RankAug: Augmented data ranking for text classification

Tiasa Singha Roy
Manipal Institute of Technology
tiasa.singharoy@gmail.com

Priyam Basu
Manipal Institute of Technology
priyambasu16@gmail.com

Abstract

Research on data generation and augmentation has been focused majorly on enhancing generation models, leaving a notable gap in the exploration and refinement of methods for evaluating synthetic data. There are several text similarity metrics within the context of generated data filtering which can impact the performance of specific Natural Language Understanding (NLU) tasks, specifically focusing on intent and sentiment classification. In this study, we propose RankAug, a text-ranking approach that detects and filters out the top augmented texts in terms of being most similar in meaning with lexical and syntactical diversity. Through experiments conducted on multiple datasets, we demonstrate that the judicious selection of filtering techniques can yield a substantial improvement of up to 35% in classification accuracy for under-represented classes.

1 Introduction

Recent advances in Large Language Models have brought along incredible progress in a wide range of NLU tasks. However, for domain specific tasks, fine-tuned models can bridge the performance gap with data Wu et al. (2023) but such domains are often low resource in nature and data collection can be quite difficult. This is where data augmentation techniques come into play, boosting model performance for a given supervised task by generating novel data points that are similar in characteristics to the available data.

There have been a large number of metrics created to evaluate data augmentation which are mostly focused on the performance of generation models Zhu et al. (2018) Kim et al. (2020) Liu et al. (2020) Sun et al. (2020). We explore various methods to evaluate and filter generated paraphrases Golovneva et al. (2022) to get high quality augmentations. Most of the prior work in this domain makes use of metrics that only take into consideration the word or embedding level similarity of the generated utterance. Popular metrics like BLEU score Papineni et al. (2002), Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Lin (2004) (Lin, 2004), and Metric for Evaluation of Translation with Explicit Ordering (METEOR) Banerjee and Lavie (2005) use n-gram based comparison. This type of evaluation is limited to a one-dimensional analysis of augmentation as high-quality data provides both contextual similarity as well as lexical diversity McCarthy et al. (2009) to the original text. To ensure this, we propose a text ranking method that outperforms other popularly used metrics to get top quality augmentations that aid in better training of models on downstream tasks. This method can be extended to any data augmentation model for evaluation and is independent of the training model as well.

Despite a variety of approaches for augmented data evaluation, there is no golden standard Bhandari et al. (2020), the real value of the generated data can be only evaluated through downstream tasks, for example by estimating how much performance improvement synthetic data can bring to the targeted supervised NLU task. In our case, we test our ranking and filtering mechanism on multiple supervised classification based scenarios for skewed datasets. It shows a consistent improvement across different experimental setups and datasets compared to the standard filtering metrics. Finally, our method [1] is also extended to a German dataset, to show that it can be applied not only to English but also to other languages.

2 Related Works

In recent years, data augmentation and generation techniques have gained significant attention in machine learning research. These techniques play a crucial role in enhancing the performance and

[1] https://github.com/whopriyam/Text-Augmentation-Filtering

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 317–323
December 6, 2023

robustness of various models across different domains. Augmentation techniques, in general, have been traditionally used in many downstream computer vision task Kingma and Welling (2013) uses Variational Autoencoders to encode the data examples to a latent representation and then new samples were generated from that latent space, which employs patch based augmentation. Alexey et al. (2016) uses rule based image transformations to generate more data for improving performance of Convolutional Neural Networks (CNNs) on feature learning tasks.

Text generation has been studied extensively leading to computational linguistics and diverse methods being suggested ever since. Sentence structures are very different and these diversities expand in different types of social media which makes text generation harder. Rule based techniques like word replacement using Finite State Transducers Rastogi et al. (2016) and synonym swap Şahin and Steedman (2019) have been some of the initial attempts at generating synthetic texts. Most such rule based methods suffer from a lack of sentence structure variation and loss of semantic context.

Multiple efforts have been made recently to use generative models too for text augmentation. Existing augmentation methods work at different granularity levels - characters, words, sentences, and documents. Yu et al. (2018) and Hou et al. (2018) use sequence-to-sequence generation for enhancing model performance in back translation and text transfer domains. Ding et al. (2020) proposes a novel approach to utilize generative augmentation for fine-grained and token-level entity tagging tasks. Pre-trained masked language models (MLMs) like BERT Devlin et al. (2018), T5 Raffel et al. (2020) and AugGPT Dai et al. (2023), which internally uses ChatGPT, can be used for contextual augmentation too. Since MLMs are pre-trained on a large number of texts, contextual augmentation can usually generate meaningful new texts.

3 Data

In our experiments, we make use of two datasets - Airline Travel Information System (ATIS) from the Microsoft Cognitive Toolkit (CNTK) Hemphill et al. (1990), an intent classification dataset, Hate Speech from a white supremacist forum de Gibert et al. (2018), a sentiment analysis dataset and Amazon Multilingual Reviews Keung et al. (2020), a product reviews corpus. All of these are standard datasets, ideal for setting benchmarks on classification tasks.

- ATIS dataset - It consists of a set of spoken utterances in the context of airline information, classified into one of 26 intents with 127 slot labels. It is important to note that the intent distribution within the ATIS dataset exhibits a significant imbalance, with over 70% of the data allocated to *atis flight* intent, while other intents contain a notably lower number of utterances.

- Hate Speech dataset - It consists user generated hate speech content from Stormfront, a white supremacist platform, manually annotated by human labellers. There is a high data imbalance here too, with 86% of the texts belonging to "no hate" and 14% belonging to "hate" sentiment.

- Multilingual Amazon Reviews Corpus - It consists of over one million product reviews in 6 languages, ranging from 1 to 5 stars, for multilingual text classification collected between November 1, 2015 and November 1, 2019. Due to the data being sufficiently large in size, we limit our experiments to 0.5% i.e 1000 samples, of the German reviews subset, while maintaining equal distribution across all 5 classes.

4 Filtering methods

4.1 Existing metrics

Existing filtering metrics are efficient in assessing the quality and relevance of text content. They are of majorly two types - word based and embedding based filtering. These methods excel in their ability to capture semantic and syntactic similarities between texts, making them a preferred choice for evaluating the performance of augmented sentences. We evaluate 5 such metrics:

- SacreBLEU: Though is primarily used for evaluating machine translation quality, it can also be applied to filter and rank text based on translation relevance by calculating the BLEU score, which measures the similarity between the reference and the candidate sentence, by measuring the linguistic similarity and fluency of the text Post (2018). The higher the number of overlapping n-grams between candidates and source sentences, the lower the score.

Dataset	# Classes	# samples	# samples before filtering	# samples after RankAug-3	# samples after RankAug-5
ATIS Intent	26	4978	38358	9985	13323
Hate Speech	2	9666	16626	11754	13146
German Reviews	5	1000	11000	4000	6000

Table 1: Benchmarking datasets

- Levenshtein distance: In order to augmentations most similar in structure and word distribution, we use this metric. It quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one text string into another Yujian and Bo (2007). The lower the score, the more similar is the reference text to the source text.

- Rouge-L: evaluates the performance of a generated text by comparing it to one or more reference texts. It considers the recall, or the ability of a generated text to capture essential information from the references, while also penalizing excessive word overlap. Lin (2004).

- Meteor: It offers a holistic evaluation by considering precision, recall, stemming, and synonymy, resulting in a more human-like assessment of text quality Banerjee and Lavie (2005). The adaptability it provides to different languages and domains makes it a good metric to rank and filter text according to its linguistic and semantic similarity to a reference.

- BERTScore: It leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. This enables us to filter sentences that might be completely different in word measure, synonym match, sentence structure, etc but could be semantically similar in meaning Zhang* et al. (2020).

4.2 RankAug

We propose RankAug, a ranking method that accounts for both similarity and diversity to filter high quality augmentations.

4.2.1 Semantic Similarity

To measure semantic similarity we utilise BERTScore which calculates similarity scores by aligning the paraphrase u_i and original sentence u on a token-level basis. This alignment process follows a greedy approach, aiming to optimize the cosine similarity between contextualized token embeddings obtained from BERT. A higher score indicates a higher semantic relevance and we denote this as R_{s_i} which represents the semantic rank of the ith paraphrase for a generated data point.

4.2.2 Diversity

Diversity as an evaluation metric is often overlooked when measuring paraphrase quality. We propose self-Levenshtein ($Self$-LD) to compute the diversity between a generated paraphrase and both the original sentence as well as the remaining paraphrases. This is derived from self-bleu Zhu et al. (2018) and computes the average word-level Levenshtein distance (LD) i.e word error rate Morris et al. (2004) across the remaining paraphrases u' and the mean is selected as the final score.

$$Self - LD = mean(Lev(u_i, u')) \qquad (1)$$

This is done for every generated paraphrase with a high score indicating a higher level of diversity and the paraphrases are ranked accordingly with R_{d_i} representing the diversity rank.

4.2.3 Final Ranking

After scores for both diversity and semantic similarity are generated for each paraphrase, we consider the ranking of each paraphrase based on these two criteria. We consider the harmonic mean of the generated scores to compute our final rank R_i.

$$R_i = \frac{2 * R_{s_i} * R_{d_i}}{(R_{s_i} + R_{d_i})} \qquad (2)$$

To utilize this final rank to filter out the best paraphrases. For our experiments, we select n=3,5 values where n denotes the number of samples ranked from top i.e. top n samples.

Filtering method	# augmentations filtered per sample (n)	Accuracy ATIS	Hate Speech	German Reviews
Baseline		98.25%	63%	50.3%
No filtering		97.35%	68.2%	48.4%
RankAug	5	**99.625**%	**74.1**%	**54.2**%
	3	98.75%	70.25%	**51.4**%
Bleu	5	99.14%	69.8%	52.1%
	3	98.60%	65%	45.2%
BertScore	5	99.00%	70.9%	50.1%
	3	98.45%	68.3%	49.4%
Levenshtein	5	99.375%	70%	47.8%
	3	**99.15**%	69%	45.6%
Rouge	5	99.12%	72%	52.2%
	3	98.70%	**70.5**%	49.4%
Meteor	5	99.00%	65.7%	46.8%
	3	98.25%	66.4%	42.8%

Table 2: Overall Accuracy for different filtering methods

5 Experiments

In this section, we describe the experimental setup for benchmark tests along with the sentence generation pipeline.

5.1 Sentence Generation

Data sparsity is a frequent problem for several NLU tasks as collecting the necessary quantities of high-quality labeled data for model training is frequently a challenging and expensive task, along with the risks of the generative model becoming too large Bender et al. (2021). We undertake the task to produce artificial data that can be utilized to enhance NLU model training. We use the original training data from the corpora as a source to the data generation model.

For augmenting the English sentences, we leveraged Google's transformer-based Pegasus model Zhang et al. (2020) for text augmentation. Pegasus internally uses self-supervised gap sentence generation for better abstraction performance by masking important tokens and applying ROGUE-n selection. It was pre-trained on the Colossal Common Crawl C4 Dodge et al. (2021) dataset. We used a pre-trained Pegasus model fine-tuned on Google Paws Yang et al. (2019), which is a paraphrasing dataset as a one-to-one sentence generator. By limiting the paraphrase token length limit, abstracting from a short sentence, the model paraphrases the text to a semantically similar sentence.

For German text augmentation, we use a pivot-based back translation pipeline Cai et al. (2021). In this process, the input texts are first translated to pivots and then paraphrases are generated. The German texts are first converted to English, which are then used as a pivot to generate the required paraphrases.

5.2 Downstream Task

For the purpose of evaluating the quality of generated text, we chose a text classification task. We use BERT-base embeddings for ATIS and Hate speech classification and, Bert-base multilingual embeddings for classifying the German Amazon reviews. Both models are trained for 4 epochs along with a batch size of 16 and a learning rate of 2e-5. Adding the corpus augmented with paraphrase improves the performance, which shows that it helps training even when fine-tuning the pre-trained language model. We average out the results from 3 runs for each scenario. Performance changes through data augmentation are significant, especially when the baseline accuracy is less, which is evident as shown in Table 5.

6 Results

For the ATIS and Hate Speech datasets, which are imbalanced datasets, we generate paraphrases for each underrepresented data point. This excludes the *atis flight* and *not hate* classes for ATIS and hate speech respectively. This gives us around 38358 for ATIS and 16626 samples in total for hate speech. However, models trained using the original data and every generated paraphrase result in a decrease in performance, highlighting the importance of using quality data points.

We then apply different filtering methods on the generated data. All our filtering methods showcase a consistent improvement on both datasets and outperform the baseline. As evident in Table 5, our method shows consistent improvement on the ATIS dataset and four other filtering methods show better results across both datasets as compared to the baseline. This shows that by utilizing less than half of the generated data we can outperform the baseline.

Intent	Train	Test	Baseline	Rank Aug-5
airfare	385	48	95.83%	97.92%
service	230	36	94.44%	100%
flight	3309	632	98.25%	99.47%
abbreviation	130	33	96.96%	100%
airline	139	38	94.74%	100%

Table 3: RankAug-5 performance on ATIS dataset on top 5 intents

Sentiment	Train	Test	Baseline	Rank Aug-5
Hate	696	500	19.02%	54.78%
Not hate	8970	500	97.60%	100%

Table 4: RankAug-5 performance on Hate speech dataset per sentiment

Rating	Train	Test	Baseline	Rank Aug-5
1 star	200	100	52%	61%
2 star	200	104	41.34%	54.81%
3 star	200	105	39.12%	49.91%
4 star	200	99	29.29%	34.34%
5 star	200	92	76.04%	75.89%

Table 5: RankAug-5 performance on German Reviews dataset per rating

Comparing against the different metrics, our method gives an increased performance for both datasets in n=5 filtered setting. For n=3, our filtering method comes close to matching the top performer. Across all 3 datasets, our method shows the best overall performance showing an increase compared to both the baseline performance and the no-filtering setting. In the case of ATIS and German review datasets, it can also be noted that no filtering augmentation setting actually reduces the performance when compared to the baseline. This essentially infers the fact that just adding augmentations does not necessarily improve performance but the generated text being of good quality is what yields good results on downstream tasks.

In Tables 2 and 3, we can see the performance improvement across different classes. We also extend our work on a balanced but a low resource German dataset. Our filtering method outperforms all other methods for both settings indicating that RankAug is adaptable to other languages as well.

7 Conclusion

When working with low resource and unbalanced datasets, data augmentation can significantly improve performance. However, it is crucial to have quality augmented data. We explored and evaluated a number of popular evaluation metrics for augmented data filtering and proposed our own method for ranking and filtering quality paraphrases. Our method, along with similarity also accounts for paraphrase diversity and achieves the best overall performance across multiple datasets while utilizing nearly half the total augmented data. Along with this, we also observe a consistent increase in performance of the underrepresented classes of the datasets explored achieving up to 35% increase in accuracy. We show that our approach can be extended to other languages as well as other varied domains to improve downstream performance and as a future work, we aim to benchmark these methods on more downstream tasks.

Limitations

While we achieve improvements in the datasets selected,wh the methods required to generate paraphrases can be very resource heavy. Along with this, BERTScore also requires GPU resources and is time consuming to use especially when a large amount of augmented data is present. For testing, we only consider downstream classification tasks which limit our evaluation as other tasks can have different requirements that our method is not able to encompass and should be explored.

References

Dosovitskiy Alexey, Philipp Fischer, Jost Tobias, Martin Riedmiller Springenberg, and Thomas Brox. 2016. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 38(9):1734–1747.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, and Pengfei Liu. 2020. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711.

Yitao Cai, Yue Cao, and Xiaojun Wan. 2021. Revisiting pivot-based paraphrase generation: Language is not the only optional pivot. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4255–4268.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6045–6057.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan, and Haiyang Yu. 2022. Task-driven augmented data evaluation.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Yanghoon Kim, Seungpil Won, Seunghyun Yoon, and Kyomin Jung. 2020. Collaborative training of gans in continuous and discrete spaces for text generation. *IEEE Access*, 8:226515–226523.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mingtong Liu, Erguang Yang, Deyi Xiong, Yujie Zhang, Yao Meng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2020. A learning-exploring method to generate diverse paraphrases with multi-objective deep reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2310–2321.

Philip M McCarthy, Rebekah H Guess, and Danielle S McNamara. 2009. The components of paraphrase evaluations. *Behavior Research Methods*, 41(3):682–690.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

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 of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers*, pages 186–191.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting finite-state transductions with neural context. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 623–633.

Gözde Gül Şahin and Mark Steedman. 2019. Data augmentation via dependency tree morphing for low-resource languages. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S Yu, and Lifang He. 2020. Mixup-transformer: dynamic data augmentation for nlp tasks. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440.

Zihao Wu, Lu Zhang, Chao Cao, Xiaowei Yu, Haixing Dai, Chong Ma, Zhengliang Liu, Lin Zhao, Gang Li, Wei Liu, et al. 2023. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *ICLR*.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.