applied sciences

MDPI

Article

Background Knowledge Based Multi-Stream Neural Network for Text Classification

Fuji Ren * and Jiawen Deng

Faculty of Engineering, Tokushima University, Tokushima 2-1, Minami josanjima-cho, Tokushima 770-8506, Japan; c501847002@tokushima-u.ac.jp
* Correspondence: ren@is.tokushima-u.ac.jp; Tel.: +81-808-043-0107

Received: 5 November 2018; Accepted: 26 November 2018; Published: 3 December 2018

Abstract: As a foundation and typical task in natural language processing, text classification has been widely applied in many fields. However, as the basis of text classification, most existing corpus are imbalanced and often result in the classifier tending its performance to those categories with more texts. In this paper, we propose a background knowledge based multi-stream neural network to make up for the imbalance or insufficient information caused by the limitations of training corpus. The multi-stream network mainly consists of the basal stream, which retained original sequence information, and background knowledge based streams. Background knowledge is composed of keywords and co-occurred words which are extracted from external corpus. Background knowledge based streams are devoted to realizing supplemental information and reinforce basal stream. To better fuse the features extracted from different streams, early-fusion and two after-fusion strategies are employed. According to the results obtained from both Chinese corpus and English corpus, it is demonstrated that the proposed background knowledge based multi-stream neural network performs well in classification tasks.

Keywords: classification algorithms; knowledge engineering; neural networks; machine learning

1. Introduction

In contemporary society, textual data are continuously increasing and have become one of the most commonly used information carriers [1]. As a kind of efficient information retrieval and data mining technology, text classification aims to get an association between the given document and one or more categories according to the features extracted. It has been widely used in many fields, such as sentiment analysis [2,3], stock analysis [4], news automatic grouping and so on.

Training corpus with enough data and accurate category labels always contribute to classification tasks. However, most of the existing corpus are imbalanced [5], and the imbalance mainly has two aspects. The first is the amount of imbalance between categories, which means that the amount of data in different categories is considerably different [6]. In classification tasks, data distribution usually will not be taken into consideration and traditional algorithms always tend their performance to the categories with more data and, at worst, the categories with less data could be ignored as outliers [7,8]. The second is the feature of imbalance within the category: for a certain category, it is difficult for the training data to include all sub-categories, especially for those categories with less data. As a result, it is easily to happen that for a certain category, the testing data and training data come from different sub-categories, and there is a huge difference of the feature words distribution between them, such as the sub-categories 'Telecommunication and networking' and 'Programming languages' of category 'Computer science'. Therefore, the incorrect judgements are easily given by classifiers.

In this information age, there is a great deal of textual data that existed in Internet while abundant information carried. Although most of them are unstructured data without category labels, this part of information can indirectly increase the amount of training data and expand the field of data coverage. Motivated by this cognition, a background knowledge based multi-stream neural network is proposed to solve the problems caused by imbalance data distribution in training corpus. The background knowledge is extracted from external corpus, which covers the data from almost all fields, and serves as priori knowledge for machine to complement the deficiency of the training data. Background knowledge mainly consist of two parts: keywords, which acquired from different categories and contain abundant category information; and co-occurred words, which co-occurred with keywords with high frequency. To better incorporate background knowledge into feature selection and extraction, we proposed a multi-stream neural network with different fusion strategies, which mainly composed by basal stream and background knowledge based streams. The basal stream takes the original word sequence as input, and retains the original sequence and semantic information. The background knowledge based streams take keywords and co-occurred words as inputs and do information supplement and reinforce for basal stream. Each stream is trained by RNN (Recurrent Neural Network) with GRU (Gated recurrent units) cell. Different fusion strategies are proposed to integrate the features extracted from different streams. Compared with basal model, the proposed method performs well in both Chinese and English corpus. The macro F1 score of Reuters 21578-R8 is up to 95.02%, which obtained 10.16% improvement, and the macro F1 score of Fudan University corpus is up to 85.03%, which obtained 8.75% improvement.

Our work makes the following contributions:

1. Background knowledge, which extracted from external unlabeled corpus, is incorporated into the classification network to indirectly enlarge the training corpus and make up for the imbalance data distribution of existing training corpus.
2. The background knowledge based streams mainly extract information according to the distribution of keywords and co-occurred words which acquired from extensive corpus, to avoid feature imbalance within a certain category.
3. A multi-stream neural network with different fusion strategies are proposed to fusion features extracted from background knowledge based streams into basal word-stream, and to realize the information supplement and information reinforce for classification task.

The remainder of this paper is organized as follows. Section 2 briefly reviews the background and summarizes the previous, related work of text classification. Section 3 introduces the proposed multi-stream neural network in which background knowledge is incorporated. Section 4 includes the experiments and discussion on both Chinese and English corpus. Finally, Section 5 concludes our work and outlines the direction of future work.

2. Related Works

In natural language processing, text classification is one of the most fundamental and important tasks with widely application. The commonly used machine learning methods, such as K-Nearest Neighbor [9], Naive Bayes [10], Decision Tree [11] and SVM [12], have achieved some achievements. However, the semantic and word order are often not taken into consideration in the referred representation, and the quality of feature selection and feature extraction directly effects the classification results.

As a branch of machine learning, deep learning has been widely researched and applied in NLP (Natural Language Processing) [13,14] after the presentation of word2vec model [15,16]. In deep learning neural networks (DNN), distributed representation is employed, and it is able to extract in-depth semantic features automatically with an efficient algorithm [17–19]. CNN and RNN, which are adept in extracting position invariant features and modeling units in sequence, respectively, are the two basal architectures widely used. In recent years, many variant architectures have been proposed, such as TextCNN [20], TextRNN [21], and the hybrid model RCNN (Recurrent Convolutional Neural Networks) [22]. The Dynamic Memory Network [23] which proposed by Ankit Kumar, casted most

NLP tasks into question answering problems over language input. The Transformation Networks [24] for target-oriented sentiment classification is proposed by Li, X., in which target information is integrated into word representation. The addition of attention mechanism can visually present the contribution of each word or sentence to the classification results [25,26], such as the hierarchical attention network proposed by Yang, Z. [27], in which the structure mirrored the hierarchical structure of documents.

The quality of training data plays a decisive role in the final classification results whether it is based on traditional machine learning methods or deep neural networks. In the task of text classification, to solve the problems caused by the incomplete coverage or imbalance of training data, there are mainly two ways. The first is processing the data self, such as data re-sampling [28]. Another is the incorporation of background knowledge as the expansion to make up for the insufficient of training data. The definition of background knowledge is different in various methods [29,30], and in most cases, it refers to the information that is essential to understand a situation or problem. In the classification task based on machine learning, background knowledge is often used as the bridge, and the classification algorithm is generally that if training data and testing data are similar to the background knowledge, then both belong to the same category. Li, C. [31] introduced external corpus to extract features based on character co-occurrence information, which served as background knowledge, to solve the problem caused by imbalanced corpus. Ren, F. et al. [32] considered the basic concept of categories and proposed a new text representation with the similarities between text and concepts, which served as general knowledge from Baidu Baike. Wikipedia is used as background knowledge by Yang, L. [33] to learn topics with respect to all target categories for short text classification. In deep learning neural networks, background knowledge is often existed in different forms in NLP tasks, such as knowledge graph, which contained a set of interconnected typed entities and their attributes [34,35]. Knowledge graph was proposed by google in 2012, which aimed to enable search engine to gain insight into the semantic information behind queries and improve the quality of answers returned. To discover latent knowledge-level connections among news, the deep knowledge-aware network is proposed by Wang, H. [36] to incorporates knowledge graph representation into news recommendation. For the task of sentiment analysis, an extension of LSTM is proposed by Ma, Y. [37], in which commonsense knowledge of sentiment-related concepts are incorporated.

The multi-stream neural network is a common architecture used in different information incorporation, and is often effected on graph and video analysis task. Simonyan, K. [38] proposed a two-stream ConvNet architecture which incorporated spatial and temporal networks, and demonstrated that good performance could be achieved in spite of limited training data. In subsequent studies, the architecture is further improved, the temporal segment network [39] is proposed to model long-term temporal structure, the importance of where a fusion layer location [40] in two-stream network is inquired to get better feature fusion, and the hidden two-stream networks [41] are proposed to capture the temporal relationships among video frames. All the above researches have proved the excellent performance with two-stream network in video action recognition. The network architecture is often universal, and those models that perform well in some tasks can often perform well in others. To increase the accuracy, the fusion of different representation models which trained by different classifiers [42] is often employed in NLP tasks. The processing procedure is similar to the multi-stream neural network, such as the deep fusion LSTMs [43] for text semantic matching and the enhanced sentence model for text categorization.

In this paper, the background knowledge extracted from external corpus is incorporated to solve the problems caused by data imbalance. To fusion the features extracted from background knowledge, multi-stream neural network with different fusion strategies is proposed to obtain better classification results.

3. Proposed Method

In this subsection, the overall methodology of background knowledge acquisition and multi-stream neural network are described in detail. The background knowledge is extracted from external corpus and is composed of keywords and co-occurred words. To better incorporate background knowledge into text classification task, a multi-stream neural network is proposed to extract in-depth features. Different fusion strategies: early-fusion and two kinds of later-fusion are employed in the feature fusion layer to better fuse the different features extracted from different single-streams.

3.1. Acquisition of Background Knowledge

For humans, it is often easy to immediately determine the category of a document by looking at the keywords or some specific words. The above decision is made based on the abundant background knowledge stored in the brain. Therefore, the assumption can be made that if background knowledge could be incorporated into the classification model, text classification task could be more accurate.

In this paper, the background knowledge is incorporated into text classification network to indirectly supplement training data and expand its coverage. The background knowledge is extracted from external corpus based on the following assumption: in the natural language, if two words w_1 and w_2 are often appeared together in the same unit window (such as paragraph, sentence, etc.), it can be assumed that they have some certain relationship, and the higher frequency of word co-occurrence, the closer relationship they have.

While word w_1 appears, the possibility of word w_2 having co-occurrence phenomenon is following:

$$R(w_2|w_1) = \frac{f(w_1, w_2)}{f(w_1)}, \tag{1}$$

where $f(w_1, w_2)$ represents the counts of word w_1 and w_2 appeared together, and $f(w_1)$ represents the counts of word w_1 appeared.

Therefore, it is reasonable to get the assumption that if there are some keywords of a certain category, the co-occurred words of these keywords with high frequency can also contribute to the text classification.

In this paper, background knowledge is composed of a set of keywords and co-occurred words. The referred keywords are extracted from labeled training corpus. The referred co-occurred words are those which appeared together within a certain word distance with keywords in a sentence. The external corpus are employed to search for the co-occurred words in a statistical unit (such as the sentence), and then the co-occurrence counts are counted. Only those co-occurred words with high frequency are remained in co-occurred words set. The flow of background knowledge acquisition is shown as Figure 1.

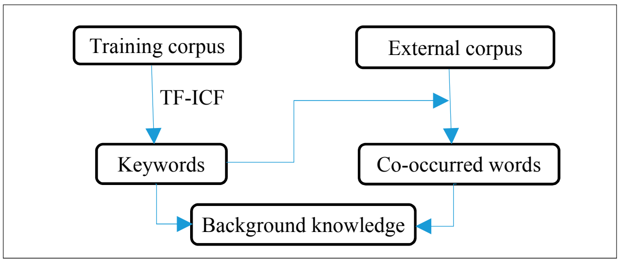

Figure 1. Extraction of background knowledge.

3.1.1. Keywords Acquisition

The training corpus have been classified, labeled with category, and pretreated to filter some useless words, such as stop words. To obtain those keywords that are important to a certain category in corpus, the TF-ICF method [44] is applied.

$$tficf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log\frac{|c|}{\left|\{j : w_i \in c_j\}\right|}, \tag{2}$$

In which $\frac{n_{i,j}}{\sum_k n_{k,j}}$ is the term frequency of word w_i occurred in category c_j, $|c|$ is the number of categories, and $\left|\{j : w_i \in c_j\}\right|$ is the number of categories which containing w_i.

For a certain category, a high weight in $tficf$ means a high term frequency in this category and a low category frequency in the whole corpus. Finally, those words with high $tficf$ weights composed the keywords set: $keywords = \{k_1, k_2, \ldots\}$. These keywords are the basis of acquisition of co-occurred words set.

3.1.2. Co-Occurred Words Acquisition

For a certain category in training corpus, it is hard to include all sub-categories, especially for those categories with less data. As a result, it is often happened that the testing data and training data come from different sub-categories, and the feature words distribution is of huge difference, thus affecting the feature extraction effect.

If two words often appear together, we can think these two words may well belong to the same category. The co-occurred words of keywords with high frequency carry abundant category information as well. The external corpus from various fields are incorporated to obtain the co-occurred words. This incorporation is devoted to covering more comprehensive data from various sub-categories and supplying sufficient information for training corpus. In order to get the word co-occurrence information, the co-occurred words and co-occurrence counts with keywords are obtained by scanning each sentence in external corpus. In this process, co-occurred words within a certain distance can also provide useful information, so the word distance is taken into consideration. For each keyword k_i, a co-occurrence matrix is obtained, which column index is co-occurred word and the value is co-occurred frequency $R(k_i|w_j)$.

$$R(k_i|w_j) = \frac{f(k_i,\ w_j)}{f(k_i)}, \tag{3}$$

In which $f(k_i|w_j)$ is the co-occurrence count of k_i and w_j in external corpus, $f(k_i)$ is the occurrence count of k_i.

Finally, the co-occurred words of each keyword with high frequency are obtained, and after duplication remove, the set of co-occurred words are obtained: $Co\_words = \{cow_1, cow_2 \ldots\}$.

3.2. Multi-Stream Neural Network

The multi-stream neural network performs well in features fusion, and is often applied in spatial and temporal networks in the task of action recognition of videos [45]. To better incorporate background knowledge based information, a multi-stream neural network is proposed, as shown in Figure 2, and different fusion strategies are used to extract comprehensive features.

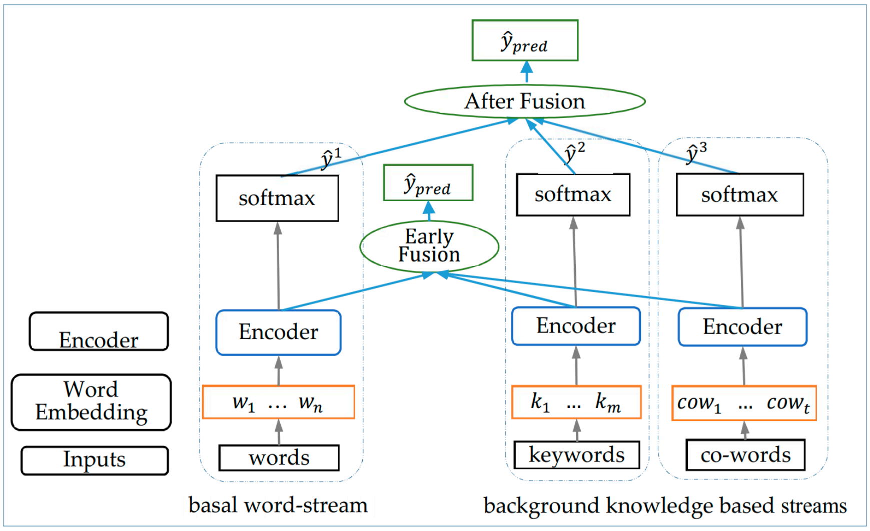

Figure 2. Multi-stream model.

The proposed multi-stream neural network consists of basal word-stream and background knowledge based aid streams, and mainly has five parts: input layer, word embedding layer, encoder layer, model training and fusion layer. Different feature sequences (detailed in Section 3.2.1) are feed into single-streams respectively. Each stream is trained on mini-batch with Adam optimizer independently. To further combine the features extracted from different streams, different fusion strategies (detailed in Section 3.3) are employed, and then the final prediction attaining.

3.2.1. Input Layer

The input of basal word-stream is original texts while the inputs of others are background knowledge based feature words. For each text s with n ordered words, the inputs of each stream in network divided into three parts: words, keywords and co-occurred word. Before the input layer, all words in texts, keywords set *keywords* and co-occurred words set *Co_words* are firstly transformed into real-valued word tokens by looking up in a pre-trained word tokens dictionary.

For each text $s = \{w_1, w_2, \dots w_n\}$, the different inputs in multi-stream are defined as follows:

1. Word-stream: words = $\{w_1, w_2, \dots w_n\}$, which consists of all words in texts.
2. Key-stream: for each word w_i in s, if w_i in keywords set *keywords*, append w_i in the input word sequences. Finally, the input of keyword-stream obtained: $key = \{k_1, k_2\ \dots\ k_m\}$.
3. Cow-stream: for each word w_i in s, if w_i in co-occurred words set *Co_words*, append w_i in the input word sequences. Finally, the input of co-occurred stream obtained: $cow = \{cow_1, cow_2 \dots cow_t\}$.

3.2.2. Word Embedding Layer

Word embedding, learned from massive unstructured text data, is widely adopted in building blocks for natural language processing. By representing each word as a fixed length vector, these embedding can group semantically similar words, while implicitly encoding rich linguistic regularities and patterns [46].

All words of corpus composed the word embedding matrix: L $\in R^{V*D}$, in which V is the vocabulary size and D is the dimension of word embedding. Each word in input is represented as $w_i \in R^{D*1}$.

3.2.3. Encoding Layer

The encoder layer with RNN is connected to extract the high-order textual and semantic features. GRU (Gated recurrent units) cells are employed in RNN and the parameters are not shared with each stream. GRU is a gating mechanism and performs well in in-depth feature extraction [47].

The final state of GRU is employed as feature sequence: $H = \{h_1, h_2 \ldots h_N\}$, in which N is the number of hidden layer in each stream, respectively.

Softmax function is connected the encoder to calculate the probability distribution, and the output is $P = \{p_1, p_2 \ldots p_n\}$, in which n is the number of categories, and p_i is the predicted probability that the text belongs to the corresponding category i, and the predicted tag $\hat{y} = argmax(P)$.

3.2.4. Model Training

Each single-stream is trained respectively, and the parameters are not shared with each other. End-to-end back propagation is employed in training and the loss function is defined as follows:

$$cross\ entropy = -\frac{1}{n}\sum yln\hat{y} + (1-y)\ln(1-\hat{y}), \tag{4}$$

The training of the model is to minimize the cross entropy in each stream respectively and AdamOptimizer is used during training.

3.3. Fusion Strategy

Different fusion strategies are employed to get the feature both considered the information of corpus self and background knowledge. After the optimal parameters are trained in each stream, early-fusion and two after fusion strategies: average pooling and soft voting, are employed to obtain the comprehensive text representation vector.

3.3.1. Early-Fusion

In early-fusion, the features output from encoder of each stream are concatenated together and then input to softmax to classify the data. The feature vector after early-fusion is:

$$H^{early} = \left[H^{word}, H^{key}, H^{cow}\right], \tag{5}$$

3.3.2. After-Fusion

Probability distribution output from softmax of each stream indicate the probabilities of the corresponding text belonging to different categories predicted by each stream. Therefore, a proper weight can be assigned to each of them, and then a comprehensive probability distribution generated and do predict. Based on above ideas, two after fusion strategies: average pooling and soft voting are proposed.

In average-pooling, a uniform weight is assigned to each stream, the final probability distribution is:

$$P^{Avg} = w_1 * P^{word} + w_2 * P^{key} + w_3 * P^{cow}, \tag{6}$$

In which P^{word}, P^{key}, P^{cow} are probability distribution of three streams respectively. w_i is the weight parameter and $\sum w_i = 1$. The final predicted tag: $\hat{y}_{pred} = argmax\left(P^{Avg}\right)$.

In the process of average pooling, a uniform weight w_i is distributed to each stream. It means that for a certain single-stream, the same value of weight is assigned on the outputs of classifier, which is the probability estimate for each category. However, in the actual case, for a certain single-stream, the features extracted for a certain category may be more discriminatory than others and the probability may be estimated more accurate. So, higher weight should be given to above category while lower weight should be given to those inaccurate estimations. Therefore, another strategy of after-fusion: soft-voting is proposed.

In soft-voting, a fully connected neural network is trained after softmax layer, as shown in Figure 2, to balance the weakness between each stream. The input is the concatenated probability distributions of all streams, and the final probability distribution is:

$$P^{soft} = w * \left[P^{word}, P^{key}, P^{cow}\right], \tag{7}$$

In which $w \in R^{3n*n}$, and n is the number of categories. The final predicted tag: $\hat{y}_{pred} = argmax\left(P^{soft}\right)$.

4. Experiments and Discussion

We conduct experiments in two Chinese copus: Fudan university corpus and the reduced version of Sougou classification corpus, and two English corpus: Reuters-21578 R8 and Reuters-21578 R52. In order to evaluate the performance of the proposed multi-stream model based on background knowledge, multiple comparison experiments are conducted to:

1. Investigate the performance of background knowledge based multi-stream neural network on text classification task;
2. Investigate the contribution of background knowledge incorporation under different fusion strategies, especially the contribution on those categories with less data;
3. Investigate the universality of the proposed multi-stream neural network in different corpus and different language environments.

4.1. Dataset and Preprocessing

Three unlabeled external corpus are used to extract background knowledge, and four training corpus are used to train classification model and do predict. All corpus have been preprocessed. Referred Chinese corpus are preprocessed by word segmentation by Stanford-Segmenter (http://nlp.stanford.edu/software/segmenter.html), part-of-speech tagging by Stanford-Postagger (http://nlp.stanford.edu/software/tagger.html), non-Chinese words removal, non-nouns removal, and stop words removal. The English corpus are preprocessed by stem and stop words removal.

While reading an article, human often can accurately judge the related area after reading some paragraphs instead of the whole. Especially in text classification task, it is able to do classification after obtaining the category information. For above common sense and to reduce the computing cost during the experiments, a fixed text length is set according to the length distribution of corpus. If the length of original text is higher than the fixed value, only the previous words are retained, otherwise, 0-padding is employed.

4.1.1. Training Corpus

There are two Chinese corpus, Fudan University text classification corpus and the reduced version of Sougou classification corpus, and two English corpus, Reuters-21578 R52 and Reuters-21578 R8 which are partial data of Reuters-21578, are used as training corpus to test the performance of the proposed method.

1. Fudan University text classification corpus (hereinafter referred as Fudan corpus, (http://www.nlpir.org/?action-viewnews-itemid-103), are provided by the natural language processing group of international database in computer information and technology department of Fudan University. It has two parts, training corpus including 9833 texts and testing corpus including 9804 texts. This corpus has 20 categories and the number of texts in each category is different, which is an imbalanced corpus. Figure 3 shows the distribution of all texts in Fudan corpus.

2. The reduced version of Sougou classification corpus (hereinafter referred as SougouC corpus, https://www.sogou.com/labs/resource/list_news.php) are balance corpus with 9 categories. There are 1990 texts in each category.
3. Reuters-21578, a collection of documents that appeared on Reuters newswire in 1987 (http://www.cs.umb.edu/~{}smimarog/textmining/datasets/). This English corpus contains 90 classes of news documents. Reuters-21578 R8 (hereinafter referred as Reuter R8), in which there are 8 classes selected from Reuters-21578. The reduced corpus contain 7674 documents, and the average number of texts in each class is 959, in which the max number of texts in a certain class is 3923 while the minimum is 51. Figure 4 shows the distribution of all texts in Reuter R8. Reuters-21578 R52 (hereinafter referred as Reuter R52), in which there are 52 classes selected from Reuters-21578 for the multiclass text classification experiments. The reduced corpus contains 9100 documents, and the average number of texts in each class is 175, in which the max number of texts in a certain class is 3923 while the minimum number is 3.

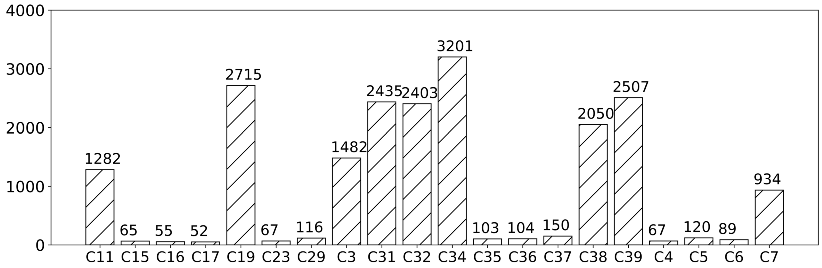

Figure 3. Document Numbers of classes in Fudan corpus.

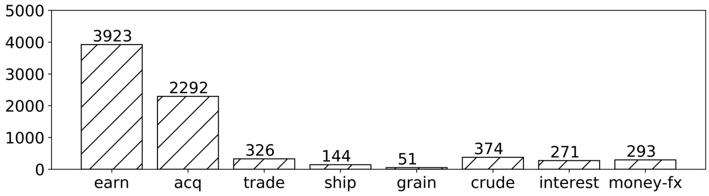

Figure 4. Document Numbers of classes in Reuter-21578 R8.

4.1.2. Background Corpus

There are two corpus are severed as external corpus to extract background knowledge:

1. Chinese corpus: People's daily news (http://paper.people.com.cn), which contain about 61 million sentences and average length is about 8 characters.
2. English corpus: Reuters Corpus, which contains about 806 thousand texts and average length is 109 words.

4.2. Experimental Setup

To investigate the contribution of background knowledge incorporation under different fusion strategies, some comparison experiments are conducted in both single-stream network and multi-stream network. Especially, the single word-stream is employed as baseline, which takes original word sequence as input and takes GRU as encoder. Some abbreviations used in this section are shown in Table 1.

Table 1. Explanatory note of Abbreviations in the experiments.

Abbreviations	Note
W	word-stream, input is *words*.
Key	key-stream, input is *key*.
Cow	cow-stream, input is *cow*.
KeyCow	keycow-stream, input is concatenate of *key* and *cow*.
W + *Key*	Fusion of word-stream and key-stream
W + *Cow*	Fusion of word-stream and cow-stream
W + *KeyCow*	Fusion of word-stream and keycow-stream
W + *Key* + *Cow*	Fusion of word-stream, key-stream and cow-stream

In our experiments, inputs of each stream are all uniformed to same length by 0-padding, the max length setting is shown as Table 2. The dimension of word level embedding is set to 128. To extract features, RNN with GRU cell is employed as encoder in each single stream, and the hidden layer is set to 256. We optimize each single stream with Adam algorithm in mini-batch, and the batch size is 256. The dropout was 0.5 and learning rate is 0.002.

Table 2. Max length setting of inputs.

	Reuter R8	Reuter R52	Fudan	Sougou
Max length of word_stream (n)	128	100	300	300
Max length of key_stream (m)	45	27	130	20
Max length of cow_stream (t)	70	56	222	92

4.3. Results and Discussion

Results of Reuter-21578 R8 are shown as Table 3. From the results of single-stream in the table, it can be seen that in the single-streams based on background knowledge: key-stream, cow-stream and keycow-stream, the macro precision, recall and F1 score have been significantly improved while the accuracy is not much different compared with the basal word-stream. In the multi-stream network, the overall macro value increased significantly while the accuracy also improved whether in early-fusion or after-fusion strategy. The highest three results under different evaluation indicators are bolded in the table. Under comprehensive consideration, the best one is obtained in three-stream network with average pooling ($P = 95.28$, $R = 94.75$, $F1 = 95.02$), which is superior than baseline (word-stream, $P = 84.71$, $R = 85.37$, $F1 = 85.04$). The improved classification results mean that the incorporation of background knowledge has enriched the text representation to a great extent. Therefore, we can infer that the background knowledge can not only make up for the insufficient information of training data, but also make up for the deep feature extraction of those data sparse categories.

The effectiveness of background knowledge incorporation in multi-stream neural network also verified in Reuters-20578 R52 and two Chinese corpus. The results of Reuters-20578 R52 are shown in Table 4, the best one acquired in three-stream network with after fusion of soft voting: $P = 81.96$, $R = 71.06$, $F1 = 76.12$ (while $P = 64.09$, $R = 63.95$, $F1 = 64.02$ in baseline). The results of SougouC corpus are shown in Table 5, the best one is acquired in three-stream network with early fusion: $P = 87.23$, $R = 87.30$, $F1 = 87.27$ (while $P = 83.72$, $R = 83.30$, $F1 = 83.36$ in baseline). The results of Fudan corpus are shown in Table 6, and the best one is acquired in three-stream network with early fusion: $P = 88.69$, $R = 81.65$, $F1 = 85.03$ (while $P = 76.70$, $R = 76.41$, $F1 = 76.55$ in baseline). In this way, we can infer that the background knowledge with multi-stream network contributes a lot for text classification task.

Table 3. Results of Reuters-21578 R8.

	Unit: %	**Macro**			**Acc**
		P	**R**	**F1**	
Single stream:	***W (Baseline)***	***84.71***	***85.37***	***85.04***	***96.12***
	Key	90.41	92.16	91.28	96.30
	Cow	91.35	89.84	90.59	95.39
	KeyCow	91.44	88.81	90.11	95.75
Early fusion:	W + Key	90.76	93.28	92.00	**97.44**
	W + Co	92.74	91.34	92.03	96.57
	W + KeyCow	**93.56**	91.03	92.28	96.67
	W +Key + Cow	93.09	92.39	92.74	96.85
After fusion:	Average pooling				
	W + Key	93.13	**93.85**	**93.49**	97.30
	W + Co	91.54	90.85	91.19	96.76
	W + KeyCow	90.24	89.63	89.94	96.67
	W + Key +Cow	**95.28**	**94.75**	**95.02**	**97.67**
	Soft voting				
	W + Key	91.75	92.52	92.14	97.08
	W + Co	91.00	89.89	90.44	96.07
	W + KeyCow	89.12	87.90	88.50	96.35
	W + Key +Cow	**95.15**	**94.30**	**94.72**	**97.67**

Table 4. Results of Reuters-21578 R52.

	Unit: %	**Macro**			**Acc**
		P	**R**	**F1**	
Single stream:	***W (Baseline)***	***64.09***	***63.95***	***64.02***	***93.30***
	Key	73.14	67.64	70.28	90.46
	Cow	68.97	63.89	66.33	91.04
	KeyCow	63.80	54.92	59.03	90.62
Early fusion:	W + Key	74.90	68.32	71.46	93.50
	W + Co	73.66	65.60	69.40	93.61
	W + KeyCow	73.31	65.84	69.38	93.69
	W +Key + Cow	**77.28**	68.40	72.57	94.00
After fusion:	Average pooling				
	W + Key	**78.10**	**69.87**	**73.76**	**94.04**
	W + Co	70.19	68.94	69.56	93.93
	W + KeyCow	71.74	65.25	68.34	**94.04**
	W + Key +Cow	76.36	**70.30**	**73.20**	**94.35**
	Soft voting				
	W + Key	76.89	67.82	72.07	93.03
	W + Co	72.27	66.64	69.34	92.91
	W + KeyCow	71.87	60.44	65.66	92.99
	W + Key + Cow	**81.96**	**71.06**	**76.12**	93.81

Table 5. Results of Sougou Corpus.

	Unit: %	**Macro**			**Acc**
		P	**R**	**F1**	
Single stream:	***W (Baseline)***	***83.72***	***83.30***	***83.36***	***83.78***
	Key	73.95	72.76	73.35	72.71
	Cow	81.86	81.76	81.81	81.71
	KeyCow	81.93	81.94	81.94	81.85
Early fusion:	W + Key	85.34	85.40	85.40	85.40
	W + Co	**87.12**	**87.15**	**87.13**	**87.13**
	W + KeyCow	**86.98**	**86.98**	**86.98**	**86.96**
	W +Key + Cow	**87.23**	**87.30**	**87.27**	**87.28**
After fusion:	Average pooling				
	W + Key	85.02	85.12	85.07	85.11
	W + Co	86.33	86.31	86.32	86.30
	W + KeyCow	86.70	86.63	86.66	86.60
	W + Key +Cow	86.62	86.65	86.64	86.64
	Soft voting				
	W + Key	84.79	84.81	84.80	84.80
	W + Co	86.05	86.07	86.06	86.04
	W + KeyCow	86.00	85.98	85.99	85.96
	W + Key +Cow	86.40	86.41	86.40	86.38

Table 6. Results of Fudan Corpus.

	Unit: %	**Macro**			**Acc**
		P	**R**	**F1**	
Single stream:	***W (Baseline)***	***76.70***	***76.41***	***76.55***	***95.43***
	Key	78.16	68.12	72.79	90.82
	Cow	83.27	78.15	80.63	95.22
	KeyCow	77.84	74.05	75.90	94.52
Early fusion:	W + Key	83.82	79.63	81.67	96.15
	W + Co	84.44	**81.73**	83.06	**96.79**
	W + KeyCow	83.02	80.80	81.90	96.53
	W +Key + Cow	**88.69**	**81.65**	**85.03**	**96.89**
After fusion:	Average pooling				
	W + Key	84.18	78.99	81.50	95.98
	W + Co	84.69	80.96	82.78	96.48
	W + KeyCow	83.90	78.66	81.19	96.13
	W + Key +Cow	85.26	**81.94**	**83.57**	**96.67**
	Soft voting				
	W + Key	83.60	76.59	79.94	95.64
	W + Co	**85.64**	79.40	82.40	96.30
	W + KeyCow	83.21	78.35	80.70	96.02
	W + Key +Cow	**88.66**	80.28	**84.26**	96.41

The comparison results between above optimal model and basal model show that the macro indicators have significantly improved while the overall accuracy growth is slight. The reason can be that in these three imbalance corpus, there are some categories with less data, and their classification results effect slightly on overall accuracy results because of the data distribution. However, during the model training, accuracy is often used as evaluation indicator and the categories with less data often be ignored. Therefore, macro evaluation can reflect the overall classification results more objective for imbalance corpus.

The background knowledge based single-streams: key-stream, cow-stream and keycow-stream, perform well in Reuter-21578 R8, as shown in Table 3. However, the results are not stable in other three corpus: the macro values have a slight rise or fall compared with word-stream, as shown in Tables 4–6. This is mainly caused by two aspects. On one hand, the contribution of background knowledge based feature extraction in single-stream has a great relationship of the quality of background knowledge and the training data distribution. On the other hand, the information extracted from background knowledge based stream are not as comprehensive as the word-stream. In background knowledge based stream, the inputs are keywords, co-occurred words or the both extracted from original text, while all words are taken as input in word-stream. The features extracted will be too sparse compared with basal model if the coverage of background knowledge is not so comprehensive, which means the case that not all keywords or those words carry abundant category information in original texts are concluded in background knowledge. Therefore, it is not certain that the classification results of background knowledge based stream can be better than basal word-stream model.

In the multi-stream network, the overall classification results are improved a lot after feature fusion. It means that, as the knowledge supplement, the background knowledge based features can make up for the problem caused by data imbalance in basal word-stream network. The imbalance referred two aspects, the first is the feature imbalance in a certain category. For example, because of the limitation of data coverage, testing data and training data may come from different sub-categories of a certain category and contain different feature words. The incorporation of background knowledge which contain almost all sub-categories information are served as bridge, create a connection between training data and testing data. The second is amount imbalance between categories. There are some categories with less data and result in less feature extracted for them. The incorporation of background knowledge can largely alleviate the limitation caused by imbalanced data distribution.

To investigate the performance of our model on certain corpus, some comparisons are conducted with other typical models. Such as GRUs [47], Multinomial NB, SVM, Bayes Network, and KNN, which are all analyzed on Reuters R8 and Reuters R52 in [48]. As shown in Table 7, the comparisons of optimum results of our proposed method against above methods are listed with the same datasets. The best results on each dataset are in bold, and they are obtained by our proposed method.

Table 7. Comparison of different methods with proposed method.

Method	**Accuracy (Unit: %)**	
	Reuters R8	**Reuters R52**
Proposed method	**97.67**	**94.35**
GRUs	96.12	93.30
Multinomial NB	96.20	86.64
SVM	95.20	90.14
Bayes Network	91.82	87.53
KNN	87.80	79.32

To investigate the contribution of background knowledge on different categories in a certain corpus, the optimal three group of experimental results obtained to compare with the basal word-stream model. The results of each category are shown as Figures 5–7 respectively. Horizontal axis refers to the categories, the bar charts refers to the classification results, and the line graph refers to the number of training texts in corresponding category.

According to the results of Reuters-21578 R8, as shown in Figure 5, the categories with relatively few training data, like 'ship' with 36 training texts and 'grain' with 10 training texts, improved a large amount in macro precision, recall, and F1 score, while the categories with relatively abundant training data, like 'earn' and 'acq', improved not very obviously. These results also verify the previous viewpoint about imbalance corpus: the incorporation of background knowledge can conspicuously make up for the insufficient information of some categories with less data, thus contribute to the overall results of classification model.

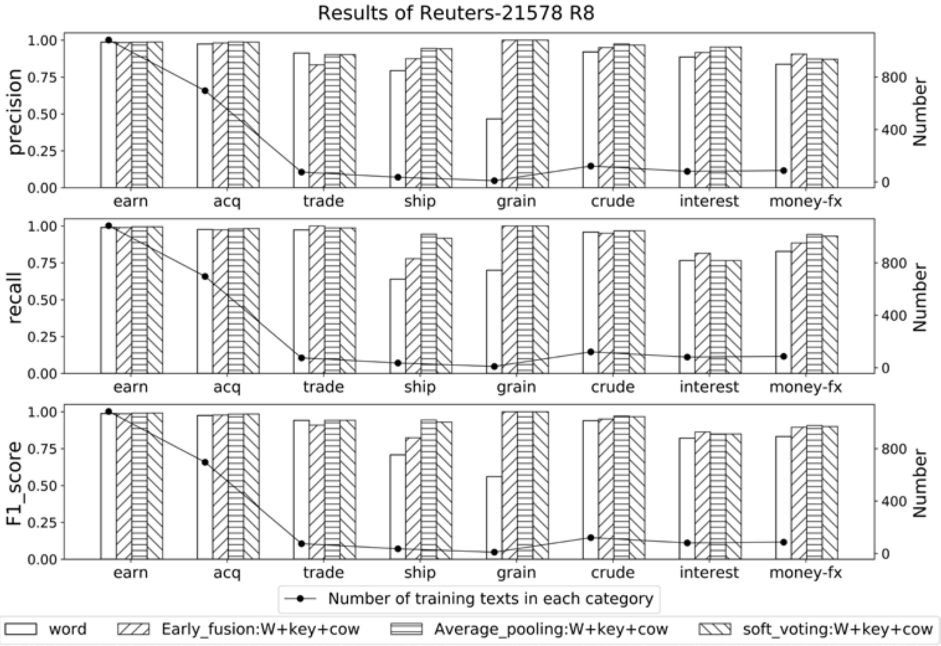

Figure 5. Results of Reuter-21578 R8.

The results of another imbalance Chinese corpus, Fudan Corpus, also verified the above assumption, as shown in Figure 6. The categories with relatively abundant training data above 600, like 'C11' and 'C19', the variation is slight. However, for other categories in which training texts are less than 100, the improvement is relatively distinct.

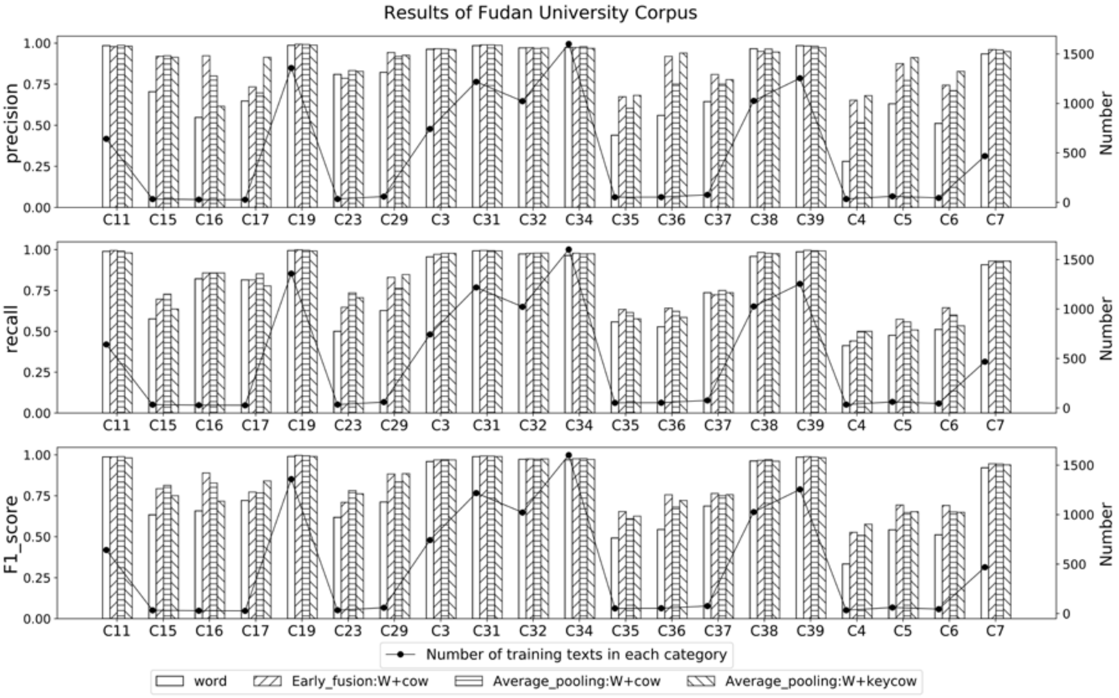

Figure 6. Results of Fudan University Corpus.

The results of the balance Sougou corpus, in which each category is composed of same number of texts, are shown as Figure 7. Compared with base model, the incorporation of background knowledge in multi-stream neural network seems contributions to each category uniformly. Although not particularly improved, the recognition accuracy of each category are all improved to some extent, which proved that the multi-stream model is not only applicable to imbalanced corpus, but also applicable to balanced corpus.

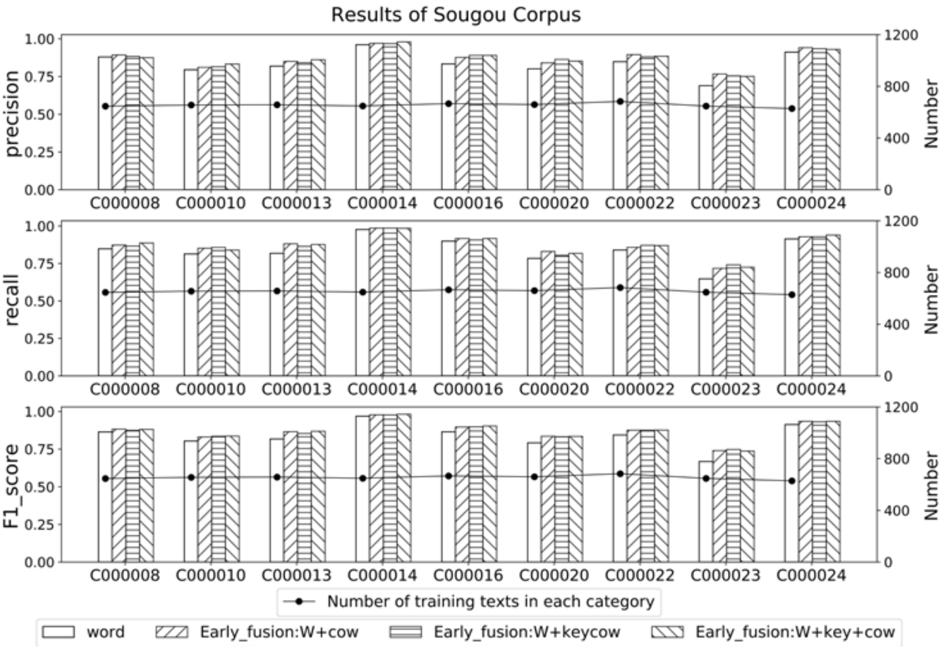

Figure 7. Results of Sougou corpus.

5. Conclusions

In this paper, we focused on the work of incorporating background knowledge into text classification task to make up for the limitation of imbalance training data. To better fuse the background knowledge based features into basal model, a multi-stream neural network with different fusion strategies was proposed.

The experimental results obtained from different corpus showed that, compared with traditional RNN based text classification model, the proposed method performed better under different evaluation indicators. The improvement of macro F1 score up to 10.16% in Reuters-21578-R8, 12.10% in Reuters-21578-R8, 8.75% in Fudan corpus, and 3.91% in Sougou Corpus. According to the results comparison of different categories in a certain category, the following conclusion can be drawn: as the knowledge supplement, the background knowledge based feature can make up for the information neglected or absented in basal text classification network, especially for imbalance corpus.

In future, the proposed work can be extended by extracting more beneficial background knowledge from more comprehensive external corpus provided by this big data era. To extract more comprehensive and in-depth feature information, more complex and effective model can serve as encoder, and different feature fusion strategy can also be further researched.

Author Contributions: F.R.: Conceptualization, Supervision and Funding Acquisition; J.D.: Methodology, Software and Writing-Review & Editing.

Funding: This research was funded by JSPS KAKENHI Grant Number 15H01712.

Acknowledgments: This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiu, L.; Lei, Q.; Zhang, Z. Advanced Sentiment Classification of Tibetan Microblogs on Smart Campuses Based on Multi-Feature Fusion. *IEEE Access* **2018**, *6*, 17896–17904. [CrossRef]
2. Valdivia, A.; Luzón, M.V.; Herrera, F. Sentiment analysis in tripadvisor. *IEEE Intell. Syst.* **2017**, *32*, 72–77. [CrossRef]
3. Bouazizi, M.; Ohtsuki, T. A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access* **2017**, *5*, 20617–20639. [CrossRef]
4. Du, X.; Deng, L.; Qian, K. Current Market Top Business Scopes Trend—A Concurrent Text and Time Series Active Learning Study of NASDAQ and NYSE Stocks from 2012 to 2017. *Appl. Sci.* **2018**, *8*, 751. [CrossRef]
5. Castellanos, F.J.; Valero-Mas, J.J.; Calvo-Zaragoza, J.; Rico-Juan, J.R. Oversampling imbalanced data in the string space. *Pattern Recognit. Lett.* **2018**, *103*, 32–38. [CrossRef]
6. Li, Y.; Guo, H.; Zhang, Q.; Gu, M.; Yang, J. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowl. Syst.* **2018**, *160*, 1–15. [CrossRef]
7. Tan, S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Syst. Appl.* **2005**, *28*, 667–671. [CrossRef]
8. Zhou, P.; Hu, X.; Li, P.; Wu, X. Online feature selection for high-dimensional class-imbalanced data. *Knowl. Syst.* **2017**, *136*, 187–199. [CrossRef]
9. Bijalwan, V.; Kumar, V.; Kumari, P.; Pascual, J. KNN based machine learning approach for text and document mining. *Int. J. Database Theory Appl.* **2014**, *7*, 61–70. [CrossRef]
10. Jiang, L.; Li, C.; Wang, S.; Zhang, L. Deep feature weighting for naive Bayes and its application to text classification. *Eng. Appl. Artif. Intell.* **2016**, *52*, 26–39. [CrossRef]
11. Zhang, Y.; Wang, S.; Phillips, P.; Ji, G. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowl. Based Syst.* **2014**, *64*, 22–31. [CrossRef]
12. HaddoudEmail, M.; Mokhtari, A.; Lecroq, T.; Abdeddaïm, S. Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowl. Inf. Syst.* **2016**, *49*, 909–931.
13. Deng, L.; Yu, D. Deep learning: Methods and applications. *Found. Trends Signal Process.* **2014**, *7*, 197–387. [CrossRef]
14. Joulin, A.; Grave, E.; Bojanowski, P. Bag of Tricks for Efficient Text Classification. Available online: https://arxiv.org/abs/1607.01759 (accessed on 6 July 2016).
15. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. Available online: https://arxiv.org/abs/1301.3781 (accessed on 7 September 2016).
16. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process Syst.* **2013**, *2*, 3111–3119.
17. Zheng, J.; Cai, F.; Shao, T.; Chen, H. Self-Interaction Attention Mechanism-Based Text Representation for Document Classification. *Appl. Sci.* **2018**, *8*, 613. [CrossRef]
18. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-attentional convolutional neural network for scene text detection. *IEEE Trans. Image Process* **2016**, *25*, 2529–2541. [CrossRef] [PubMed]
19. Zhao, R.; Mao, K. Topic-aware deep compositional models for sentence classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *25*, 248–260. [CrossRef]
20. Kim, Y. Convolutional Neural Networks for Sentence Classification. Available online: https://arxiv.org/abs/1408.5882 (accessed on 25 August 2014).
21. Liu, P.; Qiu, X.; Huang, X. Recurrent Neural Network for Text Classification with Multi-task Learning. Available online: https://arxiv.org/abs/1605.05101 (accessed on 17 May 2016).
22. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. *AAAI Conf. Artif. Intell.* **2015**, *333*, 2267–2273.
23. Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; Socher, R. Ask me anything: Dynamic memory networks for natural language processing. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1378–1387.
24. Li, X.; Bing, L.; Lam, W.; Shi, B. Transformation Networks for Target-Oriented Sentiment Classification. Available online: https://arxiv.org/abs/1805.01086 (accessed on 3 May 2018).

25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Bachrach, Y.; Zukov-Gregoric, A.; Coope, S.; Tovell, E.; Maksak, B.; Rodriguez, J.; McMurtie, C. An Attention Mechanism for Answer Selection Using a Combined Global and Local View. Available online: https://arxiv.org/abs/1707.01378 (accessed on 5 July 2017).
27. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
28. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explor. Newslett.* **2004**, *6*, 1–6. [CrossRef]
29. Annane, A.; Bellahsene, Z.; Azouaou, F.; Jonquet, C. Building an effective and efficient background knowledge resource to enhance ontology matching. *J. Web Semant.* **2018**, *51*, 51–68. [CrossRef]
30. Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 28 June–1 July 2001.
31. Li, C. Text Classification Based on Background Knowledge. Ph.D. Dissertation, Department Advance Technology Science Information, Tokushima University, Tokushima, Japan, 2016.
32. Ren, F.; Li, C. Hybrid Chinese text classification approach using general knowledge from Baidu Baike. *IEEE J. Trans. Electr. Electron. Eng.* **2016**, *11*, 488–498. [CrossRef]
33. Yang, L.; Li, C.; Ding, Q.; Li, L. Combining lexical and semantic features for short text classification. *Procedia Comput. Sci.* **2013**, *22*, 78–86. [CrossRef]
34. Chang, L.; Zhu, M.; Gu, T.; Bin, C.; Qian, J.; Zhang, J. Knowledge Graph Embedding by Dynamic Translation. *IEEE Access* **2017**, *5*, 20898–20907. [CrossRef]
35. Tan, Z.; Zhao, X.; Fang, Y.; Xiao, W. GTrans: Generic knowledge graph embedding via multi-state entities and dynamic relation spaces. *IEEE Access* **2018**, *6*, 8232–8244. [CrossRef]
36. Wang, H.; Zhang, F.; Xie, X.; Guo, M. DKN: Deep Knowledge-Aware Network for News Recommendation. In Proceedings of the 27th International Conference on World Wide Web, Lyon, France, 23–27 April 2018.
37. Ma, Y.; Peng, H.; Cambria, E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018.
38. Simonyan, K.; Zisserman, A. Two-stream Convolutional Networks for Action Recognition in Videos. Available online: https://arxiv.org/abs/1406.2199 (accessed on 12 November 2014).
39. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
40. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-stream Network Fusion for Video Action Recognition. Available online: https://arxiv.org/abs/1604.06573 (accessed on 26 September 2016).
41. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A.G. Hidden Two-stream Convolutional Networks for Action Recognition. Available online: https://arxiv.org/abs/1704.00389 (accessed on 22 October 2017).
42. Amensisa, A.D.; Patil, S.; Agrawal, P. A survey on text document categorization using enhanced sentence vector space model and bi-gram text representation model based on novel fusion techniques. In Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 19–20 January 2018.
43. Liu, P.; Qiu, X.; Chen, J.; Huang, X. Deep fusion LSTMs for text semantic matching. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016.
44. Ren, F.; Sohrab, M.G. Class-indexing-based term weighting for automatic text classification. *Inf. Sci.* **2013**, *236*, 109–125. [CrossRef]
45. Luo, X.; Li, H.; Cao, D.; Yu, Y.; Yang, X.; Huang, T. Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks. *Autom. Constr.* **2018**, *94*, 360–370. [CrossRef]

46. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014.
47. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Available online: https://arxiv.org/abs/1412.3555 (accessed on 11 December 2014).
48. Shinde, S.B.; Muzammil, M.; Thepade, S. Microblogging Comments Classification. *Int. J. Comput. Sci.* **2017**, *167*, 19–22.