Categorical Metadata Representation for Customized Text Classification

Jihyeok Kim*1 Reinald Kim Amplayo*2
Kyungjae Lee1 Sua Sung1 Minji Seo1 Seung-won Hwang1
(* equal contribution)

1Yonsei University 2University of Edinburgh
zizi1532@yonsei.ac.kr reinald.kim@ed.ac.uk
{lkj0509,dormouse,ggatalminji,seungwonh}@yonsei.ac.kr

Abstract

The performance of text classification has improved tremendously using intelligently engineered neural-based models, especially those injecting categorical metadata as additional information, e.g., using user/product information for sentiment classification. This information has been used to modify parts of the model (e.g., word embeddings, attention mechanisms) such that results can be customized according to the metadata. We observe that current representation methods for categorical metadata, which are devised for human consumption, are not as effective as claimed in popular classification methods, outperformed even by simple concatenation of categorical features in the final layer of the sentence encoder. We conjecture that categorical features are harder to represent for machine use, as available context only indirectly describes the category, and even such context is often scarce (for tail category). To this end, we propose using basis vectors to effectively incorporate categorical metadata on various parts of a neural-based model. This additionally decreases the number of parameters dramatically, especially when the number of categorical features is large. Extensive experiments on various data sets with different properties are performed and show that through our method, we can represent categorical metadata more effectively to customize parts of the model, including unexplored ones, and increase the performance of the model greatly.

1 Introduction

Text classification is the backbone of most NLP tasks: review classification in sentiment analysis (Pang et al., 2002), paper classification in scientific data discovery (Sebastiani, 2002), and question classification in question answering (Li and Roth, 2002), to name a few. While prior methods require intensive feature engineering, recent methods enjoy automatic extraction of features from text using neural-based models (Socher et al., 2011) by encoding texts into low-dimensional dense feature vectors.

This paper discusses customized text classification, generalized from personalized text classification (Baruzzo et al., 2009), where we customize classifiers based on possibly multiple different known categorical metadata information (e.g., user/product information for sentiment classification) instead of just the user information.

As shown in Figure 1, in addition to the text, a customizable text classifier is given a list of categories specific to the text to predict its class. Existing works applied metadata information to improve the performance of a model, such as user and product (Tang et al., 2015) information in sentiment classification, and author (Rosen-Zvi et al., 2004) and publication (Joorabchi and Mahdi, 2011) information in paper classification.

Towards our goal, we are inspired by the advancement in neural-based models, incorporating categorical information “as is” and injecting it on various parts of the model such as in the word embeddings (Tang et al., 2015), attention mechanism (Chen et al., 2016; Amplayo et al., 2018a) and memory networks (Dou, 2017). Indeed, these methods theoretically make use of combined features from both textual and categorical features, which make them more powerful than disconnected features. However, metadata is generated for human understanding, and thus we claim that these categories need to be carefully represented for machine use to
improve the performance of the text classifier effectively.

First, we empirically invalidate the results from previous studies by showing in our experiments on multiple data sets that popular methods using metadata categories “as is” perform worse than a simple concatenation of textual and categorical feature vectors. We argue that this is because of the difficulties of the model in learning optimized dense vector representation of the categorical features to be used by the classification model. The reasons are two-fold: (a) categorical features do not have direct context and thus rely solely on classification labels when training the feature vectors, and (b) there are categorical information that are sparse and thus cannot effectively learn optimal feature vectors.

Second, we suggest an alternative representation, using low-dimensional basis vectors to mitigate the optimization problems of categorical feature vectors. Basis vectors have nice properties that can solve the issues presented here because they (a) transform multiple categories into useful combinations, which serve as mutual context to all categories, and (b) intelligently initialize vectors, especially of sparse categorical information, to a suboptimal location to efficiently train them further. Furthermore, our method reduces the number of trainable parameters and thus is flexible for any kinds and any number of available categories.

We experiment on multiple classification tasks with different properties and kinds of categories available. Our experiments show that while customization methods using categorical information “as is” do not perform as well as the naive concatenation method, applying our proposed basis-customization method makes them much more effective than the naive method. Our method also enables the use of categorical metadata to customize other parts of the model, such as the encoder weights, that are previously unexplored due to their high space complexity and weak performance. We show that this unexplored use of customization outperform popular and conventional methods such as attention mechanism when our proposed basis-customization method is used.

Figure 1: A high-level framework of models for the Customized Text Classification Task that inputs a text with $n$ tokens (e.g., review) and $m$ categories (e.g., users, products) and outputs a class (e.g., positive/negative). Example tasks are shown in the left of the figure.
2 Preliminaries

2.1 Problem: Customized Text Classification

The original text classification task is defined as follows: Given a text $W = \{w_1, w_2, ..., w_n\}$, we are tasked to train a mapping function $f(W)$ to predict a correct class $y \in \{y_1, y_2, ..., y_p\}$ among the $p$ classes. The **customized text classification** task makes use of the categorical metadata information attached on the text to customize the mapping function. In this paper, we define categorical metadata as non-continuous information that describes the text.\(^1\) An example task is review sentiment classification with user and product information as categorical metadata.

Formally, given a text $t = \{W, C\}$, where $W = \{w_1, w_2, ..., w_n\}$, $C = \{c_1, c_2, ..., c_m\}$, $w_x$ is the $x$th of the $n$ tokens in the text, and $c_z$ is the category label of the text on the $z$th category of the $m$ available categories, the goal of customized text classification is to optimize a function $f_C(W)$ to predict a label $y$, where $f_C(W)$ is the classifier dependent with $C$. In our example task, $W$ is the review text, and we have $m = 2$ categories where $c_1$ and $c_2$ are the user and product information.

This is an interesting problem because of the vast opportunities it provides. First, we are motivated to use categorical metadata because existing work has shown that non-textual additional information, such as POS tags (Go et al., 2009) and latent topics (Zhao et al., 2017), can be used as strong supplementary supervision to improve the performance of text classification. Second, while previously used additional information is found to be helpful, they are either domain-dependent or very noisy (Amplayo et al., 2018b). On the other hand, categorical metadata are usually factual and valid information that are either inherent (e.g., user/product information) or human-labeled (e.g., research area). Finally, the customized text classification task generalizes the personalization problem (Baruzzo et al., 2009), where instead of personalizing based on single user information, we customize based on possibly multiple categories, which may or may not include user information. This consequently creates an opportunity to develop customizable virtual assistants (Papacharissi, 2002).

2.2 Base Classifier: BiLSTM

We use a Bidirectional Long Short Term Memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997) as our base text classifier as it is proven to work well on classifying text sequences (Zhou et al., 2016). Although the methods that are described here apply to other effective classifiers as well, such as convolutional neural networks (CNNs) (Kim, 2014) and hierarchical models (Yang et al., 2016), we limit our experiments to BiLSTM to cover more important findings.

Our BiLSTM classifier starts by encoding the word embeddings using a forward and a backward LSTM. The resulting pairs of vectors are concatenated to get the final encoded word vectors, as shown here:

$$ w_i \in \mathbb{W} $$

$$ h_i = LSTM_f(w_i, h_{i-1}) $$

$$ \overleftarrow{h}_i = LSTM_b(w_i, \overleftarrow{h}_{i+1}) $$

$$ h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i] $$

Next, we pool the encoded word vectors $h_i$ into a text vector $d$ using an attention mechanism (Bahdanau et al., 2015; Luong et al., 2015), which calculates importance scores using a latent context vector $x$ for all words, normalizes the scores using softmax, and uses them to do weighted sum on encoded word vectors, as shown:

$$ e_i = x^\top h_i $$

$$ a_i = \frac{exp(e_i)}{\sum_j exp(e_j)} $$

$$ d = \sum_i h_i \ast a_i $$

Finally, we use a logistic regression classifier to classify labels using learned weight matrix $W^{(c)}$ and bias vector $b^{(c)}$:

$$ y' = W^{(c)}d + b^{(c)} $$

We can then train our classifier using any gradient descent algorithm by minimizing the negative log likelihood of the log softmax of predicted labels $y'$ with respect to the actual labels $y$. 

---

\(^1\)We limit our scope to texts with categorical metadata information (product reviews, news articles, tweets, etc.), which covers most of the texts on the Web. Texts without metadata can use predicted categorical information, such as topics from a topic model, which are commonly used (Zhao et al., 2017; Chou et al., 2017). However, because the prediction may be incorrect, performance gains cannot be guaranteed. We leave the investigation of this area in future work.
2.3 Baseline 1: Concatenated BiLSTM
To incorporate the categories into the classifier, a simple and naive method is to concatenate the categorical features with the text vector \(d\). To do this, we create embedding spaces for the different categories and get the category vectors \(c_1, c_2, \ldots, c_m\) based on the category labels of text \(d\). We then use the concatenated vector as features for the logistic regression classifier:

\[
y' = W^{(c)}[d; c_1; c_2; \ldots; c_m] + b^{(c)}
\]  

(9)

2.4 Baseline 2: Customized BiLSTM
Although the Concatenated BiLSTM easily makes use of the categories as additional features for the classifier, it is not able to leverage on the possible low-level dependencies between textual and categorical features.

There are different levels of dependencies between texts and categories. For example, when predicting the sentiment of a review “The food is very sweet,” given the user who wrote the review, the classifier should give a positive label if the user likes sweet foods and a negative label otherwise. In this case, the dependency between the review and the user is on the higher level, where we look at relationships between the full text and the categories. Another example is when predicting the acceptance of a research paper given that the research area is NLP, the classifier should focus more on NLP words (e.g., language, text) rather than less-related words (e.g., biology, chemistry). In this case, the dependency between the research paper and the research area is on the lower level, where we look at relationships between segments of text and the categories.

We present five levels of Customized BiLSTM, which differ on the location where we inject the categorical features, listed here from the highest level to the lowest level of dependencies between text and categories. The main idea is to impose category-specific weights, rather than a single weight at each level of the model:

1. **Customize on the bias vector:** At this level of customization, we look at the general biases the categories have towards the problem. As a concrete example, when classifying the type of message a politician wrote, he/she can be biased towards writing personal messages than policy messages. Instead of using a single bias vector \(b^{(c)}\) in the logistic regression classifier (Equation 8), we use additional multiple bias vectors for each category, as shown below. In fact, this is in spirit essentially equivalent to concatenated BiLSTM (Equation 9), where the derivation is:

\[
y' = W_d d + b_{c_1} + \ldots + b_{c_m} + b^{(c)}
\]  

\[
y' = W_d d + W_{c_1} c_1 + \ldots + W_{c_m} c_m + b^{(c)}
\]  

\[
y' = W^{(c)}[d; c_1; c_2; \ldots; c_m] + b^{(c)}
\]

2. **Customize on the linear transformation:** At this level of customization, we look at the text-level semantic biases the categories have. As a concrete example, in the sentiment classification task, the review “The food is very sweet” can have a negative sentiment if the user who wrote the review does not like sweets. Instead of using a single weight matrix \(W^{(c)}\) in the logistic regression classifier (Equation 8), we use different weight matrices for each category:

\[
y' = W^{(c)}_1 d + W^{(c)}_2 d + \ldots + W^{(c)}_m d + b^{(c)}
\]

3. **Customize on the attention pooling:** At this level of customization, we look at the word importance biases the categories need. A concrete example is, when classifying a research paper, NLP words should be focused more when the research area is NLP. Instead of using a single context vector \(x\) when calculating the attention scores \(e\) (Equation 5), we use different context vectors for each category:

\[
e_i = x_{c_1}^T h_i + x_{c_2}^T h_i + \ldots + x_{c_m}^T h_i
\]  

\[
a = \text{softmax}(e)
\]  

\[
d = \sum_i h_i * a_i
\]

4. **Customize on the encoder weights:** At this level of customization, we look at the word contextualization biases the categories need. A concrete example is, given the text “deep learning for political message classification”, when encoding the word classification, the BiLSTM should retain the semantics of words political message more and forget the semantics of other words more when the research area is about politics. Instead of
using a single set of input, forget, output, and memory cell weights for each LSTM (Equations 2 and 3), we use multiple sets of the weights, one for each category:

\[
\begin{bmatrix}
g_t \\
i_t \\
f_t \\
o_t
\end{bmatrix} =
\begin{bmatrix}
tanh \\
\sigma \\
\sigma \\
\sigma
\end{bmatrix}
\left(\sum_{0<k \leq m} W^{(c)}_{ek}[w_t; h_{t-1}] + b\right)
\]

5. **Customize on the word embeddings**: At this level of customization, we look at the **word preference biases** the categories have. For example, a user can prefer the use of word “terribly” as a positive adverb rather than the more common usage of the word with negative sentiment. Instead of directly using the word vectors from the embedding space \( \mathbb{W} \) (Equation 1), we add a residual vector calculated based on a nonlinear transformation of the word vector using category-specific weights:

\[
r = \tanh(W^{(w)}_{c1}w_i + \ldots + W^{(w)}_{cm}w_i) + \gamma_i \]

\[
w_i = w_i + r
\]

Previous work has proposed customization on bias vectors and word embeddings (Tang et al., 2015), and on attention pooling (Chen et al., 2016). We are the first to introduce customization on the linear transformation matrix and the encoders. Moreover, we are the first to use residual perturbations as word meaning modification for customizing word embeddings, in which we saw better performance than using a naive affine transformation, proposed in Tang et al. (2015), in our prior experiments.

3 **Proposed Method**

3.1 **Problems of Customized BiLSTM**

As explained in the previous section, Customized BiLSTM should perform better than Concatenated BiLSTM. However, that is only if the optimization of category-specific weights operates properly for machine usage. Training the model to optimize these weights is very difficult for two reasons.

First, categorical information has unique properties that make it nontrivial to train. One property is that unlike texts that naturally use neighboring words/sentences as context (Lin et al., 2015; Peters et al., 2018), categorical information stands alone and thus does not have information aside from itself. This forces the learning algorithm to rely solely on the classification labels \( y \) to find the optimal category-specific weights. Another property is that some categories may contain labels that are sparse or do not have enough instances. For example, a user can be cold-start (Lam et al., 2008) or does not have enough reviews. In this case, the problem expands to few-shot learning (Li et al., 2006). Thus weights are hard to optimize using gradient-based techniques (Ravi and Larochelle, 2016).

Second, the number of weights is multiplied by the number of categories \( m \) and the number of category labels each category has, which enlarges the number of parameters needed to be trained as \( m \) increases. This magnifies the problems of context absence and information sparsity described above, since optimizing large parameters with limited inductive bias is very difficult. Moreover, because of the large parameters, some methods may not fit in commercially available machines and thus may not be practically trainable.

3.2 **Basis Customization**

We propose to solve these problems by using basis vectors to produce basis-customized weights, as shown visually in Figure 2. Specifically, we use a trainable set of \( d \ll dm \) basis vectors \( B = \{b_1, b_2, ..., b_d\} \), where \( dm \) is the dimension of the original weights. Let \( V_c \) be the vector search space that contains all the optimal customized weight vectors \( v_c \), such that \( B \) is the basis of \( V_c \). Basis vectors follow the spanning property, thus we can represent all vectors in \( v \in V_c \) as a linear combination of \( B \)—that is \( v_c = \sum \gamma_i b_i \), where the \( \gamma_i \) are the coefficients. Moreover, because we set \( d \) to a small number, we constrain the search space to a smaller vector space. Hence we can find the optimal weights in a constrained search space much faster.

To determine the \( \gamma_i \) coefficients, we first set the concatenated category vectors of the text \( q = [c_1; c_2; \ldots; c_m] \) as the query vector, and use a trainable set of key vectors \( K = \{k_1, k_2, ..., k_d\} \). We then calculate the dot product between the query and key vectors, and finally use softmax to create \( \gamma \) coefficients that sum to one:

\[
z_i = q^\top k_i
\]

\[
\gamma_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}
\]

205
We can then use the $\gamma$ coefficients to basis-customize a specific weight $v$, namely, $v = \sum \gamma_i * b_i$. In our BiLSTM classifier, we can basis-customize one of the following weights: (1) the bias vector $v = b^{(c)}$ and (2) the linear transformation matrix $v = W^{(c)}$ of the logistic regression classifier in Equation 8, (3) the context vector $v = x$ of the attention mechanism in Equation 5, (4) the BiLSTM weights $v = W^{(e)}$ in Equations 2 and 3, and (5) the nonlinear transformation matrix $v = W^{(u)}$ on the residual vector in Equation 10 to modify the word embeddings. These correspond to the five versions of Customized BiLSTM discussed earlier.

Basis-customizing weights help solve the problems of customizing BiLSTM in three ways. First, the basis vectors serve as fuzzy clusters of all the categories, that is, we can say that two sets of category labels are similar if they have similar $\gamma$ coefficients. This information can serve as mutual context information that helps the learning algorithm find optimal weights. Second, because the search space $V_c$ is constrained, the model is forced to initialize the category vectors and look for the optimal vectors inside the constrained space. This smart initialization contributes to situate vectors of sparse categorical information to a sub-optimal location and efficiently trains them further, despite the lack of instances. Finally, because we only use a very small set of basis vectors, we reduce the number of weights dramatically.

4 Experiments

We experiment on three data sets for different tasks: (1) the Yelp 2013 data set² (Tang et al., 2015) for Review Sentiment Classification, (2) the AAPR data set³ (Yang et al., 2018) for Paper Acceptance Classification, and (3) the PolMed data set⁴ for Political Message Type Classification. Statistics, categories, and properties of the data sets are reported in Table 1. Details about the data sets are discussed in the next sections.

General experimental settings are as follows. The dimensions of the word vectors are set to 300. We use pre-trained GloVe embeddings (Pennington et al., 2014) to initialize our word vectors. We create UNK tokens by transforming tokens with frequency less than five into UNK. We handle unknown category labels by setting their corresponding vectors to zero. We tune the number of basis vectors $d$ using a development set, first by sweeping across 2 to 30 with large intervals, and then by searching through the neighbors of the best configuration during the first sweep. Interestingly, $d$ tends to be very small, between values 2 to 4. We set the batch size to 32. We use stochastic gradient descent over shuffled mini-batches with the Adadelta update rule (Zeiler,

²http://ir.hit.edu.cn/~dytang.
³https://github.com/lancopku/AAPR.
⁴https://www.figure-eight.com/.
### Data Set Splits

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Splits</th>
<th>Categories</th>
<th>Properties</th>
</tr>
</thead>
</table>
| Yelp 2013 | 62,522 / 7,773 / 8,671 | • users (1.6k)  
• products (1.6k) | Categories can be sparse (i.e., there may not be enough reviews for each user/product). |
| AAPR     | 33,464 / 2,000 / 2,000 | • authors (48k)  
• research area (144) | Authors are sparse and have many category labels. Categories can have multiple labels (e.g., multiple authors, multidisciplinary fields). |
| PolMed  | 4,500 / 0 / 500 | • politician (505)  
• media source (2)  
• audience (2)  
• political bias (2) | The data set has more categories. Categories with binary labels may not be diverse enough to be useful. |

Table 1: The data sets, the split sizes (train, dev, test), and the available categories and their properties. Numbers inside parentheses are the number of unique category labels.

### Models

<table>
<thead>
<tr>
<th>Models</th>
<th>Yelp 2013</th>
<th>AAPR</th>
<th>PolMed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>RMSE</td>
<td>Param</td>
</tr>
<tr>
<td>Base: BiLSTM</td>
<td>63.7</td>
<td>0.687</td>
<td>442k</td>
</tr>
<tr>
<td>bias vector (concat) cust basis-cust</td>
<td>66.3</td>
<td>0.661</td>
<td>1.3m</td>
</tr>
<tr>
<td>linear transformation* cust basis-cust</td>
<td>67.1</td>
<td>0.662</td>
<td>655k</td>
</tr>
<tr>
<td>attention pooling cust basis-cust</td>
<td>65.4</td>
<td>0.674</td>
<td>1.3m</td>
</tr>
<tr>
<td>encoder weights* cust basis-cust</td>
<td>66.0</td>
<td>0.671</td>
<td>652k</td>
</tr>
<tr>
<td>word embedding* cust basis-cust</td>
<td>58.4</td>
<td>0.767</td>
<td>294m</td>
</tr>
</tbody>
</table>

Table 2: Accuracy, RMSE, and parameter values of competing models for all data sets. An asterisk (*) indicates customization methods first introduced in this paper. A dash (-) indicates the model is too big to be trained in an NVIDIA 1080 Ti GPU. **Boldface** indicates that the performance of basis-customization is significantly better ($p < 0.05$) than that of a simple customization. Values colored red are performance weaker than that of the BiLSTM model, thus customization hurts the performance in those cases.

2012) with $l_2$ constraint of 3. We do early stopping using the accuracy of the development set. We perform 10-fold cross-validation on the training set when the development set is not available. Data set–specific settings are described in their corresponding sections.

We compare the performance of the following competing models: the base classifier BiLSTM with no customization, the five versions (i.e., bias, linear, attention, encoder, embedding) of Customized BiLSTM, and our proposed basis-customized versions. We report the accuracy and the number of parameters of all models, and additionally report the root mean square error (RMSE) values for the sentiment classification task. We also compare with results from previous papers whenever available. Results are shown in Table 2, and further discussion is provided in the following sections.

#### 4.1 Review Sentiment Classification

Review sentiment classification is a task of predicting the sentiment label (e.g., 1 to 5 stars) of a review text (Pang et al., 2002). We use users and products as categorical metadata. One main characteristic of the categorical information here is that both user and product can be cold-start entities (Amplayo et al., 2018a). Thus issues on sparseness may aggravate. We use 256 dimensions for the hidden states in the BiLSTM encoder and the context vector in the attention mechanism, and 64 dimensions for each of the user and product category vectors.

The results in Table 2 show that when using Customized BiLSTM, customizing on the bias vector (i.e., Concatenated BiLSTM) performs the best compared to customizing on other parts of the model with lower dependencies, which is counter-intuitive and contrary to previously reported
models perform better, even though we only use a single BiLSTM encoder. **Boldface** correspond to the best values for each block. Underlined values correspond to the best values across the board.

Table 3: Performance comparison of previous and our best models in the Yelp 2013 dataset. Our best models perform better, even though we only use a single BiLSTM encoder. **Boldface** correspond to the best values for each block. Underlined values correspond to the best values across the board.

<table>
<thead>
<tr>
<th>Models</th>
<th>Acc</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPNN (Tang et al., 2015)</td>
<td>59.6</td>
<td>0.784</td>
</tr>
<tr>
<td>UPDMN (Dou, 2017)</td>
<td>63.9</td>
<td>0.662</td>
</tr>
<tr>
<td>NSC (Chen et al., 2016)</td>
<td>65.0</td>
<td>0.692</td>
</tr>
<tr>
<td>HCSC (Amplayo et al., 2018a)</td>
<td>65.7</td>
<td><strong>0.660</strong></td>
</tr>
<tr>
<td>PMA (Zhu and Yang, 2017)</td>
<td>65.8</td>
<td>0.668</td>
</tr>
<tr>
<td>DUPMN (Long et al., 2018)</td>
<td>66.2</td>
<td>0.667</td>
</tr>
<tr>
<td>CMA (Ma et al., 2017)</td>
<td><strong>66.4</strong></td>
<td>0.677</td>
</tr>
<tr>
<td><strong>Our best models</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM + encoder-basis-cust</td>
<td>66.1</td>
<td>0.665</td>
</tr>
<tr>
<td>BiLSTM + bias-basis-cust</td>
<td>66.9</td>
<td><strong>0.654</strong></td>
</tr>
<tr>
<td>BiLSTM + linear-basis-cust</td>
<td><strong>67.1</strong></td>
<td>0.662</td>
</tr>
</tbody>
</table>

The comparison in Table 3 shows that our methods outperform previous models, even though (1) we only use a single BiLSTM encoder rather than more complicated ones (UPDMN and DUPMN use deep memory networks, NSC, PMA, and CMA use hierarchical encoders) and (2) we only customize on one part of the model rather than on multiple parts (UPNN customizes on bias vectors and word embeddings).

### 4.2 Paper Acceptance Classification

Paper acceptance classification is a task of predicting whether the paper in question is accepted or rejected (Yang et al., 2018). We use the authors and the research area of the papers as categorical metadata. Both authors and research field information accept multiple labels per instance (e.g., multiple authors, multidisciplinary field), hence learning the category vector space properly is crucial to perform vector operations (Mikolov et al., 2013). We use 128 dimensions for both the hidden states in the BiLSTM encoder and the context vector in the attention mechanism and 32 dimensions for each of the categorical information. We use the paper abstract as the text. To handle multiple labels, we find that averaging the category vectors works well.

The results in Table 2 show similar trends from the sentiment classification results. First, we obtain better performance when using Concatenated BiLSTM than when using Customized BiLSTM.
Models & Accuracy using full text (Yang et al., 2018)

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>60.5</td>
</tr>
<tr>
<td>MHCNN</td>
<td>67.7</td>
</tr>
</tbody>
</table>

using abstract and categories (our setting)

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>60.6</td>
</tr>
<tr>
<td>MHNN</td>
<td>63.7</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>61.7</td>
</tr>
<tr>
<td>BiLSTM + word-basis-cust</td>
<td>65.8</td>
</tr>
<tr>
<td>BiLSTM + attention-basis-cust</td>
<td>65.9</td>
</tr>
<tr>
<td>BiLSTM + encoder-basis-cust</td>
<td>66.2</td>
</tr>
</tbody>
</table>

Table 4: Performance comparison of models using full texts and our implemented models using paper abstracts (and authors and research areas as categories for basis-customized models) as inputs in the AAPR data set.

Second, incorporating metadata information on the attention mechanism does not perform as well as previously reported. Third, when customizing on encoder weights and word embedding, the model parameters are too big to be trained on a commercial GPU. Finally, we see significant improvements in all levels of customization when using our proposed basis-customization method, except on the bias vectors where we obtain comparable results. Overall, a BiLSTM basis-customized on the encoder weights, the attention pooling, and the word embedding perform the best among all the models. We also see at least 3.7x reduction of parameters when comparing Customized BiLSTM and Basis-Customized BiLSTM.

We also compare our results from previous literature (Yang et al., 2018), where they proposed a modular and hierarchical CNN-based encoder (MHNN), and used the full text (i.e., from the title and authors up to the conclusion section), rather than just the abstract, the author and the research area information. Results are reported in Table 4, although full text and abstract results are not directly comparable because the original authors did not release the train/dev/test splits of their experiments. We instead re-run MHCNN using our settings and compare with our models. The results show that using either full text or abstract as input to LSTM produces similar results, thus using just the abstract can give us similar predictive bias when using the full text, at least in this data set. Moreover, our best models (1) perform significantly better ($p < 0.5$) than MHNN when restricted to our settings, and (2) are competitive with the state-of-the-art, even though we use a simple BiLSTM encoder and only have access to the abstract, authors, and research area information.

4.3 Political Message Type Classification

Political message type classification is a task of predicting the type of information a message written by a politician is conveying, with the following nine types: attack, constituency, information, media, mobilization, personal, policy, support, and others. Two characteristics of this data set different from others are (a) that it has four kinds of categorical information: the audience (national or constituency), bias (neutral or partisan), politician, and the source (Twitter or Facebook) information, and (b) that the category types of three categories are not diverse as they only have binary category labels. Because all of these categories may not give useful information biases to the classifier, models should be able to select which categories are informative or not. We use 64 dimensions for the hidden states in the BiLSTM encoder and the context vector in the attention mechanism, and 16 dimensions for the category vectors of each of the categorical information.

The results in Table 2 also show similar trends from the previous task, but because the data set is smaller, we can compare the performance of the model when customizing on encoder weights. We show that Customized BiLSTM on linear transformation matrix and encoder weights shows weaker performance than the base BiLSTM model, Basis-Customized BiLSTM on the same levels shows significantly improved performance, and Basis-Customized BiLSTM on linear transformation matrix performs the best among the competing models. The parameters also decreased dramatically, especially on encoder weights and on word embedding where we see at least 100x difference in parameter size.

5 Analysis

5.1 Semantics of Basis Attention Vectors

We investigate how basis vectors understand word-level semantics through the lens of the attention vectors they create. Previous models either combine user/product information into a single attention vector (Chen et al., 2016) or entirely separate them into distinct user and product attention vectors (Amplayo et al., 2018a). On the other hand, our model creates a single
attention vector, but through the \( k \) basis attention vectors, which are vectors containing fuzzy semantics among users and products. Figure 3 shows two examples of six attention vectors regarding a single text in the Yelp 2013 data set using the following: (1) the original user, product pair \((u, p)\); (2–3) a sampled user/product paired with the original product/user \((u', p)\) and \((u, p')\); and (4–6) the basis vectors. We can see in the first example that the first basis vector focuses on “cheap” and the third basis vector focuses on “delicious.” An interesting output is by user \( u \), such that they want cheaper food in product \( p \) yet care more about the taste in product \( p' \).

5.2 Document-level Customized Dependencies

Previous literature only focused on the analysis (Amplayo et al., 2018a) and case studies (Chen et al., 2016) of word-level customized dependencies, usually through attention vectors. In this paper, we additionally investigate the document-level customized dependencies, namely, how our basis-customization changes the document-level semantics when a category is different. Table 5 shows two examples, one from the AAPR data set and one from the Political Media data set, with a variable category research area and political bias, respectively. In the first example, the abstract refers to a study on bi-sequence classification problem, a task mainly studied in the natural language processing domain, and thus is classified as accepted when the research area category is cs.CL. The model also classifies the paper as accepted when the research area is cs.IR because the two areas are related. However, when the research area is changed to an unrelated area like cs.CR, the paper is rejected. In the second example, the classifier predicts that when a politician with a neutral bias posts a Christmas greeting and mentions people who work on holidays, he is conveying a personal message. However, when the politician is biased towards a political party, the classifier thinks that the message is to offer support to those workers who are unable to be with their families.

5.3 Learning Strategy of Basis-customized Vectors

We argue that because the basis vectors \( B \) limit the search space into a constrained vector space \( V_c \), finding the optimal values of the basis-customized vectors is faster. We show in Figure 4 the difference between the category vector space of Customized BiLSTM and of Basis-Customized BiLSTM. We see that the vector space of Customized BiLSTM looks random, with very few noticeable clusters, even when we iterate with four epochs. On the other hand, the basis-customized vector space starts as a cluster of one continuous spiral line, then starts to break down into smaller clusters. Multiple clusters of vectors in the vector space are clearly seen when the epoch is 4. Therefore, using the basis vectors makes optimization more efficient by following the learning strategy of starting from one cluster and dividing into smaller coherent clusters. This can also be shown in the visualization of the \( \gamma \) coefficients (also shown in the figure), where the coefficient values that are clumped together gradually spread out to their optimal values.

5.4 Performance on Sparse Conditions

We look at the performance of three models, BiLSTM, Customized BiLSTM, and Basis-Customized BiLSTM, per review frequency of user or product. Figure 5 shows plots of the accuracy of the models over different user review frequency and product review frequency on the Yelp 2013 data set. We observe that naive customization drops the performance of the BiLSTM model as the frequency of user/product review decreases. This means that the model is heavily reliant on large amounts of data for optimization. On the other hand, because basis customization can learn the optimal weights of category vectors more intelligently, it improves the performance of the model across all ranges of review frequency.
Several tasks in argumentation mining and debating, question-answering, and natural language inference involve classifying a sequence in the context of another sequence (referred as bi-sequence classification). For several single sequence classification tasks, the current state-of-the-art approaches are based on recurrent and convolutional neural networks. On the other hand, for bi-sequence classification problems, there is not much understanding as to the best deep learning architecture. In this paper, we attempt to get an understanding of this category of problems by extensive empirical evaluation of 19 different deep learning architectures (specifically on different ways of handling context) for various problems originating in natural language processing like debating, textual entailment and question-answering. Following the empirical evaluation, we offer our insights and conclusions regarding the architectures we have considered. We also establish the first deep learning baselines for three argumentation mining tasks.

<table>
<thead>
<tr>
<th>Research Area</th>
<th>cs.CL (Computation and Language)</th>
<th>cs.IR (Information Retrieval)</th>
<th>cs.CR (Cryptography and Security)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Accept</td>
<td>Accept</td>
<td>Reject</td>
</tr>
</tbody>
</table>

Message | &lt;UNK&gt; christmas and happy holidays from my family to yours. wishing special &lt;UNK&gt; to those first responders and military personnel working to ensure our safety who are unable to be with their families this holiday season. we are all thank you for your service and dedication.

<table>
<thead>
<tr>
<th>Political Bias</th>
<th>Neutral</th>
<th>Partisan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Personal</td>
<td>Support</td>
</tr>
</tbody>
</table>

Table 5: Example texts from the AAPR data set (upper) and Political Media data set (lower) with a variable category label (research field and political bias) that changes the classification label.

![Figure 4: TSNE Visualization of the category vectors of Customized BiLSTM (first row) and Basis-Customized BiLSTM (middle row), and the \( \gamma \) coefficients of the latter model (last row), when epoch is equal to 1, 2, 4, and when training has finished (left to right).](image)

We finally examine the performance of our models when data contain cold-start entities (i.e., users/products may have zero or very few reviews) using the Sparse80, subset of the Yelp 2013 data set provided in Amplayo et al. (2018a). We compare our models with three competing models: NSC (Chen et al., 2016), which uses a hierarchical LSTM encoder coupled with customization on the attention mechanism, BiLSTM+CSAA (Amplayo et al., 2018a), which uses a BiLSTM encoder with customization on a CSAA mechanism, and HCSC (Amplayo et al., 2018a), which is...
Figure 5: Accuracy per user/product review frequency on Yelp 2013 data set. The review frequency value $f$ represents the frequencies in the range $[f, f + 10)$, except when $f = 100$, where it represents the frequencies in the range $[f, \infty)$.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSC</td>
<td>51.1</td>
</tr>
<tr>
<td>BiLSTM + CSAA</td>
<td>52.7</td>
</tr>
<tr>
<td>HCSC</td>
<td>53.8</td>
</tr>
<tr>
<td>BiLSTM + encoder-basis-cust</td>
<td>50.4</td>
</tr>
<tr>
<td>BiLSTM + linear-basis-cust</td>
<td>50.8</td>
</tr>
<tr>
<td>BiLSTM + bias-basis-cust</td>
<td>51.9</td>
</tr>
<tr>
<td>BiLSTM + word-basis-cust</td>
<td>51.9</td>
</tr>
<tr>
<td>BiLSTM + attention-basis-cust</td>
<td>53.1</td>
</tr>
</tbody>
</table>

Table 6: Performance comparison of competing models in the Yelp 2013 Sparse80 data set.

a combination of CNN and the BiLSTM encoder with customization on CSAA. Results are reported in Table 6, which provide us two observations. First, the BiLSTM model customized on the linear transformation matrix, which performs the best on the original Yelp 2013 data set (see Table 3), obtains a very sharp decrease in performance. We posit that this is because basis customization is not able to handle zero-shot cold-start entities, which are amplified in the Yelp 2013 Sparse80 data set. We leave extensions of basis for zero-shot or cold-start, studied actively in machine learning (Wang et al., 2019) and recommendation domains (Sun et al., 2012), respectively. Inspired by CSAA (Amplayo et al., 2018a), using similar review texts for inferring the cold-start user (or product), we expect to infer meta context, similarly based on similar meta context, which may mitigate the zero-shot cold-start problem. Second, despite having no zero-shot learning capabilities, Basis-Customized BiLSTM on the attention mechanism performs competitively with HCSC and performs better than BiLSTM + CSAA, which is Customized BiLSTM on attention mechanism with cold-start awareness.

6 Conclusion

We presented a new study on customized text classification, a task where we are given, aside from the text, its categorical metadata information, to predict the label of the text, customized by the categories available. The issue at hand is that these categorical metadata information are hardly understandable and thus difficult to use by neural machines. This, therefore, makes neural-based models hard to train and optimize to find a proper categorical metadata representation. This issue is very critical, in such a way that a simple concatenation of these categorical information provides better performance than existing popular neural-based methods. We propose solving this problem by using basis vectors to customize parts of a classification model such as the attention mechanism and the weight matrices in the hidden layers. Our results show that customizing the weights using the basis vectors boosts the performance of a basic BiLSTM model, and also effectively outperforms the simple yet robust concatenation methods. We share the code and data sets used in our experiments here: https://github.com/zizi1532/BasisCustomize.

Acknowledgments

This work was supported by Microsoft Research Asia and IITP/MSIT research grant (no. 2017-0-01779).
References


Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016.


