Data Augmentation for Fake Reviews Detection

Ming Liu and Massimo Poesio
Queen Mary University of London
Mile End Road
London E1 4NS
{acw661,m.poesio}@qmul.ac.uk

Abstract

In this research, we studied the relationship between data augmentation and model accuracy for the task of fake review detection. We used data generation methods to augment two different fake review datasets and compared the performance of models trained with the original data and with the augmented data. Our results show that the accuracy of our fake review detection model can be improved by 0.31 percentage points on DeRev Test and by 7.65 percentage points on Amazon Test by using the augmented datasets.

1 Introduction

Online communication has increased the speed and quantity of information sharing between people. While this change has brought a number of benefits, it also increased the opportunities for unscrupulous individuals to deceive (Newman et al., 2003; Hancock et al., 2007; Vrji, 2008). Research shows that on average people tell 1 or 2 lies a day; now, lying has migrated from face-to-face communication to online (Hancock et al., 2004; Mihalcea and Strapparava, 2009). Fake reviews are a particularly problematic type of deceptive online communication, given our reliance on online reviews to guide our purchases (Ott et al., 2011; Fornaciari and Poesio, 2014; Fornaciari et al., 2020). About 1% to 6% positive hotel reviews are estimated to be fake (Ott et al., 2012).

Automatic deception detection methods rely on stylometric methods extracting from text hundreds of linguistic features (Newman et al., 2003; Hancock et al., 2007; Mihalcea and Strapparava, 2009; Fornaciari and Poesio, 2014). More recently, Deep learning has been used (Girgis et al., 2018; Kaliyar et al., 2021; Fornaciari et al., 2021; Salminen et al., 2022). This research also led to the creation of several datasets for training such models (Ott et al., 2011; Fornaciari and Poesio, 2014; Amazon, 2018; Fornaciari et al., 2020). However, these datasets have a number of limitations. They tend to be small, and very domain dependent (a dataset of TripAdvisor reviews is not suitable for training models to detect fake reviews on Amazon and vice-versa). Even more crucially, few of them consist of genuine fake reviews; most were artificially created using crowdsourcing. But crowdsourced fake reviews are known to be different from genuine fake reviews (Fornaciari et al., 2020). In this paper, we focus on the issue of creating suitable datasets for fake review detection research.

Data augmentation techniques using text generation would appear to be a potential solution to the problem of generating datasets for fake review detection when we only have a small amount of fake or genuine reviews. And given that modern text generation methods appear to be able to create artificial texts extremely similar to model texts used to prompt them, these methods might be more likely than crowdsourcing of creating artificial fake reviews similar to real fake reviews. In proposals such as (Shehnepoor et al., 2022; Aghakhani et al., 2018), generators were used to augment data to improve discriminator performance. Salminen et al. (2022) firstly uses the GPT-2 model to expand the existing data to obtain a larger data set and then applies the new data set to fake news detection. However, the Amazon dataset used by Salminen et al is very noisy, as discussed below.

In this paper, we discuss a study based on the hypothesis that data augmentation can improve the performance of deception detectors. We followed an approach similar to Salminen et al. (2022) but also used the cleaner dataset of Amazon reviews introduced by (Fornaciari et al., 2020) and present evidence that the performance of a fake review detector can be improved by augmenting an existing dataset with artificially generated reviews. Using

Proceedings of Recent Advances in Natural Language Processing, pages 673–680
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_073

our augmented datasets we achieved 0.31 and 7.65 percentage points improvements on DeRev Test and Amazon Test respectively.

2 Background

2.1 Deception Detection

There is a crucial difference between fake reviews detection and fake news detection: because reviews express subjective judgments, in fake reviews detection it is not possible to use external knowledge sources to identify deception, except perhaps for metadata (Fornaciari and Poesio, 2014).

One alternative source of evidence is the language used in the review (Newman et al., 2003). Many psychologists argue that language used while lying is different from language used in a sincere way (Vrji, 2008). To make just one example, it has been claimed that liars use second and third-person pronouns such as *you*, *her*, and *him* because they are trying to avoid using first-person pronouns and bringing unfamiliar content into themselves. Using second and third-person pronouns will shift the conversation to other people in an effort to keep themselves away from lies (Hancock et al., 2007; Mihalcea and Strapparava, 2009). However, there is consensus that there are no silver bullets - single cues that can be relied on (Fornaciari et al., 2020). The idea is that it is possible to classify deceptive reviews by looking at hundreds of cues using machine learning. This hypothesis that a liar's behaviour is reflected in his language led to the use of stylometric techniques to recognize deception – the analysis of the linguistic characteristics of deceptive language to distinguish between deception and truth (Newman et al., 2003; Hancock et al., 2007; Mihalcea and Strapparava, 2009; Fornaciari and Poesio, 2014).

Deep Learning Approaches With the development of deep learning, a whole range of new approaches have been tested. One line of research involves using Generative Adversarial Networks (GANs) for deception detection (Aghakhani et al., 2018). The FakeGAN model proposed by Aghakhani et al. (2018), its ability to detect fake reviews has reached the level of state-of-the-art models. The results demonstrate that the GANs model can be applied to the task of fake review detection. Using GANs for semi-supervised learning can effectively improve the effect of the classifier, because unlabeled samples can be added through the generator, which effectively expands the training set, thereby improving the performance of the classifier. Recently, Transformer models such as RoBERTa have also been used to identify genuine and fake reviews (Liu et al., 2019). In fact, Salminen et al. (2022) argued that the fakeRoBERTa model based on RoBERTa can more accurately distinguish between true and false reviews than human judges.

2.2 Datasets

One of the key issues for deception detection is finding suitable datasets. Some of the datasets used in research on deception detection are listed in Table 1. The methods used to collect these datasets can be distinguished into: (i) collected in the lab (e.g. Newman et al. (2003)); (ii) crowdsourced (e.g. Mihalcea and Strapparava (2009); Ott et al. (2011)); (iii) collecting reviews known as being false (e.g. DeRev (Fornaciari and Poesio, 2014; Fornaciari et al., 2020), Amazon (Amazon, 2018) recent). We discuss each method in turn.

Lab-collected datasets A popular approach in deception detection involves asking subjects to produce deceptive text in the lab. Newman et al. (2003) collected 568 writing samples from 287 students based on 5 different topics. Subjects were asked to give feedback on true and false opinions, true and false descriptions or true and false feelings based on different topics. The key issue with this approach is that it's not clear how well such datasets reflect real deceptive text. Also, students are typically used as subjects, which does not provide a good sample of typical user populations.

Crowdsourcing Another widely used approach is to create datasets using crowdsourcing. For example, Ott et al. (2011) released a hotel review dataset created in this way which is one of the most widely used datasets for studying deceptive reviews detection. However, this dataset has a number of limitations. First of all, it is pretty small: it only contains 1600 reviews, which is too small for training. Secondly, Fornaciari et al. (2020) team found that crowdsourced data is different from real data, and using crowdsourced data in the real world may lead to bias. Like with lab-created data, the key issue with such datasets is that there is no guarantee that the data thus collected reflects genuine deceptive language.

Datasets of genuinely true and false reviews A third line of research is to attempt to collect datasets

Dataset	Size	Category	Details
Stories (Newman et al., 2003)	568 writing samples	lab	Collected from 5 studies
Hotel Reviews (Yoo and Gretzel, 2009)	42 fake and 40 truthful reviews	lab	Hotel reviews
3 Topics (Mihalcea and Strapparava, 2009)	300 fake and 300 truthful reviews	crowd	Collected through Amazon Mechanical Turk
Hotel Reviews (Ott et al., 2011)	800 fake and 800 truthful reviews	crowd	Collected from TripAdvisor and Amazon Mechanical Turk
Sandulescu and Ester (Sandulescu and Ester, 2015)	9000 reviews	genuine	Shared by Trustpilot but not public
Amazon Reviews (Amazon, 2018)	10500 fake and 10500 truthful reviews	genuine	Published by Amazon
DeRev 2018 (Fornaciari and Poesio, 2014)	8311 reviews	genuine	Book reviews

Table 1: Datasets for Deception Detection.

of genuinely fake and genuine reviews. Examples of datasets created out of genuinely fake and real reviews are *DeRev 2018* (Fornaciari and Poesio, 2014; Fornaciari et al., 2020) and the *Amazon Customer Reviews Dataset* (Amazon, 2018), which were used in this experiment.

Using real data is obviously the best method for creating datasets for studying deceptive reviews, but it's very difficult to create such datasets on a large scale except for big companies that run platforms collecting reviews like Amazon or TripAdvisor. These issues motivate the search for another way of creating large-scale datasets for studying deceptive review detection.

3 Experimental Design

In this section, we discuss the datasets and the generator and classifier models we used.

3.1 Data

In our experiments, two fake reviews datasets were used: the Amazon dataset used in (Salminen et al., 2022) and DeRev used in (Fornaciari and Poesio, 2014; Fornaciari et al., 2020)– the two datasets of authentic fake reviews and authentic reviews we are aware of. The Amazon dataset is large, but it is also very noisy. DeRev is smaller than the Amazon dataset, but the quality of the data is higher.

DEREV (Fornaciari and Poesio, 2014; Fornaciari et al., 2020) consists of Amazon book reviews produced by individuals that confessed to writing fake reviews for financial gain, as well as reviews for which there is strong evidence that are genuine. Fornaciari and Poesio also collected a variety of meta information ('clues') about these reviews. Fornaciari et al. (2020) created a cleaned-up and larger version of DEREV, which we used in this study. Figure 1 illustrates the DeRev dataset, where the *gold2016* attribute is used to distinguish between deceptive(0) and genuine(1). It contains 8311 items. In addition to labelling true and false, the dataset also provides some deception clues.

The Amazon dataset Figure 2 is a sample of the Amazon dataset. The LABEL column of the Amazon dataset contains *__label1__* and *__label2__*, representing fake and real respectively. The Amazon reviews dataset contains user review data that were identified by the Amazon customer team as being clearly true or false. It contains 21,000 items, categorized into 30 classes, each of which contains 700 reviews.

Use of the datasets in our study Our experiment involves two phases. The first part of the experiment is concerned with creating a data generator to generate review data. In this process, the entire Amazon dataset is used to train the model. In the second part, we train a classifier to identify real and fake reviews. DeRev 2018 and the Amazon dataset are used in this process. Since DeRev 2018 only contains reviews about books, only a subset of the Amazon test set was used for this evaluation.

3.2 Models

In this subsection, we introduce the two types of models involved in experiments: the generator, that generates reviews, and the classifier, trained and tested using the data.

3.2.1 Generator

The primary purpose of the generator is to generate coherent text by providing an appropriate prompt. In a series of pilots, we tried to use the GPT-2 model directly to generate sentences, but that didn't work well. In order to improve the coherence and relevance of the sentences generated by the model,

```
<review ID="1" title="ADubiousPlan" writer="GeraldKubicki" author="SandraParker" date="July 20, 2012
        serialdate="735084" stars="5" found="1" fold2014="1" fold2016="1" gold2014="0" gold2016="0"
        silverMaj4="0" silverMaj3="0" silverRay4byMaj4="0" silverRay4byMaj3="0" silverRay4byRand="1"
        silverRay3byMaj4="1" silverRay3byMaj3="0" silverRay3byRand="1" silverWhi4="0" silverWhi3="0"
        clueTot="3"  clueSB="1" clueCL="0" clueNN="1" clueUP="1">
    <object>A Dubious Game-Another Kubicki Masterpiece</object>
    <body>Gerald Kubicki has done it again. A Dubious Plan, the fifth installment of the Colton Bany
        delivers on it's promise of adventure, mystery and plenty of sex in yet another engaging and
```

Figure 1: Example of DeRev 2018. Each comment is in XML document format, which contains the title, author, time and content of the comment. It also contains tokens generated by comments.

DOC_ID	LABEL	RATING	REVIEW_TEXT	VERIFIED_PURCHASE
1	__label1__	4	When least you think so, this product wil…	N
2	__label1__	4	Lithium batteries are something new intro…	Y
3	__label1__	3	I purchased this swing for my baby. She i…	N
4	__label1__	4	I was looking for an inexpensive desk cal…	N
5	__label1__	4	I only use it twice a week and the result…	N

Figure 2: Sample of Amazon Customer Reviews Dataset with tags, review text, user ratings and product categories.

we adopted instead the Interpolation model proposed by Wang et al. (2020) to generate narrative.

Wang et al. (2020)'s model consists of two parts, one dedicated to generating sentences using GPT-2, whereas the other part of the model calculates coherence scores. The generator takes two prompt sentences as input and produces an intermediate sentence. For example, sentence 1 and sentence 5 are used to generate sentence 3; then sentence 1 and sentence 3 are used to generate sentence 2, and so forth. The Coherence Ranker proposed in (Moon et al., 2019) is then used to calculate the coherence between the generated sentence and the input to select the sentence with the highest score as the result. Human judgements are used to evaluate the model, the only reliable way for assessing the quality of story generation (See et al., 2019).

In order to get a better result, we replaced the GPT-2 model with the newer OPT model (Zhang et al., 2022). According to the Meta team, OPT-175B is comparable to GPT-3, while requiring only 1/7th of the carbon footprint to develop (Zhang et al., 2022). Due to the limitation of the available hardware, we were not able to fine-tune OPT-175B, but only OPT-1.3B. Figure 3 is the generator pipeline–essentially the same as the pipeline in (Wang et al., 2020). The input is the first and last sentence of an existing comment. 10 candidate sentences are output through the fine-tuned OPT model. Then the Coherence Ranker is used to select the most coherent sentence with the input. Loop the entire generation process until the desired length of comments is generated. In this experiment, we choose 5 as the review length.

3.2.2 Classifier

In our classifier experiments, we verify whether adding the data generated as discussed earlier improves the model's performance. In this experiment, we used two classifiers, SVM (Boser et al., 1992) and RoBERTa (Liu et al., 2019), to facilitate the comparison with previous results. The classifier experiments are based on those in (Salminen et al., 2022), but there are two key differences between the present study and that work. First, Salminen et al. (2022) generated reviews using pure GPT-2. In this work, we used a text generation model that in our experiments produced much better text. Because the quality of the generated dataset cannot be directly assessed, the quality of the generated dataset can only be indirectly judged by the classification performance of the classifier. If the classification preference of the classifier with the added data is better than the original model, it means that data augmentation can improve the performance of the model. Likewise, the quality of the generated datasets is also good. Two classifier models, SVM and RoBERTa, were used in the paper when evaluating the generated dataset.

A second difference between this experiment and those in Salminen et al. (2022) is that we used two different datasets. Salminen et al. (2022) only used the Amazon dataset–but, as we will see, this dataset is problematic in a number of ways. In addition, using two datasets allowed us to compare adding 'real' data with adding artificial data.

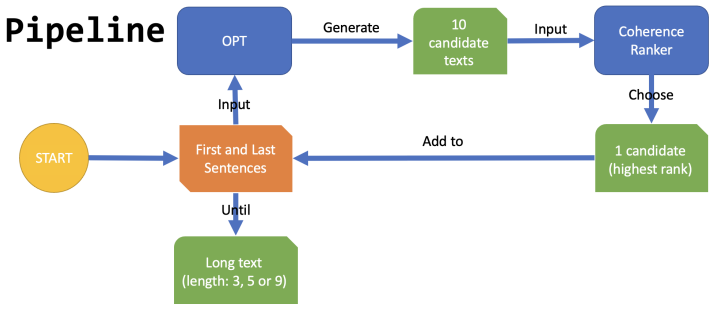

Figure 3: The generator pipeline. It contains OPT generator, coherence ranker and interpolation. It generates text of length 3, 5 or 9 after multiple iterations.

4 Experiments

We ran two series of experiments. In the first series, a part of the DeRev dataset is used for testing. In the other series, a part of the Amazon dataset is used for testing. In both series, the difference between experiments is which combination of datasets is used for training. Only book reviews were used in our experiments, as this is the domain of the reviews in DeRev.

4.1 Experiments Details

Firstly, we fine-tuned an OPT model with the Amazon dataset, which contains 13786 reviews. Then we generate reviews for the Book category of the Amazon dataset. The generated dataset contains 312 generated 'real' reviews and 325 generated 'fake' reviews.

4.2 Test on DeRev

The full list of variants of training datasets used in the experiments testing on DeRev is shown in Figure 4. But only experiment A, B, C, D, E, F, G, and G_B are included in Test on DeRev. In this first set of experiments, the DeRev dataset is the test set. DeRev Train is 80% of DeRev; Amazon Train is 100% of the Amazon datasets. 20% of DeRev is treated as the test set.

In both the DeRev and the Amazon experiments, Experiment A is the baseline: training and testing on in-domain data only. Experiment B tests whether adding human-generated data from a different dataset in the same domain can improve the accuracy of the model. Experiment C tests whether adding both additional human data *and* generated data can improve model accuracy. Experiments D, E and F verify whether the generated data are best used as real or fake data. Experiment G assesses the quality of the generated data–only generated data are added to the in-domain data. Finally, Experiment G_B is used to test whether imbalance in the data has a significant impact on the experimental results.

Specifically, in the DeRev Test experiments, in Experiment A, the models are trained on DeRev only. In Experiment B, we train on DeRev and Amazon. In Experiment C, the model is trained on DeRev, Amazon and generated data, but the generated data is divided into 'fake data' generated using the fake reviews in Amazon as seed, and 'real data' generated from the real reviews in Amazon. In Experiment D, we also train our models using DeRev, Amazon and generated data, but the data, generated from all reviews in Amazon, are all treated as 'fake data'. In Experiment E, we train again on DeRev, Amazon and generated dataset, but the data are generated from the real reviews in Amazon only, and again treated as fake. Experiment F means training on DeRev, Amazon and generated dataset, but the generated data, treated again as fake, are only generated using the fake reviews in Amazon as seeds. In both Experiment G_B and Experiment G only the generated data are added to the DeRev training set; but in Experiment G_B the number of generated and real reviews is balanced.

4.3 Test on Amazon

The full list of variants of training datasets used in the experiments testing on Amazon Test is shown in Figure 4. But only experiment A, B, C, D, E, F, G, H and I are included in Test on Amazon. In these experiments, the models are tested on the Amazon dataset. 100% of DeRev and 80% of Amazon are used as the training set. 20% Amazon dataset is treated as the test set. Experiments from A to G are identical to those with DeRev test, but using Amazon Test. In addition, in Experiment H, we

Datasets \ Experiment IDs	A	B	C	D	E	F	G	G_B	H	I
DeRev(Book)	✓	✓	✓	✓	✓	✓	✓	✓		
Amazon(Book)		✓	✓	✓	✓	✓		✓	✓	✓
generated(Amazon(real) as real)			✓				✓	✓		✓
generated(Amazon(fake) as fake)			✓			✓	✓	✓		✓
generated(Amazon(all) as fake)				✓						
generated(Amazon(real) as fake)					✓					
Notes								Balanced		

Figure 4: Test On Amazon and DeRev

only train on Amazon data, and in Experiment I, we train on Amazon + the same generated data as in Experiment G ('real' generated from real, 'fake' generated from fake).

5 Results and Discussion

5.1 DeRev Test

Figure 5 illustrates the result with DeRev Test. First of all, we can see that the performance of the SVM model is always lower than the neural network. Therefore, the discussion will focus on the RoBERTa model.

We find that the accuracy of training configuration A is slightly higher than those obtained with training configurations B, C, D, E, and F. This means that adding the Amazon dataset does not improve performance, even if the generated dataset is also added. However, adding only the generated dataset does slightly improve the performance of the classifier: compare configuration A with configurations G and G_B. We believe the result is caused by the quality of the dataset. Evidence for this is the following review from the Amazon dataset. First and most obviously, this review is not in English. Then, the sentences are clearly not part of a book review. In other words, while we are very confident that the DeRev dataset is of very high quality, the Amazon dataset was not carefully selected, which is part of the reason why adding such data to the training set does not necessarily result in an improvement in classifier performance.

Example of Amazon review :

```
[[VIDEOID:mo3LVVAW0LVYN8Y]][[ASIN:1481
976850 Libera Tu Poder Creativo: Guia
Espiritual para Prosperar y Trabajar
<br /><br />Realmente Teresa me enseño
paso apaso como manejar una entrevista
```

Comparing Experiment C with Experiments D, E, and F show that the generated datasets are also more similar to their corresponding categories. 'Fake data' generated from fake data are more like fake reviews. Likewise, a 'Real data' set of reviews generated from a true dataset is more like a real dataset. Comparing A and G show that adding additional data can improve the performance of the classifier. However, this is not the case when adding training data from the other dataset (Experiment B).

This result suggests that data augmentation techniques outperform our experiments adding an equivalent amount of data from similar datasets. Because the data obtained through data enhancement technology is controllable, the generated data seem to preserve the original features of the seed data better than similar data from another domain. However, there are still some problems. In the Amazon example just mentioned, the first few sentences of the long sentence seem disconnected from the review. This causes problems because the prompt to the generator is the first and last sentence of the review. This issue needs to be addressed in subsequent experiments.

In experiment G_B, a balanced dataset is used: the number of generated reviews and DeRev reviews are the same. The result in this setting is similar to experiments A and G. Finally, the experimental results show that adding an augmented dataset can improve the performance of the classifier, but not by much.

5.2 Amazon Test

Figure 6 indicates the result of the Amazon Test. First of all, the performance of machine learning models is not always lower than the neural network. But the RoBERTa model is able to achieve higher performance than SVM. So this discussion still focuses on the RoBERTa model. In this group of experiments, experiment H is the benchmark experiment, and its accuracy can reach 70%.

The results in experiment A (DeRev training only) are poor for the obvious reason that the training set and test are from different datasets. This difference is further confirmed by comparing B (DeRev + Amazon) and H (Amazon only), where adding DeRev to training makes performance

Experiments_results (Test set: DeRev)

Experiment ID	Classifier	Accuracy	F1-Score	Precision	Recall	eal Da	Fake Data
A	RoBERTa	0.9582	0.9571	0.9540	0.9603	495	470
B	RoBERTa	0.9421	0.9427	0.9080	0.9801	845	820
C	RoBERTa	0.9550	0.9548	0.9308	0.9801	1157	1145
D	RoBERTa	0.9100	0.9041	0.9362	0.8742	845	1457
E	RoBERTa	0.9389	0.9360	0.9521	0.9205	845	1132
F	RoBERTa	0.9486	0.9463	0.9592	0.9338	845	1145
G	RoBERTa	0.9614	0.9615	0.9317	0.9934	807	795
G_Balanced	RoBERTa	0.9582	0.9568	0.9600	0.9536	637	637
A	SVM+TfIdf	0.9325	0.9333	0.8963	0.9735	495	470
B	SVM+TfIdf	0.9100	0.9079	0.9020	0.9139	845	820
C	SVM+TfIdf	0.8778	0.8766	0.8599	0.8940	1157	1145
D	SVM+TfIdf	0.8746	0.8602	0.9375	0.7947	845	1457
E	SVM+TfIdf	0.8778	0.8652	0.9313	0.8079	845	1132
F	SVM+TfIdf	0.8842	0.8767	0.9078	0.8477	845	1145
G	SVM+TfIdf	0.8971	0.8974	0.8696	0.9272	807	795
G_Balanced	SVM+TfIdf	0.8682	0.8682	0.8438	0.8940	637	637

Figure 5: Results on DeRev Test

Experiments_results (Test set: Amazon)

Experiment ID	Classifier	Accuracy	F1-Score	Precision	Recall	eal Da	Fake Data
A	RoBERTa	0.4824	0.5769	0.4878	0.7059	495	470
B	RoBERTa	0.5824	0.4496	0.6591	0.3412	760	735
C	RoBERTa	0.8294	0.8415	0.7857	0.9059	1072	1060
D	RoBERTa	0.6412	0.5960	0.6818	0.5294	760	1372
E	RoBERTa	0.5471	0.2376	0.7500	0.1412	760	1047
F	RoBERTa	0.6471	0.6386	0.6543	0.6235	760	1060
G	RoBERTa	0.8294	0.8324	0.8182	0.8471	807	795
H	RoBERTa	0.7000	0.6752	0.7361	0.6235	265	265
I	RoBERTa	0.7765	0.7865	0.7527	0.8235	364	361
A	SVM+TfIdf	0.5294	0.6262	0.5194	0.7882	495	470
B	SVM+TfIdf	0.6059	0.6417	0.5882	0.7059	760	735
C	SVM+TfIdf	0.7647	0.7701	0.7528	0.7882	1072	1060
D	SVM+TfIdf	0.5471	0.2667	0.7000	0.1647	760	1372
E	SVM+TfIdf	0.5235	0.3415	0.5526	0.2471	760	1047
F	SVM+TfIdf	0.6765	0.6154	0.7586	0.5176	760	1060
G	SVM+TfIdf	0.7412	0.7634	0.7030	0.8353	807	795
H	SVM+TfIdf	0.6118	0.6207	0.6067	0.6353	265	265
I	SVM+TfIdf	0.6765	0.6821	0.6705	0.6941	364	361

Figure 6: Results on Amazon Test

worse. The results of experiments C, D, E and F are similar to those with DeRev Test.

The best results are again obtained using only in-domain data (Amazon in this case) and the generated data. However, in this series of studies, the results obtained with C (also including DeRev) are very close.

6 Conclusion

Our experimental results show that the Roberta-based classifier model achieves 0.31% and 7.65% accuracy improvements on the DeRev and Amazon test sets, respectively. This shows that the accuracy of the classifier model can be improved to a certain extent by adding generated data. But our current experiments are limited to a single language and single domain. In future work, we plan to apply our data augmentation method to multiple languages and domains.

7 Limitations

Our new generator can provide better data than our previous generator, and we have evidence that the data already helps, but there are still minor problems such as the problem of repeated sentences. In order to solve this problem, the OPT model needs to be fine-tuned to make the generated sentences more diverse. At the same time, the Coherence Ranker selection process needs to be optimized to avoid selecting the same sentence.

The Amazon dataset needs to be cleaned-up. The non-English data have to be eliminated. It will also be necessary to separate the review data and book information, and only keep the review data. This should also improve the quality of the generated data.

Finally, and most importantly, we need to apply the methods to a broader range of reviews than just books, as done here.

References

Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 89–95. IEEE.

Amazon. 2018. Amazon customer reviews corpus.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. Bertective:language models and contextual information for deception detection. In *Proc. of EACL*. Association for Computational Linguistics.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 54(4):1019–1058.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287. Association for Computational Linguistics.

Sherry Girgis, Eslam Amer, and Mahmoud Gadallah. 2018. Deep learning algorithms for detecting fake news in online text. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pages 93–97.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

Jeffrey T Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. 2004. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–134.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 309–312.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Xu Chi. 2019. A Unified Neural Coherence Model. *arXiv:1909.00349 [cs, stat]*. ArXiv: 1909.00349.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, pages 201–210.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon gyo Jung, and Bernard J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771.

Vlad Sandulescu and Martin Ester. 2015. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*, pages 971–976.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Saeedreza Shehnepoor, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2022. Scoregan: A fraud review detector based on regulated gan with data augmentation. *IEEE Transactions on Information Forensics and Security*, 17:280–291.

Aldert Vrji. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2nd edition. Wiley.

Su Wang, Greg Durrett, and Katrin Erk. 2020. Narrative Interpolation for Generating and Understanding Stories. *arXiv:2008.07466 [cs]*. ArXiv: 2008.07466.

Kyung-Hyan Yoo and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. In *Information and communication technologies in tourism 2009*, pages 37–47. Springer.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.