Figure 2: Frequency of Insertion Errors by Characters

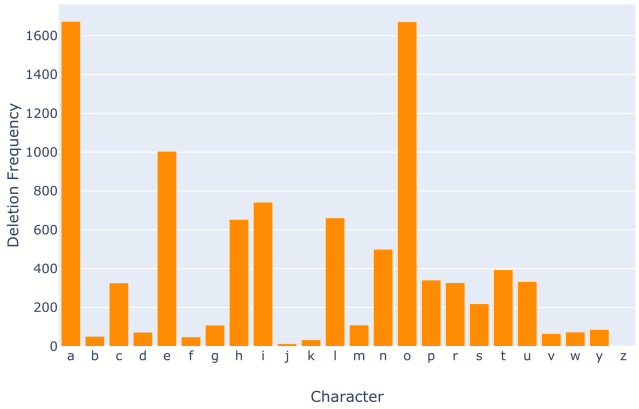

Figure 3: Frequency of Deletion Errors by Characters

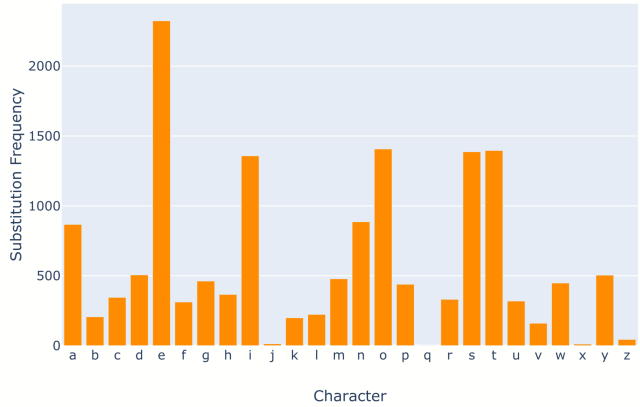

Figure 4: Frequency of Substitution Errors by Characters

to be missed. Figure 4 depicts the frequencies of different characters being replaced by another character, i.e., substitution errors. This type of error also exhibits correlation with the occurrence frequencies of characters. This statistic is used to determine the probability of a substitution error at a given character during error induction. Figure 5 shows the probabilities of particular characters being replaced by particular other characters, given that a substitution error is

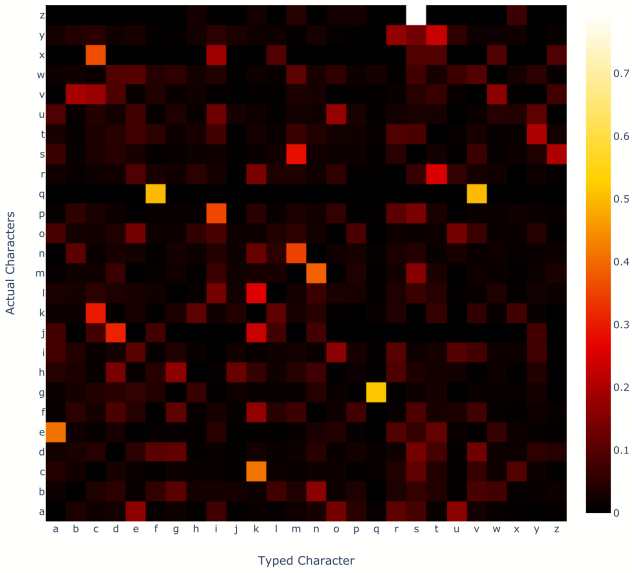

Figure 5: Substitution Probabilities between Characters

Error Level	Corrupted Words (%)	
(% of corrupted char)	Amazon	IMDB
Low (3.75%)	15.73	15.75
Medium (7.5%)	28.66	27.27
High (15%)	48.27	42.07

Table 1: Generated Error Percentage

present. These probabilities are independent of the error rate and are only applied once a substitution error is determined.

5. Experiments

Input Data. We rely on the statistics derived from the Twitter Typo Corpus (Aramaki, 2010), described in Section 4 to introduce errors into two corpora, a food review corpus, and a large movie review one. For food reviews, we consider the Amazon fine food review dataset, which consists of 568,454 food reviews collected from Amazon, along with metadata (McAuley and Leskovec, 2013). The large movie review dataset (Maas et al., 2011) contains 50,000 labeled and 50,000 unlabeled movie reviews collected from IMDB. We only use the text from both of these datasets.

Error Corpus Induction. First of all, to remove pre-existing typographical errors from these corpora, we construct a dictionary V from the Enchant spell checker, enhanced by the vocabulary of GloVe embeddings trained on 6 billion words from Wikipedia and Gigaword (Pennington et al., 2014).[2] On the fine food review dataset, we also remove reviews that are outliers in terms of their length.[3] This leads to 254,638 samples in the fine food review dataset, from

[2]We do not rely on embeddings trained on CommonCrawl, as Web data contains substantially more misspelling forms.

[3]Specifically, those with a character length three standard deviations above or below mean. Hence, we filter out reviews longer than 1,715 characters, but no reviews with shorter length, as three standard deviations below mean is less than 0.

which we sample 160,000 for training, 40,000 as a held-out validation set, and 50,000 as test data. The total size of the dataset is 1.13 GB. From the movie review data, we only exclude reviews with out-of-vocabulary words. The dataset already provides training and test splits, and we reserve 20% of the former for validation, leading to 42,452 training, 10,613 validation, and 17,915 test samples. Despite having fewer sentences, the average sentence length is longer, resulting in a size of 1.27 GB.

We use our proposed method to introduce three levels of error rates, independent of one another. At the highest error rate setting, we induce errors in 15% of the total characters. The other two levels are 7.5% and 3.75%, respectively. For generating confused words, we selected the the highest ranked incorrect suggestions by the Enchant spell-checker (Lachowicz, 2018). This allow us to consistently obtain challenging real-word errors. Table 1 shows the percentage of words with spelling errors corresponding to each level.

For analysis purposes, we evaluate separate neural models as baselines to detect and correct these errors. This allows us to evaluate the detection and correction independently and use the most suitable model for each task.

Hyperparameter	Value
Number of Epochs	50
Learning Rate	0.005
Optimizer	Adam
Embedding Dimensions	100
Recurrent Units	100
Dropout	0.5
Mini-batch Size	512

Table 2: Hyperparameters for Error Detection

Hyperparameter	Value
Number of Steps	5,000
Number of Layers	6
Number of Units	512
Number of Heads	8
FFN Dimensions	2,048
Attention Dropout	0.1
FFN Dropout	0.1

Table 3: Hyperparameters for Error Correction

Spelling Error Detection. For context-aware spelling error detection, we evaluate the effectiveness of a Bidirectional LSTM based sequence labeling model to predict the probability of every word's label being positive, with a sigmoid activation function. We initialize our embeddings with 100-dimensional GloVe vectors trained on Wikipedia and Gigaword (Pennington et al., 2014) and make it further trainable. Out-of-vocabulary words are initialized randomly with the same center and scale as the GloVe vectors. The recurrent layer contains 100 LSTM units. We also employ Dropout with a rate of 0.5 to avoid overfitting. For training, we use Adam optimization with a learning rate of 0.005 for 50 epochs on mini-batches of size 512, with early stopping enabled. To address the class imbalance and to achieve maximum effectiveness, the threshold for positive prediction is fit on the validation set to maximize the F_1 score. Hyperparameters are given in Table 2. Table 4 shows the effectiveness of the detection model on the test data. We observe that with higher error rates, the F_1 score is also higher, in part because the F_1 measure is not completely resilient to class imbalance. We also observe a difference in effectiveness between the two datasets, which may stem from the smaller training size for the IMDB dataset combined with its longer sequence lengths.

Dataset (Error Level)	Recall	Precision	F_1 score
Amazon (3.75%)	0.8867	0.9298	0.9077
Amazon (7.50%)	0.9105	0.9448	0.9273
Amazon (15.00%)	0.9329	0.9560	0.9433
IMDB (3.75%)	0.8172	0.8576	0.8369
IMDB (7.50%)	0.8673	0.8901	0.8786
IMDB (15.00%)	0.9105	0.9155	0.9130

Table 4: Effectiveness of Bidirectional LSTM for Context-Aware Error Detection

	BLEU Score		
Dataset (Error Level)	Noisy	Corrected	Gain
Amazon (3.75%)	0.6491	0.8685	0.2194
Amazon (7.50%)	0.4263	0.8107	0.3844
Amazon (15.00%)	0.1878	0.6661	0.4783
IMDB (3.75%)	0.6378	0.6852	0.0474
IMDB (7.50%)	0.4122	0.4169	0.0047
IMDB (15.00%)	0.1778	0.4806	0.3028

Table 5: Effectiveness of Transformer Network for Context-Aware Spelling Error Correction

Context-Aware Spelling Error Correction. To evaluate our data on error detection as well as correction, we rely on a Transformer model (Vaswani et al., 2017) to transform the noisy sequence with errors to the correct sequence. The intuition is that the model will consider the wider context to detect and correct errors. We use the base configuration described by Vaswani et al. (2017) but only train our model for 5,000 steps, as this problem is simpler than machine translation. We use separate vocabularies for source and target, and for efficiency consider only words in the corpus with multiple occurrences (others as <UNK>). Hyperparameters for this model are given in Table 3. As an evaluation metric, we rely on BLEU scores (Papineni et al., 2002) using the correct sentence as reference. We also report the BLEU scores for the original noisy data (i.e., a baseline that does not make any correction) to show the improvement. The results are provided in Table 5. We observe substantial improvements on the Amazon dataset. However, the performance suffers on IMDB owing to its longer sequence lengths and limited training data. This suggests that our novel datasets will be useful in encouraging further research on this task.

6. Qualitative Analysis of Generated Data

Table 6 compares the different levels of introduced errors. With a low error rate, the sentence is still intelligible with

a few spelling errors. However, the quality quickly deteriorates as the error rate increases. At 15% error rate by character, the sentence becomes very hard to understand even for a human. One can observe an increase in the length of the sentence as the error increases, as the rate of insertion errors is higher than for deletion errors in the statistics computed from the real-world data.

Type	Sentence
Original	These coffee k cups have an exceptionally bold flavor. The value is great. We bought a second box and will continue to enjoy more in the future.
Low Error Rate (3.75%)	These coffee k cups hate an exceptionallly bold flavor. The value is great. We bought a second box and will kontinue to enjoy more in thye futurt.
Medium Error Rate (7.5%)	Thsse coffee k cups ahbe an exceeptionall bold falvor. Thye value is great. We bought a second box and wgll contikre ot enjos more in the future.
High Error rate (15%)	Thkese cvfffe uk dups ave an etxcepiionallyy bolg fladvor. The value ics great. We beught a second box and whll continnue yo renjy mre un the futere.

Table 6: Levels of Introduced Error

Type	Sentence
Original	These coffee k cups have an exceptionally bold flavor. The value is great. We bought a second box and will continue to enjoy more in the future.
Medium Error Rate (7.5%)	Thsse coffee k cups ahbe an exceeptionall bold falvor. Thye value is great. We bought a second box and wgll contikre ot enjos more in the future.
Enchant Spell Corrector (without confusion enforced)	These coffee k cups ah an exceptional bold flavor. Tye value is great. We bought a second box and well continent OT enjoys more in the future.
Enchant Spell Corrector (with confusion enforced)	Those coffee k cups ah an exceptional bold flavor. Tye value is great. We bought a second box and well continent OT enjoys more in the future.

Table 7: Result from Dictionary-based Spelling Correction

Table 7 shows the output of the Enchant spell corrector on a sentence corrupted with medium error rate. For the given sentence, we get the same result for both with and without enforcing confusion. This happens when Enchant does not suggest the correct word for any of the errors. These errors

Type	Sentence
Original	my baby eats these like they are going out of style they are the perfect size for her to grasp with her fingers so nutritional too
Medium Error Rate (7.5%)	my baby eats these likx thre are going oht of style tehy ared the perfect sjze for her yo grasp wivh herv fingers so nutriitional to
Enchant Spell Corrector (with confusion enforced)	my baby eats these alix thee are going hot of style thy ares the perfect saxe for her yew grasp wive herb fingers so malnutrition to
Corrected with Transformer (from Enchant correction)	my baby eats these like they are going hot of style they are the perfect size for her to grasp with her fingers so nutritional too
Original	don't the sellers read these reviews and say something to the manufacturer it is a terrible that amazon can sell this product to the public
Medium Error Rate (7.5%)	do't the sellers read these revies nad say something to he manufacturer pit is a rerrible that amazon can ysle ths prxduct to hte public
Enchant Spell Corrector (with confusion enforced)	dot the sellers read these revise bad say something to eh manufacturer pit is a revertible that amazon can isle th duct to ht public
Corrected with Transformer (from Enchant correction)	got the sellers read these reviews and say something to the manufacturer it is a terrible that amazon can ship the product to the public
Original	i bought this for my niece for christmas she loves it the bamboo showed up intact and looking gorgeous and green in the middle of december
Medium Error Rate (7.5%)	i booqhtt htis for my niece for chrustmas she loves it the bmbou showed up intact and looking gorgeous antd creen in the wmiddle fo dacember
Enchant Spell Corrector (with confusion enforced)	i booth hits for my niece for christmas she loves it the bomb showed up intact and looking gorgeous ants screen in the middle few december
Corrected with Transformer (from Enchant correction)	i bought this for my niece for christmas she loves it the combo showed up intact and looking gorgeous and screen in the middle of december

Table 8: Randomly Selected Examples of Context-Aware Spelling Correction