8-12-2019

A Review of Text Corpus-Based Tourism Big Data Mining

Qin Lin

Shaobo Li

Sen Zhang

Jie Hu

Jianjun Hu
jianjunh@cec.sc.edu

Publication Info

Published in *Applied Sciences*, Volume 9, Issue 16, 2019, pages 1-27.

applied sciences

MDPI

Review

A Review of Text Corpus-Based Tourism Big Data Mining

Qin Li [1,2], Shaobo Li [3,4,*], Sen Zhang [1,2], Jie Hu [5] and Jianjun Hu [3,6,*]

1 Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
4 Guizhou Provincial Key Laboratory of Public Big Data (Guizhou University), Guiyang, Guizhou 550025, China
5 College of Big Data Statistics, GuiZhou University of Finance and Economics, Guiyang, Guizhou 550025, China
6 Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
* Correspondence: lishaobo@gzu.edu.cn (S.L.); jianjunh@cse.sc.edu (J.H.); Tel.: +01-803-777-7304 (J.H.)

Received: 25 June 2019; Accepted: 6 August 2019; Published: 12 August 2019

Abstract: With the massive growth of the Internet, text data has become one of the main formats of tourism big data. As an effective expression means of tourists' opinions, text mining of such data has big potential to inspire innovations for tourism practitioners. In the past decade, a variety of text mining techniques have been proposed and applied to tourism analysis to develop tourism value analysis models, build tourism recommendation systems, create tourist profiles, and make policies for supervising tourism markets. The successes of these techniques have been further boosted by the progress of natural language processing (NLP), machine learning, and deep learning. With the understanding of the complexity due to this diverse set of techniques and tourism text data sources, this work attempts to provide a detailed and up-to-date review of text mining techniques that have been, or have the potential to be, applied to modern tourism big data analysis. We summarize and discuss different text representation strategies, text-based NLP techniques for topic extraction, text classification, sentiment analysis, and text clustering in the context of tourism text mining, and their applications in tourist profiling, destination image analysis, market demand, etc. Our work also provides guidelines for constructing new tourism big data applications and outlines promising research areas in this field for incoming years.

Keywords: tourism big data; text mining; NLP; deep learning

1. Introduction

Text is an effective and widely existing form of opinion expression and evaluation by users, as shown by the large number of online review comments over tourism sites, hotels, and services. As a direct expression of users' needs and emotions, text-based tourism data mining has the potential to transform the tourism industry. Indeed, tourists' decision-making is dramatically influenced by the travel experience of other individuals [1] in forms of tourism reviews or blogs, etc. These texts can give valuable insight for potential tourists, and assist them in optimizing destination choices and exploring travel routes, or for tourism practitioners to improve their services. Tourism platforms such as TripAdvisor and Ctrip now routinely provide an explosive amount of text data, which makes it possible to use deep learning [2], NLP [3], and other machine learning [4,5] and data mining techniques for tourism analysis. Studies [6,7] have shown that the competitiveness of the tourism industry dramatically relies on tourists' sentiment and opinions about the events occurring during

the travel. In order to utilize this user-generated content properly and further to meet the needs of tourists and promote the tourism industry, we need to analyze and exploit tourists' needs and opinions, and then identify the problems of tourism services or destinations, which has become a new path for tourism development. Besides, as tourism needs become increasingly personalized, visitors begin to pursue self-likeness, self-worth, and diversified travel experiences, and they are no longer willing to endure delays or waits. How to recognize and respond to visitors' behaviors and needs quickly and identify potential customers have become essential factors for the success of tourism stakeholders. By exploiting the subjective information contained in tourism text data, we can assist tourism stakeholders to provide better services for tourists.

A large number of text mining techniques have been proposed and applied to tourism text data analysis for creating tourist profiles [8–15] and making effective market supervision [16–25]. These approaches exploit a variety of text representation strategies [26–32] and use different NLP techniques for topic extraction [33], text classification [34], sentiment analysis [35], and text clustering [36]. Moreover, while aiming to make computers understand human language, NLP has become the essential tool for text data analysis and is undergoing fast-pace growing based on the applications of deep learning in word embedding, syntax analysis, machine translation, and text understanding. Machine learning-based NLP techniques have been widely used in tourism text analysis, with superior results [19,25]. In addition, due to its high capability for extracting selective and invariant features from texts, and its independency of prior knowledge and linguistic resources, deep learning has been reported to achieve higher performance than other approaches on many NLP tasks [32]. A range of deep learning algorithms such as deep neural networks (DNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs), along with special DNN techniques such as memory strategies and attention mechanisms [37], have been successfully applied to NLP tasks.

Due to the significance of tourism text data mining and the emergence of a large number of recent applied text mining techniques with diverse design strategies and methodologies, this paper aims to give a systematic review of these tourism text data mining techniques and applications. This paper divides NLP techniques into two types, in which NLP based on language scenes requires knowledge rather than text, such as knowledge of the domains and common sense, while text corpus-based big data analysis requires only an amount of text data for macro analysis. Figure 1 shows the basic process of applying text corpus-based NLP techniques to tourism analysis.

The structure of the paper is as follows: Section 2 introduces the recent text representation strategies and summarizes the basic applications of text corpus-based NLP techniques and their applications to tourism text data analysis in recent years. Section 3 provides a global analysis of the special tourism application techniques from the perspective of tourists and market. Section 4 also provides guidelines to be followed in the design of tourism value analysis or tourism recommendation systems, and outlines the most promising areas in the future of text corpus-based tourism big data mining. Section 5 summarizes our work from the exploration of the existing techniques.

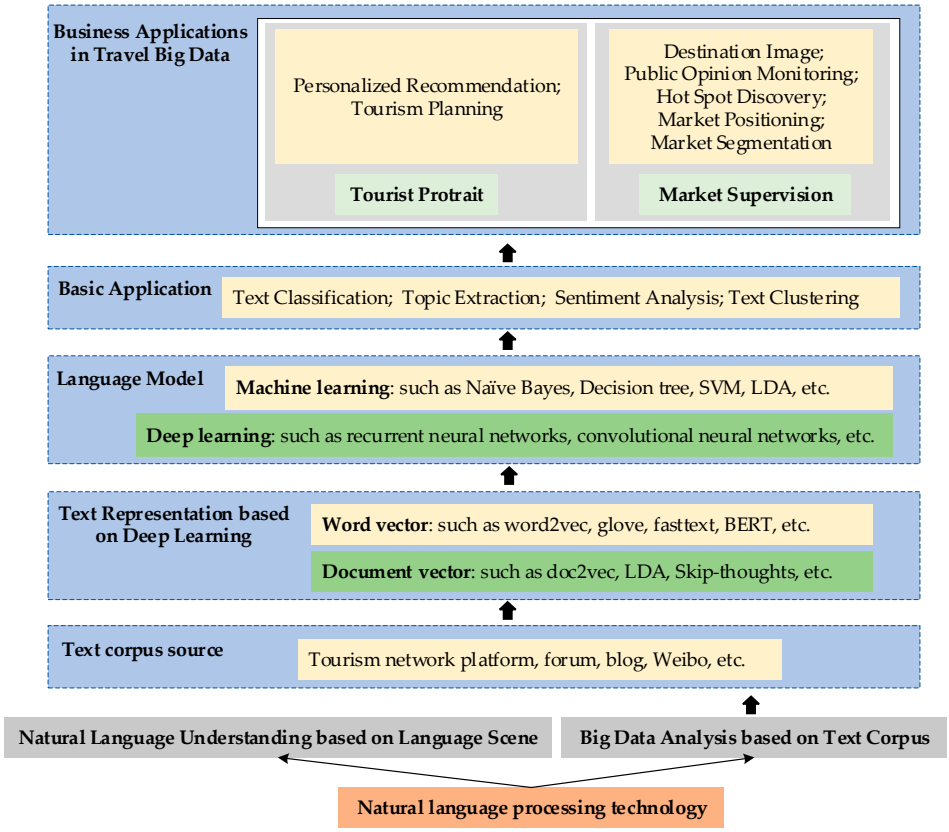

Figure 1. Text corpus-based tourism big data mining.

2. Review Protocol Used in This Review

We have followed a systematic review protocol [38,39] to write this paper, to reduce bias in literature included and to improve the comprehensiveness in our review activities. It is suggested that following an explicit review protocol can help define the source selection and search processes, quality criteria, and information synthesis.

We have used the following digital libraries to search for primary related studies:

- Google Scholar;
- Science Direct;
- ACM Digital Library;
- Citeseer Library;
- Springer Link;
- IEEE explore; and
- Web of Science.

The following queries have been created to conduct the keywords based search:

Text mining; topic extraction, text classification, sentiment classification; text clustering; tourist profile; destination image; market supervision/demand; transfer learning; meta-learning; sentiment aspect; aspect-base sentiment analysis; target-dependent sentiment analysis; NLP; deep learning; machine learning; text representation; word embedding; document vector; supervised learning, semi-supervised learning, unsupervised learning; short text; keyword; attention mechanism; tourism; tourism recommendation; tourism hotspot; crisis events or emergency analysis; tourism dataset; domain adaptability; tourism review. Logical OR, AND, and NOT descriptors have been used also during the literature search.

We have also listed our inclusion criteria in Table 1.

Table 1. Inclusion criteria.

Include	Exclude
Studies focused on text mining techniques based on macro corpus analysis, including topic extraction, text classification, sentiment classification, and text clustering.	NLP based on language scenes which requires knowledge rather than text, such as knowledge of domains and common sense.
Tourism related studies in which text is the main research object, and other data structures can be the auxiliary means. Text corpus-based tourism big data mining related to tourist profiling and market supervision.	Tourism related studies in which other data structures (pictures, videos, etc.) are the main research object, while text can be the auxiliary means. Text corpus-based tourism big data mining related to question answering, or others.
English texts.	All other languages.
Studies published between 2014 and 2019.	Studies published before 2014.
Peer reviewed academic journals or books, conference proceedings.	Dissertations, non-peer reviewed sources.
In person context.	Online context.

Study selection procedure: We first checked the paper titles and then reviewed the abstracts, keywords, results, and conclusions to obtain the first list of studies. Then, we screened the remaining list by the inclusion criteria in Table 1. We then double checked the reference list to identify additional studies that were relevant to our review topic. Finally, we evaluated the quality of remaining list of related studies using the checklist as suggested in [40].

3. Text Corpus-Based Tourism Big Data Mining Techniques

3.1. Text Representations

Recently, text data mining techniques have been mainly based on machine learning and deep learning, which play a decisive role for the improvement of NLP. In order to transform NLP problems into the problem of machine learning, symbols such as text need to be digitized firstly; that is, text representation must be obtained. In 2003, Bengio et al. [41] proposed a word vector model based on N-gram statistical language, which pioneered the neural network as a language model. Since then, a lot of word representations in low-dimensional space have been proposed, which are pre-trained on a large set of unlabeled text corpus. In contrast to high-dimensional space, these word representations or word embeddings can be compared in sematic distance and can be easily applied to other models. Traditionally, word embedding methods such as Word2vec [26,27] and Glove [28] were proposed to create a global word representation which considers the word in all sentences. Currently, more and more works have started to notice the different semantics of a word in different contexts. For example, contextual word vectors (CoVe) [29] capture contextual information by an encoder in an attentional seq-to-seq machine translation model; and embeddings from language model (ELMO) [30] extracts context-sensitive features from a bidirectional language model (biLM). Subsequently, with the proposal of Transformer network, generative pre-training Transformer (OpenAI-GPT) [31] and large-scale pre-training language model based on bidirectional Transformer (BERT) [32] pre-trained language models from Transformers for extracting contextual word embeddings and showed better performance than ever on many tasks [32]. Table 2 provides a directory of existing main pre-trained models which can be used for generating embeddings as an input of the next model training. In addition, the supervised methods for generating word vectors [42–44] can also promote word representations, such as the methods for improving the topics extraction or sentiment classification and so on.

Table 2. Pre-trained models for word embeddings.

Time	Pre-Trained Model	URL	Accessed Data
2013	Word2vec	https://radimrehurek.com/gensim/models/word2vec.html	24 July 2019
2014	Glove	https://nlp.stanford.edu/projects/glove/	24 July 2019
2016	FastText [45]	https://fasttext.cc/	24 July 2019
2016	WordRank [46]	https://radimrehurek.com/gensim/models/wrappers/wordrank.html	24 July 2019
2017	CoVe	https://github.com/salesforce/cove	24 July 2019
2018	ULMFiT [47]	http://nlp.fast.ai/ulmfit	24 July 2019
2018	ELMO	https://github.com/allenai/allennlp/blob/master/tutorials/how_to/elmo.md	24 July 2019
2018	OpenAI-GPT	https://openai.com/blog/language-unsupervised/	24 July 2019
2018	BERT	https://github.com/google-research/bert	24 July 2019

A word is a basic unit of a sentence, paragraph, or document. To better represent a text, converting the word vectors into a short text or long text representation is an efficient operation. Early typical text representations have bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF) models, but these document vector models are usually too simple and lack context information and word-to-word associations. They often perform poorly on some complex tasks. Latent Dirichlet Allocation (LDA) computes the distribution of topics for documents, which is commonly used for representation of documents. Doc2vec [48] was proposed based on Word2vec, which considered the context information and semantic information, while it is only slightly better than the simple average word vectors of the document in the classification task. Aiming at it, Arora et al. [49] encoded sentences by a linear weighted combination of word vectors for improvement. For a long time, researchers have focused on unsupervised sentence learning, such as Skip-thoughts [50] and its improved Quick-thought [51] afterwards. When the supervised task is suitable for sentence embedding training, such as the natural language inference (NLI) task, the quality of sentence vectors learned by supervised methods can achieve higher performance [52], especially under the framework of multi-task learning [53].

The vector representation for the granularity of words, sentences, documents, etc., is the basis for related machine learning, and the pre-trained models of these vectors provide the premise for the input of other models. In contrast to randomly initialized vectors, the vectors provided by these pre-training models can reduce the data demand and save training time for deep learning, and the captured features of vectors can also significantly improve the performance of the model.

3.2. Text Corpus-Based NLP Techniques in Tourism Data Mining

This section analyzes four basic technical applications of text corpus-based NLP techniques: Topic extraction, text classification, sentiment analysis, and text clustering. These basic technical applications are the basis of related tourism business applications. In order to illustrate them and their applications in tourism, we firstly analyze recent tourism business works based on these techniques and then outline the recent techniques of the basic technical applications.

3.2.1. Topic Extraction

Topic extraction is a technique for extracting topics or aspects from large-scale text data, which can give the decision-maker some insights and help them identify associated sentiments. In tourism, topic extraction can be used to capture tourism public concerns and track hot events, by which tourists can track travel trends and important events and tourism practitioners can find business opportunities or travel crisis in real time, accordingly, to take measures to guide public opinion or promote related fields. Hot topics are often reviewed and reprinted more frequently. Based on this idea, the study [54] filtered out information with high evaluation and reload times, and introduced time distribution

calculation methods into LDA to detect hot topics. In the study [55], Structure Topic Model (STM) [56] was adopted to detect negative reviews and how public concerns vary over different hotel grades. In addition, topic extraction can also be used to find tourism characteristics of attractions or build tourism product and user profiles. For example, the study [4] defined the global topic as "scenic spot" to filter noise semantics from tourism corpus, to explore the local topics of "scenic spot" by LDA, and then display the distribution of the attraction types or the local topics in the form of attraction maps. In this study [5], a season topic model based on LDA (STLDA) was proposed to explore the topic characteristics of attractions with seasonal features, which was of great significance for mining related topics with seasons and for personalized recommendations. In this study [57], a topic model was utilized to detect explicit interests by interactions between users for the following users' implicit interest profile building.

Topic extraction has been studied a lot recently and is an important issue for tourism analysis. There are many text mining techniques for topic extraction. Traditionally, keywords are usually used for expressing the topics of documents, and keyword extraction has played an important role in topic extraction for many years. According to existing studies, most of keyword extraction techniques are based on unsupervised learning [58], such as algorithms based on statistical word features (e.g., TF-IDF, Kullback–Leibler divergence, chi-square test, etc.), algorithms based on topic models (e.g., LDA, etc.), and algorithms based on network graphs (e.g., TextRank, Rapid Automatic Keyword Extraction (RAKE) [59], TopicRank [60]). With the appearance of word-distributed representations, some works have also tried to make use of Word2vec to improve keyword extraction, while only considering the similarity calculation of words [61].

Keywords can express topics, but currently, their relevance with document topics mainly depends on the improvement of existing probability topic models [33], such as the LDA model. As a probability generation model, LDA can be easily extended to other probability models. For example, the study [62] considered that different topics in the same corpus are often related, and used this idea as a research perspective to promote topic extraction of the LDA model; some probability topic models for specific tasks such as topic sentiments, time changes, document authors, etc. are also expected to extract the topics related to target task [63]. To extract more relevant topics, some experts have also considered using word-distributed representation to improve the semantics and syntactics of text information. For example, the study [64] introduced the similarity of word embedding into LDA to calculate the relevance of the acquired words; the study [65] improved the way of context acquisition by introducing LDA into Word2vec; the study [66] introduced Word2vec to calculate the distance between the topic vector and the document vector to correct the topics of LDA mining. Moreover, the combination of deep learning and LDA has become another topic extraction method. For example, in this study [67], a novel neural topic model was proposed to acquire the N-gram topic by the deep neural network and then used LDA to obtain the topic representation of the document; the study [68] integrated LDA into language model LSTM for joint training. In addition to the LDA and its extensions, there are also methods for generating subject-related vectors by the neural attention model, such as Attention-based Aspect Extraction (ABAE) [69].

In summary, researches about topic extraction are mostly dependent on the topic probability model such as LDA, and improved aiming at different text structures. Topic extraction in short text often suffers from data sparseness due to the insufficient word co-occurrence and lack of context information. Using long text to assist short text by importing external-related information from Wikipedia and WordNet, etc. [70], or aggregating short text based on the posterior probability of words in original documents [71] can help with the short text task to some extent. Another approach of improving short text task is to enhance the interpretability or the semantical coherence of the topic model, such as informative words rewarding [72]. Besides, as above analysis, LDA ignores the order structure of texts and the meaning of words, so it is one of the research directions that scholars have focused on by exploring the features of words and sentences, etc., to enhance the ability of topic extraction, and will be show great potential in the future. Except for the issue of short text, in business

application, users or practitioners may concern different aspects of a topic or aspect related information; the context information of the topic aspect is often used to explore topic-sensitive content [73]. In order to understand a topic more granularly, the structural relation among topics is also a problem, which is widely concerned such as the exploring the hierarchical structure among topics or the global and local relation [74].

3.2.2. Text Classification

Text classification is a process in which the computer automatically classifies the input texts according to their content, and it includes spam filtering, information retrieval, topic classification, sentiment classification, and so on; sentiment classification of which will be discussed in the Section 3.2.3. In tourism, text classification is generally about topic classification, and the topics of texts in travel mostly involve various aspects of the travel process, such as transportation, accommodation, food, entertainment, and so on. In addition, tourism texts such as tourism comment reviews are often short but contain a lot of information. They usually contain multiple topics but cannot be attributed to one certain aspect or target. Therefore, text classification in the current tourism field is usually carried out by topic extraction, to extract all aspects of the text for targeted analysis [22,75].

Traditionally, text classifications are mainly based on machine learning such as Naive Bayes [76], maximum entropy [77], Support Vector Machine [78], K-nearest neighbor algorithm [78], etc. They usually use keywords or topics to reflect the feature of documents and realize text classification automatically [79], which still plays an important role currently [80]. At present, the mainstream feature extraction techniques characterized by keywords include TF-IDF and information divergence, and more advanced deep learning approaches [81]. Besides, there are some classic feature representation models for words such as Vector Space Model, N-grams [82], etc., but these models have some disadvantages, such as semantic information being missing. With word-distributed representations, text classification no longer relies on the keywords only and begins to pay more attention to the semantics of words themselves. Following this trend, there have emerged a lot of text classification algorithms based on the improvement of word embedding features [45,65,83].

Benefitting from deep learning, various researches on text classification techniques based on deep neural networks have also made significant progress. For example, the representative and innovative algorithms of CNN in text classification include the CNN proposed by Kim [34], character-level convolutional network (ConvNets) proposed by Zhang [84], the CNN-based classification algorithm applied to patent classification [85], the RNN-based text classification algorithms which includes bidirectional RNNs [86], hierarchical attention mechanism (HAN) [35], and the recurrent convolutional network (RCNN) [87]. In addition, the connection between CNNs and RNNs [88], the introduction of multi-task learning frameworks [89], and the increased depth in deep learning models [90] will also improve the performance of text classification to a certain extent.

At present, text classification models are mostly based on supervised learning. Supervised text classification often relies on the integrity of the domain corpus, as well as its annotations, so researchers are more concerned with the improvement of theory but lack of practical applications. In order to solve the issue of few in-domain labeled data, many following studies have been made; Table 3 shows four different strategies: Co-training, training samples extension, meta-learning, and transfer learning. Among them, co-training and training samples extension need auxiliary training with unlabeled data or external knowledge, but as the training samples increase, the noise in the auto-tagging instance will continue to accumulate. Meta-learning or transfer learning attempt to learn general representation or meta knowledge among tasks, but still have to make further improvement, such as in the issue of negative transfer.

Table 3. Strategies of few in-domain labeled data in 2018.

Author-Study	Contribution	Basic Language Model/Classifier
[91]	It proposes a novel co-training algorithm which uses an ensemble of classifiers created in multiple training iterations, with labeled data and unlabeled data trained jointly and with no added computational complexity.	Naïve Bayes; Support Vector Machine
[92]	It uses the knowledge of Wikipedia to extend the training samples, which is realized by network graph construction.	Naïve Bayes; Support Vector Machine; Random Forest
[93]	It introduces an attentive meta-learning method for task-agnostic representation and realizes fast adaption in different tasks, thus having the ability of learning shared representation across tasks.	Temporal Convolutional Networks (TCN)
[47]	It proposes a transfer learning method of universal language model fine-tuning (ULMFiT), which trains on three common text classification tasks; it can prevent overfitting, even with few labeled data in classification tasks by novel fine-tuning techniques.	Averaged stochastic gradient descent Weight-Dropped LSTM (AWD-LSTM) [94]

3.2.3. Sentiment Analysis

In tourism, the application of sentiment classification techniques can help manage obtain tourist sentiment tendency and opinions in real time, thus making appropriate measures. For example, the study [95] proposed a tourist destination recommendation system by analyzing and evaluating the user's sentiment tendency; the study [96] explored the sustainable tourism development path through the sentiment analysis of the user reviews of the shared bicycle system in Spain; and in this study [97], a visual analysis system was designed to analyze regional trends and sentiment changes in visitors.

Text sentiment analysis is the process of automatically classifying the polarity of a given text with subjective sentiments by computer. It includes many tasks, such as sentiment classification, opinion extraction, and so on [98]. We only discuss sentiment classification here. The method of sentiment classification can be divided into two categories: The first one is based on machine learning methods such as neural networks, and the other is based on dictionary-based methods that use pre-defined sentiment dictionaries such as WordNet, HowNet, LIWC, etc., which have sentiment-related terms, and their polarity values. Dictionary-based methods are based on grammatical rules of text analysis, relying on the quality of the sentiment dictionary and its continuous updating, involving more work refinement, such as the extraction and discrimination of evaluation words, and the consideration of the influence of the word contexts, etc. [99]. In addition to the two types of sentiment classification methods, sentiment classification through the integration of machine learning and dictionary-based methods also shows great potential. For example, this research [100] used the sentiment polarity and part of speech in the sentiment dictionary to extract the feature of the text representation combined with convolutional neural networks.

Sentiment classification with machine learning methods is mostly based on supervised learning and relies on the completeness of the labeled training corpus, which is a classification method about features. This research [101] pointed out that feature extraction, feature weight, and sentiment classifier are three essential design elements that affect the accuracy of text sentiment classification. Based on this, sentiment classification with machine learning is mainly carried out around these elements. No matter what kind of design element, the optimal vector representation is sought to achieve better precision and speed for model training. With the development of deep language models and their superiority, sentiment classifiers are more improved and optimized based on recurrent neural networks and convolutional neural networks [102]. Researchers have proposed different feature improvement strategies. For example, for short texts presented on social networks, the study [103] used a two-layer convolutional network to jointly train characters, words, and sentence features. For the long-distance dependence problem of long text, in the study [104], a TopicRNN model was proposed to get the global semantic information, which introduced the topic model to RNN model training to obtain unsupervised

feature of document global semantic information. Aiming at the respective characteristics of the classifiers, such as the dependence of the CNN model on window size and step size [105], and the long-distance dependence problem of the mechanism of the RNN model, the study [106] proposed to use LSTM as the pooling layer in CNN to promote sentiment classification. In view of the fact that there are similar contexts in the representation of sentiment words and the opposite of the emotional tendency of words, such as "good" and "bad", the research [107] proposed to introduce sentiment information into the word vector, thus promoting the learning and classification of sentiment words.

In recent years, the attention mechanism has become one of the mainstream techniques of text processing due to its superior performance. In previous text research, the attention mechanism was mainly applied in the recurrent network structure. For example, in this study [35], a hierarchical attention mechanism (HAN) was proposed to achieve sentiment classification, which constructed the sentence representation with the word-level attention mechanism and the document representation with sentence-level attention mechanism. However, the recurrent network is a sequence-dependent structure, which has a disadvantage in training speed and memory consumption. Aiming at these problems, the Transformer model [37] was proposed, which consists entirely of attention mechanisms and applied self-attention mechanism, and has a complete advantage over the structure of recurrent and convolutional sentiment classification [108].

In the last few years, most of the research on sentiment classification focuses on how to improve the classification accuracy of the entire text, but rarely analyzes the sentiment polarity based on the aspects or targets appearing in the text. In tourism analysis, not only is knowing the overall sentiment tendency of the tourists' comments needed, but also knowing the various sentiments of each entity in the tourism or each aspect of the tourist comments is required, so as to better self-evaluate and propose more targeted solutions. Due to the complexity of the process and the lack of related corpora, most of the works are unable to achieve an effective evaluation of aspect extraction and sentiment classification. The research [3] considered the possibility of describing the topic words by considering the distance between the topic words and the sentiment words, and exploring the preferences of tourists for tourism products. This method is simple, but the accuracy of the result is low due to the existence of the virtual target and the implicit evaluation object [109]. The study [22] extracted topics from the destination reviews based on LDA and then analyzed the sentiment state of each topic in more detail or for finer gain. The study [110] used text mining and sentiment analysis techniques to analyze hotel online reviews to explore the characteristics of hotel products that visitors were more concerned about. Most of these methods only make use of the model, while the adaptability of the model to the domain is not well explained.

At present, sentiment analysis based on specific targets or aspects has become a research hotspot for scholars. This method no longer separates the topic model from the sentiment analysis, but unites them into a single model [109,111]. Most works for aspect-based or target-dependent sentiment classification are based on supervised learning and achieve good results, as shown in Tables 4 and 5. While manually labeling data for the supervised model is usually insufficient and costly, unsupervised or semi-supervised models can utilize unlabeled samples for training, which can resolve the problem of insufficient resources. For example, the study [63] incorporated sentiment distribution and sentiment-oriented local topic distribution into the topic probability distribution model, so that the mining of sentiments and local topics was carried out simultaneously. The study [112] proposed to learn specific aspects of word embeddings based on the Topic Word Embeddings model (TWE2), and used the semi-supervised variational autoencoder (SSVAE) to perform aspect-level sentiment analysis.

Table 4. Common aspect-based sentiment analysis (ABSA) and target-dependent sentiment analysis on SemEval 2014 Task 4 in restaurant domain.

Time	Model	Basic Idea	Accuracy (%)
2016	AE-LSTM [113]	The target words given in each sentence of the training corpus are vectorized and added to the LSTM model as input for training together.	76.20
2016	AT-LSTM [113]	An attention mechanism is proposed to capture key parts of a sentence related to a given aspect.	77.90
2018	AF-LSTM [114]	A new association layer that defines two correlation operators, circular convolution and cyclic correlation, is introduced to learn the relationship between sentence words and aspects.	75.44
2015	TD-LST [115]	Two LSTM networks are adopted to model separately, based on context before and after target words for target-dependent sentiment analysis tasks.	75.63
2015	TC-LSTM [115]	On the basis of TD-LSTM, target word information is added as an input.	76.01
2016	ATAE-LSTM [113]	On the basis of TD-LSTM, aspect information is introduced in two parts of the model: Input part and hidden part	77.20
2017	IAN [116]	It learns attentions in target and context words interactively, and generates the representations for targets and contexts separately.	78.60
2016	MemNet(k) [117]	It uses deep memory network with multiple computational layers (hops) to classify sentiments at the aspect level, where k is the number of layers.	(k = 2) 78.61 (k = 3) 79.06
2018	Coattention-LSTM Coattention-MemNet(3) [118]	A collaborative attention mechanism is proposed to alternately use target-level and context-level attention mechanisms.	78.8 79.7
2017	BILSTM-ATT-G [119]	Based on the Vanilla Attention Model, this model is extended to differentiate left and right contexts, and uses the gate method to control the output of the data stream.	79.73
2018	TNet-LF TNet-AS [120]	The CNN is used to replace the attention-based recurrent neural network (RNN) to extract the classification features, and the context-preserving transformation (CPT) structure such as lossless forwarding (LF) and adaptive scaling (AS) is used to capture the target entity information and the retention context information.	80.79 80.69
2018	AE-DLSTMs [121]	On the basis of AE-LSTMs, this model captures contextual semantic information in both forward and backward directions in aspect words.	79.57
2018	AELA-DLSTMs [121]	Based on AE-DLSTMs, this model introduces the context position information weight of the aspect word.	80.35
2018	StageI+StageII [122]	It introduces a position attention mechanism based on position context between aspect and context, and also considers the disturbance of other aspects in the same sentence.	80.10
2018	DMN+AttGRU (k = 3) [123]	A dynamic memory network which uses multiple attention blocks of multiple attention mechanisms is proposed to extract sentiment-related features in memory information, where k stands for attention steps.	81.41
2018	MGAN [124]	This model designs an aspect alignment loss to depict aspect-level interactions among aspects with the same context, and to strengthen the attention differences among aspects with the same context and different sentiment polarities.	81.25

In the training process, the pre-trained word vectors in these models were all initialized by 300-dimension Glove embeddings and the sentiment classification was performed in a three-way classification.

Table 5. Common aspect-based sentiment analysis (ABSA) and target-dependent sentiment analysis on subtask 1 (slot 2) SemEval 2016 Task 5 in restaurant domain.

Time	Model	Basic Idea	Accuracy (%)
2019	BERT [125]	Bidirectional Transformer (BERT) is extended with an additional task-specific layer and fine-tuned on each end task	81.54
2019	BERT-PT [125]	On the basis of BERT, two pre-training objectives are used: Masking language model (MLM) and next sentence prediction (NSP), to post-train domain knowledge, and else task (MRC) knowledge.	84.95

The sentiment classification is performed in a three-way classification.

As shown in Tables 4 and 5, the sentiment classification algorithms are all trained on existing data sets from SemEval2014 and SemEval2016 on which they achieved excellent performance, while the limited data sets making them not universally applicable in other new domains. An effective way to enhancing the ability of automatic labeling for target domain is learning shared features from source domain or transferring knowledge from source domain into target domain. Aiming at it, some works have been done attempting to enhance transfer learning, such as importing of domain knowledge into the training process expecting to contribute to the knowledge transfer for realizing cross-domain aspect sentiment classification [126], but which still need human intervention or processing in a semi-supervised manner. As a result, in tourism application, few studies about aspect-based or target-dependent sentiment classification have been made. The study [2] considered using lda2vec to explore the focuses of tourist reviews and also as an input knowledge to enhance the sentiment analysis of these focuses, but still had room for fine-grained sentiment analysis for aspects. The study [127] proposed a novel probability model to judge user sentiment and topic sentiment in an unsupervised manner, which provides a direction for tourism recommendation due to its introduction of user information, but an effective evaluation method is needed.

In summary, the sentiment classification model, improved by the attention mechanism, will be the mainstream trend in the future. In addition, based on the study of sentiment targets or sentiment aspects, the sentiments can be more fine-grained and interpretable, which will be more conducive to the practical application analysis of tourism.

3.2.4. Text Clustering

Text clustering mainly involves unsupervised algorithms that can discover potential knowledge and rules from large-scale text data sets, facilitating the effective organization, abstracting, and navigation of texts. In the field of tourism, text clustering is mainly applied in the research of tourist hotspots or emergencies. For example, the study [128] performed co-occurrence clustering analysis by constructing a high-frequency word co-occurrence matrix in order to acquire hot things, and measured the weights of the connection between the regions within or outside the Tibet by degree centrality in the social network map; the tourist hotspots and their interrelationships were then obtained and tourism planning was further promoted. In this study [129], cohesive hierarchical clustering methods were used to detect the emergencies by using bursty topics to represent texts. Besides, text clustering can also be applied in the subdivision problem of the tourism market in which a clustering method is used to obtain different characteristics of the group for targeted analysis and self-improvement. In this study [130], the collaborative clustering algorithm was used to cluster the five-star hotels and hotel reviews in Rome from two dimensions, which not only solved the feature clustering of each hotel, but also solved the description of the features. As mentioned above, text clustering can be an efficient method for tourism analysis. Next we will review it from the technical perspective.

The object for text clustering can be documents, sentences, paragraphs, and so on. Similarity is the basis of text clustering. At present, the mainstream text similarity calculation method is mainly based on vector space method, including cosine similarity, Manhattan distance, Euclidean distance,

and so on. With the development of deep learning, word vectors generated by neural networks such as Word2vec can make words closer in semantic distance, and are more suitable for similarity calculation of various text granularities.

Document clustering techniques include multiple types, such as agglomerative hierarchical clustering algorithm, partitioning clustering algorithm, density-based clustering algorithm, etc. Different clustering algorithms have different requirements for application scenarios, and the prior knowledge of specific tasks [131] must be considered. Currently, there are few researches on text clustering algorithms. The main reason is that the computational overhead of clustering algorithms tends to be large. When the amount of data rises to a certain extent, most clustering algorithms cannot be used, so the time complexity of most clustering algorithms needs to be considered [132]. K-means, which belongs to the partitioning clustering algorithm, is a commonly used text clustering algorithm whose disadvantage is that it cannot effectively determine the number of clusters and select the initial clustering point, and has poor performance on high dimensional data, etc. While compared to other clustering algorithms, K-means is fast and easy to implement on a huge database [36]. Consequently, there are many researches carried out based on K-means, such as through optimization or dynamic definition of the initial clustering point [133,134], improvement of text representation by genetic algorithm, graph structure, deep learning, etc. [135–137], and optimization of algorithm objective function [138].

Text clustering can also be a semi-supervised algorithm, which mainly uses text labels like potential topics as prior knowledge to guide the clustering process, and is a bridge connecting unsupervised clustering and supervised classification problems [139]. The semi-supervised clustering algorithm is usually a combination of unsupervised clustering algorithm and supervised model, which is mainly used to find the similarity between texts and label data samples currently, and can alleviate the demand for data volume and improve the performance of supervised models. With the research and application of large-scale knowledge graph, text clustering algorithms will play more of a role—for example, clustering algorithms can be used to build hierarchical ontological relationships, discover semantic relationships between domain concepts [140], and gradually form large-scale semantic network diagrams, etc.

4. Applications of Text Corpus-Based Tourism Big Data Mining

The tourism process mainly includes five stages: Imagination, planning, scheduling, experience, and sharing [141]. The sharing stage is the most critical stage in the tourism process. Whether tourism behavior occurs or not depends on whether the plan is successfully completed, and the tourism plan depends on other completed visitor shares. If a tourism stakeholder or destination wants to improve their services and attract more visitors, it must know what the tourists are thinking and needs to understand their preferences, needs, and purposes. As tourists, they hope that the journey will involve "zero" conflict, and they can get useful information from the travel network platform. The recommended destination is based on their preferences, and the tour route is greatly optimized.

Tourists share the experience based on their own experience, which can not only reflect their preferences, but also the problems of tourism stakeholders and destinations in time. From the tourism innovative applications based on the techniques of text corpus-based tourism big data mining, this section analyzes the two main tourism application scenarios: Tourist profile and market supervision.

4.1. Tourist Profile

Tourist profile is used to abstract the specific labels from the attribute information of a tourist. The attributes usually include: Demographic characteristics (individual or organization, age, gender, location), mental state and lifestyle (education, profession, purchasing ability, family, property, emotional attitude, interest, fear, etc.), travel preferences, and travel purposes. The tourist attributes are diverse, which leads to that the demand is also diverse, driven by the consumption upgrade [142]. Therefore, tourist market segmentation is of great significance to tourism destinations such as the

discovery of market opportunities, the planning of right marketing and competition strategies, and the realization of personalized recommendations.

By dividing their natural attributes, tourists can be divided into female and male groups, youth and old age groups, single and married groups, local and foreign groups, etc. By the analysis of preferences and behaviors, tourists can be divided into more groups such as travel buyers, the decompressions, and so on. Different groups have different characteristics, and individuals with the same attributes may have similarities in tourism behavior [143]. Study [73] has confirmed that differences in tourist attributes can also lead to differences in tourist perceptions. Through the study of the Cape Town tourism market [144], it was found that visitors' age, place of residence, destination stay time, return visits, etc., had an important influence on the perception of tourists, and the sentiments they conveyed [145] also had different characteristics.

Tourist profile are an important means of understanding tourist behavior and meeting the tourist expectation. Since the content of the tourists' comments often reflects their subjective thinking, we can extract information such as preferences, concerns, and purposes of different tourists from the texts. By obtaining their relevant attributes, tourists' profiles can be effectively created. While how to generate user profiles through practical text analysis is still a hot and challenging issue for scholars, through literature research, it has been found that user profiles are mainly obtained by supervised learning, or realized by the feature recognition from data labeled gender, age, occupation, ratings, and so on [14]. In addition, the user attributes for profiles are always treated as isolated in feature recognition; in other words, the relationship between user attributes is ignored, while the attributes are often interrelated. Aiming at this problem, multiple attributes joint learning can be efficient to improve the user attribute prediction [8]. However, although the supervised learning for tourist profiles has achieved good results, it still has limitations because its performance depends entirely on the number of data and its domain. Taking some sample data as the research object, the study [9] extracted the "co-words" from different users in the sample to obtain a universal judgment criterion for each user, but the viewpoints or conclusions derived from the sample were often one-sided due to the limited numbers. By using the text information from a large number of existing users on social media, a unified user vector learning model can be obtained to fill the knowledge gap between the source social media and the target social media, and then the problem of uncertainty of user labels for the target media can be solved [10]. Similar works [15] were also done, which considered matching user accounts on different social networks to build user profiles by user identification based on User Generated Content (UGC) in a supervised manner. These methods are all based on this assumption that the data for the same attribute or the same person has common features, such as commonality of the same gender [13], to resolve the problem of the limited labeled data. In supervised learning, current methods for tourist profiles are usually around gender, age, and other explicit feature predictions. The topic-based model is an unsupervised algorithm which can extract user preferences or hobbies, etc., and is an efficient method for the acquisition of the user's attribute information, except for explicit feature classification [11]. Furthermore, the unsupervised aspect-based or target-dependent sentiment analysis, which is studied a lot currently, can recognize user preference for aspects or the target, and provide a more fine-grained analysis for user profiles [12].

Tourist profile is a feature extraction process and a vital step of the personalized recommendation in tourism big data mining. The personalized recommendation system is a process of intelligent recommending for users according to their preferences, habits, and individual needs. In the field of tourism, the recommendation system is more complicated, because we not only need to consider personal attributes, but also need consider travel characteristics. These two considerations jointly determine tourist decision-making behavior [146]. Travel characteristics include destination type, travel distance, traffic mode, travel expenses, and so on. Different tourist personal attributes have different features in the performance of travel characteristics [147], and tourist characteristics directly influence the choice of travel characteristics, such as the choice of destination.

Mastering the tourist psychological characteristics in travel planning is the critical procedure for a good personalized recommendation system design, and the text reviews become an important supplement to the data sparsity in the tourism recommendation process. By mining user reviews, user preferences, and travel destination, reputation or features can be gained and introduced to the travel recommendation system for final recommendations [148]. Because the topic model can detect tourists' preferences, frequent behaviors, or new travel trends in an unsupervised manner, it has become a hot research direction for scholars. For example, the study [149] mined the feature information of tourists and locations by the topic model, and used knowledge-based filtering techniques to achieve destination recommendations for tourists by semantic similarity. The study [150] proposed a Topic Criterion (TC) model by improving the topic model and the Topic Sentiment Criterion (TSC) model to calculate tourist profiles and item profiles, as well as their matching degrees to achieve project recommendations for potential tourists. In addition, some scholars have also considered the context of travel in the recommendation process, such as seasons, holidays, etc. In the study [95], a text mining technique was used to calculate the user's sentiment tendency toward the destination, and the influence of time elements such as seasons and holidays on the tourists' sentiments were considered comprehensively to promote the tourism recommendation system greatly. Based on the literature research of tourism recommendation system [151,152], we summarize the general framework of the tourism recommendation system based on text mining (shown in Figure 2).

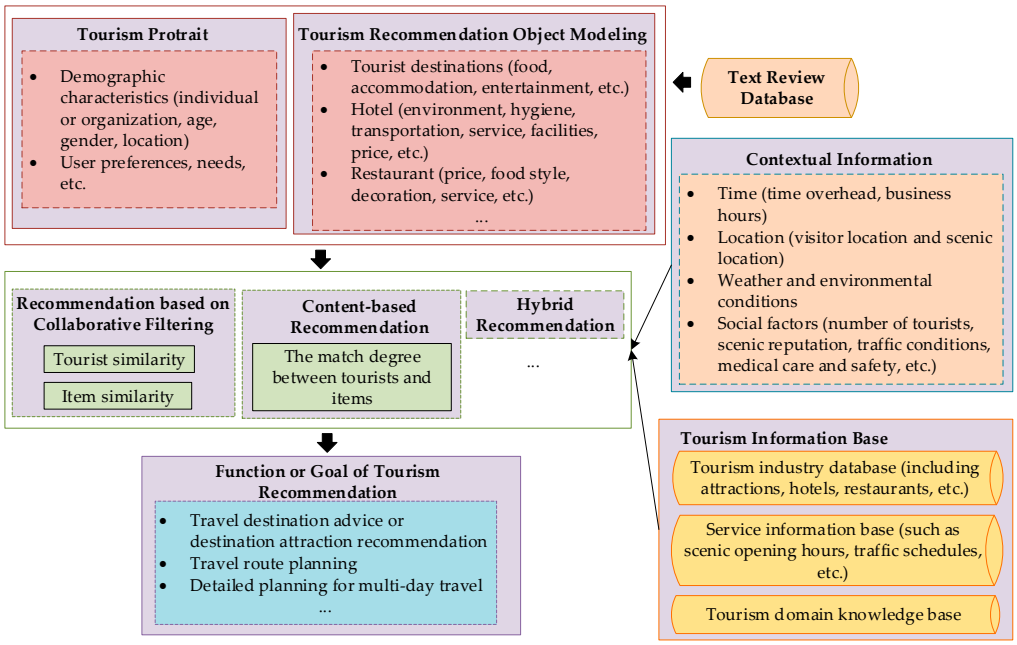

Figure 2. Tourist recommendation system framework based on text mining.

4.2. Market Supervision

The tourism market is the basis for tourism to survive in. Research on the tourism destination market has important theoretical and practical significance for tourism development [50]. The existence of the tourism system depends on the existence of tourist demand, which is always related to aspects of the tourism process such as "food, accommodation, transport, sightseeing, purchase, entertainment", and is diverse due to the difference of tourist natural and social attributes. By the analysis of tourist demand and preferences, researchers or practitioners can assess the market composition of tourism destinations and adjust the tourism market resource allocation or make marketing strategies to maximize the degree of satisfaction of tourists.

In the context of big data, the online tourism market is gradually driven by user data. Texts, as a main component of user-generated content, can accurately reflect the needs of visitors. From

the perspective of market or tourism stakeholders, this paper summarizes five aspects of text content analysis: Target topic, dimension and weight of concerns, satisfaction evaluation or preference, the reason for sentiment, and new trend, to assist market strategy planning. The analysis of specific "target topic" of the text can analyze the tourist needs more specifically; "dimension and weight of concerns" is to analyze the tourist demand from their attention to various aspects and compute their weights on the attention; "satisfaction evaluation or preference" is to analyze the sentiment orientation of comments to obtain tourists' satisfaction with the travel experience; "the reason for sentiment" uses sentiment cause detection techniques to detect the cause of sentiment in order to find the reason behind the sentiment; and the "new trend" analyzes the emergence and developing process of new things from the perspective of time series. Next, we will explain with examples in detail. The study [17] explored the key elements of hotel customer comment by the topic model LDA and analyzed the significance of their influence through the perception map of hotel reviews, which are important for the analysis of customer satisfaction. The study [18] combined the three elements of tourist market share, tourist sentiment orientation, and potential tourist awareness to define and calculate the competitiveness of the tourism destination market, the tourists' sentiment orientation of which is obtained through the sentiment analysis model based on text comments. In this study [19], sentiment analysis on a specific topic, "traffic", was conducted to analyze the causes of negative sentiments by using the method of co-occurrence of words and evaluation objects.

The destination image is a reflection of the tourist market, including the national country image, the city image, the scenic spot image, etc. Research on the destination image was first proposed by Gunn [153]. As a basis of market positioning, the destination image has attracted a lot of attention from scientists. Compared with the traditional customer survey method, the method of user-generated text analysis can reflect the various dimensions of the destination image more accurately [154], and the real travel experience of the tourists can effectively improve the accuracy of the destination image evaluation [155]. This research [20] was the early study of using online text content for destination image analysis. With the rapid process of "Internet Tourism", more and more researches have begun to explore text analysis as a means of destination image analysis [16,21].

In the process of evaluating the destination image, it is necessary to recognize the tourists' sentiments and the sentiments for all aspects in their tourism in order to assess the satisfaction of the destination components [21,22]. Specifically, the components which compose the destination image are extracted by text mining techniques, and the sentiment analysis is performed for each component to obtain the satisfaction evaluation. Figure 3 shows the key determinants of tourist satisfaction and their impact from a macro-causal perspective. Customer satisfaction is determined by a combination of customer expectations, perceived quality, and value; it has a direct effect on customer loyalty, which is essential for destinations in gaining competitive advantage [156,157]. Besides, because the determinants of tourist satisfaction may be different for different destinations, it is also a feasible method for studying the constituent variables of the image and their weights [23] to promote the evaluation of destination image.

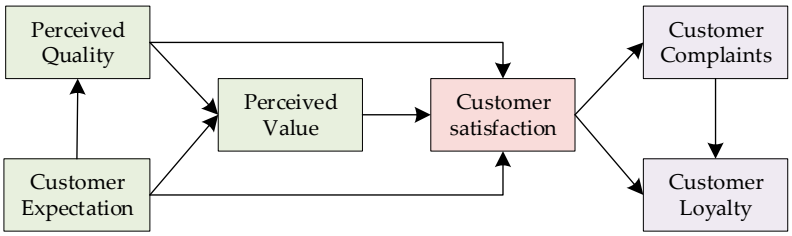

Figure 3. Tourist satisfaction model [18].

From the perspective of tourist behavior, market supervision also includes public sentiment analysis, such as tourism hotspots, crisis events, or emergency analysis, among which tourism hotspots include popular attractions, popular tourist routes, and hot topics. People often show strong concern

about current hot spots, hot issues, or public opinion crisis. Especially in the era of highly developed online media, these concerns are often presented in the form of text on the network platform and show high-frequency characteristics. Some scholars take this as a point of view, using statistical word frequency and word frequency co-occurrence to explore tourism hotspots. In the study [24], the location of popular attractions and tourist routes were obtained by mining frequent geographic patterns in travel journals. The study [25] used the maximum confidence and frequent mining patterns to capture neighborhood relationships of the attractions in the tourist log, and further to obtain the most famous sights and frequent tourist routes. Some scholars also use the method of keyword extraction or text clustering to explore the common concerns, get tourist hotspot events, and use sentiment analysis techniques to obtain public opinion orientation with the event [158]. For emergencies or crisis events, due to their real-time characteristics—that is, sudden bursts of growth in a short time—the timing changes of words need to be considered.

For the convenience of readers, we summarize the main contributions, benefits, and main methods of a selected list of informative articles in Table 6.

Table 6. Informative articles about text corpus-based tourism big data mining between 2015 and 2019.

	Contributions	Benefits	Methods
[5]	The topic features of attractions in the context of seasons are firstly explored, which are precisely at the fine-grained season levels.	The proposed a season topic model based on LDA (STLDA) model can distinguish attractions with different seasonal feature distributions, which helps improve personalized recommendations.	Latent Dirichlet Allocation (LDA)
[12]	This paper proposes a sentiment-aspect-region model with the information of Point of Interests (POIs) and geo-tagged reviews to identify the topical-region, topical-aspect, and sentiment for each user; it also proposes an efficient online recommendation algorithm and can provide explanations for recommendations.	POI recommendation, user recommendation, and aspect satisfaction analysis in regions can be achieved by this model.	Probability generative model; expectation-maximization (EM)
[25]	It firstly divides tourism blog contents into semantic word vectors and creatively uses the frequent pattern mining and maximum confidence to capture the neighborhood relationships of the attractions in the tourist log.	Popular attractions and frequent travel routes from massive blog data analysis can be extracted, and thus potential tourists can schedule their travel plans efficiently.	Term Frequency (TF); frequent pattern mining; maximum confidence
[55]	It proposes a negative review detection method by adapting Structure Topic Model (STM); the variation of document-topic proportions with different level of covariates can be easily determined.	It enhances our understanding of the aspects of dissatisfaction in text reviews.	STM
[95]	It employs text mining techniques to access sentiment tendency which is incorporated into an enhanced Singular Value Decomposition (SVD++) model for model amendment also with the temporal influence, such as seasons and holidays on the tourists' sentiments.	It can help alleviate the cold-start problem effectively and thus improve the tourism recommendation system.	SVD++
[127]	It proposes a topic model which can judge users' sentiment distribution and topic sentiment distribution in a topical tree format.	It offers a general model for practitioners to determine why users like or dislike the topics.	Hierarchical probability generative model
[150]	It proposes a Topic Criterion (TC) model and the Topic Sentiment Criterion (TSC) model to calculate tourist profiles and item profiles, as well as their matching degrees to achieve recommendations.	It can be beneficial to tourism recommendation and provide an interpretation of users and item profiles.	LDA; JST (Joint Sentiment-Topic model)

5. Outlook

Tourism big data in the form of text plays an important role in tourism applications. First of all, tourism is a service system, emphasizing the sentiment or value experience of tourism individuals. Text mining techniques have become indispensable to the sentiment judgment and value-oriented analysis in modern tourism applications. Secondly, text mining techniques are experiencing a period of rapid development and are achieving much improvement. Benefit from deep learning techniques such as text classification and sentiment analysis have made many breakthroughs [159].

However, text mining techniques based on deep learning are often less practical in tourism due to the requirements of deep learning for data volume and labeled data, and most of them only use existing data to explore future tourism trends. Aiming at the problem of lack of existing standard tourism corpus and the limitations of deep learning such as interpretability, this paper makes a detailed analysis and puts forward some major trends of future tourism text data mining.

(1) Lack of domain corpus. The languages of the existing tourism corpora are mostly English and the limited multi-language categories make the existing tourism corpora not universally adaptable. In addition, the annotation of the tourism corpus often relies on manual labor, lack of system and formativeness, and the scale of the corpus is usually small. How to automatically and effectively construct a standardized large-scale multi-language tourism corpus has become one of the keys to the successful application of tourism big data. Given the impact of publicly annotated data sets on tourism big data mining and for the convenience of research, we summarize some of the relevant publicly available text data sets currently in the tourism domain, with the data sets described and the dataset sources listed in Table 7.

Table 7. Publicly available tourism data sets.

Name	Description	Source	Accessed Date
Hotel_Reviews 515,000	It contains 515,000 customer reviews with positive and negative aspects and ratings of 1493 luxury hotels across Europe, as well as the location of the hotel.	https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe	24 July 2019
TripAdvisor Hotel Review Dataset	It contains 20,490 hotel customer reviews and related review ratings.	https://zenodo.org/record/1219899#.XFSETygzY2w	24 July 2019
Citysearch Restaurant Review Dataset	It contains 35,000 food reviews and lists representative words related to the attributes of each entity. It also includes 3418 sentences with labeled sentiment polarity for the attributes of each entity.	http://dilab.korea.ac.kr/jmts/jmtsdataset.zip	24 July 2019
OpinRank Dataset	It contains 259,000 reviews of 10 different cities (Dubai, Beijing, London, New York, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, and Chicago), each city of which has approximately 80–700 hotels, with dates, comment title, and full comment included.	https://github.com/kavgan/OpinRank/blob/master/OpinRankDatasetWithJudgments.zip	24 July 2019
SemEval ABSA Restaurant Reviews-English (2014–2016) SemEval ABSA Hotels Domain-English (2015–2016)	It includes multiple English data sets for restaurants and hotels which are composed of comments, with the attributes (E#A pairs) and the target and the corresponding sentiment polarities marked.	http://metashare.ilsp.gr:8080/repository/search/?q=SemEval	24 July 2019
SentiBridge: A Knowledge Base for Entity-Sentiment Representation	The dictionary contains a total of 300,000 entity-sentiment pairs, currently from the three domains of news, travel, and catering.	https://github.com/rainarch/SentiBridge	24 July 2019
ChnSentiCorp-Htl-unba-10000	It contains 7000 positive and 3000 negative hotel reviews in Chinese.	https://download.csdn.net/download/sinat_30045277/9862005	24 July 2019

Table 7. *Cont.*

Name	Description	Source	Accessed Date
TourPedia	It contains two main data sets: Places, and reviews about places. Places contains accommodations, restaurants, attractions, and points of interest, and each place is descripted with address, location, polarity, etc. Reviews about places has some auxiliary information such as rating, time, polarity, place, etc.	http://tour-pedia.org/about/datasets.html	24 July 2019
Museum reviews from TripAdvisor	It contains 1600 museum data including address, category, review, rating, popularity, etc.	https://www.kaggle.com/annecool37/museum-data	24 July 2019

Recently, knowledge transfer has attracted a lot of attention, with attempts to transfer the knowledge learned from the existing large-scale data to the target task to reduce the demand for the target data, and plays a decisive role in putting future research of tourism big data mining into practical applications. At present, transfer learning has made good progress in cross-domain feature extraction and sentiment analysis [160]. For the lack of training data, scholars have proposed some transfer learning models such as few-shot, one-shot and zero-shot, which can learn the relevant features of the data for classification prediction if the training samples in the target area are not provided or only provided a small, and they have achieved good results in the text data mining [161,162]. Recently, the pre-trained BERT model shows great advantages in multi-language and multi-task transfer learning, without substantial task-specific architecture modifications, which makes transfer learning widely applicable to the text mining.

Meta-learning also achieves outstanding results in response to the problem of too little training data in the target domain. The goal of meta-learning is to train a model on multi-task and to obtain common attributes that adapt to a new task, reducing the need for pre-trained data. For example, for the problem of fuzzy learning tasks with the small sample in NLP field, an adaptive metrics meta-learning method has been proposed, which automatically captures the best-weighted combination of metrics from the meta-training task [161].

In addition, the use of semi-supervised and unsupervised learning methods can also reduce the dependence of labeled data. Studies [163] have pointed out that when multilingual transfer learning such as BERT, unsupervised models, and meta-learning are combined, the areas with fewer data resources are very promising.

(2) Limitations of deep learning. The first one is poor interpretability. For a long time, deep learning has been lacking in rigorous mathematical theory, and it is impossible to explain the quality of the results and the variables that lead to the results. In the tourism domain, the interpretable performance of deep learning is more conducive to discover knowledge and understand the nature of the problem, thus the practitioners can make operational service adjustments. The use of attention mechanisms in deep learning also provides an interpretable channel for deep learning. However, for deep learning itself, it still seems to be a black box problem. Some scholars also consider using the knowledge graph to eliminate the semantic gap between NLP and deep learning, which will provide vital support for deep learning interpretability in the future.

The second is the limited expression capability. Text information extraction can be realized by multiple feature learning layers of deep learning model. However, as the complexity of the model increases, the learning ability strengthens, but there exists an over-fitting problem. This problem can be solved by acquiring massive data to a certain extent, while the lack of labeled data and the finiteness of hierarchy in deep learning models restricts the learning ability of deep learning. Currently, transfer learning is an effective way to solve this problem. Besides, some scholars, such as Professor Zhou

Zhihua, have considered the use of non-differentiable models to enhance the expressive ability of deep learning aiming at simulate the diversity of the real world, which is a great challenge for deep learning.

(3) The future trend of text corpus-based tourism application. Tourism has a high degree of social nature. It uses the text information shared in social network media to explore the new vitality of tourism services or develop products by feedbacks from tourists, which is the general way of tourism text data mining currently. Tourism personalized recommendation is a significant and potential direction because it caters to current social needs. However, in tourism recommendation, the cold start problem for tourists or tourism items has always been a difficult task for scholars to explore, and thus fail to solve. Combined with other enriched multimedia content such as videos, photographs, text, links to websites, etc., text-based recommendation will be enhanced [164,165], which is also a supplement for addressing the cold start problem. Besides, how to dynamically explore tourist preferences and how to explore the unknown or unfamiliar tourism area or travel style for tourists will become a hot spot for future research.

This paper mainly reviews the automatic text mining techniques in NLP, which can assist people to acquire information from a large of text data. Text generation techniques will also be necessary for future tourism development and application. From tourism recommendations, NLP is transitioning to the process of assisting people in understanding, which will provide a decisive or interpretable way for tourism. In the future, with the improvement of interactive NLP [166], the machine will be able to understand human language more accurately and communicate more naturally with users, thus providing tourists with real-time intelligent answers and suggestions. In the future, reinforcement learning (RL) [167] will give a powerful impetus to tourism big data because it can adapt to the instant changes in the environment.

For the convenience of readers, we summarize the main key methodologies of the text mining techniques in the surveyed papers in Table 8:

Table 8. Main take-aways for the reader.

Main Take-Aways
Topic probability model is a basic model used in most topic extraction algorithms, which can be improved by enhancing topic coherence of short texts or exploiting the sematic feature of words and text enabled by deep learning.
Language models based on deep learning models such as CNN and RNN, etc., are widely applied in text classification. Focusing on their requirement for abundant labeled data for supervised learning, many strategies have been proposed such as co-training, training samples extension, meta-learning, and transfer learning.
One of the mainstream trends in sentiment classification is to exploit the attention mechanism in deep learning. Based on the study of sentiment targets or sentiment aspects, the sentiments can be more fine-grained and interpretable, which is more conducive to practical application analysis.
K-means is a method often commonly used in text clustering due to its small time complexity. Optimization of initial clustering points, improvement of text representation, and optimization of objective functions are all popular aspects of improvements to K-means-based text clustering.

6. Conclusions

Big data analysis is changing the operating mode of the global tourism economy, providing tourism managers with deeper insights, and infiltrating into all aspects of tourist travels, while driving tourism innovation and development [168]. Tourism text big data mining techniques have made it possible to analyze the behaviors of tourists and realize real-time monitoring of the market. As the key technique of text analysis, NLP is experiencing a period of vigorous development. Both machine learning and current deep learning with high achievements have been greatly applied in NLP. The deep learning language model provides a general learning framework, which can flexibly represent the

text, and can be easily extended to different network models—such as standard methods CNN, LSTM, GRU, and various variants of standard methods—which laid the foundation for the deepening of the deep learning theory in the NLP field, and thus provided a solid theoretical basis for the improvement of the text corpus-based tourism big data mining.

This paper systematically summarizes current and potential applications of big data text mining techniques in Internet tourism economy and provides some guides for further research in tourism big data analysis. At present, most of the existing studies on tourism big data mining tend to be driven by data and algorithm innovation. However, tourism data analysis and service evaluations without considering the subjective nature of tourists may be inherently biased. Personalized subjective analysis and evaluation methods, such as Kansei engineering, widely use product evaluation [169], and thus have big potential in tourism big data analysis. Combining data-driven methods with tourism domain knowledge, such as the considering of domain-specific words [170], is also another direction that needs exploration in the future.

Author Contributions: Q.L. and S.L. conceived the conception; Q.L. conducted literature collection and manuscript writing; S.L., J.H. (Jianjun Hu), J.H. (Jie Hu) and S.Z. revised and polished the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant No. 91746116, National Science and Technology Support Program under Grant No. 2014BAH05F02, Science and Technology Program of Guizhou Province Nos. [2015]4011, Nos. [2016]5103, Nos. [2017]5788, and the Science and Technology Project of Guizhou Province under Grant Talents Nos. [2018] 5774-034.

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

References

1. Ye, Q.; Law, R.; Gu, B.; Chen, W. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput. Hum. Behav.* **2011**, *27*, 634–639. [CrossRef]
2. Li, Q.; Li, S.; Hu, J.; Zhang, S.; Hu, J. Tourism Review Sentiment Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism and Topic-Enriched Word Vectors. *Sustainability* **2018**, *10*, 3313. [CrossRef]
3. Marrese-Taylor, E.; Velásquez, J.D.; Bravo-Marquez, F.; Matsuo, Y. Identifying Customer Preferences about Tourism Products Using an Aspect-based Opinion Mining Approach. *Proc. Comput. Sci.* **2013**, *22*, 182–191. [CrossRef]
4. Xu, J.; Fan, Y.; Bai, B. Knowledge mining and visualizing for scenic spots with probabilistic topic model. *J. Comput. Appl.* **2016**, *36*, 2103–2108.
5. Huang, C.; Wang, Q.; Yang, D.; Xu, F. Topic mining of tourist attractions based on a seasonal context aware LDA model. *Intell. Data Anal.* **2018**, *22*, 383–405. [CrossRef]
6. Al-Horaibi, L.; Khan, M.B. Sentiment analysis of Arabic tweets using text mining techniques. In Proceedings of the First International Workshop on Pattern Recognition, Tokyo, Japan, 11–13 May 2016; p. 100111F.
7. Okazaki, S.; Andreu, L.; Campo, S. Knowledge sharing among tourists via social media: A comparison between Facebook and TripAdvisor. *Int. J. Tour. Res.* **2017**, *19*, 107–119. [CrossRef]
8. Wang, J.; Li, S.; Zhou, G. Joint Learning on Relevant User Attributes in Micro-blog. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 4130–4136.
9. Gu, H.; Wang, J.; Wang, Z.; Zhuang, B.; Su, F. Modeling of User Portrait Through Social Media. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018.
10. Wang, J.; Li, S.; Jiang, M.; Wu, H.; Zhou, G. Cross-media User Profiling with Joint Textual and Social User Embedding. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 21–25 August 2018; pp. 1410–1420.
11. Pennacchiotti, M.; Popescu, A. A Machine Learning Approach to Twitter User Classification. In Proceedings of the International Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.

12. Zhao, K.; Cong, G.; Yuan, Q.; Zhu, K.Q. SAR: A Sentiment-aspect-region Model for User Preference Analysis in Geo-tagged Reviews. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 13–17 April 2015; pp. 675–686.
13. Teso, E.; Olmedilla, M.; Martínez-Torres, M.; Toral, S. Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. *Technol. Forecast. Soc. Chang.* **2018**, *129*, 131–142. [CrossRef]
14. Škrlj, B.; Martinc, M.; Kralj, J.; Lavrač, N.; Pollak, S. tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification. *arXiv* **2019**, arXiv:1902.00438.
15. Li, Y.; Zhang, Z.; Peng, Y.; Yin, H.; Xu, Q. Matching user accounts based on user generated content across social networks. *Fut. Gen. Comput. Syst.* **2018**, *83*, 104–115. [CrossRef]
16. Költringer, C.; Dickinger, A. Analyzing destination branding and image from online sources: A web content mining approach. *J. Bus. Res.* **2015**, *68*, 1836–1843. [CrossRef]
17. Yue, G.; Barnes, S.J.; Jia, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichletallocation. *Tour. Manag.* **2017**, *59*, 467–483. [CrossRef]
18. Wang, Y. *More Important than Ever: Measuring Tourist Satisfaction*; Griffith Institute for Tourism, Griffith University: Queensland, Australia, 2016.
19. Kim, K.; Park, O.; Yun, S.; Yun, H. What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technol. Forecast. Soc. Chang.* **2017**, *123*. [CrossRef]
20. Govers, R.; Go, F.M. Projected destination image online: Website content analysis of pictures and text. *Inf. Technol. Tour.* **2005**, *7*, 73–89. [CrossRef]
21. Chi, T.; Wu, B.; Morrison, A.M.; Zhang, J.; Chen, Y.C. Travel blogs on China as a destination image formation agent: A qualitative analysis using Leximancer. *Tour. Manag.* **2015**, *46*, 347–358. [CrossRef]
22. Ren, G.; Hong, T. Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach. *Sustainability* **2017**, *9*, 1765. [CrossRef]
23. Rodrigues, A.I.; Correia, A.; Kozak, M.; Tuohino, A. Lake-destination image attributes: Content analysis of text and pictures. In *Marketing Places and Spaces*; Emerald Group Publishing Limited: Bingley, UK, 2015; pp. 293–314.
24. Yuan, H.; Xu, H.; Qian, Y.; Ye, K. Towards Summarizing Popular Information from Massive Tourism Blogs. In Proceedings of the IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014; pp. 409–416.
25. Yuan, H.; Xu, H.; Qian, Y.; Li, Y. Make your travel smarter: Summarizing urban tourism information from massive blog data. *Int. J. Inf. Manag.* **2016**, *36*, 1306–1319. [CrossRef]
26. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
27. Mikolov, T.; Sutskever, I.; Chen, K. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, CA, USA, 5–10 December 2013; pp. 3111–3119.
28. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
29. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in translation: Contextualized word vectors. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6294–6305.
30. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *N. Am. Chapter Assoc. Comput. Linguist.* **2018**, *1*, 2227–2237.
31. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/languageunderstandingpaper.pdf (accessed on 7 June 2018).
32. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
33. Xu, G.; Wang, H. The Development of Topic Models in Natural Language Processing. *Chin. J. Comput.* **2011**, *34*, 1423–1436. [CrossRef]

34. Kim, Y. Convolutional Neural Networks for Sentence Classification. *Empir. Methods Nat. Lang. Process.* **2014**, 1746–1751. [CrossRef]
35. 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, 13–15 June 2016; pp. 1480–1489.
36. Suyal, H.; Panwar, A.; Negi, A.S. Text Clustering Algorithms: A Review. *Int. J. Comput. Appl.* **2014**, *96*, 36–40. [CrossRef]
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in neural information processing systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
38. Genc-Nayebi, N.; Abran, A. A systematic literature review: Opinion mining studies from mobile app store user reviews. *J. Syst. Softw.* **2017**, *125*, 207–219. [CrossRef]
39. Moher, D.; Shamseer, L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.A. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **2015**, *4*, 1. [CrossRef] [PubMed]
40. Keele, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report Ver. 2.3 EBSE Technical Report; EBSE: Durham, UK, 2007.
41. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
42. Qiao, C.; Huang, B.; Niu, G.; Li, D.; Dong, D.; He, W.; Yu, D.; Wu, H. A New Method of Region Embedding for Text Classification. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
43. Xiong, S.; Lv, H.; Zhao, W.; Ji, D. Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing* **2018**, *275*, 2459–2466. [CrossRef]
44. Xiong, S. Improving Twitter Sentiment Classification via Multi-Level Sentiment-Enriched Word Embeddings. *arXiv* **2016**, arXiv:1611.00126. [CrossRef]
45. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *Conf. Eur. Chapter Assoc. Comput. Linguist.* **2017**, *2*, 427–431. [CrossRef]
46. Ji, S.; Yun, H.; Yanardag, P.; Matsushima, S.; Vishwanathan, S.V.N. WordRank: Learning Word Embeddings via Robust Ranking. *Comput. Sci.* **2015**, 658–668. [CrossRef]
47. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146.
48. Le, Q.V.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196.
49. Arora, S.; Liang, Y.; Ma, T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
50. Kiros, R.; Zhu, Y.; Salakhutdinov, R.R.; Zemel, R.; Urtasun, R.; Torralba, A.; Fidler, S. Skip-thought vectors. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; pp. 3294–3302.
51. Logeswaran, L.; Lee, H. An efficient framework for learning sentence representations. *arXiv* **2018**, arXiv:1803.02893.
52. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 31 October–4 November 2018; pp. 670–680.
53. Subramanian, S.; Trischler, A.; Bengio, Y.; Pal, C. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In Proceedings of the International Conference on Learning Representations, Vancouver, Canada, 30 April–3 May 2018.
54. Liu, G.; Xu, X.; Zhu, Y.; Li, L. An Improved Latent Dirichlet Allocation Model for Hot Topic Extraction. In Proceedings of the IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, Australia, 3–5 December 2014; pp. 470–476.
55. Hu, N.; Zhang, T.; Gao, B.; Bose, I. What do hotel customers complain about? Text analysis using structural topic model. *Tour. Manag.* **2019**, *72*, 417–426. [CrossRef]

56. Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Lederluis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. Structural topic models for open ended survey responses. *Am. J. Polit. Sci.* **2014**, *58*, 1064–1082. [CrossRef]
57. Zarrinkalam, F.; Kahani, M.; Bagheri, E. Mining user interests over active topics on social networks. *Inf. Process. Manag.* **2018**, *54*, 339–357. [CrossRef]
58. Rana, T.A.; Cheah, Y.N. Aspect extraction in sentiment analysis: Comparative analysis and survey. *Artif. Intell. Rev.* **2016**, 1–25. [CrossRef]
59. Rose, S.; Engel, D.; Cramer, N.; Cowley, W. Automatic Keyword Extraction from Individual Documents. *Text Min. Appl. Theory* **2010**, 1–20. [CrossRef]
60. Bougouin, A.; Boudin, F.; Daille, B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In Proceedings of the International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October 2013; pp. 543–551.
61. Ning, J.; Liu, J. Using Word2vec with TextRank to Extract Keywords. *New Technol. Libr. Inf. Serv.* **2016**, *32*, 20–27. [CrossRef]
62. Xun, G.; Li, Y.; Zhao, W.X.; Gao, J.; Zhang, A. A Correlated Topic Model Using Word Embeddings. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, Australian, 19–25 August 2017; pp. 4207–4213.
63. Alam, M.H.; Ryu, W.-J.; Lee, S. Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Inf. Sci.* **2016**, *339*, 206–223. [CrossRef]
64. Yao, L.; Zhang, Y.; Chen, Q.; Qian, H.; Wei, B.; Hu, Z. Mining coherent topics in documents using word embeddings and large-scale text data. *Eng. Appl. Arti. Intell.* **2017**, *64*, 432–439. [CrossRef]
65. Moody, C.E. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv* **2016**, arXiv:1605.02019.
66. Wang, Z.; Ma, L.; Zhang, Y. A hybrid document feature extraction method using latent Dirichlet allocation and word2vec. In Proceedings of the 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), Changsha, China, 13–16 June 2016; pp. 98–103.
67. Cao, Z.; Li, S.; Liu, Y.; Li, W.; Ji, H. A novel neural topic model and its supervised extension. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 2210–2216.
68. Lau, J.H.; Baldwin, T.; Cohn, T. Topically driven neural language model. *arXiv* **2017**, arXiv:1704.08012.
69. He, R.; Lee, W.S.; Ng, H.T.; Dahlmeier, D. An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 388–397.
70. Qiu, L.; Yu, J. CLDA: An effective topic model for mining user interest preference under big data background. *Complexity* **2018**, *2018*. [CrossRef]
71. Zheng, C.T.; Liu, C.; San Wong, H. Corpus-based topic diffusion for short text clustering. *Neurocomputing* **2018**, *275*, 2444–2458. [CrossRef]
72. Li, X.; Zhang, A.; Li, C.; Ouyang, J.; Cai, Y. Exploring coherent topics by topic modeling with term weighting. *Inf. Process. Manag.* **2018**, *54*, 1345–1358. [CrossRef]
73. Liang, Y.; Liu, Y.; Chen, C.; Jiang, Z. Extracting topic-sensitive content from textual documents—A hybrid topic model approach. *Eng. Appl. Artif. Intell.* **2018**, *70*, 81–91. [CrossRef]
74. Xu, Y.; Yin, J.; Huang, J.; Yin, Y. Hierarchical topic modeling with automatic knowledge mining. *Expert Syst. Appl.* **2018**, *103*, 106–117. [CrossRef]
75. Afzaal, M.; Usman, M.; Fong, A.C.M.; Fong, S.; Zhuang, Y. Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews. *Adv. Fuzzy Syst.* **2016**, *2016*, 1–14. [CrossRef]
76. Tang, B.; Kay, S.; He, H. Toward optimal feature selection in naive Bayes for text categorization. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2508–2521. [CrossRef]
77. Hamzah, A.; Widyastuti, N. Opinion classification using maximum entropy and K-means clustering. In Proceedings of the 2016 International Conference on Information & Communication Technology and Systems (ICTS), Surabaya, Indonesia, 12 October 2016; pp. 162–166.
78. Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Syst. Appl.* **2016**, *66*, 245–260. [CrossRef]
79. An, J.; Chen, Y.P. Keyword extraction for text categorization. In Proceedings of the Active Media Technology, Kagawa, Japan, 19–21 May 2005; pp. 556–561.

80. Hu, J.; Li, S.; Yao, Y.; Yu, L.; Yang, G.; Hu, J. Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy* **2018**, *20*, 104. [CrossRef]
81. Hu, J.; Li, S.; Hu, J.; Yang, G. A Hierarchical Feature Extraction Model for Multi-Label Mechanical Patent Classification. *Sustainability* **2018**, *10*, 219. [CrossRef]
82. Ogada, K.; Mwangi, W.; Cheruiyot, W. N-gram Based Text Categorization Method for Improved Data Mining. *J. Inf. Eng. Appl.* **2015**, *5*, 35–43.
83. Zhang, H.; Zhong, G. Improving short text classification by learning vector representations of both words and hidden topics. *Knowl. Based Syst.* **2016**, *102*, 76–86. [CrossRef]
84. Zhang, X.; Zhao, J.J.; Lecun, Y. Character-level convolutional networks for text classification. *Neural Inf. Process. Syst.* **2015**, 649–657. [CrossRef]
85. Li, S.; Hu, J.; Cui, Y.; Hu, J. DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics* **2018**, *117*, 721–744. [CrossRef]
86. Zhou, P.; Qi, Z.; Zheng, S.; Xu, J.; Bao, H.; Xu, B. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. *Int. Conf. Comput. Ling.* **2016**, 3485–3495.
87. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
88. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, QLD, Australia, 19–24 April 2015; pp. 4580–4584.
89. Liu, P.; Qiu, X.; Huang, X. Recurrent neural network for text classification with multi-task learning. *arXiv* **2016**, arXiv:1605.05101.
90. Conneau, A.; Schwenk, H.; Barrault, L.; Lecun, Y.; Conneau, A.; Schwenk, H.; Barrault, L.; Lecun, Y. Very Deep Convolutional Networks for Text Classification. *Comput. Sci.* **2016**, 1107–1116. [CrossRef]
91. Katz, G.; Caragea, C.; Shabtai, A. Vertical Ensemble Co-Training for Text Classification. *ACM Trans. Intell. Syst. Technol. TIST* **2018**, *9*, 21. [CrossRef]
92. Zhu, W.; Liu, Y.; Hu, G.; Ni, J.; Lu, Z. A Sample Extension Method Based on Wikipedia and Its Application in Text Classification. *Wirel. Pers. Commun.* **2018**, *102*, 3851–3867. [CrossRef]
93. Jiang, X.; Havaei, M.; Chartrand, G.; Chouaib, H.; Vincent, T.; Jesson, A.; Chapados, N.; Matwin, S. On the Importance of Attention in Meta-Learning for Few-Shot Text Classification. *arXiv* **2018**, arXiv:1806.00852.
94. Merity, S.; Keskar, N.S.; Socher, R. Regularizing and optimizing LSTM language models. *arXiv* **2017**, arXiv:1708.02182.
95. Zheng, X.; Luo, Y.; Sun, L.; Ji, Z.; Chen, F. A tourism destination recommender system using users' sentiment and temporal dynamics. *J. Intell. Inf. Syst.* **2018**, 1–22. [CrossRef]
96. Serna, A.; Gerrikagoitia, J.K.; Bernabe, U.; Ruiz, T. A Method to Assess Sustainable Mobility for Sustainable Tourism: The Case of the Public Bike Systems. In Proceedings of the Enter Conference | Etourism: Sustaining Culture & Creativity Organized by International Federation for Information Technology & Travel & Tourism, Rome, Italy, 24–26 January 2017.
97. Li, Q.; Wu, Y.; Wang, S.; Lin, M.; Feng, X.; Wang, H. VisTravel: Visualizing tourism network opinion from the user generated content. *J. Vis.* **2016**, *19*, 489–502. [CrossRef]
98. Zong, C. *Statistical Natural Language Processing*; Tsinghua University Press: Beijing, China, 2013.
99. Zhao, Y.; Qin, B.; Liu, T. Sentiment Analysis. *J. Softw.* **2010**, *21*, 1834–1848. [CrossRef]
100. Chen, Z.; Xu, R.; Gui, L.; Lu, Q. Combining Convolutional Neural Networks and Word Sentiment Sequence Features for Chinese Text Sentiment Analysis. *J. Chin. Inf. Process.* **2015**, *29*, 172–178.
101. Fu, Y.; Hao, J.-X.; Li, X.; Hsu, C.H. Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News. *J. Travel Res.* **2018**, 0047287518772361. [CrossRef]
102. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
103. Santos, C.N.D.; Gatti, M.A.D.C. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In Proceedings of the International Conference on Computational Linguistics, Dublin, Ireland, 23–29 August 2014; pp. 69–78.
104. Dieng, A.B.; Wang, C.; Gao, J.; Paisley, J. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv* **2016**, arXiv:1611.01702.

105. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. *Meet. Assoc. Comput. Ling.* **2014**, 655–665.
106. Hassan, A.; Mahmood, A. Deep Learning approach for sentiment analysis of short texts. In Proceedings of the International Conference on Control and Automation, Ohrid, Macedonia, 3–6 July 2017; pp. 705–710.
107. Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; Qin, B. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In Proceedings of the Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 1555–1565.
108. Ambartsoumian, A.; Popowich, F. Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers. *Empir. Methods Nat. Lang. Process.* **2018**, 130–139. [CrossRef]
109. Jiang, T.; Wan, C.; Liu, D. Extracting Target-Opinion Pairs Based on Semantic Analysis. *Chin. J. Comput.* **2017**, *40*, 617–633.
110. He, W.; Tian, X.; Tao, R.; Zhang, W.; Yan, G.; Akula, V. Application of social media analytics: A case of analyzing online hotel reviews. *Online Inf. Rev.* **2017**, *41*, 921–935. [CrossRef]
111. Hu, C.; Liang, N. Deeper attention-based LSTM for aspect sentiment analysis. *Appl. Res. Comput.* **2019**, *36*.
112. Fu, X.; Wei, Y.; Xu, F.; Wang, T.; Lu, Y.; Li, J.; Huang, J.Z. Semi-supervised Aspect-level Sentiment Classification Model based on Variational Autoencoder. *Knowl. Based Syst.* **2019**, *171*, 81–92. [CrossRef]
113. Wang, Y.; Huang, M.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
114. Tay, Y.; Tuan, L.A.; Hui, S.C. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018.
115. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for target-dependent sentiment classification. *arXiv* **2015**, arXiv:1512.01100.
116. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. *arXiv* **2017**, arXiv:1709.00893. [CrossRef]
117. Tang, D.; Qin, B.; Liu, T. Aspect level sentiment classification with deep memory network. *arXiv* **2016**, arXiv:1605.08900. [CrossRef]
118. Yang, C.; Zhang, H.; Jiang, B.; Li, K. Aspect-based sentiment analysis with alternating coattention networks. *Inf. Process. Manag.* **2019**, *56*, 463–478. [CrossRef]
119. Liu, J.; Zhang, Y. Attention modeling for targeted sentiment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2, Short Papers), Valencia, Spain, 3–7 April 2017; pp. 572–577.
120. Li, X.; Bing, L.; Lam, W.; Shi, B. Transformation networks for target-oriented sentiment classification. *arXiv* **2018**, arXiv:1805.01086.
121. Shuang, K.; Ren, X.; Yang, Q.; Li, R.; Loo, J. AELA-DLSTMs: Attention-Enabled and Location-Aware Double LSTMs for aspect-level sentiment classification. *Neurocomputing* **2019**, *334*, 25–34. [CrossRef]
122. Ma, X.; Zeng, J.; Peng, L.; Fortino, G.; Zhang, Y. Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. *Fut. Gen. Comput. Syst.* **2019**, *93*, 304–311. [CrossRef]
123. Zhang, Z.; Wang, L.; Zou, Y.; Gan, C. The optimally designed dynamic memory networks for targeted sentiment classification. *Neurocomputing* **2018**, *309*, 36–45. [CrossRef]
124. Fan, F.; Feng, Y.; Zhao, D. Multi-grained attention network for aspect-level sentiment classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3433–3442.
125. Xu, H.; Liu, B.; Shu, L.; Yu, P.S. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. *arXiv* **2019**, arXiv:1904.02232.
126. Yang, M.; Yin, W.; Qu, Q.; Tu, W.; Shen, Y.; Chen, X. Neural Attentive Network for Cross-Domain Aspect-level Sentiment Classification. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
127. Almars, A.; Li, X.; Zhao, X. Modelling user attitudes using hierarchical sentiment-topic model. *Data Knowl. Eng.* **2019**, *119*, 139–149. [CrossRef]
128. Li, J.; Yujie, C.; Zhao, Z. Tibetan Tourism Hotspots: Co-word Cluster Analysis of English Blogs. *Tour. Trib.* **2015**, *30*, 35–43.

129. Ding, S.; Gong, S.; Li, H. A New Method to Detect Bursty Events from Micro-blog Posts Based on Bursty Topic Words and Agglomerative Hierarchical Clustering Algorithm. *New Technol. Libr. Inf. Serv.* **2016**, *32*, 12–20. [CrossRef]
130. Celardo, L.; Iezzi, D.F.; Vichi, M. Multi-mode partitioning for text clustering to reduce dimensionality and noises. In Proceedings of the 13th International Conference on Statistical Analysis of Textual Data, Nice, France, 7–10 June 2016.
131. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Min. Text Data* **2017**. [CrossRef]
132. Huang, L.-J.; Cheng, M.-Z.; Xiao, Y. Text Clustering Algorithm Based on Random Cluster Core. In Proceedings of the ITM Web of Conferences, Julius, France; 2016; p. 05001.
133. Xiong, C.; Hua, Z.; Lv, K.; Li, X. An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers. In Proceedings of the International Conference on Cloud Computing & Big Data, Macau, China, 16–18 November 2016.
134. Huan, Z.; Pengzhou, Z.; Zeyang, G. K-means Text Dynamic Clustering Algorithm Based on KL Divergence. In Proceedings of the 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore, 6 June 2018; pp. 659–663.
135. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering. In Proceedings of the 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–15 July 2016; pp. 1–6.
136. Jin, C.X.; Bai, Q.C. Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence. In Proceedings of the International Conference on Information System and Artificial Intelligence, Hangzhou, China, 14–16 July 2017; pp. 497–502.
137. Wang, B.; Liu, W.; Lin, Z.; Hu, X.; Wei, J.; Liu, C. Text clustering algorithm based on deep representation learning. *J. Eng.* **2018**, *2018*, 1407–1414. [CrossRef]
138. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Multi-objectives-based text clustering technique using K-mean algorithm. In Proceedings of the International Conference on Computer Science & Information Technology, Amman, Jordan, 13–15 July 2016.
139. Aggarwal, C.C.; Zhai, C. A survey of text clustering algorithms. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 77–128.
140. Yu, J.C.X. Ontology Concepts Clustering Based on Encyclopedia Entr. *J. Univ. Electron. Sci. Technol. China* **2017**, *46*, 636–640.
141. Horner, S.; Swarbrooke, J. *Consumer Behaviour in Tourism*; Routledge: London, UK, 2016.
142. Alén, E.; Losada, N.; Domínguez, T. The Impact of Ageing on the Tourism Industry: An Approach to the Senior Tourist Profile. *Soc. Indic. Res.* **2016**, *127*, 1–20. [CrossRef]
143. Liu, Y.; Huang, K.; Bao, J.; Chen, K. Listen to the voices from home: An analysis of Chinese tourists' sentiments regarding Australian destinations. *Tour. Manag.* **2019**, *71*, 337–347. [CrossRef]
144. Ezeuduji, I.O.; November, K.L.; Haupt, C. Tourist Profile and Destination Brand Perception: The Case of Cape Town, South Africa. *Acta Univ. Danub. Oeconomica* **2016**, *12*, 115–132.
145. Padilla, J.J.; Kavak, H.; Lynch, C.J.; Gore, R.J.; Diallo, S.Y. Temporal and Spatiotemporal Investigation of Tourist Attraction Visit Sentiment on Twitter. *PLoS ONE* **2018**, *13*, e0198857. [CrossRef]
146. Pan, M.H.; Yang, X.X.; Pan, Z. Influence Factors of the Old-age Care Tourism Decision Making Behavior based on the Life Course Theory: A Case of Chongqing. *Hum. Geogr.* **2017**, *6*, 154–160. [CrossRef]
147. Qi, S.; Wong, C.U.I.; Chen, N.; Rong, J.; Du, J. Profiling Macau cultural tourists by using user-generated content from online social media. *Inf. Technol. Tour.* **2018**, 1–20. [CrossRef]
148. Zheng, X.; Luo, Y.; Xu, Z.; Yu, Q.; Lu, L. Tourism Destination Recommender System for the Cold Start Problem. *KSII Trans. Internet Inf. Syst.* **2016**, *10*. [CrossRef]
149. Leal, F.; González–Vélez, H.; Malheiro, B.; Burguillo, J.C. Semantic profiling and destination recommendation based on crowd-sourced tourist reviews. In Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence, Porto, Portugal, 21–23 June 2017; pp. 140–147.
150. Rossetti, M.; Stella, F.; Cao, L.; Zanker, M. *Analysing User Reviews in Tourism with Topic Models*; Springer International Publishing: Lugano, Switzerland, 2015; pp. 47–58.
151. Borràs, J.; Moreno, A.; Valls, A. Intelligent tourism recommender systems: A survey. *Expert Syst. Appl.* **2014**, *41*, 7370–7389. [CrossRef]

152. Qiao, X.; Zhang, L. Overseas Applied Studies on Travel Recommender System in the Past Ten Years. *Tour. Trib.* **2014**. [CrossRef]
153. Batat, W.; Phou, S. *Building Understanding of the Domain of Destination Image: A Review*; Springer International Publishing: Atlanta, GA, USA, 2016.
154. Dickinger, A.; Költringer, C.; Körbitz, W. *Comparing Online Destination Image with Conventional Image Measurement—The Case of Tallinn*; Springer: Vienna, Austria, 2011; pp. 165–177.
155. Gunn, C.A. *Vacationscape: Designing Tourist Regions*; Van Nostrand Reinhold: New York, NY, USA, 1988.
156. Castro, J.C.; Quisimalin, M.; de Pablos, C.; Gancino, V.; Jerez, J. Tourism Marketing: Measuring Tourist Satisfaction. *J. Serv. Sci. Manag.* **2017**, *10*, 280. [CrossRef]
157. San Martín, H.; Herrero, A.; García de los Salmones, M.d.M. An integrative model of destination brand equity and tourist satisfaction. *Curr. Iss. Tour.* **2018**, 1–22. [CrossRef]
158. Kim, J.; Bae, J.; Hastak, M. Emergency information diffusion on online social media during storm Cindy in US. *Int. J. Inf. Manag.* **2018**, *40*, 153–165. [CrossRef]
159. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
160. Liao, X.; Wu, X.; Gui, L.; Huang, J.; Chen, G. Cross-Domain Sentiment Classification Based on Representation Learning and Transfer Learning. *Beijing Da Xue Xue Bao* **2019**, *55*, 37–46. [CrossRef]
161. Yu, M.; Guo, X.; Yi, J.; Chang, S.; Potdar, S.; Cheng, Y.; Tesauro, G.; Wang, H.; Zhou, B. Diverse Few-Shot Text Classification with Multiple Metrics. *arXiv* **2018**, arXiv:1805.07513. [CrossRef]
162. Lampinen, A.; Mcclelland, J.L. One-shot and few-shot learning of word embeddings. *arXiv* **2017**, arXiv:1710.10280.
163. Gu, J.; Wang, Y.; Chen, Y.; Li, V.O.K.; Cho, K. Meta-Learning for Low-Resource Neural Machine Translation. *Empir. Methods Nat. Lang. Process.* **2018**, 3622–3631.
164. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimed. Tools Appl.* **2018**, *77*, 283–326. [CrossRef]
165. Pouli, V.; Kafetzoglou, S.; Tsiropoulou, E.E.; Dimitriou, A.; Papavassiliou, S. Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience. In Proceedings of the 13th International Conference on Telecommunications (ConTEL), Graz, Austria, 13–15 July 2015; pp. 1–8.
166. Zhang, H.; Yu, H.; Xu, W. Listen, interact and talk: Learning to speak via interaction. *arXiv* **2017**, arXiv:1705.09906.
167. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
168. Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [CrossRef]
169. Quan, H.; Li, S.; Hu, J. Product Innovation Design Based on Deep Learning and Kansei Engineering. *Appl. Sci.* **2018**, *8*, 2397. [CrossRef]
170. Li, W.; Guo, K.; Shi, Y.; Zhu, L.; Zheng, Y. DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowl. Based Syst.* **2018**, *146*, 203–214. [CrossRef]