Unsupervised Deep Learning for Fake Content Detection in Social Media

Jie Tao
Fairfield University
Fairfield, CT, USA, 06824
jtao@fairfield.edu

Xing Fang
Illinois State University
Normal, IL, USA, 61790
xfang13@ilstu.edu

Lina Zhou
The University of North Carolina at Charlotte
Charlotte, NC, USA, 28223
lzhou8@uncc.edu

Abstract

Fake content is ever increasing in the online environment, driven by various motivations such as gaining commercial and political advantages. The interactive and collaborative nature of social media further fuels the growth of fake content by exerting fast and widespread influence. Despite growing and interdisciplinary efforts in detecting fake content in social media, some common research challenges remain to be addressed such as humans' cognitive bias and scarcity of labeled data for training supervised machine learning models. This study aims to tackle both challenges by developing unsupervised deep learning models for the detection of fake content in social media. In view that traditional linguistic features fail to capture context information, our proposed method learns feature representations from the context in social media content. The empirical evaluation results with fake comments from YouTube demonstrate that our proposed methods not only outperform baseline models with traditional unsupervised machine learning techniques, but also achieve comparable performance to the state-of-the-art supervised models. The proposed analytical pipeline provides an end-to-end solution to detecting fake social media contents, which largely reduce the human labor required in collaborative data science teams (i.e., particularly the data labeling). The findings of this study can be used to facilitate collaboration in data science by reducing humans' cognitive bias and improve the collaboration efficiency.

1. Introduction

The interactive and collaborative nature of social media has lent itself to the efficient and widespread reach of information diffusion. Social media use has become ever popular during the global pandemic. However, along with the prevalent use of social media in our personal and professional lives come the concerns about the quality of social media content such as fake content [1]. Fake content in social media may include, but is not limited to, fake news, fake reviews, comment spam, fake engagement, hoax, rumor, and propaganda. Much of the fake content can result from illegal marketing practices to create commercial advantages [2], such as promoting certain products or damaging the reputations of competing products [3]. According to the 2020 Sprout Social Index, 77% of consumers would purchase from, and 75% consumers would increase their spending with, the brands they follow on social media. Those fake content, if left unaddressed, would mislead social media users into forming misinformed opinions or making incorrect purchase decisions, which further compromises the trustworthiness of social media content [3], [4]. Therefore, the detection of fake content in social media has immense practical value.

Despite a growing amount of interdisciplinary effort toward detecting fake content in social media, some common research challenges remain. First, the speed of social media content generation significantly outpaces humans' cognitive capacity. There were 3.8 billion social media users worldwide in January 2020 [5]. As a result, manual validation or fact checking does not meet the practical demand of coping with the big social media data. Second, even if there was sufficient manpower and related expertise to scrutinize all social media content, humans suffer from cognitive biases in detecting fake social media content [3] and online deception in general [6], which result in poor performance. For instance, the average accuracy of three human reviewers in identifying spam comments is 57.33% [7]. Third, extant studies on developing automated methods for online fake content detection and deception detection rely heavily on supervised learning techniques, which in turn require labelled datasets. The preparation of labelled datasets, a typically manual process, is subject to the cognitive bias of individual coders, as mentioned above. It is particularly difficult to identify fake content. One way to alleviate the cognitive bias is by collaborative data labeling. More advanced supervised learning techniques (e.g., deep neural networks) usually require a larger sized data for parameter learning or model training to reap their benefits. Despite that researchers have sought to alleviate the above issues using semi-supervised learning techniques (e.g., [8]), these techniques still require some labeled data to begin with, and these data are context-

URI: https://hdl.handle.net/10125/70643
978-0-9981331-4-0

specific. This research primarily aims to address the above challenges by developing unsupervised machine learning models for detecting fake content in social media.

Building machine learning models for the detection of fake content in social media has employed a variety of input features [9], [10]. However, these feature representations are either too coarse-grained (e.g. text statistics) [11] or word-based (e.g., TF-IDF, bag-of-words) [12]. They have largely overlooked the context embedded in social media content to support fake content detection. Recent studies exploring the context in social media content beyond the word level to detect fake contents have demonstrated great promise (e.g., [13]). Thus, a second aim of the study is to investigate the impacts of context information on the performance of fake content detection in social media.

To achieve the two research aims, we propose an unsupervised deep learning based analytical pipeline for the detection of fake content in social media. Specifically, the proposed pipeline utilizes two Deep Neural Networks (DNNs) based unsupervised learning techniques, namely Stacked AutoEncoders (SAEs) and Generative Adversarial Active Learning (GAAL), in analyzing the textual content of social media. We performe an empirical evaluation using a YouTube dataset created using a collaborative tagging tool. The evaluation results demonstrate that the proposed analytical pipeline not only outperformed baseline unsupervised learning techniques, but also achieved comparable performances to supervised learning models.

This study makes multi-fold research contributions to data science, specifically knowledge discovery from collaborative data in social media. First, this is among the first studies that apply unsupervised learning techniques to fake content detection in social media. Deep learning based unsupervised methods, particularly Generative Adversarial Networks, have not been used for the target problem except for one very recent study [14], As discussed above, fake social media content datasets remain scarce, and it is resource-intensive to acquire labeled data. By addressing the reliance on labelled data, unsupervised learning techniques can make the fake content detection models more generally applicable. Second, it highlights the important role of context in social media content in identifying fake content. Unlike previous studies that have focused on the linguistic features at the word level (e.g. the number of words, ratio of emotional words) [3], [15], [16], this study learned feature representations from the context of social media content. Particularly, we employed a document-level representation that is capable of capturing the inter-sentence transitions. Third, this study combined context features with linguistic features identified in the previous studies and analyzed their relative impacts on the detection performance. The proposed analytical pipeline can provide an end-to-end support for detecting fake social media content.

2. Background and Related Work

In this section, we discuss related studies that led to the two research gaps outlined earlier, as well as the techniques employed in this study.

2.1. Fake Content Detection Methods and Input Features

The detection of fake social media contents has become one of the most important applications in the field of text analytics. Researchers perform word-level analysis so that context, such as information of the document level, is usually ignored. Luo et al. [17] proposed a multi-aspect based neural network model to distinguish fake contents, where the two word-level features, similarity and sentiment, obtained via TF-IDF, are used as the input of a feed-forward neural network for classification. The authors mentioned that the document-level features are preferred for capturing the context, as one of their future study directions. Specifically, within the context of spam content detection, a recent study analyzed the textual contents of YouTube comments for the purpose of spam filtering, and then included them in the training processes of supervised machine learning models [11]. However, the features used in this study are at the word level, the context among the words are neglected. Similarly, another recent study analyzing the toxic comments on YouTube videos also employed topic modeling at the word level as the main analytical vehicle [18]. One of the biggest drawbacks of topic modeling, as an unsupervised learning method, is the quality control of the extracted topics. It is partially attributed to the word level nature of the method. Additionally, Nisha [12] selected several word level text representation models (TF-IDF, bag of words) with supervised machine learning techniques (e.g. Naïve Bayes and Support Vector Classifier) to detect spam comments on YouTube. A similar study was reported in [10].

Due to the predominant supervised nature of fake content detection, researchers have relied much on labeled data to construct their classification models. Although spam detection/filtering is one of the most popular applications in text analytics, there are limited labeled datasets from different social media sites (e.g. on YouTube comments) in training of supervised machine learning models [11]. As a result, researchers have relied on manual labeling to create labeled datasets for classification [12]. To address the limited labeled data for fake review detections, Juuti et al. [19] presented an

advanced neural machine translation (NMT) based technique, where the model can generate fake reviews. The model essentially consisted of a couple of character-level recurrent neural networks (RNNs) that function as a sequence-to-sequence model. The NMT model served as the generator of labeled data to train a fake review detection model (i.e. adaptive boosting) in the study. In view of subtle linguistic patterns of fake reviews, it remains difficult to control the quality of the generated fake contents, which will in turn affect the model performance on real reviews.

Prior studies have relied on a wide range of features in context for the purpose of fake content detection. The first group of features is termed as behavioral features, which measure the actions and social media interactions of users from different platforms. The most popular behavioral features across different platforms include the rating (i.e. star rating), number of posts from the same users, and information about the users (e.g. registered users, or profile pictures) [4]. Other behavioral features, which capture the human interactions on social media, have dominated previous studies analyzing the spam contents in social media. For instance, Ammari et al. [9] proposed a network analysis based approach to filter the spam information within YouTube comments. Moor et al. [20] conducted an analysis on data collected via a customized survey to understand the phenomenon of 'flaming comments' (defined as hostile or insulting comments) on YouTube. Sureka [21] studied the behavioral features (e.g. time difference in comments) and linguistic features (e.g. comment redundancy) and the relationship to spam contents on YouTube. Moreover, in the broader context of social media, researchers have examined the features extracted from other types of social media contents. Kumar et al. [16] presented a hierarchical approach to increase the likelihood of detecting anomalies. Their approach exclusively analyzed several nonverbal behavioral features and then engineered them to capture the collective behaviors for review manipulation detection purposes, which is a form of fake reviews. The features are then used in several supervised-learning techniques for classifying review manipulations. To better capture the effect of the review contents, Zhang et al. [3] identified a variety of linguistic features, in addition to the nonverbal behavioral features, and examined their relative significance for the detection of fake reviews. The authors discovered that solely relying on the linguistic features limits the performance of fake review detection models. Further, they also acknowledge the difficulty of extracting meaningful features from the review contents. Thus, an efficient method for extracting features from review contents are deemed necessary in the context of fake review detection. Advanced text representation methods, such as word2vec and doc2vec, can be used for extracting linguistic features from the textual contents.

2.2. Unsupervised Deep Learning Methods

In this study, we employ two types of unsupervised deep learning methods, namely SAE and GAAL. An autoencoder [22] is a neural network designed to reproduce its input by its output. The network consists of two parts: an encoder function $h = f(x)$, and a decoder function $r = g(h)$, where r is a reconstruction of x. Autoencoders can be trained for feature learning by minimizing the training error.

Assume that there are m samples in the training set, let $x^{(i)} \epsilon R^n$ and $r^{(i)} \epsilon R^n$ denote the i^{th} training data and its reconstruction, respectively. The training process is described as minimizing the reconstruction error, which is the mean squared error between $x^{(i)}$ and $r^{(i)}$:

$$L(x, r) = \frac{1}{2m}\sum_{i=1}^{m}\left\|x^{(i)} - r^{(i)}\right\|_2^2 \qquad \textbf{(1)}$$

Reproducing the input may seem useless, however, the focus here is to let h capture useful features from x by training the network. If h has a lower dimension than x, the network is then known as an under-complete autoencoder. An under-complete representation, h, is able to capture most salient features from x. To improve the performance of a simple autoencoder, a stacked autoencoder (SAE) extends both the encoder and decoder to include multi-layered structures. Since SAEs can be used to learn features, researchers have employed them in the context of generative models. For instance, Liu et al. [23] proposed a GAAL-based model to detect outliers in data. This unsupervised anomaly detection approach uses two single-layered perceptrons, one is known as the generator, G, and the other is known as the discriminator, D. In order to train the discriminator, G, which can be recognized as an autoencoder, generates outliers (fake data) based on a set of uniformly distributed noise data, z. The fake data, $G(z)$, is then mixed with some real data, x, to train D. To train the generator, z is reused to generate $G(z)$ first, which is in turn used as the only input to D, and the label of $G(z)$ is marked as authentic rather than fake. This is a crucial step, since the key idea to train G is to make it generate fake data that are hard to distinguish by D. The training process stops when the parameters of D converge.

3. Methodologies

In this section, we discuss the proposed analytical pipeline for detecting fake contents on social media, the experiment setup for evaluating the pipeline, and the experiment results. The proposed pipeline addresses the

research limitations identified in Section 2 with several novel design elements as the following.

- It alleviates the dependence on labeled data in developing models;
- It is able to capture the context beyond individual words, within the textual content of each post;
- It synthesizes the results from different unsupervised deep learning methods in a variety of ensemble models.

3.1. The Analytical Problem

The overarching analytical problem in this study is "are context features extracted from the social media contents indicative of the fake contents, using purely unsupervised learning methods?" As discussed in Section 2, previous studies have touched upon the textual contents of the online consumer reviews, relying on traditional text analytics techniques such as term frequency-inverse document frequency (TF-IDF), Latent Dirichlet Allocation (LDA), word2vec, and sentiment classifications [15]. Other related techniques include clustering on the document similarity between the social media contents [17]. These attempts prove that researchers are shifting focuses to the textual contents, since they are valuable in terms of fake content detection. In order to better organize this study, we develop two sub research questions based on the aforementioned overarching research question.

The first sub research question is "does combining context features with linguistic features improve the performance of fake content detection?" Prior studies (e.g., [3]) have indicated that the spammers (who compose the fake contents) put in an effort to edit and rehearse the text messages. Thus, solely relying on the linguistic or the context features for fake content detection purposes is inadequate. Furthermore, aforementioned methods for analyzing the textual contents are either word based (e.g. TF-IDF, LDA, word2vec), which are difficult to capture the interrelationships between words (i.e. the context), or too coarse grained (e.g. document similarity, document-level sentiment analysis), which cannot capture the finesse patterns (e.g. choice of words, writing styles). Thus, we examine the context features extracted by doc2vec, in combination with the linguistic features suggested in prior related studies, for their effectiveness in the fake content detection analysis.

The second sub research question is "does the proposed analytical pipeline perform better than other unsupervised learning techniques, in terms of fake content detection purposes?" As discussed in Section 2, the majority of the prior related studies employed supervised learning techniques. It is also evidential that datasets containing fake social media contents labeled with authentic/fake information are scarce, and manually labeling the (fake) contents are too tedious and error-prone. Thus, advanced unsupervised learning based techniques, as included in the proposed analytical pipeline, are able to relieve the reliance on labeled data. Additionally, the unsupervised learning methods are often used to derive implicit patterns from unlabeled data, which can be used for enhancing supervised learning methods. The key to address this sub research question is to evaluate the proposed unsupervised learning technique(s), ensuring that it outperforms extant unsupervised learning techniques for detecting fake contents in social media. Moreover, we need to demonstrate that the results from our proposed analytical pipeline are comparable to the results from the supervised learning counterparts. Additionally, the unsupervised models can serve as complements to human knowledge in collaborative data science teams.

3.2. The Proposed Analytical Pipeline

The proposed analytical pipeline is sketched in Figure 1, which consists of four components. This pipeline takes social media contents texts as the main input, along with the labels (i.e., authentic/fake) for model evaluation. It produces the classification results of whether each text is fake or authentic as the output.

Feature Extraction. The first two components extract features from social media contents. In this study, we focus on two types of features for fake content detection: linguistic features and context features.

Linguistic features. Linguistic features have been widely used in fake content detection studies, which include a variety of text statistics (normally at the word level and in a coarse fashion) extracted from the texts. Drawing from related prior studies (e.g. [3]), we select the following linguistic features in this study.

- Content length: number of words in a piece of social media content;
- Average sentence length: average number of words per sentence;
- Noun ratio: ratio of nouns in a piece of social media content;
- Part-of-Speech (POS) ratios: ratios of verbs, adjectives, adverbs, and personal pronouns in a piece of social media content;
- Unique POS ratio: ratio of unique POS-word pairs in a piece of social media content;
- Content Diversity: ratio of unique nouns and verbs to total number of nouns and verbs in a piece of social media content;
- Capital words count: counts of words with all capital letters in a piece of social media content;

- Emotiveness: ratio of adjectives and adverbs to nouns and verbs in a piece of social media content;
- Self-reference ratio: ratio of first person pronouns to all pronouns in a piece of social media content;
- Title mention: number of mentions of the title in the content of a piece of social media;
- Subjectivity: ratio of subjective to objective words in a piece of social media content;
- Average sentiment scores: overall (average) sentiment scores of a piece of social media content.

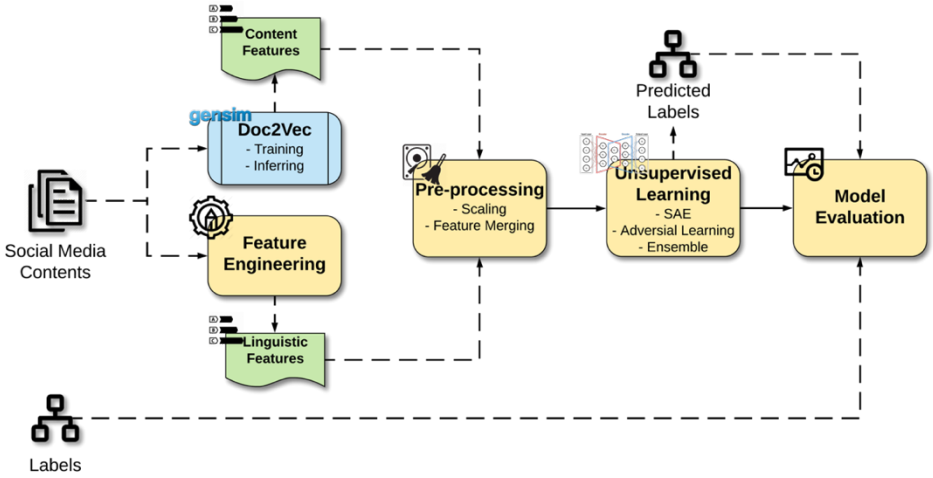

Figure 1. The Proposed Analytical Pipeline

Context features. Context features can be extracted from different grain levels of social media contents such as word-, or document-level. In order to better encapsulate the contextual information embedded in the social media contents, we use a deep neural network based method, namely doc2vec [24], to map the values of the context features to the document embedding. More specifically, we use the distributed memory (PV-DM) method in the doc2vec model to represent the textual contents as vectors in a document embedding model, as shown in Figure 2. Compared to the distributed bag-of-words method in doc2vec, which ignores the context words and forces the model to predict words randomly sampled from the document, the PV-DM method is better at capturing implicit context in the contents. Similar to the popular word2vec model, the PV-DM model is trained and evaluated on a pseudo classification problem, namely predicting a center word (e.g. "this") with the surrounding words as the context (e.g. "watch", "video", "now") and a paragraph identifier as the input. The context words and the paragraph identifier are first represented as vectors of arbitrary length, and then aggregated into a document vector of a predefined length. If the pseudo classification performance reaches a satisfactory level, the document vectors are used to infer *unseen* documents with stochastic gradient descent optimization. Therefore, the inferred document vectors represent each piece of social media content as a real-valued vector, which captures not only the words, but also the context of the words. Through this step, we extend the extant methods by incorporating the context (at the document level) in the text representation model. To the best of our knowledge, this is the first study incorporating document-level context for the purpose of fake content detection.

Preprocessing. This component involves two substeps: scaling and feature merging. During this step, the linguistics and context based features are combined. Since the two types of features tend to have different ranges of values, they need rescaling to the same range before we aggregate them, which can potentially improve the performance of the unsupervised learning models.

Unsupervised learning. To evaluate the proposed method, we employ three modeling techniques: SAE-based, GAAL-based, and ensemble-based techniques. The unsupervised learning models we use in this study are fundamentally different from traditional unsupervised learning models. In the traditional unsupervised learning models (e.g., on text data), the unlabeled documents are categorized into different groups where the intra-group similarity between the documents is high, and the inter-group similarity is low. There is no training/evaluation split in the modeling process. However, in our SAE based models (i.e. SAE-MLP, SAE-LSTM, SAE-GRU), we train the models using a subset of the data, and then using the rest as testing data. By splitting the data, we can assess whether the models are overfitting. The decision rule for the classification is: for a dataset containing both fake and authentic contents, the models are trained using the authentic contents only;

consequently, the reconstruction error on the fake contents is relatively higher. Hence, we can use the reconstruction error the decision criterion in the unsupervised models for classifying fake contents from all the contents. The threshold for the reconstruction error is determined heuristically. In our GAAL based models, a generator G and a discriminator D work in a mini-max game, with G generating outliers (i.e. fake contents) and D classifying them with a decision boundary, in an iterative fashion.

We incorporate a variety of ensemble techniques to boost the performance of the unsupervised learning models. The assumption of ensemble learning is that each base model captures different facets of the patterns that can be used to distinguish fake contents from their authentic counterparts; by merging these facets in a systematic way, the ensemble models are able to capture more comprehensive patterns. We select the following ensemble methods: average, maximization, average-of-maximization, maximum-of-average, and majority vote based methods [25]. It is worth noting that the design of our voting mechanism is different from the mainstream methods. Instead of directly using the predicted class labels (i.e. fake/authentic) for voting, we first collect the predicted probabilities of each social media text being fake or authentic from all the member models, and then conduct a two-tailed t-test to compare the means of the probabilities from the two groups.

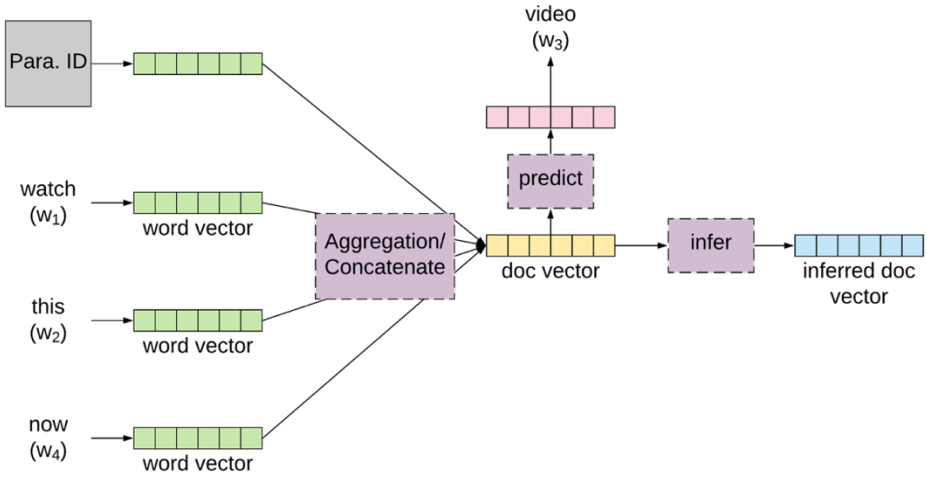

Figure 2. The PV-DM Method for Document Embedding

Model Evaluation. In this step, the modeling results are compared against the actual labels of the contents. We extend common evaluation metrics for supervised learning techniques, such as the Area Under the Receiver Operating Characteristic Curve (AUC) and f1-score, are used for the modeling results of unsupervised learning. In addition to the proposed analytical pipeline, we also evaluate other unsupervised learning techniques such as clustering and PCA and supervised learning techniques as baselines. In this study, we choose f1-score and AUC as evaluation metrics for two reasons: 1) these metrics are less biased when the dataset is imbalanced, compared to other metrics (e.g. classification accuracy); and 2) they allow us to compare the performances of our proposed pipeline to other supervised learning models. If the results are comparable, it would provide strong evidence for the effectiveness of our proposed pipeline.

The proposed analytical pipeline is implemented using Keras [26] and Gensim [27]. In the SAE-based models, the hidden layers are either Multi-Layer Perceptron (MLP) layers, or recurrent layers (e.g., Long Short-Term Memory, Gated Recurrent Unit). We adopt Single Objective Generative Adversarial Active Learning (SO-GAAL) [23] as the GAAL-based technique in our proposed pipeline. All available models are used for the ensemble purposes. Despite that the unsupervised learning nature of our pipeline does not require splitting the data into training and test sets, we reserve 10% of the data for optimizing the proposed models.

3.3. Experiment Data and Configuration

Given the scarcity of labeled data for fake content detection, previous studies have relied on either manually labeled or synthetic datasets. It is also difficult to control the extent to which synthesized contents are similar to or different from the authentic contents. Therefore, we use a dataset collected from YouTube.com, including 1,956 comments on 5 different videos. The dataset was manually labeled as fake or authentic with a collaborative tagging tool [10]. Among them, 1,000 comments are manually labeled as fake, and the rest as authentic. Previous studies (e.g. [7]) indicated that a fairly small

sample size (e.g., 400 genuine and 400 fake consumer reviews) is sufficient for developing detection models.

We extract linguistic and context features from the social media content. All the contents are prepared using standard text preprocessing steps. In terms of linguistic features, each content is converted into a document vector of 300 dimension after a 150-step inference from the trained PV-DM model. The descriptive statistics of some linguistic features is reported in Table 1.

Table 1. Summary Statistics of the Linguistic Features

Feature	Authentic		Fake		t-statistics
	Mean	STDEV	Mean	STDEV	
Content Length	9.58	10.79	23.23	27.89	**8.24***
Average Sentence Length	7.12	6.30	11.41	12.09	**6.75***
Noun Ratio	0.46	0.29	0.47	0.22	1.01
POS Ratio - NN	0.48	0.34	0.50	0.26	1.21
POS Ratio - VB	0.14	0.13	0.14	0.09	0.67
POS Ratio - JJ	0.08	0.12	0.05	0.08	**5.92***
POS Ratio - RB	0.06	0.09	0.02	0.04	**-6.84***
POS Ratio - PR	0.07	0.11	0.08	0.08	0.96
Unique POS Ratio	0.97	0.08	0.96	0.10	-0.61
Content Diversity	0.97	0.12	0.94	0.13	-1.45
Capital Words Count	0.25	0.27	0.36	0.26	**3.44***
Emotiveness	0.30	0.40	0.15	0.20	**8.65***
Self-reference Ratio	0.04	0.08	0.06	0.08	**1.95****
Title Mention	1.18	0.39	1.17	0.38	-1.25
Subjectivity	0.19	0.33	0.05	0.08	**-8.18***
Average Sentiment Scores	0.22	0.38	0.14	0.25	**2.94***

Note: *: $p < 0.01$, **: $p < 0.05$

Based on the t-test results between the authentic and fake contents, as shown in Table 1, fake contents have longer content length, and longer sentences. There are two possible explanations for the observations. One is that the spammers may write longer contents and sentences to disguise their deceptive intentions. The other is that the spammers may reuse the contents they previously wrote for other products/services, as suggested by prior studies (e.g., [15]). The results also show that the emotiveness appears to be higher for the authentic contents, compared to their fake counterparts. This is also confirmed by the ratio of different POSs (e.g. adjective (JJ) and adverb (RB)) across the contents. Since the emotional words indicate the spammers' feelings or mental reactions toward the target products/services, authentic contents express stronger emotions from some actual experiences than the spammers. The explanation is confirmed by the fact that the subjectivity of authentic contents is approximately 3.8 times that of the fake contents (which disclose less sentiment signals). Additionally, we select the compound scores from TextBlob as the metric for the average sentiment score for a piece of social media contents. The compound scores from TextBlob show that the genuine contents are more positive compared to the fake contents. This finding is aligned with prior related studies, which have indicated that malicious negative contents, which attack competitors' products and services, is a common phenomenon [3], [19].

On the other hand, it is also noted from the table that the several linguistic features, such as the ratio of nouns, the ratios of unique word-POS pairs, and the number of mentions of video titles (i.e. analogous to the product/service names) are fairly similar between fake and authentic contents. These features represent the amount of information expressed in the social media contents. Thus, these observations point to the ineffectiveness of detecting fake contents based on the amount of information content at a coarse granularity. Similarly, little difference was detected in capital words counts and self-reference ratio between fake and authentic social media contents. Above findings indicate that additional features (i.e., context features) need to be included in the models, in order to achieve better effectiveness for the purpose of fake content detection.

For the SAEs with MLP layers (SAE-MLP), we consider different model configurations, including four MLP layers consisting of 64, 32, 32, and 64 neurons in each layer, respectively (SAE-MLP1), five layers with

64, 32, 16, 32, and 64 neurons in each layer, respectively (SAE-MLP2), and seven layers with 128, 64, 32, 16, 32, 64, and 128 neurons in each layer, respectively (SAE-MLP3). For the SAEs with recurrent layers (SAE-LSTM and SAE-GRU), each model contains four hidden layers, with 128, 64, 64, and 128 neurons in each layer, respectively. It is also worth noting that all the recurrent layers are made bi-directional. For the GAAL-based technique, we adopt the model structure from the original study [23], with grid search based hyperparameter optimization. As a result, we use *tanh* instead of *ReLU* as the activation function in all the hidden layers in the generator, and the Adam optimizer for training both the generator and the discriminator. We also tune the GAAL based models against different values of the contamination ratios (i.e. the ratio of possible fake contents in the sample) in search for the optimal performance.

4. Results and Discussion

4.1. Experiment Results

Table 2 reports the performances of the proposed analytical pipeline, including all the member models. Among the individual models in the proposed analytical pipeline, SO-GAAL achieved the best performance.

Table 2. Performance of Fake Content Detection

MODEL	COMBINED FEATURES		CONTEXT FEATURES		LINGUISTIC FEATURES	
	AUC	F1	AUC	F1	AUC	F1
SAE-MLP1	.6643	.6451	.6007	.5992	.5102	.5033
SAE-MLP2	.6658	.6469	.6011	.5989	.5087	.5001
SAE-MLP3	.6701	.6488	.6089	.5807	.4962	.4988
SAE-LSTM	.6607	.6441	.6121	.6055	.5088	.5069
SAE-GRU	.6712	.6557	.6155	.6314	.5093	.5100
SO-GAAL	*.7689*	*.7758*	*.7200*	*.7244*	*.5211*	*.5293*
ENS-AVG	.7253	.7189	.6911	.6934	.5003	.5060
ENS-MAX	.7481	.7309	.7008	.6888	.5211	.5339
ENS-MED	.7549	.7214	.7096	.6922	.5063	.5117
ENS-MOA	.7483	.7199	.7040	.6931	.5079	.5184
ENS-AOM	.7470	.7307	.7022	.6899	.5080	.5199
ENS-VOTE	**.7808**	**.7995**	**.7102**	**.7251**	**.5279**	**.5403**

Thus, compared with the SAE based models that rely on latent features extracted between the encoders and decoders in each model to reconstruct the data (features), the generator in the SO-GAAL model is more effective in generating synthesized data for the discriminator to identify the decision boundary. Additionally, it is interesting to observe from the results that the SAE-GRU model yields the highest f1-score for the fake class - indicating that it is the most effective model in detecting fake contents; however, the overall (weighted average) f1-score is relatively low due to its poor performance for the authentic class. Table 2 also shows that the voting based model (ENS-VOTE) performs the best among all ensemble models. The voting is performed via statistical tests, whereas the null hypothesis is that the means of the two groups of probabilities are the same. If the null hypothesis is rejected, the class label with the higher mean probability is assigned to the instance; otherwise, a class label is randomly assigned to the instance. Voting based ensemble methods enable learning from different facets of the data from different member models, thus, the ENS-VOTE model performs better than other bagging-based ensemble models. Additionally, it is shown in Table 2 that the performances across all models using the linguistic and context features combined are superior, compared to the models using just linguistic or context features, respectively. We also observe that the models using the context features outperform the counterparts using the linguistic features, which suggests that the extracted context features as indicated in the proposed approach are indicative for the fake content detection purpose.

4.2. Discussion

Given its pure unsupervised nature of the proposed analytical pipeline in this study, we also compare its performances with other popular unsupervised learning techniques, inlcuding clustering methods, T-distributed Stochastic Neighbor Embedding (T-SNE), and Principal Component Analysis (PCA). It is worth noting that we perform these analyses on the whole dataset.

(a) T-SNE

(b) PCA

Figure 3. Visualization Results

For the clustering technique, we select Birch clustering since it is widely adopted in the domain of outlier detection. With the context features only (where the distance metric is the document similarity of the textual contents), the clustering model (with a silhouette score of 0.3410) yields an AUC score of 0.5405 and an f1-score of 0.5369. With the linguistic features only, the clustering model (with a silhouette score of 0.1326) yields an AUC score of 0.5133 and an f1-score of 0.5664. With the combined features of document vectors and the linguistic features, the clustering model (with a silhouette score of 0.2969) yields an AUC score of 0.6108 and a f1-score of 0.6315.

For the results of T-SNE and PCA models, we conduct visual analyses. To support the visualization, we limit the dimensionality of both models to 2. The visualizations of T-SNE (with a Kullback-Leibler divergence of 1.5246) results and PCA (with a combined explained variance ratio of 0.3549) results are plotted in Figure 3. It is evident from the figure that, although the T-SNE model shows better delineation between the fake and authentic contents, both models show that the two types of contents are not linearly separable.

Given that the experiment dataset contains labels, we compare the performance of the proposed pipeline with some supervised learning models. The best performing supervised learning model is eXtreme Gradient Boosting (XGBoost), with an AUC score of 0.8104 and an f1-score of 0.7986. The results demonstrate that our proposed analytical pipeline, which is purely unsupervised in nature, is comparable to the supervised learning models in the performance of fake content detection.

5. Conclusion and Future Work

As social media become an important data source for the support of various decisions, there is an increasing attention to the problem of fake content. Previous studies toward fake content detection mainly focused supervised machine learning methods using linguistic features. While some studies have incorporated context features into their detection models, the extraction of those features was conducted at the word level, limiting their abilities to capture the textual pattern formed by multiple words. In this study, we propose an analytical pipeline that aim to tackle the above-mentioned research limitations. The proposed methods extract context features at the document level, and combine the context features with linguistic features. More importantly, the methods employ deep learning based unsupervised learning techniques, namely SAEs and GAAL, which are capable of generating potential outliers to better train the generator-discriminator models. Furthermore, we design ensemble modeling techniques to boost the model performance in fake content detection. In particular, the customized voting based ensemble method yields the best results. The experiment results on a collaboratively tagged dataset demonstrate that the proposed analytical pipeline achieved superior performance to unsupervised learning baselines, and comparable performance to prior studies employing supervised learning techniques. The findings of this study can be used to facilitate collaboration in data science by reducing humans' cognitive bias and improve the collaboration efficiency.

Given the exploratory nature of this study, it has limitations that point to the directions for future extensions. First, to understand the generalizability of the proposed analytical pipeline, it needs to be tested on different datasets or social media content collected from different platforms (e.g., online consumer reviews). For instance, comparative studies on detecting fake reviews between experiential and non-experiential goods, or virtual and physical goods, can be an interesting research direction. Secondly, based on our inspection of the immediate results from the proposed pipeline, the latent features learned from the encoder can be used to engineer new features to enhance the performance in fake content detection. In addition, combining unsupervised learning techniques as proposed in this study (i.e., as feature representation

learning) and advanced supervised learning techniques (e.g. attention based neural networks and transfer learning) holds great promise for further enhancing the model performances. Last but not least, it would be helpful to provide interpretable classification results by understanding what feature(s) contributes to fake content detection outcomes, and to what extent. Given that the advanced classification models are typically "black boxes", the interpretation of the classification results can lead to decision rules for fake content detection.

6. References

[1] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," Expert Systems with Applications, vol. 153. Elsevier Ltd, p. 112986, 2020.

[2] Y. R. Chen and H. H. Chen, "Opinion spam detection in web forum: A real case study," WWW 2015 - Proceedings of the 24th International Conference on World Wide Web, no. 1, pp. 173–183, 2015.

[3] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews," Journal of Management Information Systems, vol. 33, no. 2, pp. 456–481, 2016.

[4] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," Management Science, vol. 62, no. 12, pp. 3412–3427, 2016.

[5] S. Kemp, "Digital 2020: Global Digital Overview." [Online]. Available: https://datareportal.com/reports/digital-2020-global-digital-overview.

[6] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication," Group Decision and Negotiation, vol. 13, no. 1, pp. 81–106, 2004.

[7] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, pp. 497–501, 2013.

[8] S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169–175, 2018.

[9] A. Ammari, V. Dimitrova, and D. Despotakis, "Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models," in Proceedings of 19th International Conference on User Modeling, Adaptation and Personalization, 2011.

[10] C. Alberto, J. V Lochter, and T. A. Almeida, "TubeSpam : Comment Spam Filtering on YouTube," in Proceedings of 2015 IEEE 14th International Conference on Machine Learning and Applications, 2015, pp. 138–143.

[11] A. K. Uysal, "Feature Selection for Comment Spam Filtering on YouTube," DATA SCIENCE AND APPLICATIONS, vol. 1, no. 1, pp. 4–8, 2018.

[12] P. Nisha, "N-Gram Assisted Youtube Spam Comment Detection," Procedia Computer Science, vol. 132, no. Iccids, pp. 174–182, 2018.

[13] L. Li, B. Qin, W. Ren, and T. Liu, "Document representation and feature combination for deceptive spam review detection," Neurocomputing, vol. 254, pp. 33–41, 2017.

[14] M. Gong, Y. Gao, Y. Xie, and A. K. Qin, "An attention-based unsupervised adversarial model for movie review spam detection," IEEE Transactions on Multimedia, vol. 9210, no. c, pp. 1–1, 2020.

[15] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," Information Processing and Management, vol. 56, no. 4, pp. 1234–1244, 2019.

[16] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," Journal of Management Information Systems, vol. 35, no. 1, pp. 350–380, 2018.

[17] N. Luo, H. Deng, L. Zhao, Y. Liu, X. Wang, and Z. Tan, "Multi-aspect feature based neural network model in detecting fake reviews," Proceedings - 2017 4th International Conference on Information Science and Control Engineering, ICISCE 2017, pp. 475–479, 2017.

[18] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying Toxicity Within YouTube Video Comment," in Proceedings of the 12th International Conference on Social, Cultural, and Behavioral Modeling, 2019, pp. 214–222.

[19] M. Juuti, B. Sun, T. Mori, and N. Asokan, "Stay on-topic: Generating context-specific fake restaurant reviews," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11098 LNCS, pp. 132–151, 2018.

[20] P. J. Moor, A. Heuvelman, and R. Verleur, "Flaming on YouTube," Computers in Human Behavior, vol. 26, no. 6, pp. 1536–1546, 2010.

[21] A. Sureka, "Mining User Comment Activity for Detecting Forum Spammers in YouTube," 2011.

[22] C. Wu, F. Wu, S. Wu, Z. Yuan, J. Liu, and Y. Huang, "Semi-supervised dimensional sentiment analysis with variational autoencoder," Knowledge-Based Systems, vol. 165, pp. 30–39, 2019.

[23] Y. Liu et al., "Generative Adversarial Active Learning for Unsupervised Outlier Detection," IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2019.

[24] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," Information Sciences, vol. 477, pp. 15–29, 2019.

[25] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python Toolbox for Scalable Outlier Detection," Journal of Machine Learning Research, vol. 20, no. 96, pp. 1–7, 2019.

[26] F. Chollet and others, "Keras." 2015.

[27] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.