5 Sources of Clickbaits You Should Know! Using Synthetic Clickbaits to Improve Prediction and Distinguish between Bot-Generated and Human-Written Headlines

Thai Le*, Kai Shu†, Maria D. Molina*, Dongwon Lee*, S. Shyam Sundar*, Huan Liu†

*The Pennsylvania State University, USA †The Arizona State University, USA
*{tql3, mdm63, dongwon, sss12}@psu.edu †{kai.shu, huanliu}@asu.edu

***Abstract*—Clickbait is an attractive yet misleading headline that lures readers to commit click-conversion. Development of robust clickbait detection models has been, however, hampered due to the shortage of high-quality labeled training samples. To overcome this challenge, we investigate how to exploit human-written and machine-generated *synthetic* clickbaits. We first ask crowdworkers and journalism students to generate clickbaity news headlines. Second, we utilize deep generative models to generate clickbaity headlines. Through empirical evaluations, we demonstrate that synthetic clickbaits by human entities and deep generative models are *consistently* useful in improving the accuracy of various prediction models, by as much as 14.5% in AUC, across two real datasets and different types of algorithms. Especially, we observe an improvement in accuracy, up to 8.5% in AUC, even for top-ranked clickbait detectors from Clickbait Challenge 2017. Our study proposes a novel direction to address the shortage of labeled training data, one of fundamental bottlenecks in supervised learning, by means of synthetic training data with reinforced domain knowledge. It also provides a solution for distinguishing between bot-generated and human-written clickbaits, thus aiding the work of moderators and better alerting news consumers.**

1. Introduction

In Feb. 2018, US President Donald Trump posted on Twitter: *"NEW FBI TEXTS ARE BOMBSHELLS!"*, which has drawn much attention from the public and media. This tweet exhibits many characteristics of *clickbaits*–i.e., catchy social posts or sensational headlines that attempt to lure readers to click. Other examples of clickbaits can be found in Table 1. Clickbaits often hide critical information or fabricate the contents on the landing pages by using exaggerated or catchy wording. Yet, social media has made it possible for clickbaits to quickly go viral, thus a potential means of spreading misinformation.

ASONAM'19, August 27-30, 2019, Vancouver, BC, Canada

ACM 978-1-4503-6868-1/19/08 $15.00
https://doi.org/10.1145/3341161.3342875

TABLE 1: Human-written & machine-generated clickbaits

Human-written clickbaits
Pregnant mother of 12 accused of keeping kids in waste-filled... *54 facts that will change the way you watch disney movies*
Machine-generated clickbaits (by $\mathcal{G}_{\mathrm{VAE}}$)
5 ways criminals will try to scam you about tax this summer *We know your personality based on which cat you choose*
Machine-generated clickbaits (by $\mathcal{G}_{\mathrm{infoVAE}}$)
29 things every student has while in college *29 ridiculously posts about the damn disasters*

Indeed, a Facebook media analysis [1] shows that a clickbait post receives more attention via shares and comments than a non-clickbait one. Viewed as one type of *fake news* in a broad sense [2], clickbaits not only frustrate readers, but also violate the journalistic code of ethics [3]. Scholars have argued that the current trend toward merging commercial and editorial interests by means of clickbaits is severely detrimental to the overall information ecosystem, particularly posing a threat to societal/democratic values [4].

Therefore, it has become critically important to develop proactive solutions to the use of clickbaits. By and large, existing approaches have tended to focus on the "*postmortem*" approach–i.e., assuming that clickbaits are out there, how to develop computational solutions to best detect them (e.g., [1], [5], [6], [7]). While these prior works are important and effective, their performance is highly dependent on the quantity and quality of training datasets available. However, there is a paucity of high-quality labeled training datasets that are heterogeneous in sources and large in quantity. This is because annotating labels is expensive, and most of existing datasets are passively collected from external sources, instead of actively generated. To overcome this problem at large, we propose a research question: *how to generate new headlines and titles that resemble real-life clickbaits and how to use them as additional training samples to improve clickbait detection models?*

To systematically study this question, we commission various human entities (e.g., crowdworkers and journalism students) and deep generative models in simulated experiments to generate clickbaits from scratch. We refer to these clickbaits that are generated under *simulated intent,*

i.e., to generate attractive headlines for news articles, as **synthetic clickbaits**. Furthermore, by simulating a similar intent in generating attractive headlines, we hypothesize that synthetic clickbaits might share some similarities in the use of language (e.g., writing style, word choice, grammar patterns) with clickbaits collected in real life, which might be different from that of non-clickbait. We tested this proposition by examining whether generated synthetic clickbaits can be used as additional training examples to strengthen classification performance. We also want to compare these synthetic clickbaits with synthetic data sampled by Synthetic Minority Over-sampling Technique (SMOTE) in terms of predictability, how they capture the original NLP features in terms of distribution, robustness, and interpretability.

Considering that generated clickbaits can resemble various characteristics of real clickbaits, malicious publishers might take advantage of different entities (e.g., machines) to generate vast amount of clickbaits to disseminate low-quality content just to attract online traffic. As a proactive defense against this potential, we further examine the use of Machine Learning (ML) models to verify news sources, in order to differentiate various types of clickbaits. Formally, we propose the following research questions:

RQ1 How do we generate synthetic clickbaits from raw training samples as additional training samples to improve supervised-learning clickbait detection models?

RQ2 What are the differences among synthetic clickbaits generated by humans, generative models, and statistical method in terms of predictive power, NLP feature encapsulation, robustness and interpretability?

RQ3 How can we differentiate clickbaits based on sources (e.g., machine versus human-written clickbaits)?

By answering these research questions, our paper makes the following contributions:

- Overcoming the lack of labeled training samples by exploiting human and deep generative models, we generate diverse types of synthetic clickbaits.
- We demonstrate that using both raw and synthetic clickbait samples (generated from raw samples) consistently improve clickbait detection models by up to 14.5% in AUC, even outperforming SMOTE and two top-performed clickbait detection algorithms from Clickbait Challenge 2017.
- Leveraging deep learning models, we generate more interpretable oversampling data that also better capture the distribution of NLP-based domain knowledge from original clickbaits compared with SMOTE.
- We demonstrate that the clickbaits generated by different entities have significant differences in features so that ML models can differentiate them with an accuracy of 20%–39% higher than random guesses.

2. Literature Review

Collecting, Generating Clickbaits. Researchers have attempted to collect and build labeled clickbait datasets, by using the following approach. First, recognizing that certain online news media outlets frequently use catchy headlines, researchers collect headlines from such sites as candidate clickbaits. Second, as the definition of clickbaits is often fuzzy and subjective, researchers tend to rely on the voted labels of candidate clickbaits from human judges or crowdworkers (e.g. [5], [7], [8]). However, the generation aspect of clickbaits was never a focus in these works.

A recent attempt to "generate" clickbaits is found in Click-O-Tron[1] that trains the RNN with millions of articles from sites such as BuzzFeed, Huffington Post, and Upworthy. Algorithmically, this line of work can be derived from the task of language modeling and text generation in AI. There has been considerable progress in generating realistic text, either in randomized (e.g., [9]) or controllable fashion (e.g., [10], [11]). Leveraging these developments, our work adopts VAE-based generative models [12], [13] to demonstrate the generation of realistic clickbaits. Unlike these works, however, in Section **RQ1**, we also illustrate experimental designs to generate clickbaits by different human creators (e.g., crowdworkers and journalism students). Throughout this paper, we adopt the two public datasets curated by [7] collected from *Professional* publishing websites, and by [5] collected from *Social Media Twitter* as datasets $\mathcal{P}$ and $\mathcal{M}$ respectively.

TABLE 2: Statistics of five types of synthetic clickbaits

Statistics	$\mathcal{P}$	$\mathcal{M}$	$\mathcal{C}$	$\mathcal{S}$	$\mathcal{A}$
Avg # words	10.27	11.00	11.83	8.61	10.18
Std. Dev.	4.34	4.39	4.44	2.79	2.94
Avg # chars	46.81	51.95	56.6	42.32	44.69
Std. Dev.	12.61	19.96	21.12	20.32	12.48

Postmortem: Detecting Clickbaits. Clickbait detection has attracted increasing attention in recent years. Most of existing clickbait detection approaches explore engineering features in a supervised ML framework (e.g. [7], [14], [15]). More recently, researchers have employed the deep neural framework to automatically learn latent features from clickbaits (e.g. [1], [8], [16]). Many of these attempts focus on extracting different features and building a predictive model to approach the problem, yet they are bounded by the availability and quality of existing labeled training datasets. Note that our paper is *not* aiming at directly comparing against these existing works. Rather, our ideas in **RQ1** explore the potentials of generating synthetic clickbaits and utilizing them in improving detection models further.

3. Generating Synthetic Clickbaits (RQ1)

We begin with two raw datasets, $\mathcal{P}$ and $\mathcal{M}$, representing two dominant sources of clickbaits prevalent today– i.e., mainstream news media and general social media, respectively [1]. Then, to generate synthetic clickbaits from raw datasets, we explore two types of human sources (i.e., crowdworkers as novice users and journalism majors as domain experts) and VAE-based generative models. Table 2 compares the lengths of synthetic clickbaits from all five entities.

1. http://clickotron.com/

TABLE 3: Experiment datasets

Dataset	Description	#Pos	#Neg
$\mathcal{P}_{train}$	Training set from $\mathcal{P}$	2,239	11,201
$\mathcal{P}_{test}$	Testing set from $\mathcal{P}$	960	4,800
$\mathcal{M}_{train}$	Training set from $\mathcal{M}$	3,681	11,337
$\mathcal{M}_{test}$	Testing set from $\mathcal{M}$	1,578	4,859
$\mathcal{C}$	Training set from workers	778	0
$\mathcal{S}$	Training set from Students	785	0
$\mathcal{A}^{\mathcal{P}}_{\text{VAE}}$	$\mathcal{G}_{\text{VAE}}$ trained on $\mathcal{P}_{train}$	8,962	0
$\mathcal{A}^{\mathcal{M}}_{\text{VAE}}$	$\mathcal{G}_{\text{VAE}}$ trained on $\mathcal{M}_{train}$	7,656	0
$\mathcal{A}^{\mathcal{P}}_{\text{infoVAE}}$	$\mathcal{G}_{\text{infoVAE}}$ trained on $\mathcal{P}_{train}$	8,962	0
$\mathcal{A}^{\mathcal{M}}_{\text{infoVAE}}$	$\mathcal{G}_{\text{infoVAE}}$ trained on $\mathcal{M}_{train}$	7,656	0
$\mathcal{O}^{\mathcal{P}}$	SMOTE on $\mathcal{P}_{train}$	8,962	0
$\mathcal{O}^{\mathcal{M}}$	SMOTE on $\mathcal{M}_{train}$	7,656	0

Crowdworkers-Generated Clickbaits $\mathcal{C}$: To collect clickbaits generated by crowdworkers, we utilize the Amazon MTurk (AMT) platform. From the articles used in the Clickbait Challenge 2017, we first filtered out very short articles with less than 50 words in content. In addition, for very long articles with more than 500 words in content, we presented only the first 500 words to reduce the amount of reading for workers. As the first 3-4 paragraphs of news articles often summarize the content, the first 500 words sufficiently captured the gist of the articles. Then, we recruited AMT workers located in US (who are more likely to be familiar with the topics of the articles) with approval rates > 0.95.

In the MTurk task, next, we first showed a Wikipedia link[2] with the definition of clickbait, but did not provide additional information that might influence the way workers generated clickbaits. Second, for each article shown, we asked workers to read the article and write a clickbait headline, with no more than 25 words. In the end, 85 workers generated 778 clickbait headlines for 200 selected articles. This provided us a total of 62 articles with 3 different clickbait headlines, 113 articles with 4 clickbaits, 10 articles with 5 clickbaits, and 15 articles having 6 clickbaits.

Student-Generated Clickbaits $\mathcal{S}$: Another source for headline creation was undergraduate students who are being trained to learn about the art and craft of journalistic writing and reporting. We recruited participants from 8 different classes at a large northeastern university in US. Participants received extra course credit for their participation. Because we wanted to include participants with different levels of expertise, we recruited from 3 lower-level classes and 5 upper-level classes. Participants in the lower-level classes represent *amateurs* who are beginning to learn about the journalistic style of writing, whereas those from the upper-level classes represent students who are *semi-experts* and have an advanced understanding of the principles of reporting and headline creation. A total of 125 students participated (i.e., 76.8% and 23.2% from lower- and upper-level classes, respectively).

The design principle and articles used to generate these headlines were the same as the ones used for AMT participants. We first provided students with a definition of clickbait, without providing additional information, and asked them to generate a clickbait headline, with no more than 25 words. Each student completed an average of 6 headlines, ranging from 1 to 22. The students generated 785 clickbaits in total.

Algorithm-Generated Clickbaits $\mathcal{A}$: Due to recent advancements in generative models, next, we turn to machine-generated clickbaits. We utilize different variations of VAE-based generative models in the task of generating synthetic clickbaits. VAE-based generative models are selected because the latent code z learned from the model appears to encapsulate information about the number of tokens, and their parts of speech (POS) and topics [12], all of which are shown to be effective predictive NLP features in differentiating clickbaits from non-clickbaits (e.g., [3], [7], [14]). We utilized the two generative models, namely VAE and infoVAE, as introduced in [12], [13] to generate synthetic clickbaits. While the first model uses original VAE loss function [12], the second model uses Maximum Mean Discrepancy with a Gaussian Kernel [13] to replace the original KL divergence term. We denote the synthetic datasets generated by the two models $\mathcal{A}_{\text{VAE}}$ and $\mathcal{A}_{\text{infoVAE}}$ respectively. We refer the readers to their original papers for objective functions formulation and as well as optimization techniques. Table 1 lists some the examples of clickbaits generated by the two models trained on a subset of clickbaits drawn from $\mathcal{M}$ and $\mathcal{P}$.

4. Assessing Synthetic Clickbaits (RQ2)

In this section, we seek to differentiate *synthetic clickbaits* generated by humans, generative models, and SMOTE with four analytic questions (AQs) as follows:

AQ1 Predictive Power: How much do synthetic clickbaits help improve ML models in detecting clickbaits?

AQ2 NLP Encapsulation: How well do synthetic clickbaits encapsulate NLP feature distribution from the training dataset?

AQ3 Robustness: What is the minimum amount of synthetic clickbaits needed to improve ML clickbait detection model? Is such an improvement proportional to the increase in synthetic clickbaits?

AQ4 Interpretability: Are synthetic clickbaits interpretable to humans?

4.1. Predictive Power of Synthetic Clickbaits (AQ1)

We examine how much generated synthetic clickbaits described in **RQ1** can improve ML clickbait detection models. We name this process of generating a large amount of synthetic data to enhance supervised learning tasks as *Synthesized Supervised Learning (SSL)*. We further compare SSL with SMOTE and top-2 performed clickbait detectors from Clickbait Challenge 2017 (CBC)[3]. For classical ML algorithms, we use NLP-based features as input. Since our

2. https://en.wikipedia.org/wiki/Clickbait

3. https://clickbait-challenge.org

TABLE 4: ***AQ1**:Mean AUC scores and their relative changes (%)* on $\mathcal{P}_{test}$ using different *oversampling* methods.

	Baseline	Synthesized Supervised Learning (SSL)				SMOTE
Algorithms	$\mathcal{P}_{train}$	$\mathcal{P}_{train} \cup \mathcal{C}$	$\mathcal{P}_{train} \cup \mathcal{S}$	$\mathcal{P}_{train} \cup \mathcal{A}^{\mathcal{P}}_{\text{VAE}}$	$\mathcal{P}_{train} \cup \mathcal{A}^{\mathcal{P}}_{\text{infoVAE}}$	$\mathcal{P}_{train} \cup \mathcal{O}^{\mathcal{P}}$
AdaBoost	0.88	0.88 (-0.11%)	0.88 (-0.11%)	**0.91 (+2.79%)**	0.90 (+1.98%)	0.89 (+1.23%)
Bagging Clf	0.88	0.89 (+0.74%)	0.89 (+0.67%)	**0.91 (+3.05%)**	0.90 (+1.98%)	0.88 (+0.22%)
Decision Tree	0.86	0.87 (+0.71%)	0.87 (+0.66%)	**0.89 (+2.95%)**	0.87 (+0.62%)	0.86 (+0.33%)
GradientBoosting	0.89	0.89 (-0.32%)	0.89 (-0.16%)	**0.92 (+3.13%)**	0.91 (+1.94%)	0.90 (+1.23%)
KNeighbors Clf	0.83	0.83 (+0.08%)	0.83 (+0.08%)	**0.88 (+6.89%)**	0.86 (+4.64%)	0.87 (+5.06%)
Logistic Regression	0.91	0.90 (-0.69%)	0.90 (-0.69%)	**0.92 (+1.48%)**	0.92 (+1.22%)	0.92 (+1.14%)
Naive Bayes	0.85	0.82 (-2.74%)	0.82 (-2.74%)	0.86 (+2.14%)	**0.87 (+3.11%)**	0.86 (+2.18%)
Random Forest	0.87	0.88 (+0.80%)	0.87 (+0.44%)	**0.91 (+4.20%)**	0.89 (+2.89%)	0.88 (+0.72%)
SVM	0.86	0.86 (+0.38%)	0.86 (+0.25%)	**0.92 (+6.85%)**	0.91 (+5.49%)	0.92 (+6.84%)
albacore (#1 in CBC)	0.95	–	–	**0.97 (+1.49%)**	0.95 (+0.27%)	–
zingel (#2 in CBC)	0.93	–	–	**0.95 (+1.86%)**	0.94 (+1.29%)	–

TABLE 5: ***AQ1**:Mean AUC scores and their relative changes (%)* on $\mathcal{M}_{test}$ using different *oversampling* methods.

	Baseline	Synthesized Supervised Learning (SSL)				SMOTE
Algorithms	$\mathcal{M}_{train}$	$\mathcal{M}_{train} \cup \mathcal{C}$	$\mathcal{M}_{train} \cup \mathcal{S}$	$\mathcal{M}_{train} \cup \mathcal{A}^{\mathcal{M}}_{\text{VAE}}$	$\mathcal{M}_{train} \cup \mathcal{A}^{\mathcal{M}}_{\text{infoVAE}}$	$\mathcal{M}_{train} \cup \mathcal{O}^{\mathcal{M}}$
AdaBoost	0.68	0.69 (+1.12%)	0.69 (+1.12%)	**0.74 (+8.60%)**	0.71 (+4.74%)	0.72 (+5.80%)
Bagging Clf	0.67	0.67 (+0.09%)	0.67 (+0.42%)	**0.71 (+7.11%)**	0.68 (+2.84%)	0.67 (+0.93%)
Decision Tree	0.64	0.65 (+1.15%)	0.65 (+1.67%)	**0.67 (+3.66%)**	0.66 (+3.39%)	0.65 (+1.41%)
GradientBoosting	0.69	0.69 (+0.88%)	0.69 (+0.70%)	**0.74 (+7.93%)**	0.71 (+3.69%)	0.71 (+3.04%)
KNeighbors Clf	0.64	0.64 (+0.35%)	0.64 (+0.40%)	**0.69 (+8.11%)**	0.68 (+6.22%)	0.66 (+3.72%)
Logistic Regression	0.70	0.70 (+0.04%)	0.70 (+0.04%)	**0.74 (+6.02%)**	0.72 (+3.15%)	0.75 (+6.66%)
Naive Bayes	0.66	0.63 (-4.74%)	0.63 (-4.74%)	0.70 (+5.64%)	0.67 (+1.02%)	**0.72 (+7.96%)**
Random Forest	0.65	0.66 (+0.76%)	0.66 (+0.67%)	**0.71 (+9.42%)**	0.69 (+6.66%)	0.65 (+0.47%)
SVM	0.65	0.66 (+1.55%)	0.66 (+1.53%)	**0.75 (+14.56%)**	0.71 (+8.1%)	0.75 (+14.51%)
albacore (#1 in CBC)	0.71	–	–	**0.77 (+8.5%)**	0.75 (+6.0%)	–
zingel (#2 in CBC)	0.71	–	–	**0.76 (+6.9%)**	0.74 (+4.55%)	–

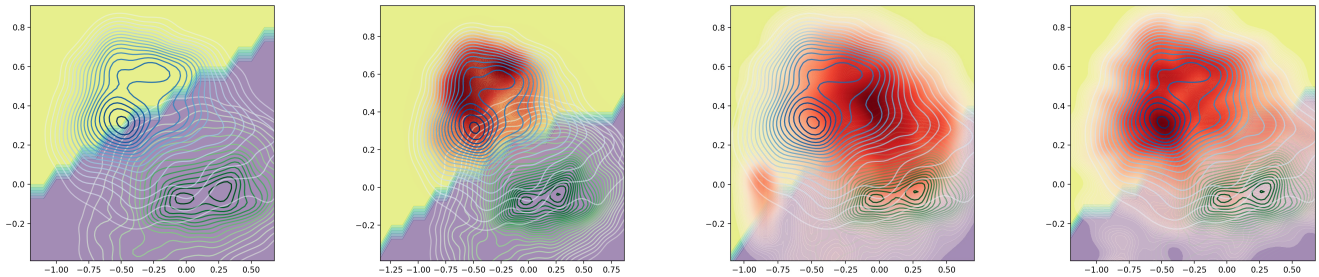
work does not aim to develop new features for predicting clickbaits, simply, we have selected several features from the literature that manifests different nuances in the use of language for writing headlines. They are selected because of their reported effectiveness in detecting clickbait or misleading headlines across different published works (e.g., [3], [5], [6], [7], [14]). Except for general POS-N-gram and Word-N-grams features, Table 6 lists all of the selected features. Being deep learning based, two top-performing models from CBC (albacore and zingel) automatically learn feature representation from a large amount of raw text data. Due to limited number of human-written synthetic clickbaits, we only examine this with machine-generated clickbaits. Since SMOTE cannot over-sample on raw data space, it is not applicable for the deep learning based detectors. We used open-source implementations published on the CBC website for two deep learning based models.

We first constructed training and testing sets from $\mathcal{P}$ and $\mathcal{M}$ in the ratio of 3:1, resulting in $\mathcal{P}_{train}, \mathcal{P}_{test}$ and $\mathcal{M}_{train}, \mathcal{M}_{test}$ respectively. Then, to see if synthetic clickbaits are useful to improve the detection of clickbaits, when they are added as additional labeled training samples, we first used *only* the positive training data in each of the datasets $\mathcal{P}_{train}$ and $\mathcal{M}_{train}$ to train generative models $\mathcal{G}_{\text{VAE}}$ and $\mathcal{G}_{\text{infoVAE}}$ as described in **RQ1**. The trained models were subsequently used to generate four respective synthetic datasets $\mathcal{A}^{\mathcal{P}}_{\text{VAE}}$. $\mathcal{A}^{\mathcal{M}}_{\text{VAE}}$, $\mathcal{A}^{\mathcal{P}}_{\text{infoVAE}}$, $\mathcal{A}^{\mathcal{M}}_{\text{infoVAE}}$. Next, we combined these with the original training sets, $\mathcal{P}_{train}$ and $\mathcal{M}_{train}$, to train different predictive models, and tested against the original testing sets, $\mathcal{P}_{test}$ and $\mathcal{M}_{test}$, respectively. Table 3 summarizes the datasets in this research.

TABLE 6: *NLP Features Descriptions*

Type	Feature Description
Summary Statistics	Average word length, Stop-words ratio Counts of words, POS tags Length of the longest word
Sentiment	Intensity Score
Forward References	Pattern: (this/these/etc.) + Noun
Linguistic Patterns	Pattern: Number + Noun + That ? Pattern: Number + Noun + Verb ? Starting with a number, 5WH?
Informality	Flesch-Kincaid score Counts of Internet Slangs
Special Indicators	".", "!", "?", "@"; "http", "#"; "***"

Area-Under-the-Curve (*AUC*) is selected as the main evaluation measure for their robustness toward skewed labels distribution [17] of testing sets $\mathcal{P}_{test}$ and $\mathcal{M}_{test}$, where there exist 3–5 times more non-clickbaits than clickbaits. Such imbalanced distribution reflects real challenges, where we usually have many more non-clickbait than clickbait text.



(a) *Baseline* (b) $\mathcal{G}_{\text{VAE}}$ (c) $\mathcal{G}_{\text{infoVAE}}$ (d) SMOTE

Figure 1: **AQ2**:Decision boundary of a trained SVM classifier on $\mathcal{P}_{train}$ changed with and without different additional synthetic clickbaits. Blue contour, green contour, and red shade depict the density of positive, negative, and synthetic clickbaits, respectively.

For each algorithm, we reported relative changes of *AUC score* between models trained *with* and *without* additional synthetic clickbait datasets (baseline).

Tables 4 and 5 summarize experimental results. Numbers in **Bold** and underline indicate best of each row and column respectively. We showed that our framework helped improve on both algorithm-wise and dataset-wise. Particularly, from the same original training set, our approach of generating and using synthetic clickbaits was able to enhance the detection performance of both NLP-based and deep-learning based algorithms. Especially, $\mathcal{A}^{\mathcal{P}}_{\text{VAE}}$ and $\mathcal{A}^{\mathcal{M}}_{\text{infoVAE}}$ *consistently* improved *AUC scores* across all algorithms. Interestingly, performance of NLP-based algorithms with the proposed data-enhancement approach achieved comparable ($\mathcal{P}_{test}$) or even better ($\mathcal{M}_{test}$) than top-ranked deep learning models without the need of collecting any additional *real* data. Furthermore, clickbaits generated by generative models outperformed over-sampled data synthesized by SMOTE on all predictive algorithms for $\mathcal{P}_{test}$, and 8 out of 9 cases for $\mathcal{M}_{test}$. Noticeable, generative models outperformed SMOTE significantly on all of the ensemble-based classifiers across the testing sets on $\mathcal{M}_{test}$. The proposed framework can even further improve performance of top-ranked models from CBC by as much as 8.5%, achieving the best performance overall in both datasets.

In summary, we demonstrated that both models $\mathcal{G}_{\text{VAE}}$ and $\mathcal{G}_{\text{infoVAE}}$ could generate synthetic clickbaits (learned from training data) that, when added to training data, significantly improved domain-engineered predictive models. The fact that these NLP features have been built in many non-computational domains (e.g., journalism, communication, social science) illustrates that one may leverage the capability of generative models to model complex natural language distribution that reinforce our domain knowledge.

4.2. NLP Feature Encapsulation (AQ2)

From the strong results reported in **AQ1**, we then examine whether synthetic clickbaits share the same distribution of NLP features as real positive clickbaits. We achieved this by both (1) visual examination and (2) analytic testing. For visual examination, we used $\mathcal{P}$ as an illustration. We first trained an Isomap dimension reduction model [18] on $\mathcal{P}_{train}$, and used the trained model to project the features extracted from $\mathcal{A}_{\text{VAE}^{\mathcal{P}}}$, $\mathcal{A}_{\text{infoVAE}^{\mathcal{P}}}$ and $\mathcal{O}^{\mathcal{P}}$ into a 2D feature space. Next, we trained an SVM classifier with new features and plotted its decision boundary between two class samples, resulting in Figure 1. Even though SMOTE over-sampled data directly on NLP features space, many new samples are mis-located in the original negative samples' area. In fact, without directly learning from feature set, NLP feature distributions of $\mathcal{A}_{\text{VAE}^{\mathcal{P}}}$ and $\mathcal{A}_{\text{infoVAE}^{\mathcal{P}}}$ are highly overlapped with the original positive samples. Especially, that of $\mathcal{A}_{\text{VAE}^{\mathcal{P}}}$ neatly concentrated around the center of original positive samples, while that of $\mathcal{A}_{\text{infoVAE}^{\mathcal{P}}}$ is located near the boundary between two original classes. Therefore, we can see that Figure 1 confirms predictive results of SVM on $\mathcal{P}_{test}$ in Table 4.

To analytically test, next, for each clickbait x_i in a synthetic dataset $\mathcal{Q}$, we extracted different NLP features listed in Table 6 and used a K Nearest-Neighbor (KNN) searching model to find its k nearest samples from the original training set $\mathcal{T}$ ($\text{NN}_{\mathcal{T}}(k, x_i)$) in the feature space, and calculated the ratio between the number of positive samples found over k. For each generated synthetic dataset, we averaged all the ratios to total $\mathcal{N}$ number of data points in $\mathcal{Q}$ to calculate a statistic:

$$Overlap_{\text{NLP}}(k, \mathcal{Q}, \mathcal{T}) = \frac{1}{\mathcal{N}} \sum_{x_i \in \mathcal{Q}} \frac{|\text{NN}_{\mathcal{T}}(k, x_i) \cap \mathcal{T}_{pos}|}{k}. \tag{1}$$

This statistic captures on average how likely generated samples of a synthetic dataset $\mathcal{Q}$ will be close to original positive clickbaits in the feature space. We calculated such measure for each generated synthetic dataset and illustrated the result in Table 7. This result shows that $\mathcal{G}_{\text{VAE}}$-generated clickbaits are the one most overlapping with the original positive samples in the features space, which coincides with our visual examination in Figure 1. Overall, even though SMOTE directly generated data on the set of NLP predictive features, both $\mathcal{G}_{\text{VAE}}$ and $\mathcal{G}_{\text{infoVAE}}$ were better in capturing similar NLP structures of original clickbait data, resulting in better prediction of ML models than those NLP features,

TABLE 7: ***AQ2**:$Overlap_{\mathrm{NLP}}$ score with $k = 5$ of synthetic datasets on $\mathcal{P}_{train}$ and $\mathcal{M}_{train}$ (the higher, the better)*

Statistic	$\mathcal{A}_{\mathrm{VAE}}$	$\mathcal{A}_{\mathrm{infoVAE}}$	$\mathcal{S}$	$\mathcal{C}$	SMOTE
$\mathcal{P}_{train}$	**0.7**	0.44	0.49	0.49	0.4
$\mathcal{M}_{train}$	**0.5**	0.44	0.38	0.38	0.35

as reported in **AQ1**.

4.3. Robustness of Synthetic Clickbaits (AQ3)

In **AQ2**, we illustrated that different types of synthetic clickbaits improved clickbait detection models to different extents. In this section, we examine the robustness of them. Because of limited data samples generated by human users, we focus on the comparison between generative models and SMOTE. We only demonstrate on NLP-based models due to limited computational resources. We measure the robustness of an oversampling method by answering: (1) does a method improve predictive models with a small amount of additional generated samples? and (2) is such an improvement consistent as more data is added to the training set?

Figures 2 and 3 plot the relations between the amount of additional positive clickbait samples generated by $\mathcal{G}_{\mathrm{VAE}}$, $\mathcal{G}_{\mathrm{infoVAE}}$, SMOTE, and their improvements in absolute AUC for all of the examined algorithms. Regarding the performance on $\mathcal{P}_{test}$, generative models only needed 20% of total additional training data until balanced to outperform the baseline across all of the algorithms, while SMOTE needed as much as 25% to achieve the same result. However, such performance of generative models showed a much larger improvement margin compared to SMOTE. Especially, as we add more synthetic clickbaits generated by $\mathcal{G}_{\mathrm{VAE}}$, the improvement was more consistent, showing smoother improvement lines in absolute AUC, compared to the cases of $\mathcal{G}_{\mathrm{infoVAE}}$ and SMOTE. The same outcome was also observed in the case of $\mathcal{M}_{test}$. In fact, only 30% of total clickbaits generated by $\mathcal{G}_{\mathrm{VAE}}$ was needed to outperform 100% of data sampled by SMOTE (balanced training set) in most algorithms in both datasets.

Overall, generative algorithms generated more robust synthetic clickbaits than SMOTE, showing consistent and continuous improvements while adding more training data.

4.4. Interpretability of Synthetic Clickbaits (AQ4)

Humans and deep generative models, $\mathcal{G}_{\mathrm{VAE}}$ and $\mathcal{G}_{\mathrm{infoVAE}}$, clearly have advantages over SMOTE in terms of interpretability. Algorithms-wise, SMOTE samples data only on feature space, i.e., numerical features extracted from text domain, the results of which cannot be converted back to the data space, i.e., natural text. However, $\mathcal{G}_{\mathrm{VAE}}$ and $\mathcal{G}_{\mathrm{infoVAE}}$ learn and generate natural sentences that are interpretable to humans (e.g., Table 1). This shows that while the samples generated by SMOTE are task-independent, i.e., they are represented only on a pre-defined set of features, the sentences produced by generated models can transfer to other tasks or domains such as misinformation analysis.

5. Differentiate Clickbaits per Sources (RQ3)

Next, we ask if entity-cross differences among synthetic clickbaits are consistent and identifiable by ML models. This type of study can be also useful in a security scenario– e.g., malicious publishers take an advantage of different entities to generate clickbaits to propagate low-quality news content, or to attract more traffic. A demo system such as Click-O-Tron[4] and Link Bait Title Generator[5] illustrates the possibility of such an attack scenario to mass-generate clickbaity headlines with malevolent intents.

From the synthetic clickbait datasets in Table 3, we aim to achieve the following specific objectives:

Obj1 Can we distinguish among clickbaits in $\mathcal{P}$, $\mathcal{M}$, $\mathcal{C}$, $\mathcal{S}$, $\mathcal{A}$?

Obj2 Can we distinguish among clickbaits by trained writers ($\mathcal{P} \cup \mathcal{S}$), general public ($\mathcal{M} \cup \mathcal{C}$), and machine ($\mathcal{A}$)?

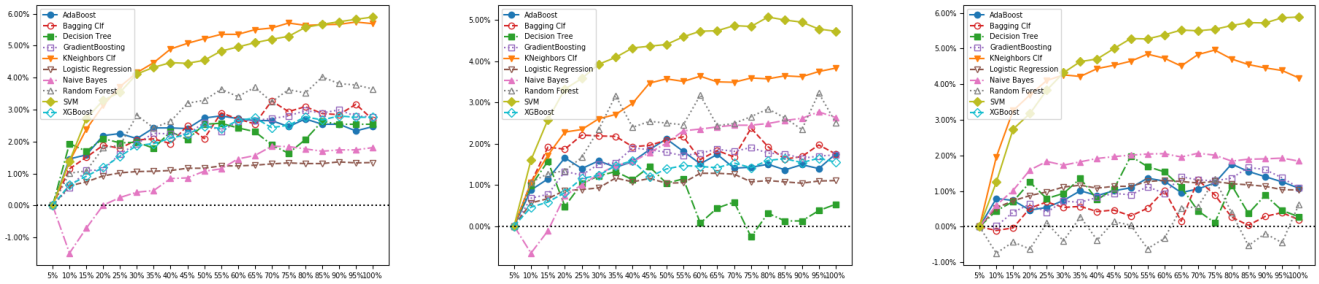
Obj3 Can we distinguish clickbaits by humans ($\mathcal{P} \cup \mathcal{M} \cup \mathcal{C} \cup \mathcal{S}$) vs. machine ($\mathcal{A}$)?

These tasks can be modeled as three different multinomial classification problems. Since the nature of these tasks is similar to the ones in the previous section, we re-use some of the introduced algorithms by changing the ground-truth labels accordingly. From analysis in section **RQ2**, we select $\mathcal{A}_{\mathrm{VAE}}$ as the representative synthetic clickbait set generated by machine $\mathcal{A}$ because it better captures characteristics of real clickbaits than $\mathcal{A}_{\mathrm{infoVAE}}$.

Table 8 summarizes the experimental results, where baselines (i.e., random guess) have accuracies of 20% for **Obj1** in differentiating clickbaits of five different entities, 33.3% for **Obj2** in distinguishing between trained writers, general public and machine, and 50% for **Obj3** in classifying between human-written and machine-generated clickbaits. Note that all three objectives can be achieved with accuracy as high as 59%, 61% and 70%, all of them considerably higher than those of baselines. Overall, while it is challenging to differentiate clickbaits written by different sources, we achieve reasonable results on *accuracy* and *average F1 score* measures. It is especially encouraging that we can distinguish clickbaits generated by machine from those generated by human with as high as 65% in averaged F1 score. This further demonstrates the utility of our synthetic clickbaits in developing models that have strong potential for empirical use.

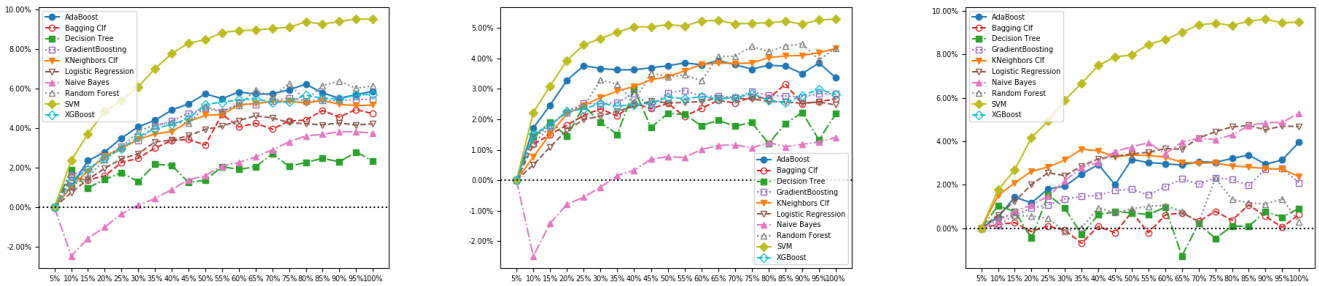
Table 9 illustrates top predictive features resulting from the Gradient Boosting classifier trained for the three objectives. Overall, since we are grouping some entities in **Obj1** to examine **Obj2** and **Obj3**, the result shows many repeated top features across all three tasks. Among the five groups of clickbait headlines, average word length is the most distinguishable feature. Journalism students use longer words in their clickbaits compared to other entities. We find that crowdworkers and students use significantly lower number of Wh-determiners (which, that, etc. as determiners) in their headlines compared to other sources. Also, professional

4. http://clickotron.com/
5. http://www.contentrow.com/tools/link-bait-title-generator



(a) $\mathcal{G}_{\text{VAE}}$ (b) $\mathcal{G}_{\text{infoVAE}}$ (c) SMOTE

Figure 2: **AQ3**:Proportion of additional synthetic clickbaits versus absolute AUC score improvement from baseline on $\mathcal{P}_{test}$



(a) $\mathcal{G}_{\text{VAE}}$ (b) $\mathcal{G}_{\text{infoVAE}}$ (c) SMOTE

Figure 3: **AQ3**:Proportion of additional synthetic clickbaits versus absolute AUC score improvement from baseline on $\mathcal{M}_{test}$

TABLE 8: Clickbaits' Source Verification Benchmark

Alg	Obj1		Obj2		Obj3	
	Acc	F1_avg	Acc	F1_avg	Acc	F1_avg
LogReg	0.54	0.54	0.59	**0.58**	0.61	0.62
NBayes	0.52	0.50	0.56	0.54	0.57	0.59
DTree	0.47	0.47	0.50	0.50	0.61	0.61
RForest	0.54	0.53	0.56	0.55	0.67	**0.65**
XGBoost	0.58	0.52	0.61	0.55	0.70	0.59
AdaBoost	0.55	0.50	0.59	0.55	**0.70**	0.62
SVM	0.57	**0.56**	0.60	0.58	0.61	0.57
GradBoost	**0.59**	0.55	**0.61**	0.56	0.70	0.60
Bagging	0.54	0.53	0.57	0.56	0.66	0.64
KNeighbor	0.51	0.49	0.53	0.51	0.66	0.64

TABLE 9: Top distinguishing features

Top	Obj1	Obj2	Obj3
1	avg word length	# of end mark	# of end mark
2	# of words	avg word length	avg word length
3	% of stop words	# of words	% of stop words
4	# of end mark	% of stop words	# of POS tags
5	# of POS tags	start with number?	# of JJ-NN

writers use personal pronouns (I, you, he, she, etc.) much more often than other entities. We also see differences in other writing strategies among the five entities. Specifically, professional writers are more likely to start their clickbaits with numbers (e.g. "20 things to do before 20"), and media users are more more likely to use question and exclamation marks and more than single sentence in their headlines.

Despite the fact that the generative algorithms can be biased towards the type of clickbaits having the majority of training samples (professional writers), the fact that it still generates clickbaits that simulate human behaviors, which eventually makes it very challenging for us to differentiate, is intriguing. As indicated in Table 9, many of the features that best distinguish the two groups of clickbaits are counts of various POS tags and their combinations. The generative algorithm's strategy might have been learning to replicate different collocations from human-written clickbaits. Moreover, it also learns the relative position of those phrases.

6. Discussion

Utility of Synthetic Text. There have been prior works in generating natural-looking, realistic, and human-readable synthetic text (e.g., [10], [11], [19]). However, few of them have explored the characteristics and utility of synthetic text for downstream machine learning tasks such as prediction and clustering. In fact, to perform well in these machine learning tasks, synthetic text does not have to be realistic and coherent, but must capture certain characteristics or domain knowledge of original text. By using clickbait domain as a case study, in this work, we have demonstrated that synthetic text (generated by diverse methods) can help improve classification tasks and introduce insights into domain specific problems. Generalizing the findings to other domains and applications will be our future work.

Implications for Combating Misinformation. Our findings highlight the promise in using generative algorithms to

detect misinformation (spam, fake news, etc.), a domain that usually lacks high-quality labeled data. **RQ2** illustrates that the aggregation of synthetic clickbaity text by both humans and machines can be beneficial to improve clickbait detection accuracy by as much as 14.5% in *AUC scores*. Moreover, machine-generated clickbaits (**RQ1**) can be used to develop a defense mechanism to battle against mass propagation of false information initiated by malicious bots in social networks, which would help human fact-checkers focus more on detecting intentional misinformation.

Our paper suggests features that are useful not only for developing algorithms that both effectively detect and discriminate various types of clickbaits, but also for training humans to become more aware and sensitive to potential misinformation by attaching a source label to flagged clickbaits. The outcomes also provide insights on the potential presentation of clickbaity headlines. To illustrate, **RQ1** shows that formally-trained journalism students often present clickbaity headlines with political context even for non-political target content, while such behaviors are not observed among social media users. Similar behavior and its influence has been studied in detail by [20], [21]

Limitation and Future Direction. A more thorough study of text generation as an oversampling method by other models (e.g., GAN-based) is of future interest. Insights gained therefrom will enable us to frame a better oversampling method that can better generate useful samples. Even though we are not trying to generate realistic text clickbaits, we plan to carry out a field survey to analyze and compare how users would perceive and react to synthetic clickbaits generated by different entities, and to answer the question: "how clickbaity are they?" Finally, we plan to apply the framework in other domains where collecting training data is either challenging or limited (e.g., rumor detection, writing-based Alzheimer detection).

7. Conclusion

We explored the utility of synthetically generated text in the context of clickbaits, and demonstrated that synthetic clickbaits can be useful as additional labeled training samples to train regular ML models to detect clickbaits better, by as high as 14.5% in AUC. We showed that VAE-based generative algorithms can generate high quality text that captures the most similar NLP feature distribution as the real ones among all synthetic sources. Even though such an overlap in NLP feature distribution does not directly make synthetic clickbaits as meaningful as real clickbaits, the outcomes demonstrated a promising track in using machines to generate realistic text in general. This framework can, thus, present a novel direction toward solving the problem of insufficient training data in supervised learning.[6]

8. Acknowledgement

This work was in part supported by NSF awards #1742702, #1820609, and #1915801, ORAU-directed R&D program in 2018, and ONR under grant N000141812108. We appreciate reviewers for all of their constructive comments.

6. All codes and datasets used in this work are available for public access at: http://pike.psu.edu/download/asonam19/clickbait/

References

[1] M. M. U. Rony, N. Hassan, and M. Yousuf, "Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?" in *IEEE/ACM ASONAM 2017*, 2017, pp. 232–239.

[2] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading Online Content: Recognizing Clickbait As "False News"," in *ACM WMDD 2015*, New York, NY, USA, 2015, pp. 15–19.

[3] X. Cao, T. Le, Jason, and Zhang, "Machine Learning Based Detection of Clickbait Posts in Social Media," *arXiv preprint arXiv:1710.01977*, oct 2017.

[4] N. Couldry and J. Turow, "Big Data, Big Questions— Advertising, Big Data and the Clearance of the Public Realm: Marketers' New Approaches to the Content Subsidy," *International Journal of Communication*, vol. 8, no. 0, 2014.

[5] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *European Conference on Information Retrieval*. Springer, 2016, pp. 810–817.

[6] W. Wei and X. Wan, "Learning to Identify Ambiguous and Misleading News Headlines," in *IJCAI 2017*. AAAI Press, 2017, pp. 4172–4178.

[7] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *IEEE/ACM ASONAM 2016*, 2016, pp. 9–16.

[8] A. Agrawal, "Clickbait detection using deep learning," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, 2016, pp. 268–272.

[9] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *ICML 2016*, 2016, pp. 1727–1736.

[10] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Controllable text generation," *arXiv preprint arXiv:1703.00955*, 2017.

[11] K. Shu, S. Wang, T. Le, D. Lee, and H. Liu, "Deep headline generation for clickbait detection," in *IEEE ICDM 2018*. IEEE, 08 2018.

[12] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," *CoRR*, vol. abs/1511.06349, 2015.

[13] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[14] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, "" 8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality." in *AAAI*, 2016, pp. 94–100.

[15] A. Elyashar, J. Bendahan, and R. Puzis, "Detecting Clickbait in Online Social Media: You Won't Believe How We Did It," *arXiv preprint arXiv:1710.06699*, 2017.

[16] S. Gairola, Y. K. Lal, V. Kumar, and D. Khattar, "A Neural Clickbait Detection Engine," *arXiv preprint arXiv:1710.01507*, 2017.

[17] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 245–251.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for non linear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] Y. Miao and P. Blunsom, "Language as a Latent Variable: Discrete Generative Models for Sentence Compression," in *EMNLP 2016*, 2016, pp. 319–328.

[20] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a Social Botnet: Growth, Content and Influence in Twitter," in *ACM CSCW 2015*, New York, NY, USA, 2015, pp. 839–851.

[21] J. G. Geer and K. F. Kahn, "Grabbing attention: An experimental investigation of headlines during campaigns," *Political Communication*, vol. 10, no. 2, pp. 175–191, 1993.