Bengali Fake Review Detection using Semi-supervised Generative Adversarial Networks

Md. Tanvir Rouf Shawon, G. M. Shahariar, Faisal Muhammad Shah, Mohammad Shafiul Alam, Md. Shahriar Mahbub

Paper ID – CN094

Presenter

Md. Tanvir Rouf Shawon

ICNLP

Natural Language Processing

Contents of Presentation

Introduction

Fake Reviews

A fake review is a dishonest review that is written with the intention of misleading consumers into making uninformed decisions. It can be written by someone who has not actually used the product or service being reviewed or has a bias or conflict of interest.


This case is really good, it looks very nice and doesn't add bulk ...

By [redacted] on January 16, 2016

Color: Black & Black

The case is lightweight and slick. This case is really good, it looks very nice and doesn't add bulk or excessive size to my phone.It is so easy to get in and out of my pocket and the color is nice. It fits snuggly and provides good protection. It's so pretty and it seems durable. I receive this product to give honest review.

2 comments | 0 of 8 people found this helpful. Was this review helpful to you? Yes No Report abuse

Figure 1: An example of fake product review posted by an user. *[source-internet]*

Research Gap

- Language used in reviews depend on the user’s background and cultural context.
- Accurately labeling fraudulent reviews manually is exceedingly difficult.
- While there are several research works available for languages like English, Persian, and Roman-Urdu, there are none for Bengali.

Problem Statement

Create a system that can detect fake reviews written in Bengali language using a dataset that has been manually labeled, by using a semi-supervised generative adversarial network architecture that incorporates five pretrained language models.

Related Work

GAN Based Text Classification

- Raihan et al. [1] employed semi-supervised GAN-BERT architecture in order to categorize Bengali texts with a few labeled examples. They evaluated how well GAN-Bangla-BERT performed on two downstream Bengali tasks (hate speech and fake news detection) in comparison to Bangla-Electra and Bangla BERT Base.

- Ta et al. [2] identified violent and abusive social media posts in Spanish. However, noise vectors were tweaked using random rate different from the original SS-GAN architecture before being fed to the generator network.

Dialect Identification

- Zaharia et al. [3] utilized Romanian BERT with SS-GAN for Romanian dialect identification.

- Yusuf et al. [4] fine-tuned ARBERT and MARBERT with SS-GAN for Arabic dialect identification.

Sentiment & Compliant Classification

- Colo´n-Ruiz and Segura-Bedmar [5] employed a BERT model followed by Bidirectional Long Short Term Memory (Bi-LSTM) network with SS-GAN for sentiment analysis on drug reviews.

- Auti et al. [6] utilized BioBERT with SS-GAN which performed best to classify pharmaceutical compliant and non-compliant texts.

Background Study

Generative Adversarial Networks

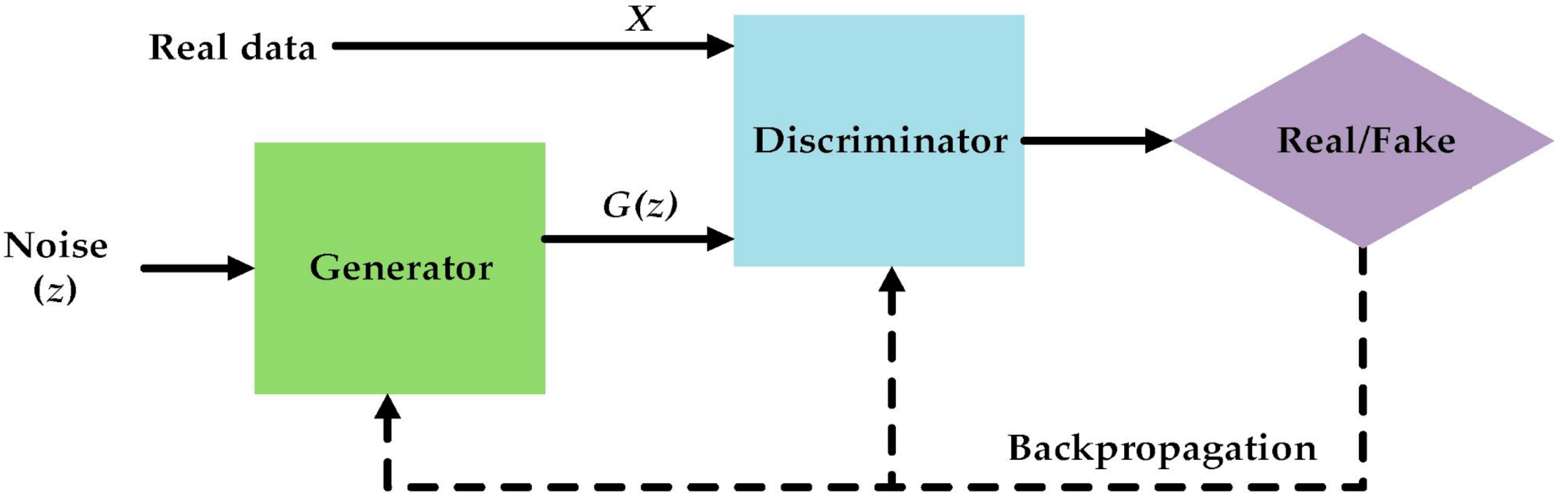


Figure 2: Generative Adversarial Network architecture [7]

Semi-supervised Generative Adversarial Network

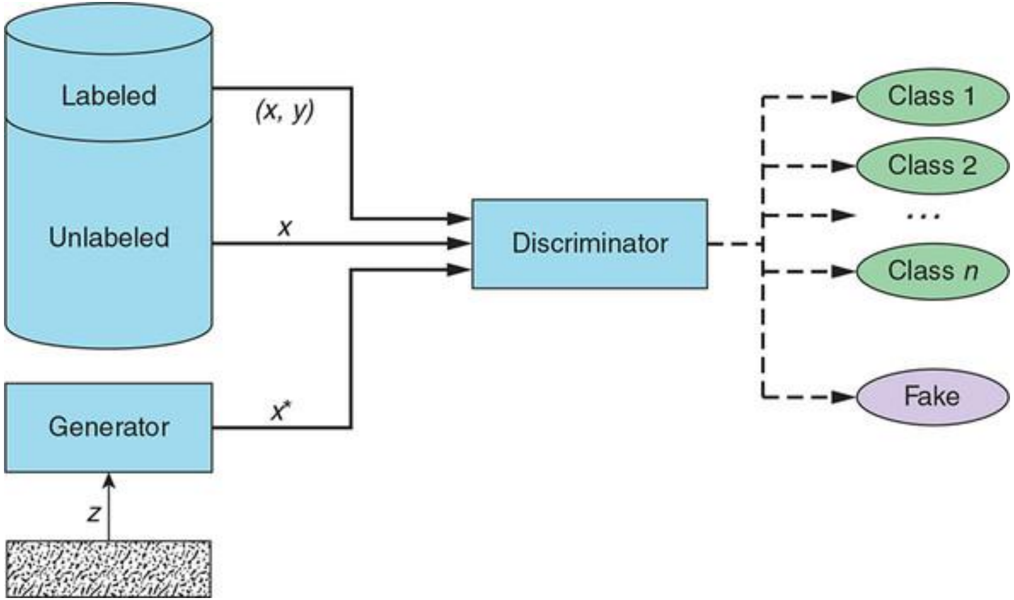


Figure 3: Semi-supervised Generative Adversarial Network (SS-GAN) architecture [8]

Pretrained Language Models for Bengali

Pretrained Language Models	Type	Parameters	Embedding Size
Bangla BERT Base[1]	BERT-base	110M	768
BanglaBERT[2]	ELECTRA-base	110M	768
BanglaBERT generator[3]	ELECTRA-base	34M	768
sahajBERT[4]	ALBERT-large	18M	128
Bangla-Electra[5]	ELECTRA-small	14M	128

Table 1: Bengali pre-trained language models with configurations.

1. https://huggingface.co/sagorsarker/bangla-bert-base
2. https://huggingface.co/csebuetnlp/banglabert
3. https://huggingface.co/csebuetnlp/banglabert_generator
4. https://huggingface.co/neuropark/sahajBERT
5. https://huggingface.co/monsoon-nlp/bangla-electra

Dataset

Dataset Description

- The dataset used in this study comprises of total 6014 fake and authentic reviews written in Bengali language.
- Collected from selected publicly accessible Facebook groups.
- Manually gathered the posts made by the members of the groups regarding the food items and services of various restaurants and labeled them as authentic or fake.
- After majority voting on the annotations performed by three different individuals, 871 fake and 5015 authentic reviews were gathered respectively.

Data Distribution

No. of Labeled Samples	Unlabeled Samples	Testing Samples
32	512	512
64	512	512
128	512	512
256	512	512
512	512	512
1024	512	128

Table 2: Data distribution for the experimentations of this study

Methodology

Proposed Approach

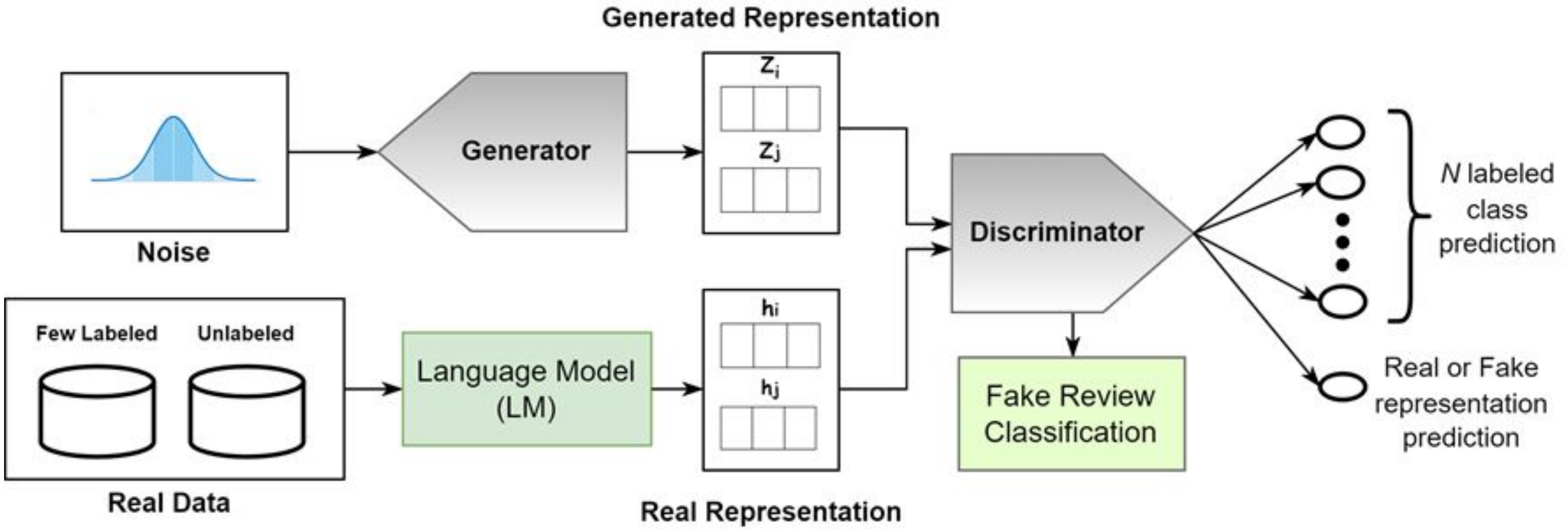


Figure 4: Proposed methodology for fine-tuning language models with semi-supervised GAN architecture.

Experimental Results

Experiments

- Binary Classification using SS-GAN with the following five pretrained Bengali language models:

    1. Bangla BERT Base
    2. BanglaBERT
    3. BanglaBERT generator
    4. sahajBERT
    5. Bangla-Electra

Hyper Parameter Settings

Pretrained Language Models	No. of Epoch	Batch Size	Loss Function	Learning Rate	Optimizer
Bangla BERT Base	7	16	Binary Cross Entropy	5e-5	AdamW
BanglaBERT	18				
BanglaBERT generator	25				
sahajBERT	13				
Bangla-Electra	26				

Table 3: Hyperparameter used in different language models

Experimental Results

Table 4: Performance comparison of different GAN-LM model

Model	No. of Labeled Samples	Accuracy	Precision	Recall	F1 score
Fine Tuned Bangla BERT	1024	**0.65625**	0.68000	0.54838	0.60714
GAN Bangla BERT	32	**0.69531**	0.71179	0.64427	0.67635
	64	**0.71093**	0.76650	0.59684	0.67111
	128	0.73047	0.73469	0.71146	0.72289
	256	0.75391	0.71381	0.83795	0.77091
	512	0.76563	0.78298	0.72727	0.75410
	1024	**0.83593**	0.84286	0.85507	0.84892

Experimental Results

Table 5: Performance comparison of different GAN-LM model

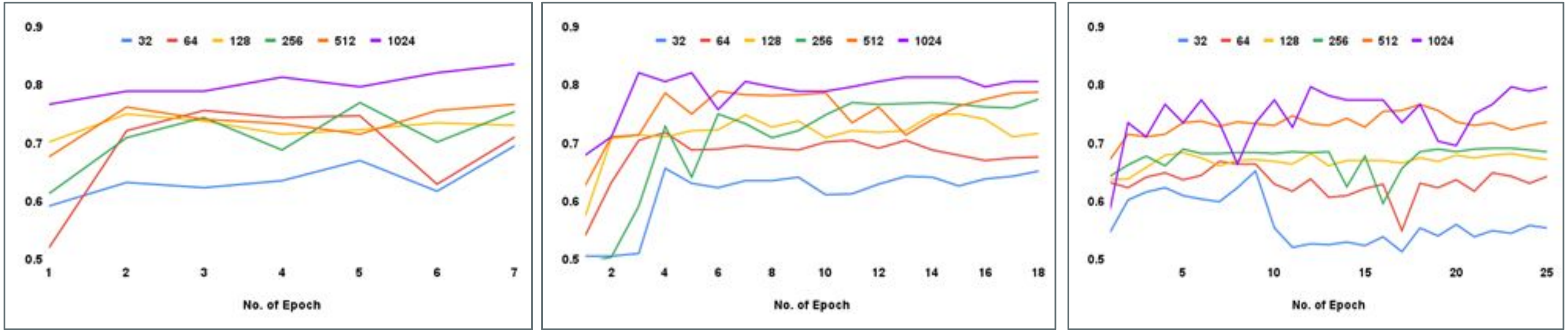
Model	No. of Labeled Samples	Accuracy	Precision	Recall	F1 score
GAN Bangla BERT Generator	32	**0.65234**	0.6506	0.64032	0.64542
	64	0.67578	0.6654	0.6917	0.6783
	128	0.7168	0.7093	0.72332	0.71624
	256	0.77539	0.75746	0.80237	0.77927
	512	0.78711	0.78571	0.78261	0.78416
	1024	**0.80468**	0.82353	0.81159	0.81752
GAN Bangla BERT Base	32	0.55469	0.5576	0.47826	0.51489
	64	0.64258	0.62868	0.67589	0.65143
	128	0.67188	0.66798	0.66798	0.66798
	256	0.68555	0.69328	0.65217	0.6721
	512	0.73633	0.72519	0.75099	0.73786
	1024	**0.79687**	0.79452	0.84058	0.8169

Experimental Results

Table 6: Performance comparison of different GAN-LM model

Model	No. of Labeled Samples	Accuracy	Precision	Recall	F1 score
GAN Bangla-Electra	32	0.55273	0.55405	0.48617	0.5179
	64	0.58398	0.60638	0.45059	0.51701
	128	0.65234	0.64591	0.65613	0.65098
	256	0.65234	0.6161	0.78656	0.69097
	512	0.68359	0.64444	0.80237	0.71479
	1024	0.73437	0.72727	0.81159	0.76712
GAN sahaj BERT	32	0.51172	0.50319	0.93676	0.6547
	64	0.6875	0.72683	0.58893	0.65066
	128	0.66016	0.67873	0.59289	0.63291
	256	0.72852	0.70956	0.76285	0.73524
	512	0.73438	0.7191	0.75889	0.73846
	1024	0.46094	N/A	N/A	N/A

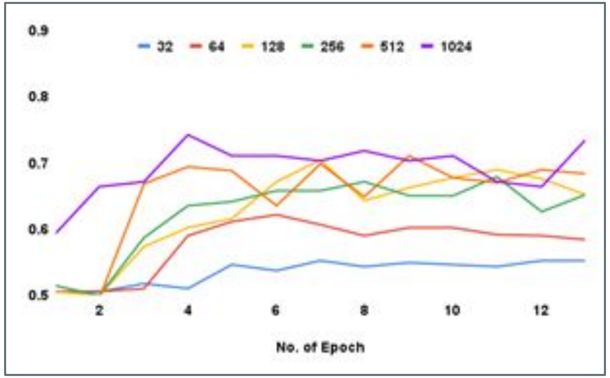
Comparison of Testing accuracy of Different GAN-LM Models



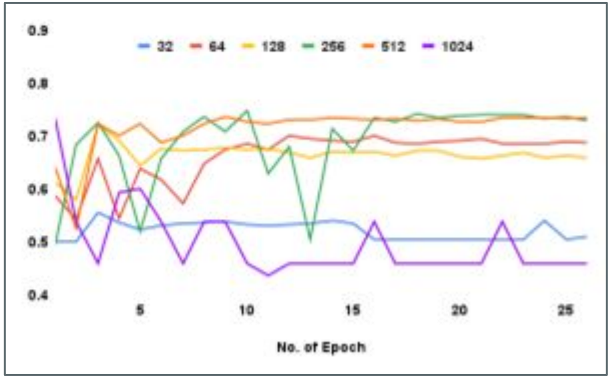
(a) BanglaBERT (b) BanglaBERT Generator (c) Bangla BERT Base

Figure 5 : Test accuracy vs. #epochs of the experimental models with 32, 64, 128, 256, 512, and 1024 labeled samples

Comparison of Testing accuracy of Different GAN-LM Models



(a) Bangla-Electra



(b) sahajBERT

Figure 6 : Test accuracy vs. #epochs of the experimental models with 32, 64, 128, 256, 512, and 1024 labeled samples

Conclusion and Future Works

- Attempt to create a fake review detection system
- Proposed a semi supervised GAN architecture that incorporates 5 language models
- Achieved decent results using very few data
- A more robust annotated dataset on Bengali fake review can be a great contribution
- The exploration of how GANs may be utilized to generate Bengali reviews will be an intriguing development.

Acknowledgement

This research work is conducted under "Bengali Fake Reviews: Development of Benchmark Dataset and Deep Learning-based Detection System" project funded by AUST Internal Research Grant.

References

[1] Tanvir, Raihan, Md Tanvir Rouf Shawon, Md Humaion Kabir Mehedi, Md Motahar Mahtab, and Annajiat Alim Rasel. "A GAN-BERT Based Approach for Bengali Text Classification with a Few Labeled Examples." In International Symposium on Distributed Computing and Artificial Intelligence, pp. 20-30. Springer, Cham, 2023.

[2] Ta, Hoang Thang, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. “GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media.” In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS. org. 2022.

[3] Zaharia, George-Eduard, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. “Dialect identification through adversarial learning and knowledge distillation on romanian bert.” In Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 113-119. 2021.

[4] Yusuf, Mahmoud, Marwan Torki, and Nagwa M. El-Makky. “Arabic Dialect Identification with a Few Labeled Examples Using Generative Adversarial Networks.” In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, pp. 196-204. 2022.

References

[5] Col o´n-Ruiz, Cristo´bal, and Isabel Segura-Bedmar. “Semi-Supervised Generative Adversarial Network for Sentiment Analysis of drug re- views.” (2021).

[6] Auti, Tapan, Rajdeep Sarkar, Bernardo Stearns, Atul Kr Ojha, Arindam Paul, Michaela Comerford, Jay Megaro, John Mariano, Vall Herard, and John Philip McCrae. “Towards Classification of Legal Pharmaceutical Text using GAN-BERT.” In Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference, pp. 52-57. 2022.

[7] Feng, Jie, Xueliang Feng, Jiantong Chen, Xianghai Cao, Xiangrong Zhang, Licheng Jiao, and Tao Yu. 2020. "Generative Adversarial Networks Based on Collaborative Learning and Attention Mechanism for Hyperspectral Image Classification" Remote Sensing 12, no. 7: 1149. https://doi.org/10.3390/rs12071149

[8] “Chapter 7. Semi-Supervised Gan · Gans in Action: Deep Learning with Generative Adversarial Networks.” · GANs in Action: Deep Learning with Generative Adversarial Networks, https://livebook.manning.com/book/gans-in-action/chapter-7/42.

THANKS!

Any questions?