Data Augmentation for Multiclass Utterance Classification – A Systematic Study

Binxia Xu[†] Siyuan Qiu[†] Jie Zhang[†] Yafang Wang[†*] Xiaoyu Shen[‡] Gerard de Melo[§]

Ant Financial Services Group[†]
Max Planck Institute for Informatics[‡]
Hasso Plattner Institute, University of Potsdam[§]

Abstract

Utterance classification is a key component in many conversational systems. However, classifying real-world user utterances is challenging, as people may express their ideas and thoughts in manifold ways, and the amount of training data for some categories may be fairly limited, resulting in imbalanced data distributions. To alleviate these issues, we conduct a comprehensive survey regarding data augmentation approaches for text classification, including simple random resampling, word-level transformations, and neural text generation to cope with imbalanced data. Our experiments focus on multi-class datasets with a large number of data samples, which has not been systematically studied in previous work. The results show that the effectiveness of different data augmentation schemes depends on the nature of the dataset under consideration.

1 Introduction

In modern conversational systems, classifying incoming user utterances is among the most crucial processes. This is particularly evident for automated customer service systems: if the underlying demand can successfully be classified based on the user's description, a known solution can directly be provided to them. A weak classifier may miscategorize a request, resulting in customer dissatisfaction. A considerable cause for low-performing classification is a lack of sufficient training data for certain categories, which manifests as the problem of *imbalanced data* distributions. Classifying utterances is particularly challenging, as people may choose various different forms of expressing their ideas and thoughts. Thus, very different utterances may reflect the same underlying intent. Recent findings on query variation in search systems suggest substantial potential for considering language diversity in NLP applications (Koopman et al., 2017; Scells et al., 2018; Sultan et al., 2020). Motivated by these observations, we study how to improve utterance classification results by drawing on utterance variation. Unfortunately, soliciting clean human-generated data can be expensive and difficult to scale. In this study, we instead consider automated utterance generation schemes to augment the original dataset during the training process, in order to: (1) mitigate the data imbalance, and (2) improve the classification effectiveness. Since automatically generated data is cheap and easy to obtain, we can use it as a way of augmenting existing data, which may improve our classification model's effectiveness and robustness.

In this study, we conduct a thorough investigation of current data augmentation approaches to mitigate the imbalanced data problem, including simple resampling, word-level transformations, and neural text generation. Subsequently, an optimal balanced spot is sought to achieve a better classification result.

Contributions: The main contributions of this paper are twofold. First, we conduct a comprehensive survey regarding state-of-the-art data augmentation approaches for multi-class text

*corresponding author, email: yafang.wyf@antfin.com

Proceedings of the 28th International Conference on Computational Linguistics, pages 5494–5506
Barcelona, Spain (Online), December 8-13, 2020

classification. Second, our extensive empirical analysis compares the effectiveness of such data augmentation approaches, showing the value of adopting variational autoencoding techniques along with hybrid combinations of oversampling and undersampling.

2 Data Augmentation Strategies

Empirically, it is well-known that the scarcity of data can hamper the effectiveness of machine learning models, particularly data-hungry deep neural networks. Hence, data augmentation techniques have become ubiquitous in certain fields to alleviate this issue. A substantial number of studies have focused on applying data augmentation to facilitate classification tasks in computer vision. In the vision domain, very simple data augmentation tricks are able to engender significant performance gains. Some operations, such as rotation, translation, and flipping have come to be routinely invoked for both handwritten document recognition (LeCun et al., 1998) as well as regular image classification tasks (Simonyan and Zisserman, 2014). In addition to these macroscopic operations, Krizhevsky et al. (2012) further leverage RGB channel intensity alterations to reduce overfitting caused by data insufficiency. For speech recognition tasks, data augmentation is as well commonly invoked to train more robust models, and is mainly applied at the audio signal level (Cui et al., 2015; Ko et al., 2015).

Unlike images and speech, text cannot naturally be regarded as a continuous signal that can be perturbed arbitrarily, given its discrete units of semantic meaning. To faithfully retain these features, paraphrases are an intuitive and in some sense ideal way to expand a dataset by incorporating alternative expressions for the existing sentences. However, human rephrasing is too expensive and unrealistic, whereas machine paraphrasing currently has its limitations, e.g., it only works on specific tasks (Wang and Yang, 2015; Hou et al., 2018) and a specific paraphrase corpus may be required (Fader et al., 2013; Qiu et al., 2020).

To cope with the generalization issue of data augmentation, a number of universal approaches have been proposed, and we thoroughly evaluate several such universal data augmentation approaches, which can be classified into three types: simple re-sampling, word-level transformations, and neural text generation. In the following, we explicitly introduce each method, and for VAE-based neural generation modeling, we additionally introduce some new adaptions to facilitate the task of learning a classification.

2.1 Simple Resampling

Simple resampling is the most effortless and convenient method to deal with the data imbalance issue and has been invoked in numerous studies. For instance, Japkowicz and Stephen (2002) investigated the performance of oversampling and undersampling strategies applied to a binary classification task.

We thus consider two different simple resampling operations: (1) We refer to the resampling procedure as *Undersampling* when data samples for each of the majority classes are randomly dropped, thus undersampling such classes, such that the amount of data in all classes becomes the same as the smallest one. (2) *Oversampling* instead refers to the opposite scenario, i.e., minority classes are oversampled to increase their size.

2.2 Word-level Transformations

Word-level transformations can be leveraged to produce new sentences while preserving the semantic features of the original texts to a certain extent. The most intuitive way is synonym replacement (SR) (Kobayashi, 2018), which entails replacing a random word in a data sample with one of its synonyms to construct a new sentence. This can be a promising way of obtaining likely paraphrases of the original sentences, especially for classification tasks (Kobayashi, 2018). Easy Data Augmentation (EDA) is another universal data augmentation technique for NLP (Wei and Zou, 2019), in which one of a set of possible operations, including synonym replacement, random insertion, random swapping, and random deletion, are randomly chosen

and applied to a given sentence. Although the authors show a promising performance gain of EDA over SR, EDA's effectiveness has only been demonstrated on small datasets.

2.3 Neural Text Generation

Text generation is a widely explored yet still very challenging task in NLP. The application of neural networks to text generation has achieved great success in a sizeable number of works (Bowman et al., 2015; Shen et al., 2017; Radford et al., 2019; Shen et al., 2020).

This raises the question of whether neural text generation can also serve as a data augmentation technique. The first neural language model was proposed by Bengio et al. (2003), and the superiority of applying neural network models to text generation tasks has been validated in subsequent work, such as recurrent neural network language models (RNNLM) (Mikolov et al., 2010) and long short-term memory (Hochreiter and Schmidhuber, 1997). Compared with conventional language models, generative adversarial nets (GANs) (Goodfellow et al., 2014) and variational autoencoding (VAE) (Kingma and Welling, 2013) along with its variants, are capable of producing more diverse results (Su et al., 2020). Currently, GAN-based models have excelled primarily in image generation (Radford et al., 2015; Denton et al., 2015) rather than in language tasks. Although a number of attempts regarding text generation have been made (Yu et al., 2017; Su et al., 2018), the training process is known to be extremely unstable and the model requires very careful tuning to find a sweet pot between diversity and quality.

In this work, we propose to exploit standard Seq2Seq neural generation as well as VAE-based models that inject additional variation with stochastical latent variables for data augmentation. Specifically, we consider the following models.

Seq2Seq text generation: OpenNMT (Klein et al., 2017) is a neural machine translation system that can also be used to generate text (Hou et al., 2018). MASS (Song et al., 2019) is another Seq2Seq neural generative model, including pre-training procedure for a language model within a masked encoder–decoder framework and further fine-tuning process for other downstream tasks, such as text summarization and conversational response generation. Since Seq2Seq models require both source and target texts, this option is usually applied for conversational text inputs and may not be suitable for arbitrary ordinary text classification tasks.

VAE models: In contrast to vanilla language modelling, VAE assumes a two-step text generation process: **(1)** A latent code $\mathbf{z}$ is first sampled from a prior distribution $p(\mathbf{z})$, **(2)** The corresponding text is then generated based on the conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$. By introducing the additional latent code trained to be distributed in a stochastic prior space, the generated text is able to demonstrate superior diversity compared with a vanilla language model, where the only stochasticity comes from the output softmax layer. The increased diversity is often considered desirable in data augmentation. As the conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$ is often parametrized with deep neural networks, the exact likelihood cannot be derived analytically. VAE circumvents this problem by resorting to a lower bound of the real log-likelihood, known as the evidence lower bound (ELBO):

$$\begin{aligned} \mathcal{L}(\theta, \phi) &= \log \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} p_\theta(\mathbf{x}|\mathbf{z}) = \log \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \\ &\geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})), \end{aligned} \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is an encoder trained to map each text into its posterior latent code space. To maintain the training efficiency, we define $p(\mathbf{z})$ as a standard Normal distribution and parametrize $q_\phi(\mathbf{z} \mid \mathbf{x})$ as a Gaussian distribution with a diagonal covariance matrix. θ and ϕ are simultaneously trained to minimize $\mathcal{L}(\theta, \phi)$ by gradient descent. The reparametrization trick (Kingma and Welling, 2013; Rezende et al., 2014) is used to backpropagate gradients through sampled stochastic latent variables.

Eq. 1 can also be extended to a label-dependent form. This essentially turns it into a conditional variational autoencoder (CVAE) (Sohn et al., 2015; Zhao et al., 2017; Shen et al., 2019a).

Dataset	Category Sizes						Average Size	# Minority Classes
CoQA	Wikipedia	RACE	Gutenberg	CNN	MCTest	-	18,661	1
	21,127	21,615	21,488	21,819*	7,255†	-		
ICS	Class A	Class B	Class C	Class D	Class E	-	65,183	3
	62,739	59,693	49,084†	77,680*	76,716			
NEWS	Politics	Wellness	Entertainment	Style&Beauty	Travel	Parenting	11,813	4
	23,728*	14,292	10,756	7,672	7,482	6,945†		

Table 1: Class distribution of training data in CoQA, ICS, and NEWS datasets, where * indicates the class with the largest number of data samples and † denotes the smallest.

The objective function is changed accordingly to condition on an extra label l:

$$\begin{aligned} &\mathcal{L}(\theta, \phi) = \log \mathbb{E}_{\mathbf{z} \sim p_\theta(\mathbf{z}|l)}\, p_\theta(\mathbf{x}|\mathbf{z}, l) \\ \geq &\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, l)} \log p_\theta(\mathbf{x}|\mathbf{z}, l) - KL(q_\phi(\mathbf{z}|\mathbf{x}, l) \,||\, p_\theta(\mathbf{z}|l)) \end{aligned} \tag{2}$$

$p_\theta(\mathbf{z}|l)$ is parametrized as a label-dependent Gaussian distribution with a diagonal covariance matrix (Shen et al., 2019b).

The training of VAEs often falls into a posterior collapse (Bowman et al., 2015; Yang et al., 2017; Shen et al., 2019a), where the KL term tends to be over-optimized. When the KL term reduces to zero, the entire model degenerates to the vanilla language model with the latent code losing its impact. We adopt the common practice of reserving some bits for the KL term (Kingma et al., 2016), where it is only optimized when it exceeds the reserved value.

When using the VAE to generate augmented text, we can either train separate **unconditional VAEs** (Eq. 1) for each class, or train a single **conditional VAE** (Eq. 2) by taking the class information as an additional input. As for the sampling strategy of the latent code, we also have two options: sampling from the **prior distribution** or from the **posterior distribution** for each training data point. The posterior distribution has a lower variance, and thus it usually corresponds to text semantically similar to the training data. On the contrary, samples from the prior distribution exhibit a greater diversity and can often synthesize novel text different from the training corpus (Bowman et al., 2015; Serban et al., 2017; Shen et al., 2018). By combining these variants, we arrive at the following three kinds of models for data augmentation: 1) **SentenceVAE**: Unconditional VAE + prior sampling; 2) **CVAE**: conditional VAE + prior sampling; 3) **CVAE-posterior (CVAE-p)**: conditional VAE + posterior sampling

2.4 Sweet Spot Optimization

As suggested by López et al. (2013), a blend of oversampling and undersampling might mitigate the imbalance issues in binary classification tasks. Here, we propose that in multi-class classification tasks, there exists a *sweet spot* with regard to the balance of majority and minority classes. That is to say, supplementing the data of some categories while decreasing that of others may achieve better results in some cases. Hence, in our framework, for each generation method and classifier, the most optimal balanced sweet spot is identified. The procedure of how this balanced spot is found is illustrated in Section 3.4.

3 Experiments

In this section, we describe the datasets used in the experiments, followed by an elaboration of our experimental setup and pertinent implementation details. Subsequently, we provide our quantitative experimental results as well as a detailed analysis.

3.1 Datasets

For a fair and comprehensive comparison, we conducted our experiments on three large multi-class datasets, including two conversational datasets and one news dataset. We set apart 10%

of the data as the validation set, and 10% of the data as the testing set. The number of training examples for each class in the training set is listed in Table 1.

CoQA: CoQA (Reddy et al., 2019) is a large English language conversational question answering dataset, consisting of more than 127,000 questions with answers collected from more than 8,000 conversations. We used data from five open domains and reconstruct the dataset to make it more suitable for our task: each extractive answer is a data sample in a class, and we combined the short answer and the question as source, and used the extractive answer as the target in the Seq2Seq models.

Intelligent Customer Service Dataset (ICS): We further obtained more than 400K Chinese language user–agent conversations from a real intelligent customer service system from five different domains in the financial sector of a company. Considering that the lack of diversity in automatic system responses might adversely affect the generative models, only the users' questions were retained, except for Seq2Seq generative models, which require system responses to generate user questions. Based on the domain information, each data sample is annotated with a label (Classes A to E).

News Category Dataset (NEWS): This dataset by Misra (2018) is an English news archive consisting of 200K news headlines and short descriptions along with their corresponding categories in the Huffington Post from the years 2012 to 2018. In our task, short descriptions are used to predict the labels, and, thus, data samples without such short descriptions were omitted. Given that our experiment focused on large datasets, 6 categories of news with comparatively large amounts of data samples were selected.

3.2 Experimental Settings

Text preprocessing. For ordinary text such as news or reviews, no additional preprocessing was required. For utterances in a multi-turn conversational system, we combined the user's questions within one dialogue into a sequence and separated each question by a vertical bar.

Ex.: Hello, I have a question about a 1310 bucks payment I made by credit pay just now| Why the money hasn't arrived?| Right

Data generation. To ensure that the comparison between each generation method was fair and thorough, a uniform experimental workflow with the same training and test data split was required, and in this part of experiment, the optimal balanced spot search was not included. Thus, for all of the methods, we oversampled the generated data to the size of the largest class.

The general flow was as follows: The imbalanced data D_i is oversampled to dataset D_a based on different generation methods. For simple resample methods, sentences in D_a are randomly selected from D_i, and word transformation approaches apply arbitrary simple word-level changes to sentences in D_i so as to obtain an enlarged dataset D_a. Neural sentence generation models demand a learning process to build their D_a. Thus, for each system, one or several neural generative model $M_{g,i}$ is trained using D_i to generate new texts according to the learned distributions to expand D_i. Note that for the CoQA dataset, the combination of short answer and question are sources, while the corresponding extractive answers are targets. For the ICS dataset, system responses serve as sources, whereas user questions are considered targets to be generated. We do not apply Seq2Seq models to augment the NEWS dataset.

Augmentation Models. The SentenceVAE model we used in our experiment was identical with that of Bowman et al. (2015). The CVAE model was inspired by Sohn et al. (2015)'s work, which concatenates label information with each data sample. The architecture of the CVAE-posterior model was similar to the model of Yoo et al. (2019). For each generative model, we used a word embedding of dimensionality 300 with stochastic initialization.

Classification Models. For the evaluation, we used 4 popular classification models: BiLSTM (Schuster and Paliwal, 1997), TextCNN (Kim, 2014), TextRCNN (Lai et al., 2015), and FastText

	Classifier	Original	U-Samp.	O-Samp.	SR	EDA	OpenNMT	MASS	Sent.VAE	CVAE	CVAE-p
CoQA	TextCNN	0.6387	0.5599	**0.6418***	0.6390*	0.6369	0.6119	0.6256	0.6384	0.6357	0.6386
	BiLSTM	0.6224	0.5247	0.6250*	0.6207	0.6237*	0.5977	0.6126	**0.6287***	0.6230	0.6229
	TextRCNN	0.6496	0.5831	**0.6511***	0.6389	0.6412	0.6092	0.6403	0.6414	0.6485	0.6469
	FastText	0.6592	0.5994	0.6629*	0.6574	**0.6641***	0.6226	0.6525	0.6613*	0.6565	0.6617*
ICS	TextCNN	0.8061	0.8061	0.8061	0.8058	0.8067*	0.8052	0.8025	0.8044	0.8038	**0.8074***
	BiLSTM	0.8182	0.8144	0.8174	0.8171	0.8175	0.8168	0.8167	0.8162	0.8160	**0.8189***
	TextRCNN	0.8180	0.8143	0.8163	0.8156	0.8158	0.8162	0.8147	0.8153	0.8146	**0.8192***
	FastText	**0.8180**	0.8150	0.8170	0.8168	0.8162	0.817	0.8168	0.8163	0.8166	**0.8180**
NEWS	TextCNN	0.6774	0.6605	0.6789*	**0.6908***	0.6871*	—	—	0.6714	0.6794*	0.6893*
	BiLSTM	0.6847	0.6551	0.6700	0.6854*	0.6957*	—	—	0.6916*	0.6881*	**0.6974***
	TextRCNN	0.6946	0.6689	0.6935	0.6956*	0.6985*	—	—	0.6839	0.6985*	**0.6998***
	FastText	0.6917	0.6606	0.6833	0.6865	0.6874	—	—	0.6714	0.6908	**0.6985***

Table 2: F_1 scores for classification tasks on the testing sets of the CoQA, ICS, and NEWS datasets. Here, * denotes that the corresponding augmentation approaches improve the results. Bold highlighting reflects the best result across all augmentation methods.

(Joulin et al., 2016). To counterbalance the effect of randomicity in the training process, we selected 5 different random seeds for each of the five models, and evaluated each model by computing the average F_1 scores. For each classification model, we used Adam (Kingma and Ba, 2014) optimization with the same learning rate (0.001) for parameter optimization. We padded or clipped each sentence to 32 characters for the Chinese data, and 64 words for the English datasets, with the same embedding size of 300. We trained each classifier with an early stopping strategy for at most 20 epochs to get optimal results, which are shown in the following section.

3.3 Results

In the following, we present the general experimental results comparing different augmentation strategies in Section 3.3.1 and further fine-grained analysis in Section 3.3.2.

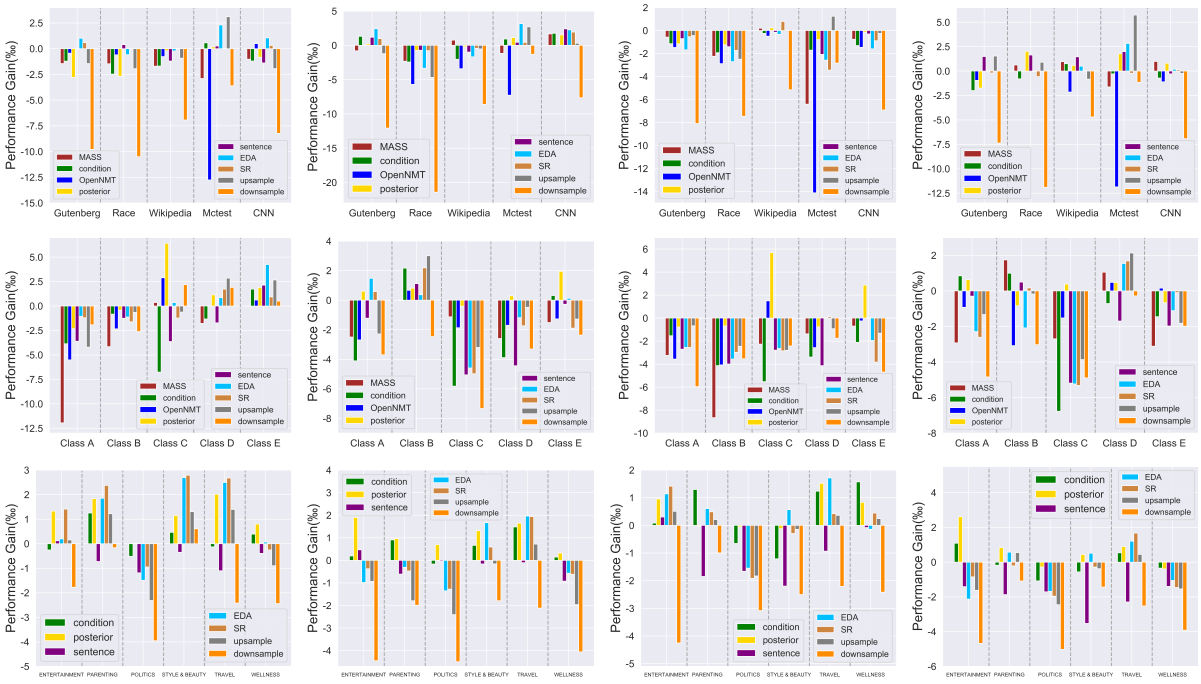
3.3.1 Comparison of Augmentation Strategies

Table 2 shows the F_1 scores for the text classification tasks with different augmentation approaches before best balanced spot optimization, over all the three datasets, which means all the classes were supplemented to the size of the largest one except for Undersampling.

From Table 2, it can be observed that some data augmentation approaches can improve the classification performance while some cannot, and for different datasets, the most effective methods vary. Simple resampling is the most effortless way to achieve data augmentation. In our experiment, Undersampling was shown to be inferior across all datasets, a result consistent with the intuition that dropping data samples from majority classes can lead to information loss (He and Garcia, 2009), thus impeding the classification results. Although Oversampling did not show improvements across all of the three datasets, it achieved a noticeable performance gain on CoQA.

Regarding word-level transformation approaches, EDA was more effective than SR, as it was able to enhance the F_1 score in more situations, and compared with Oversampling, EDA had more positive impact on ICS and NEWS.

Seq2Seq models have proven extremely useful in NLG tasks such as neural translation and dialogue generation. However, in our setting, both of the two Seq2Seq models were found to be counterproductive. Interestingly, we observed that MASS produced the most fluent sentences among all the generation methods under consideration, yet it is a far less ideal choice for data augmentation. This may stem from the fact that, during the pretraining step, to build



(a) TextCNN (b) BiLSTM (c) TextRCNN (d) FastText

Figure 1: The performance gain for each class with different augmentation approaches. From top to bottom, the plots are for CoQA, ICS, NEWS.

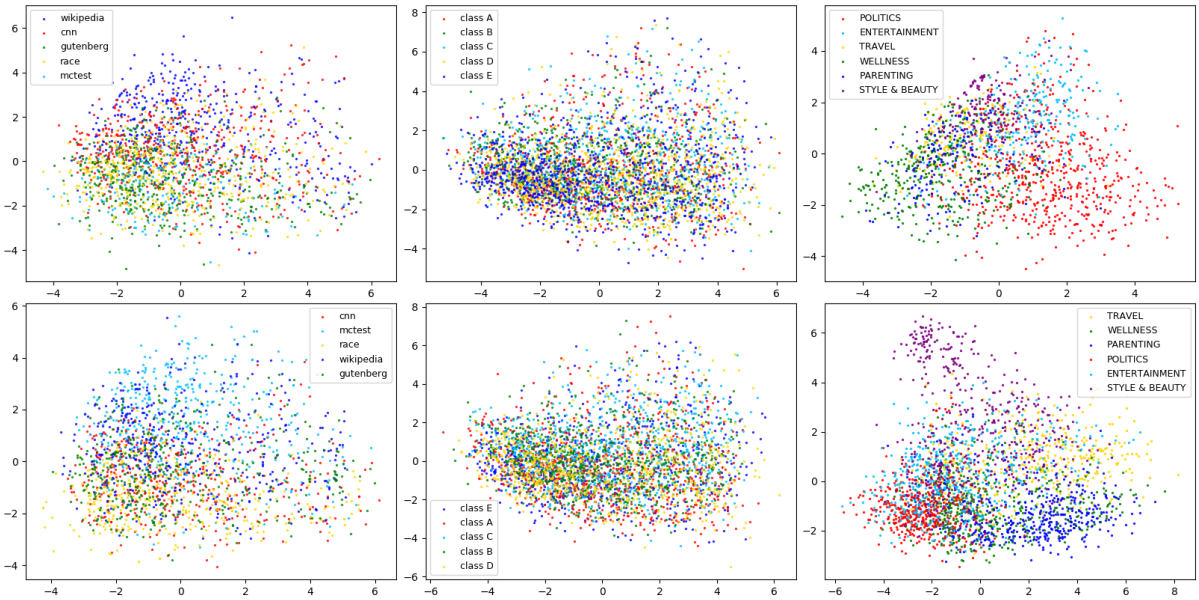
up a robust language model, we feed in all data samples for training. Although we fine-tuned the pretrained model for each class, the pretrained model still introduced a lot of common features shared across different classes, which may contradict the goal in classification tasks of recognizing the distinctive features of different categories.

Another type of neural generation model, VAE, performed much better than Seq2Seq models as data augmentation methods for text classification tasks. Comparing different variants, although conditional VAE has been shown effective in classification tasks (Sohn et al., 2015), it barely improved the results in this case. It can easily be observed from Table 2 that there is a negligible difference between F_1 scores for classification with training sets augmented by SentenceVAE and by the CVAE model with prior sampling. The CVAE model with posterior sampling, however, notably facilitated classification, as it had greatly positive influence on all of the classifiers, especially on the ICS and NEWS datasets. These considerable gains suggest that posterior sampling can help to generate data samples with better categorical characteristics.

3.3.2 Imbalance Level Evaluation and Dataset Disparity

He and Garcia (2009) argue that the imbalanced data problem cannot simply be reduced to considering the relative imbalance between the majority and minority class sizes, as the absolute sample sizes and concept complexity substantially affect the classifier's learning ability as well. Therefore, merely evaluating the overall performance of a classifier is insufficient. Instead, additional analysis of the specific changes for majority and minority classes is needed. In the following, we consider those classes with a number of data samples substantially below the average number for each category as minority classes, and the remaining ones as majority classes, as shown in Table 1.

Figure 1 illustrates the performance gain for each category compared with the original class distribution under different augmentation schemes. The expectation is that minority classes suffer more from data imbalance and augmentation strategies are typically invoked to boost



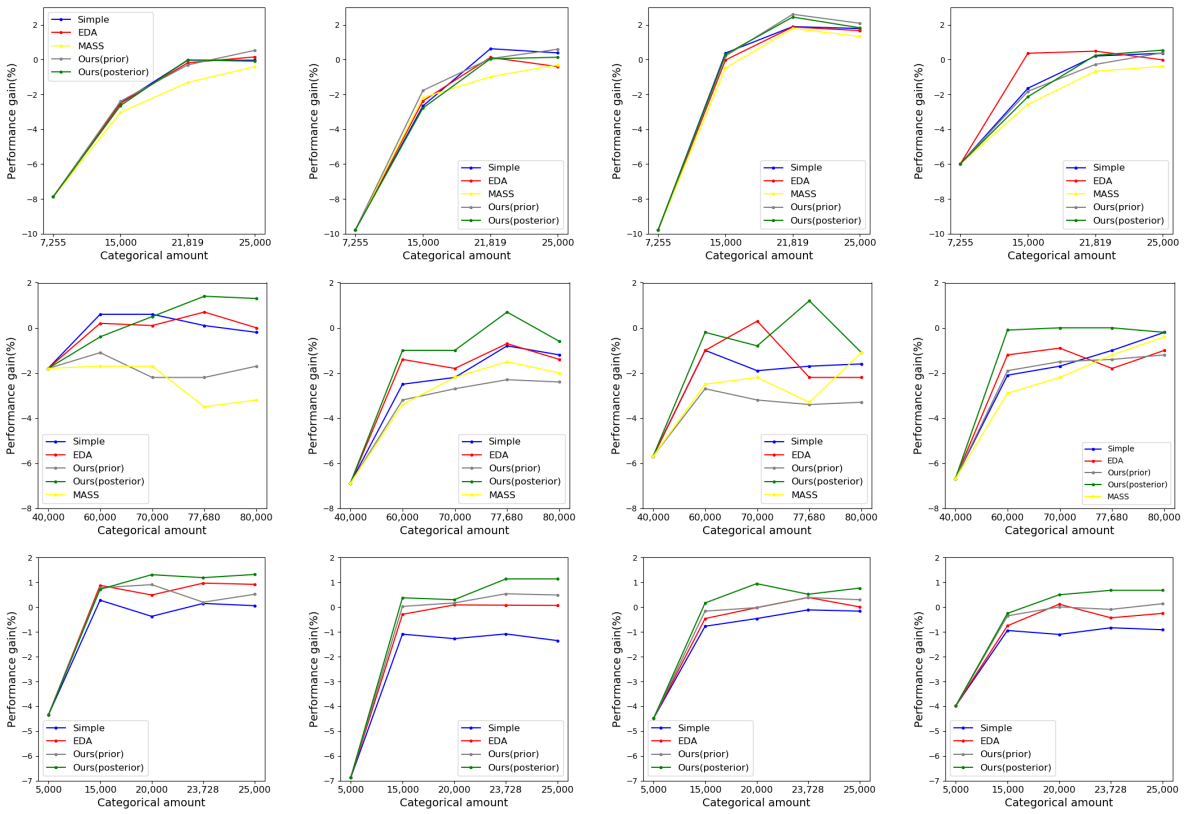
(a) CoQA Dataset (b) ICS Dataset (c) NEWS Dataset

Figure 2: Distribution of features in two-dimensional space. The top row visualizes the original imbalanced data, while the bottom row shows the data after application of the augmentation approach with the best F_1 score.

the classifier's performance on minority classes more than for the majority ones. The results on our NEWS dataset accord with this intuition: According to the Figure 1 (bottom), on the NEWS dataset, with the application of most augmentation methods, the F_1 scores of minority classes are enhanced. Regarding the majority classes in NEWS: For Politics and Wellness, the augmentation approaches had a negligible effect or even impeded the classification results, except for Wellness with TextRCNN. On the CoQA dataset, in Figure 1 (top), we find that there was only one minority class: MCTest. Compared with NEWS, the skew of the class distribution in CoQA is more severe, however, the F_1 score of MCTest was improved most with TextCNN and FastText, while when using BiLSTMs, the class with the largest number of data samples: CNN, improved the most. In addition, it can be observed from Figure 1 (middle) that the change of classification results on the ICS dataset was inconsistent with the conclusion made in previous works, as there is no conspicuous sign indicating that data augmentation can help the minority classes.

The discrepancy among the three datasets can mainly be analysed from two perspectives: 1) the absolute sample size of minority classes along with the level of relative imbalance (Japkowicz and Stephen, 2002), and 2) the feature overlap degree among the different classes. First, as shown in Table 1, the data volume of each class in ICS is much larger than that of the two English datasets. The number of samples in the smallest class of the two English datasets is around 7K, while the smallest class in ICS Class C, has around 50K samples, which is 7 times as many as for the other two datasets. Moreover, for the English datasets, the number of samples in the largest category (Politics, CNN) is more than three times as large as that of the smallest one (Travel, MCTest), and thus the relative data imbalance is much more severe.

Second, the features for each data sample within a category can affect the performance of data augmentation. Figure 2 shows the distribution of sentence vectors encoded by the BERT (Devlin et al., 2018) pretrained model, the dimensionality of which is reduced to 2 using PCA (Wold et al., 1987). Of course, a 2-dimensional visualization can only be regarded as indicative of the distribution of each class, due to the loss of information in the dimensionality reduction process. Still, comparing the data sample distribution of different classes in three datasets



(a) TextCNN (b) BiLSTM (c) TextRCNN (d) FastText

Figure 3: Performance gain of different balance spots for each dataset. From top to bottom, the plots are for CoQA, ICS, NEWS.

in the top row of Figures 2(a), (b), and (c), we observe that ICS suffers from an overlap of features more than from the relative imbalance, as the distribution of data samples within one class severely overlaps with the others and the divergence between different classes is overly small. After leveraging the CVAE-posterior model as the augmentation approach, the patterns of different classes are still hard to recognize (Figure 2(a) bottom). The two English datasets, especially NEWS, suffer more from class imbalance than the Chinese one. For instance, PARENTING and TRAVEL might be regarded as noise relative to the majority classes. After applying augmentation to eliminate the imbalance, the distribution of the minority data becomes much more distinct, so that the margin between each class is more easily recognizable, thus facilitating the classifier's learning ability in discerning minority classes.

3.4 Sweet Spot Identification

Finally, we explored the relationship between the sizes of categories chosen for augmentation and the performance gains on all of the three datasets. We argue that conducting oversampling and undersampling operations simultaneously on different categories within a dataset may achieve a better performance gain than strictly undersampling or oversampling all categories to the same amount as the smallest or largest class, respectively.

The results of the best balanced spot search are illustrated in Figures 3. It can easily be observed that for across all three datasets, in most cases, a comparatively optimal spot between the smallest and the largest categories exists regardless of the generation approach. For example, when trained with an RCNN, the classification results on CoQA dataset show the highest F_1 score at the balance spot 20,000 with the CVAE-posterior model. The results confirm that random undersampling may significantly hamper the learning ability of all clas-

sifiers, while there remains a small probability that the optimal spot is found above the largest amount, such as on NEWS with FastText. The precise location of the sweet spot depends on the specific classification model and datasets. For instance, the overall tendencies on the ICS dataset are less stable than for others. Even when a particular dataset is augmented by the same approach, the optimal spot for different classifiers can be different. Overall, these results confirm the utility of augmentation methods that more flexibly choose a hybrid form of oversampling and undersampling.

4 Conclusion

This paper presents a survey and systematic experimental framework to investigate state-of-the-art data augmentation schemes for text classification, considering both utterance classification and ordinary multi-class text categorization. We carried out a thorough set of experiments to compare the effectiveness of different strategies. Our results highlight the potential of using recent neural generative models as a method to facilitate classification for large datasets. In particular, we VAE methods were found to be comparable to if not better than previous state-of-the-art approaches such as EDA. Our experiments further show that an optimal balanced spot is able to further improve the classification results. Finally, based on our detailed analyses regarding multiclass imbalance, we argue that the imbalance issue cannot be reduced to merely considering the relative imbalance in the number of data samples. Rather, more focus should be placed on the absolute counts and the feature representations within each class.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(9):1469–1477.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv preprint arXiv:1807.01554*.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Bevan Koopman, Liam Cripwell, and Guido Zuccon. 2017. Generating clinical queries from patient narratives: A comparison between machines and humans. In *Proceedings of the 40th international ACM SIGIR conference on Research and development in information retrieval*, pages 853–856.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Rishabh Misra. 2018. News category dataset, June.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard de Melo, Chong Long, and Xiaolong Li. 2020. Easyaug: An automatic textual data augmentation platform for classification tasks. In *Companion Proceedings of the Web Conference 2020*, pages 249–252.

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Harrisen Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. 2018. Query variation performance prediction for systematic reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1089–1092.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. 2017. Estimation of gap between current language models and human performance. *Proc. Interspeech 2017*, pages 553–557.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019a. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019b. Improving latent alignment in text summarization by generalizing the pointer generator. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3753–3764.

Xiaoyu Shen, Ernie Chang, Hui Su, Jie Zhou, and Dietrich Klakow. 2020. Neural data-to-text generation via jointly learning the segmentation and correspondence. *arXiv preprint arXiv:2005.01096*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Hui Su, Xiaoyu Shen, Pengwei Hu, Wenjie Li, and Yun Chen. 2018. Dialogue generation with gan. In *AAAI*.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. *arXiv preprint arXiv:2005.04346*.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for qa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Jason W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3881–3890. JMLR. org.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7402–7409.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.