A Reverse Turing Test for Detecting Machine-Made Texts

Jialin Shao
Jialin_Shao@outlook.com
Beijing University of Technology
Beijing, China

Adaku Uchendu
azu5030@psu.edu
The Pennsylvania State University
University Park, PA, USA

Dongwon Lee
dongwon@psu.edu
The Pennsylvania State University
University Park, PA, USA

ABSTRACT

As AI technologies rapidly advance, the artifacts created by machines will become prevalent. As recent incidents by the *Deepfake* illustrate, then, being able to differentiate man-made vs. machine-made artifacts, especially in social media space, becomes more important. In this preliminary work, in this regard, we formulate such a classification task as the *Reverse Turing Test* (RTT) and investigate on the contemporary status to be able to classify man-made vs. machine-made texts. Studying real-life machine-made texts in three domains of financial earning reports, research articles, and chatbot dialogues, we found that the classification of man-made vs. machine-made texts can be done at least as accurate as 0.84 in F1 score. We also found some differences between man-made and machine-made in sentiment, readability, and textual features, which can help differentiate them.

CCS CONCEPTS

•**Computing methodologies** → *Supervised learning by classification;* •**Applied computing** → *Document analysis;*

KEYWORDS

Reverse turing test, supervised learning, machine-made text

ACM Reference format:
Jialin Shao, Adaku Uchendu, and Dongwon Lee. 2019. A Reverse Turing Test for Detecting Machine-Made Texts. In *Proceedings of 11th ACM Conference on Web Science, Boston, MA, USA, June 30-July 3, 2019 (WebSci '19),* 5 pages.
DOI: 10.1145/3292522.3326042

1 INTRODUCTION

Recent advancements in AI technologies have enabled the machine-generation of realistic artifacts that are a little different from genuine artifacts. For instance, BigGAN [2] or Deepfake[1] introduced novel synthesis methods capable of generating realistic (but fake) images or videos, respectively. In the domain of "text" that is the focus of this work, similarly, the advancement of *Natural Language Generation* (NLG) has led to the automation of realistic text generation. It has advanced from heavy rule/template-based approaches to algorithm-based automatic methods using data ontology and user inputs. Currently, for instance, news media such as Associated Press, Forbes, and LA times reportedly use machine learning methods to generate realistic-looking financial earning and weather reports [10].

Figure 1: Turing Test (left) vs. Reverse Turing Test (right)

As such novel technologies become more sophisticated, however, pitfalls and risk of technologies also rapidly increase. Adversaries may use such technologies to generate realistic artifacts to trick naive users in fraudulent activities (e.g., a fake image in a tweet to spread fake news or machine-made chatbot conversation in a phishing scam). To prepare for such a cybersecurity problem better, in this work, we ask if one can accurately distinguish machine-made texts from man-made ones by solely looking at the contents of texts.

This research question that we pose bears a similarity to the *Turing Test*, developed by Alan Turing in 1950, that determines if a human judge (A) is observing a machine (B) or human (C) in some task. If the machine (B) shows the behavior indistinguishable from a human, thus fools the human judge (A), it is said to "passed the Turing Test." In our setting, we aim to develop a machine learning model (A') that determines if the give texts in question were generated by a machine (B) or human (C). To emphasize the fact that the observing judge is a machine (A'), not a human (A), this problem is referred to as the **Reverse Turing Test (RTT)**[2]. Figure 1 illustrates the subtle but important difference.

The underlying hypothesis of this research is to ask if there exists a subtle but fundamental difference (e.g., information loss or patterns of expressions) that can differentiate man-made texts from machine-made ones. As AI technologies rapidly advance, of course, such differences will diminish, making the RTT problem harder. This preliminary work, therefore, aims to investigate on the contemporary status to distinguish machine-made vs. man-made texts. As the work [11] recently showed that the "eye blinking" could be exploited to detect AI-generated videos, we hope to find a similar finding in machine-made texts.

Studying real-life machine-made texts in three domains–e.g., financial earning reports, research articles, and chatbot dialogues, our preliminary results indicate that it is indeed possible to accurately detect machine-made texts *for now* with the classification accuracy in the range of 0.84 – 1 in F1 score. Through the lens such as sentiment analysis, readability analysis, and topic model analysis, we show that man-made texts can be distinguished from machine-made ones.

[1] https://github.com/deepfakes/faceswap

[2] https://tinyurl.com/yc62z7wk

WebSci '19, Boston, MA, USA

DOI: 10.1145/3292522.3326042

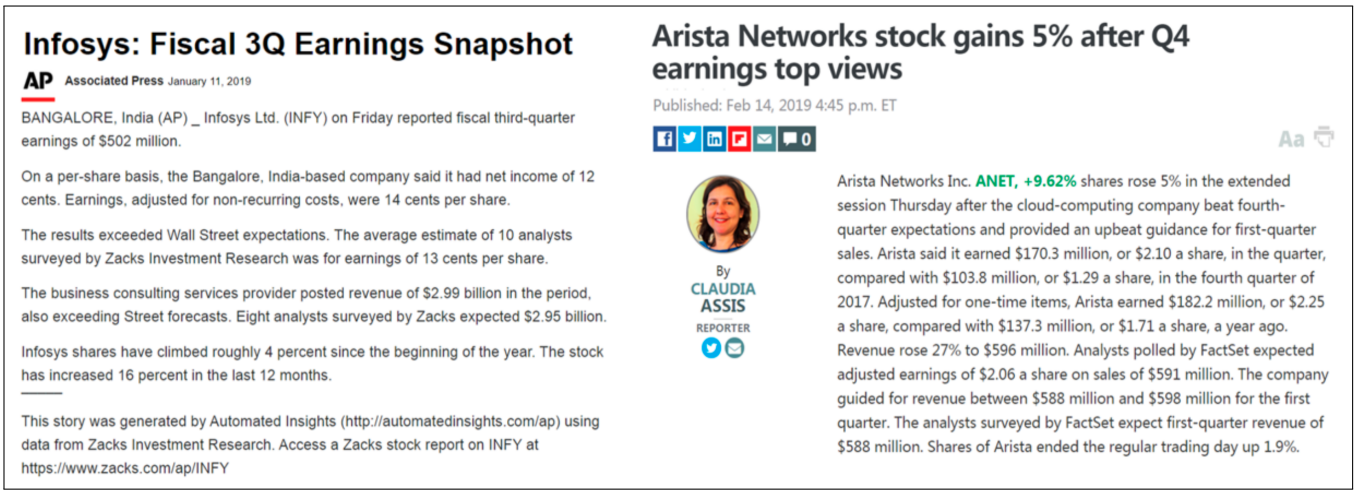

Figure 2: Examples of machine-made (left) and man-made (right) earning reports.

2 RELATED WORK

Natural Language Generation (NLG) has matured from the most basic template-based methods to coded grammatical and statistical software. It has fostered the generation of tailored reports for specific audiences [1, 6]. Companies, such as the Associated Press, Forbes and the LA Times, have adopted this NLG technology to generate weather forecast, earnings reports, and sports recap news [9].

While NLG evaluation is marked by a great deal of variety, it is hard to compare them due to the diverse inputs and different evaluation purpose and criteria. Currently, the evaluation of NLG outputs is dominated by two methods: (1) one relying on human judgments (i.e., *Turing Test*) which is subjective, and (2) the other using corpora and (i.e., *Reverse Turing Test*) [6]. The *Turing Test* mainly focuses on clarity, fluency, and readability. However, as judged by human evaluators, it may exhibit high variance across domains [3]. For the *Reverse Turing Test*, a variety of corpus-based metrics (e.g., BLEU, CIDEr and ROUGE) are used to evaluate translation, academic summarization, and image description [12, 13, 17]. While the aforementioned methods are concerned with the evaluation of the outputs from NLG, [4] has used TF-IDF and comprehensive profiles as features to build a SVM classifier to identify man-made texts. Related, [16] used meta data in a cluster model to find strong predictors for social bots in twitter.

3 DATASET

We have collected or generated man-made vs. machine-made texts in three domains as follows:

(1) **Academic Papers**: Using *SCIgen*[3], an automatic CS paper generator, developed at MIT, we have generated 908 synthetically-generated Computer Science papers. The collection is named as raw_paper$_M$. For the man-made academic papers, next, we first collected an open-source dataset from Kaggle, which contains papers published in the AAAI Neural Information Processing Systems (NIPS) conferences, and papers from the *Translation Archive* that includes papers from 52 different computer science conferences. Due to the influence from NIPS, note that man-made academic paper dataset has more AI flavour. A total of 7,876 papers in this collection is named as raw_paper$_H$.

Table 1: Summary of three datasets.

Dataset Name	# of files	AVG # of words	S.D. # of words
raw_paper$_M$	908	2,087.91	229.56
raw_paper$_H$	7,876	3,835.14	1,846.20
paper$_M$	908	2,090.11	241.99
paper$_H$	1031	2,554.87	602.76
raw_report$_M$	4,210	158.89	47.76
raw_report$_H$	2,100	139.97	57.74
report$_M$	1,450	159.00	33.20
report$_H$	1,516	140.00	27.42
raw_dialog$_M$	993	7.99	5.04
raw_dialog$_H$	993	11.18	10.59
dialog$_M$	979	6.52	2.23
dialog$_H$	955	7.87	4.90

(2) **Earnings Reports**: For machine-made news articles, we crawled and scraped data from media websites, such as Yahoo Finance and Forbes. Two leading companies, *Automated Insights* and *Narrative Science*, are in partnership with Yahoo Finance and Forbes, respectively, providing auto-generated financial earning reports. Merging the reports of these two websites together and removing each company's canned copyright message (e.g., "this story is generated by Automated Insights"), we obtained a total of 4,210 earning reports, named as raw_report$_M$. For man-made news articles, next, we chose earnings report of similar lengths and topics, written by human reporters. We collected 2,100 earnings reports from a financial website MarketWatch[4] and named it as raw_report$_H$. Figure 2

[3] https://pdos.csail.mit.edu/archive/scigen/

[4] https://www.marketwatch.com/

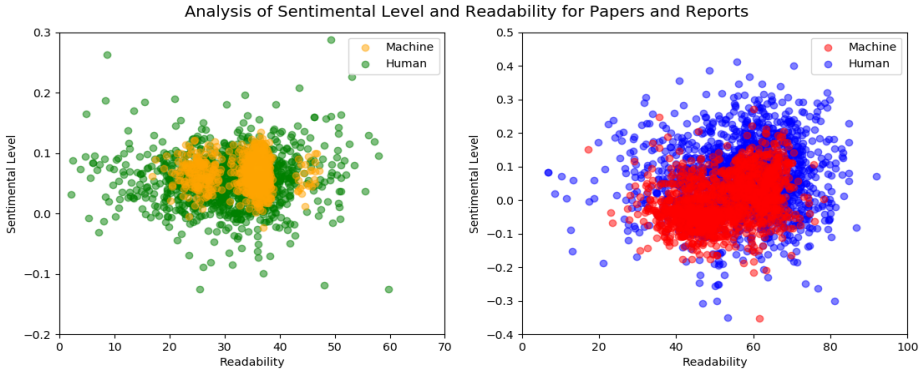

Figure 3: Readability and sentimental level analysis of papers datasets (left) and of reports datasets (right).

provides the examples of man-made vs. machine-made earnings reports in our datasets.

(3) **Chatbot Dialogues**: The chatbot dialogue data comprises of machine-made and man-made texts from a chatbot competition, known as the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB)[5]. In this competition, a human judge (A) converses with a counterpart who can be either another human (B) or a chatbot (C) (i.e., identity hidden). The response from the counterpart is generated in texts and rated by the human judge (A) for the ability to pass the Turing Test (i.e., how response is likely to be written by a human). The dataset, consisting of 993 dialogues, is named as raw_dialog$_M$ and raw_dialog$_H$, for the machine-made and man-made texts, respectively.

In pre-processing three datasets, we first eliminate outliers in each dataset according to the 3-sigma principle, which is to say that the length of texts in each dataset is within the three standard deviations of mean length. Capital letters are converted to lowercase and numbers within texts are ignored. Words shorter than 3 characters are also removed. Some stop words, which are provided by Scikit-learn [14], are taken out. To make the dataset of machine-made and man-made texts more comparable, in addition, we delete the machine-made texts whose length deviates too much from the average length of human-made dataset. At the end, we obtain three pairs of human-made vs. machine-made texts with comparable lengths. The pre-processed human-made vs. machine-made data is then named as paper$_H$, paper$_M$, report$_H$, report$_M$, dialog$_H$, and dialog$_M$, respectively. Summary of statistics is shown in Table 1. The final pre-processed datasets are available for download at GitHub[6].

4 CHARACTERIZING MAN-MADE VS. MACHINE-MADE

To understand the characteristics of man-made and machine-made texts better, we investigate the datasets from three angles: (1) sentiment analysis, (2) readability via Flesch Reading Ease [5], and (3) topic model via Latent Dirichlet Allocation (LDA) as follows:

(1) **Sentiment**: By borrowing the definition of "polarity" from *Textblob*[7], we define the sentiment level as a float value within the range $[-1.0, 1.0]$ where 0 indicates neutral, $+1.0$ indicates a very positive sentiment, and -1.0 represents a very negative sentiment. Figure 3, for instance, shows the distributions of sentiment scores in Y-axis from two datasets–i.e., earnings reports (report$_M$ and report$_H$) and academic papers (paper$_M$ and paper$_H$). In both datasets, note that the range of sentiment scores for man-made texts is wider and stronger than that for machine-made ones.

(2) **Readability**: The readability is defined by *the Flesch Reading Ease* as a formula, that generates a score usually between 0 and 100. A higher readability score means that text is more readable. In general, a score between $70-80$ is viewed as equivalent to the $7th$ grade level. Figure 3 shows the distributions of readability scores in X-axis from the same two datasets. Similar to sentiment, readability of man-made texts varies more widely. However, in general, it is not trivial to differentiate man-made from machine-made by only using readability scores. Additionally, for the chatbot dataset, due to short texts, the differences in both sentiment and readability between man-made and machine-made texts are shown to be negligible.

(3) **Topic Model**: To further explore the differences between man-made and machine-made datasets, we conducted a topic model analysis. Using LDA, we found out that even though each pair of man-made and machine-made datasets are in the same domain (such as earnings reports, computer science academic papers, and chatbot dialogues), there still exist differentiating factors in their textual expressions, attitudes, and concerns. For the academic papers datasets (paper$_M$ and paper$_H$), the most notable top topic words in paper$_H$ include: *datasets*, *model*, *learning*, and *training*. On the other hand, paper$_M$ talks more about *algorithms*, *evaluation*, *methodology*, and *results*. This is expected as paper$_M$, generated by SCIgen, covers broader and older computer science topics, algorithm and methodology, while

[5]https://www.aisb.org.uk/
[6]https://tinyurl.com/y9d4wh3j
[7]https://textblob.readthedocs.io/en/dev/

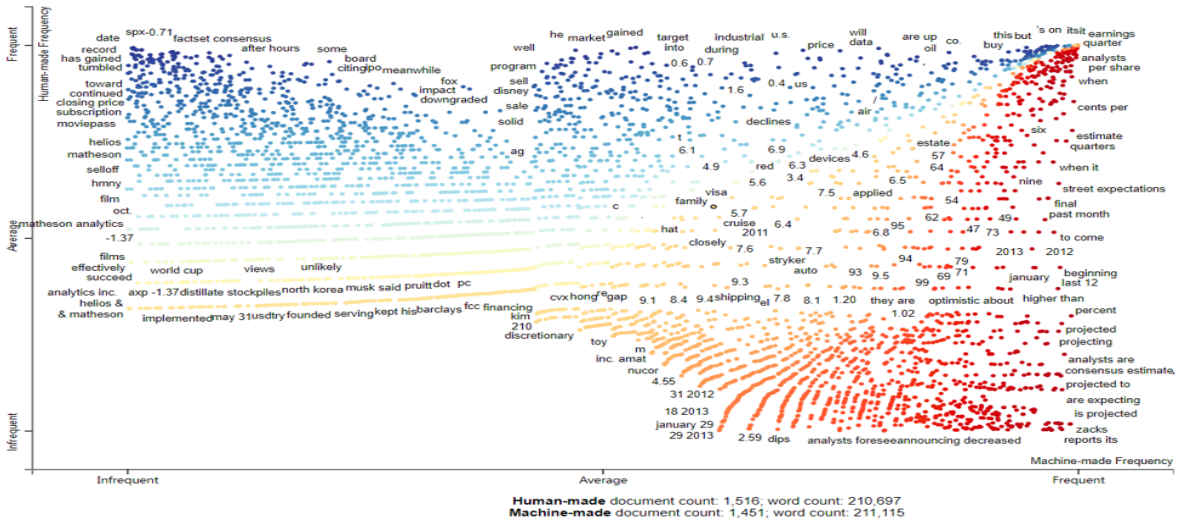

Figure 4: Term associations of earnings reports dataset.

paper$_H$ covers more recent and AI-ish topics. Second, for the earning report datasets, the top topic words of report$_M$ include *analysts*, *reported*, and *estimate*, that tend to quote analysts' point of view. In addition, report$_M$ prefers to quote numbers. On the other hand, report$_H$ has a set of notable topic words, including *gained* and *gains*, that sometimes contribute to the positive sentiment scores. In order to visualize the words and phrases that represent the characteristics of machine-made and human-made texts better, we use *Scattertext* [8] to plot the the words of the earning reports datasets, shown in Figure 4. The X-axis and Y-axis indicate the term frequency in machine-made and human-made texts, respectively. For instance, the upper-left area in Figure 4 shows the terms frequently occurring in man-made texts, while the lower-right area shows the frequent terms in machine-made texts. As to the dialogue datasets, there were no particularly interesting topic words, except that many dialogues in dialog$_H$ contain question-types as the original corpus focused on question-answering scenarios.

5 PREDICTING MAN-MADE VS. MACHINE-MADE

Based on the characterization, in this section, we test if it is possible to build a prediction model to differentiate between man-made and machine-made texts. We studied mainly two types of features in building a machine learning model: (1) syntactic frequency based feature via TF-IDF, and (2) linguistic context based word embedding via word2vec [7].

In order to achieve an accurate classification, we built a classifier using Support Vector Machines (SVM), a machine learning algorithm trained to maximize the margin of separation between positive and negative examples [15]. We tag different class labels for machine-made and man-made texts and split the training and test sets in the 8:2 ratio. Since the number of features is similar to the number of samples in the dataset, we used the Linear Kernel. We tested each model using 10-fold cross validation, with arbitrarily generated seeds to ensure the randomness as much as possible. All classifications are evaluated using F1 score which considers both precision and recall of the test. The model building was performed using Scikit-learn.

Figure 5 shows the F1 scores of classification over all three datasets–i.e., academic papers, earning reports, and chatbot dialogues. X-axis indicates % of text used in learning. For instance, 50% means that we used the half of the contents in the training data in learning. A few findings are notable. First, word2vec based features in general outperform TF-IDF based ones. Linguistic context and subtle differences in semantics captured by word2vec are able to distinguish man-made from machine-made texts much better than a simple frequency-based scheme. Second, after about 50% of training contents are used in learning, the SVM model for papers and reports datasets can differentiate man-made vs. machine-made texts with a near perfect accuracy. However, when less than 50% of contents are used in learning, the accuracy significantly drops, especially using TF-IDF features. On the other hand, it is interesting to see that using as little as 20% of contents in texts, word2vec achieves over 0.95 in F1 score. Finally, for the dialogues dataset that has much less contents than the other two datasets, the classification accuracy is comparatively lower–i.e., 0.84 in F1 score using word2vec and degrades to 0.82 in F1 score using TF-IDF.

Finally, it is noteworthy thatusing only 20% of contents in dialogues (i.e., using approximately nly 5 words), F1 score of dialogue's word2vec based SVM model has still exceeded 0.8.

6 DISCUSSION

There are several limitations in our study. First, because many commercial NLG technologies are still trade secrets, it was difficult to get sufficient amount of NLG-generated machine-made datasets. In addition, due to the limited amount of texts, classifiers took in length and genre features which could lead to a small degree of overfitting.

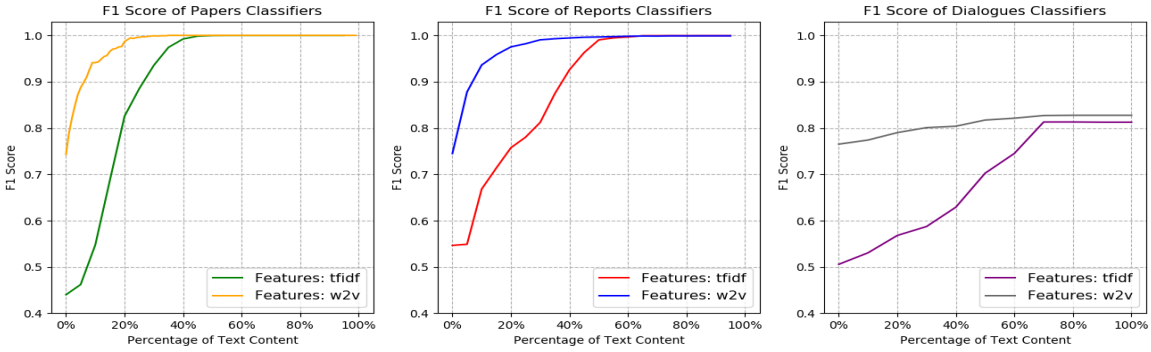

Figure 5: F1 scores of: papers classifiers (left), earning reports classifiers (middle), and chatbot dialogues classifiers (right)

Second, as the topics and genres between man-made and machine-made texts in our datasets are not completely aligned, one may not achieve the reported F1 scores for more challenging datasets. Reversely, at the same time, one could improve the prediction accuracy further by using more more powerful learning models (e.g., deep learning) or sophisticated features.

That said, from the prediction experiments, it is clear that using simple features such as TF-IDF or word2vec, it is already possible to accurately detect machine-made texts from man-made texts, especially long texts such as academic papers and earning reports. This may be due to the limited datasets but also it is possible that there is subtle but clear difference in the way machine generates texts, different from what humans write or speak. In order to generalize the findings, however, we plan to collect more datasets of man-made and machine-made in different domains.

As NLG technologies are advancing rapidly, in near future, it will become more challenging to detect machine-made texts from man-made ones. One can imagine a scenario where malicious adversaries use an NLG technology to mislead users. For instance, adversaries may use machine-made texts in a chatbot discussion based phishing attack. If a computational solution can tell naive users whether the part of chatbot dialogue is likely to be man-made or machine-made, it can warn a potential victim that she is conversing with a machine, not a human. Therefore, our research is a good starting point toward this important research direction and novel security applications.

7 CONCLUSION

We have formulated the Reverse Turing Test (RTT) problem as the binary classification to differentiate between man-made and machine-made texts. The results of the analysis of the characteristics of machine-made and man-made texts using three real datasets, suggested that machine-made texts tend to be more neutral and relatively more difficult to read than man-made texts. We also demonstrated that there are some special expressions and concerns of machine-made texts in different domains, compared to man-made texts. Further, having the F1 scores of at least 0.84 from binary classifications, we showed that it was possible to differentiate machine-made texts from man-made ones accurately. However, as the relevant technologies rapidly advance, in near future, one may not be able to distinguish machine-made texts from man-made ones effectively.

8 ACKNOWLEDGEMENT

This work was in part supported by NSF awards #1742702 and #1820609, and ORAU-directed R&D program award in 2018.

REFERENCES

[1] Elizabeth Blankespoor, Christina Zhu, and others. 2018. Capital market effects of media synthesis and dissemination: Evidence from robo-journalism. *Review of Accounting Studies* 23, 1 (2018), 1–36.

[2] Andrew Brock, Jeff Donahue, and Simonyan Karen. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[3] Matt Carlson. 2015. The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital journalism* 3, 3 (2015), 416–431.

[4] Mehmet M Dalkilic, Wyatt T Clark, James C Costello, and Predrag Radivojac. 2006. Using compression to identify classes of inauthentic texts. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. SIAM, 604–608.

[5] R.F. Flesch. 1979. *How to write plain English: a book for lawyers and consumers*. Harper & Row. https://books.google.com/books?id=-kpZAAAAMAAJ

[6] Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.

[7] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[8] Jason S Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. *arXiv preprint arXiv:1703.00565* (2017).

[9] Celeste Lecompte. 2015. Automation in the Newsroom. *Nieman Reports* 69, 3 (2015), 32–45.

[10] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-Driven News Generation for Automated Journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*. 188–197.

[11] Yuezun Li Li, Ming-Ching Chang, and Siwei Lyu. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1806.02877* (2018).

[12] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. 311–318.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[15] Bernhard Scholkopf and Alex Smola. 2002. *Support Vector Machines and Kernel Algorithms*. The Handbook of Brain Theory and Neural Networks, MA Arbib (Eds.), MIT Press.

[16] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).

[17] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.