\textsuperscript{VAE} both qualitatively and quantitatively and also internally with respect to the different input feature sets and classifier models and externally to 4 state-of-the-art competing methods.

5.3.1. Interpreting model parameters

We qualitatively evaluate the separability of features obtained by the VAE, LDA, and their concatenation on training and test sets using PCA and t-SNE. The concatenated feature matrices of size $|D_{tr}| \times (n_f + K)$ and $|D_{te}| \times (n_f + K)$ for $D_{tr}$ and $D_{te}$ respectively, are embedded down to 2 dimensions. We also consider the distribution of topics across fake and real articles. We visualize the average posterior mean of the variational parameters for $\phi_i$ across real and fake news articles with radar plots.

5.3.2. Metrics on classification outcome

We compute accuracy, F1 score, false positive rate, and false negative rates internally for $\text{LDA}^{\text{VAE}}$ using the feature sets obtained by VAE, LDA, and their concatenation. Accuracy is the ratio of the total correct labels to the size of the dataset:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FN are true positives, true negative, false positives, and false negatives, respectively. Precision is the fraction of actual fake news among all the fake detected news $\frac{TP}{TP + FP}$ and recall is the ratio of news truly detected as fake to all the fake news in the data $\frac{TP}{TP + FN}$. We present the precision and recall as their harmonic mean:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We also consider the false positive (FPR $= \frac{FP}{TN + FP}$) and false negative rates (FNR $= \frac{FN}{FN + \text{TP}}$).

5.4. Experimental Settings

Each news post is considered as a vector of $L$ words $t_i^{(1)}, \ldots, t_i^{(L)}$ where $L = \max\{L_i : i = 1, \ldots, N\}$. Using a word2vec model pretrained on the Google News dataset (Rehurek and Sojka, 2011; Mikolov et al., 2013), we produced $w$-dimensional embeddings for each word in the three datasets resulting in a $|V| \times w$ embedding matrix, where $|V|$ is the size of vocabulary set in the model. The input to the VAE is the concatenation of all word embeddings in a news post; if $L_i \leq L$, we apply zero-padding to set the input vector dimension to $L \times w$. For each dataset, the VAE is trained to minimize the reconstruction and classifier loss (Eq. 7) on 90% of the data. Architectural details, code documentation, and details for reproducing the subsequent analyses are available in the code repository\footnote{https://github.com/Marjan-Hosseini/LDA\textsuperscript{VAE}}.

LDA is trained on the same 90% training data as the VAE, however we make the assumption that a news post can be represented as a set of words in the vocabulary $V$ (i.e. bags of words). We tune the number of topics $K \in \{8, 10, 16, 32, 64\}$ using the coherence score of the trained model on a validation set. Although perplexity and predictive likelihood are also used for model evaluation, maximizing coherence has been previously associated with better interpretability (Röder et al., 2015). We concatenate the features extracted from the LDA and VAE models and apply six widely used classifiers, MLP, RF, SVM, LR, NB, and KNN.

5.5. Fake News Classification Results

5.5.1. Ablation Study

We first evaluated the usefulness of integrating neural embeddings and topic features in $\text{LDA}^{\text{VAE}}$ by comparing the news post classification accuracy using only LDA features, only VAE features, and combined LDA and VAE features. Simultaneously, we compare the accuracy, F1 score, FPR, and FNR for the six classifiers across the three datasets.

For most of the dataset and evaluation criteria configurations, classifier performance was higher when the two feature sets are concatenated (Tab. 1). While the $\text{LDA}^{\text{VAE}}$ is most
frequently the highest scoring model configuration, topic features alone (LDA) achieve a higher accuracy and F1 score in the Twitter dataset, potentially because the frequency of fake news posts vary based on topics (Torabi Asr and Taboada, 2019). In fact, the VAE only classifiers perform poorly on the Twitter dataset in general possibly due to the short and unstructured nature of Twitter news posts. This underscores the usefulness of incorporating topic features into fake news classification. Since random forest classifiers yielded the best aggregate results, we only consider random forest classifiers for the subsequent results.

5.5.2. Comparison with SOTA

Next, we compare the performance of LDA/VAE to four SOTA methods in the same category of fake news classification (content-based): MVAE (Khattar et al., 2019), FakeDetector (Zhang et al., 2020), Hybrid CNN (Wang, 2017) and RNN (Mikolov et al., 2010). To keep the analysis consistent across methods, we modified the methods to only consider textual data, e.g., we removed the layers of MVAE associated with image processing.

LDA/VAE achieves the best accuracy and F1 score in the ISOT dataset, which is the largest dataset in terms of news post quantity and is also well-curated compared to the other datasets since it includes articles from Reuters and human labeled fake news posts; it also achieves comparable results on the COVID and Twitter data (Tab. 2). Both the Twitter and COVID datasets are noisy and model training was performed without syntactic or semantic text correction. Among the three datasets, the performance of VAE-based methods is worst on the Twitter data likely due to having few samples and news posts having short length. Additionally, the VAE objective function, which requires balancing the reconstruction error, classification error, and sample generation quality, is difficult to balance during training for specific downstream tasks such as classification (Böh n and Seljak, 2020). RNNs use a simpler autoregressive architecture and the results deteriorate when the sequence length is increased.

5.6. Model Interpretability

Here, we explore the interpretability of LDA/VAE through analysis and visualization of the parameters in the LDA model.

5.6.1. News Post Topic Distributions

The variable posteriors of $\beta_i$ and $\varphi_i$ are interpretable as the distribution of words in topics and the distribution of topics in the news posts respectively. Further, posterior inference yields a distribution rather than point estimates, consequently the parameter uncertainty can be presented to users to give flexibility in interpreting variable importance and effect regarding the class label. For example, we interpret the effect of topics on class outcome by visualizing the topic distribution across news posts and separated by their labels. Specifically, we set the number of topics $K = 10$ (for visualization purposes) and plot the averaged posterior means of $\varphi_i$ for real and fake news posts (Fig. 7). The frequencies of topics in real and fake news posts are largely distinct with some topic overlap. The topics that are more discriminative can be interpreted as larger differences in average posterior means. For example, all topics besides 0 and 4 are informative in the ISOT dataset (Fig. 7 (a)).

5.6.2. Topic Distributions

For each dataset, we compute the most discriminatory topics using LDA/VAE parameters $\varphi$ and $\beta$. We consider the two topics that have the largest difference in the average posterior means of $\varphi_i$ variables across real and fake news posts. In other words, we sort the vector $(\sum_{i \in D^R} \varphi_{ik} - \sum_{i \in D^F} \varphi_{ik})_{k=1}^K$ in descending order where $D^R$ and $D^F$ are the training sample sets for fake and real news posts respectively. We then plot the normalized frequency of the top 20 most frequent words (Fig. 7 (b, c)). Word distributions

### Table 1

Comparison of the accuracy metrics after running the model on ISOT, COVID, and Twitter datasets and classifying VAE, LDA feature set (baseline), and their concatenation. The best scores are shown in bold.

| Metric | ISOT | COVID | Twitter |
|--------|------|-------|---------|
| SVM    | 0.88 | 0.90 | 0.88 |
| F-Sc.  | 0.87 | 0.91 | 0.88 |
| FNR    | 0.11 | 0.16 | 0.12 |

### Table 2

Comparison of the metrics among different methods (an RF classifier is used with LDA/VAE).

| Method       | ISOT Train (Test) | COVID Train (Test) | Twitter Train (Test) |
|--------------|------------------|--------------------|----------------------|
| SVM          | 1.00 (0.88)      | 1.00 (0.92)        | 1.00 (0.92)          |
| F-Sc.        | 0.99 (0.89)      | 1.00 (0.92)        | 0.99 (0.92)          |
| FNR          | 0.01 (0.10)      | 0.00 (0.09)        | 0.00 (0.09)          |
| MVAE (Khattar et al., 2019) | 0.87 (0.88) | 0.78 (0.79) | 0.76 (0.73) |
| FakeDetector (Zhang et al., 2020) | 0.91 (0.92) | 0.90 (0.92) | 0.94 (0.95) |
| Hybrid CNN (Wang, 2017) | 0.91 (0.92) | 0.95 (0.95) | 0.90 (0.79) |
| RNN (Mikolov et al., 2010) | 0.75 (0.74) | 0.99 (0.78) | 0.96 (0.95) |
in discriminative topics are subtly different. For example, while words associated with U.S. conservative politics and China were found in both topics associated with real and fake news, their frequencies were substantially different.

5.6.3. Dimensionality Reduction

Lastly, we qualitatively evaluated the separability of LDA, VAE, and combined LDA and VAE features in 2-dimensional embeddings for the ISOT dataset; the plots of
other datasets show similar behavior. We applied PCA and t-SNE to the features learned from the different architecture configurations with $K = 10$ and $\nu = n_f = 32$ (Figs. 8 and 9). We observe that in both PCA and t-SNE plots, the real and fake news posts are well separated by the combined VAE and LDA features.

6. Conclusion and Future Work

We proposed LDAVAE, a combined LDA, supervised BiLSTM VAE, and classifier architecture for classifying fake news text. We use a Bayesian admixture model for topic modelling to increase the interpretability of the method, add informative features, and remove the necessity of costly manual topic selection. We justified our architecture with an extensive ablation study and evaluated performance of LDAVAE by comparing with 4 SOTA baseline methods across 3 datasets and showed highly competitive performance. We then provided mechanisms to evaluate the interpretability of model parameters and class discrimination allowing for exploration of the model and features.

The main disadvantage of our modelling assumptions is that we cannot extract features from multimedia data. In addition, training a separate probabilistic model requires additional computation, however understanding the model and important features are crucial for providing explanations and establishing user trust. More sophisticated natural language architectures that use self-attention, like transformers, should provide a mechanism for improving accuracy. Integrating transformers, which typically model text as an ordered sequence, with topics that are intrinsically unordered is a priming area of future work (Wang et al., 2020). Also, alternatives to VAEs, like architectures based on normalizing flows, should be investigated in the context of fake news detection (Böhm and Seljak, 2020). For example, the Probabilistic Auto-Encoder leverages normalizing flows to achieve better reconstruction than traditional VAEs while retaining high sample quality. Future work may also consider applying feature selection techniques suitable for high dimensional data or selecting the dimension of the autoencoder latent space (Brankovic et al., 2018; Hosseini, 2018), and using the text provided in reviews (Qian et al., 2018) and conversation graphs (Brambilla et al., 2021a) of live events participants (Brambilla et al., 2021b; Javadian Sabet et al., 2021) to detect the disinformation propagation on inner circles (de Souza et al., 2020).

Declaration of Competing Interest

The authors declare that there is no conflict of interest in all aspects of this manuscript preparation and data analysis.

CRediT authorship contribution statement

Marjan Hosseini: Conceptualization, methodology, software, validation, investigation, data curation, writing – original draft, visualization. Alireza Javadian Sabet: Conceptualization, methodology, software, validation, investigation, data curation, writing – original draft, visualization. Suining He: Supervision, conceptualization, methodology, validation, investigation, writing – review & editing. Derek Aguiar: Supervision, funding acquisition, conceptualization, methodology, validation, investigation, writing – review & editing.

References

J. Posetti, A. Matthews. A short guide to the history of ‘fake news’ and disinformation, International Center for Journalists 7 (2018) 1–19.
Interpretable Fake News Detection with Topic and Deep Variational Models

M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proceedings of the National Academy of Sciences 118 (2021).

M. Chalikiadakis, A. Kornilakis, P. Papadopoulos, E. Markatos, N. Kourtellis, The rise and fall of fake news sites: A traffic analysis, in: 13th ACM Web Science Conference 2021, 2021, pp. 168–177.

H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2017) 211–36.

S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

Society of Professional Journalists, SPI Code of Ethics, 2021. URL: https://www.spj.org/ethicscode.asp.

B. Carminati, E. Ferrari, J. Girardi, Trust and share: Trusted information sharing in online social networks, in: 2012 IEEE 28th international conference on data engineering, IEEE, 2012, pp. 1281–1284.

M. Hindman, V. Barash, Disinformation, and influence campaigns on twitter, Knight Foundation: George Washington University (2018).

T. Lee, The global rise of “fake news” and the threat to democratic elections in the usa, Public Administration and Policy (2019).

A. M. Guess, D. Lockett, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, “fake news” may have limited effects beyond increasing beliefs in false claims, Harvard Kennedy School Misinformation Review 1 (2020).

D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science 359 (2018) 1094–1096.

K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovury, J. Martin, E. Morgan, et al., Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media, Political Behavior 42 (2020) 1073–1095.

M. Babaei, A. Chakraborty, J. Kulshrestha, E. M. Redmiles, M. Cha, K. P. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, et al., The principles of the truth-o-meter: Politifact’s methodology for independent fact-checking, 2020. URL: https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-1/.

Snopes, What is Snopes’ fact-checking process?, 2019. URL: https://www.snopes.com/faq/fact-checking-process/.

G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality, Proceedings of the National Academy of Sciences 116 (2019) 2521–2526.

F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning models, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

D. Khattar, G. Gupta, V. Varma, Mvaa: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

H. Li, Z. Chen, B. Liu, X. Wei, J. Shao, Spotting fake reviews via collective positive-unlabeled learning, in: 2014 IEEE international conference on data mining, IEEE, 2014, pp. 899–904.

F. Qian, C. Gong, K. Sharma, Y. Liu, Neural user response generator: Fake news detection with collective user intelligence., in: IJCAI, volume 18, 2018, pp. 3834–3840.

W. Y. Wang, “liar, liar pants on fire”: A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

J. Zhang, B. Dong, S. Y. Philip, Fakedetector: Effective fake news detection with deep diffusive neural network, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1826–1829.

S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger, Social media analytics – challenges in topic discovery, data collection, and data preparation, International Journal of Information Management 39 (2018) 156–168.

G. One, A list of text-only news sites, 2021. URL: https://greycoder.com/a-list-of-text-only-new-sites/.

M.-F. Moens, J. Li, T.-S. Chua, Mining user generated content, CRC press, 2014.

H. Ahmed, I. Traore, S. Saad, Detecting opinion spam and fake news using text classification, Security and Privacy 1 (2018) e9.

S. Banki, Covid fake news dataset, 2020. URL: https://doi.org/10.5281/zenodo.4282522.

C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, Y. Kompatiairis, Detection and visualization of misleading content on twitter, International Journal of Multimedia Information Retrieval 7 (2018) 71–86.

T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, S. Khadankur, Recurrent neural network based language model., in: Interspeech, volume 2, Makuhari, 2010, pp. 1045–1048.

W. Antoun, F. Baly, R. Achour, A. Hussein, H. Hajj, State of the art models for fake news detection tasks, in: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), IEEE, 2020, pp. 519–524.

M. D. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of Machine Learning Research 3 (2003) 993–1022.

D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

A. Graves, J. Schmidhuber, Frameworks phoneme classification with bidirectional lstm and other neural network architectures, Neural networks 18 (2005) 602–610.

G. Liu, J. Guo, Bidirectional lstm with attention mechanism and convolutional layer for text classification, Neurocomputing 337 (2019) 325–338.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, arXiv preprint arXiv:1310.4546 (2013).

J. P. Cunningham, Z. Gahramani, Linear dimensionality reduction: Survey, insights, and generalizations, The Journal of Machine Learning Research 16 (2015) 2859–2900.

R. Rospal, M. Girolami, L. J. Trejo, A. Cichocki, Kernel pca for feature extraction and de-noising in nonlinear regression, Neural Computing & Applications 10 (2001) 231–243.

M. Ringné, What is principal component analysis?, Nature biotechnology 28 (2006) 303–304.

L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

A. Narayan, B. Berger, H. Cho, Assessing single-cell transcriptional variability through density-preserving data visualization, Nature biotechnology 39 (2021) 765–774.

Z. Jin, J. Cao, Y.-G. Jiang, Y. Zhang, News credibility evaluation on microblog with a hierarchical propagation model, in: 2014 IEEE International Conference on Data Mining, IEEE, 2014, pp. 230–239.

R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, arXiv preprint arXiv:1810.01765 (2018).

X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020).

S. Afroz, M. Brennan, R. Greenstadt, Detecting hoaxes, frauds, and deception in writing style online, in: 2012 IEEE Symposium on Security and Privacy, IEEE, 2012, pp. 461–475.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proc. of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.

V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.
B. Liu, et al., Sentiment analysis and subjectivity., Handbook of natural language processing 2 (2010) 627–666.
A. Yadollahi, A. G. Shahraki, O. R. Zaiane, Current state of text sentiment analysis from opinion to emotion mining, ACM Computing Surveys (CSUR) 50 (2017) 1–33.
J. Ito, J. Song, H. Toda, Y. Koike, S. Oyama, Assessment of tweet credibility with lda features, in: Proceedings of the 24th WWW, 2015, pp. 953–958.
J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 16, AAAI Press, 2016, p. 3818–3824.
A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38–55.
M. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference., Journal of Machine Learning Research 14 (2013).
N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
K. Zhao, Z. Xu, M. Yan, Y. Tang, M. Fan, G. Catolino, Just-in-time defect prediction for android apps via imbalanced deep learning model, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021, pp. 1447–1454.
Detection, visualization of misleading content on Twitter, Boididou, christina and papadopoulos, syeneon and zampoglou, markos and apostolidis, lazaros and papadopoulos, olga and kompasiasiari, yiannis, International Journal of Multimedia Information Retrieval 7 (2018) 71–86.
U. Can, B. Alatas, A new direction in social network analysis: Online social network analysis problems and applications, Physica A: Statistical Mechanics and its Applications 535 (2019) 122372.
L. Verdoliva, Media forensics and deepfakes: an overview, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 910–932.
R. Rehurek, P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3 (2011) 2.
M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, 2015, pp. 399–408.
F. Torabi Asr, M. Taboada, Big data and quality data for fake news and misinformation detection., Big Data & Society 6 (2019) 205395119843310.
V. Böhme, U. Seljak, Probabilistic auto-encoder, arXiv preprint arXiv:2006.05479 (2020).
Z. Wang, Z. Duan, H. Zhang, C. Wang, L. Tian, B. Chen, M. Zhou, Friendly topic assistant for transformer based abstractive summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 485–497.
A. Brankovic, M. Hosseini, L. Piroddi, A distributed feature selection algorithm based on distance correlation with an application to microarrays, IEEE/ACM transactions on computational biology and bioinformatics 16 (2018) 1802–1815.
M. Hosseini, Feature selection for microarray classification problems, 2018.
M. Brambilla, A. Javadian, A. E. Sulistiawati, Conversation graphs in online social media, in: Web Engineering, Springer International Publishing, Cham, 2021a, pp. 97–112. URL: https://doi.org/10.1007/978-3-030-74296-6_8, doi: 10.1007/978-3-030-74296-6_8.
M. Brambilla, A. Javadian Sabet, M. Hosseini, The role of social media in long-running live events: The case of the big four fashion weeks dataset, Data in Brief 35 (2021b) 106840.
A. Javadian Sabet, M. Brambilla, M. Hosseini, A multi-perspective approach for analyzing long-running live events on social media. a case study on the “big four” international fashion weeks, Online Social Networks and Media 24 (2021) 100140.
J. V. de Souza, J. Gomes Jr, F. M. de Souza Filho, A. M. de Oliveira Julio, J. F. de Souza, A systematic mapping on automatic classification of fake news in social media, Social Network Analysis and Mining 10 (2020) 1–21.

Marjan Hosseini received her baccalaureate in computer engineering from Shomal University, Iran, and her master’s degree in computer science and engineering from Politecnico di Milano (POLIMI) in 2018. Afterwards, she joined DEEP-SE Lab at POLIMI and worked in natural language processing and machine learning. Currently, she is working towards her PhD at the University of Connecticut in the Computer Science and Engineering Department. Her research interests lie in Bayesian and probabilistic methods in computational biology, machine learning and bioinformatics.

Ali reza Javadian Sabet is a PhD student in the Department of Informatics and Networked Systems at the University of Pittsburgh. He received his master’s degree in computer science and engineering from POLIMI. He worked as a research fellow at the Data Science Lab and DEpendable Evolvable Pervasive Software Engineering (DEEP-SE) group at POLIMI. His research interests lie in computational social science, knowledge extraction from social networks, business analysis, and context-aware recommender systems.

Sui ning He is currently working as an assistant professor at the Department of Computer Science and Engineering, University of Connecticut. He received his PhD degree in computer science and engineering at the Hong Kong University of Science and Technology, and worked as a postdoctoral research fellow at the Real-Time Computing Lab (RTCL) at the University of Michigan, Ann Arbor. His research interests include ubiquitous and mobile computing, crowdsourcing, and big data analytics.

Derek Aguilar is an assistant professor in the Computer Science and Engineering Department at the University of Connecticut. He graduated from the University of Rhode Island with baccalaureates in computer engineering and computer science, received his PhD in computer science from Brown University, and completed his postdoctoral work at Princeton University. His research aims to develop Bayesian machine learning models, scalable inference methods, and graph theoretic algorithms to better understand high-dimensional data, particularly in the areas of genomics and genetics.