information

MDPI

Article

Aspect Term Extraction Based on MFE-CRF

Yanmin Xiang, Hongye He and Jin Zheng *

School of Information Science and Engineering, Central South University, Changsha 410000, China; xymcsu@foxmail.com (Y.X.); csumm@foxmail.com (H.H.)

* Correspondence: 164612192@csu.edu.cn; Tel.: +86-0731-8887-9628

Received: 4 July 2018; Accepted: 2 August 2018; Published: 3 August 2018

Abstract: This paper is focused on aspect term extraction in aspect-based sentiment analysis (ABSA), which is one of the hot spots in natural language processing (NLP). This paper proposes MFE-CRF that introduces Multi-Feature Embedding (MFE) clustering based on the Conditional Random Field (CRF) model to improve the effect of aspect term extraction in ABSA. First, Multi-Feature Embedding (MFE) is proposed to improve the text representation and capture more semantic information from text. Then the authors use kmeans++ algorithm to obtain MFE and word clustering to enrich the position features of CRF. Finally, the clustering classes of MFE and word embedding are set as the additional position features to train the model of CRF for aspect term extraction. The experiments on SemEval datasets validate the effectiveness of this model. The results of different models indicate that MFE-CRF can greatly improve the Recall rate of CRF model. Additionally, the Precision rate also is increased obviously when the semantics of text is complex.

Keywords: sentiment analysis; aspect term extraction; MFE; CRF; kmeans++ algorithm

1. Introduction

Sentiment analysis is one of the hot spots of natural language processing (NLP), which aims to mine opinions, sentiments, and emotions based on observations of people's actions that can be captured using their writing [1]. The early sentiment analysis focuses on the coarse-grained emotion analysis at the document-level and sentence-level, such as the sentiment polarity of documents, or the subjectivity and opinion mining of sentences. The coarse-grained emotion analysis tries to detect the overall polarity (negative or positive) of a sentence (or a document) regardless of the target entities and their aspects [2]. Deepening of the research allows for aspect-based sentiment analysis (ABSA) development [3]. Its target is to analyze the emotion of text from multiple angles, extract the emotion entities and assess the emotion polarity of the target entities. ABSA can capture more precise information in the text [4]. In restaurant reviews, the document-level sentiment analysis only makes a judgment on the whole document, but ABSA requires identifying the entities that are endowed with reviews, such as food taste and service quality. Thus, ABSA has a great reference value to create a recommendation system for e-commerce services and commodities classification [5].

Improving text representation is an important way to improve the effect of NLP tasks. In 2013, Mikolov put forward the word embedding models CBOW and Skip-Gram [6,7], then the researchers of text representation turned to word embedding. One-hot encoding is a $1 \times N$ matrix (vector) used to distinguish each word in a vocabulary. The vector consists of 0s in all cells with the exception of a single 1 in a cell used uniquely to identify the word [8], which will cause the space of word vector to become sparse when there are more features. Due to the sparseness of the one-hot encoding and the neglect of the semantic information, the method of traditional natural language processing based on Bag of Words hits a bottleneck in NLP tasks. Word embedding compresses words into a low-dimensional continuous space and overcomes the above disadvantages, which causes the

related studies concerning word embedding to play important roles in the NLP conferences. However, word embedding still has some disadvantages, such as the loss of context semantic meaning. Therefore, the further improvement of word vector has become a research focus. The improvement of ABSA in this paper is also based on strengthening the text representation.

To get more granular text emotional information, Choi et al. first extracted the opinion source by Conditional Random Field (CRF) in 2005 [9], which laid the foundation for aspect-based sentiment analysis (ABSA). ABSA was divided into aspect term extraction and aspect term polarity. Aspect term extraction is the problem of sequence labeling. However, CRF can count word frequency and extract contextual information of sentences, which makes CRF superior to ABSA. Jakob et al., in 2010, proposed a model that used CRF to extract opinion entities from data sets in the single domains and across domains [10]. Miao et al. used a probability model to reinforce the effect of CRF on the basis of domain information [11]. The DLIREC system won the best grade of aspect term extraction in competition, which used word embedding clustering as the position features of CRF. The ability of CRF to recognize aspect term has been demonstrated [12]. Since internet text is constantly expanding, ABSA applications in specific tasks are frequent [13,14] and the challenges in ABSA are also gradually displayed [15], such as the low Recall ratio of recognition in aspect term and difficulty in conjunction recognition.

The purpose of this paper is to study aspect term extraction of ABSA. The main innovation points are improved text representation and additional position features of CRF. The main works of this paper are as follows:

- Put forward Multi-Feature Embedding (MFE) model: this model strengthens the text based on distributed word embedding to overcome the weakness of insufficient contextual information in word embedding.
- Put forward MFE-CRF: MFE clustering based on word embedding is used to strengthen the effect of CRF in aspect term extraction. MFE clustering classes are obtained by Kmeans++ algorithm and the clustering classes are set as the additional position features to strengthen the effect of CRF. It proves that MFE clustering can significantly improve the effect of CRF in aspect term extraction through experiments.

2. Related Work

2.1. Aspect Term Extraction

The aspect-based sentiment analysis (ABSA) project joined the SemEval international semantic competition in 2014. There are two subtasks that are aspect term extraction (ATE) and term polarity classification (TPC). ATE is significantly meaningful to aspect-based sentiment analysis, which aims to identify the aspect terms in the given domain [16].

The influence of SemEval has led to the high attention on ABSA in recent years. Poria et al. used the rule of dependent syntax to extract the dependency of words in the text to improve the effect of ABSA [17]. Khan et al. pointed out the effectiveness of the topic model in solving ABSA tasks [18]. Poria et al. also took advantage of the idea of topic model and made topic model clustering as position features to identify emotional entities [19]. Schouten et al. analyzed in detail the semantic features that are beneficial to aspect term extraction and proposed an improvement scheme based on external dictionaries [20]. H. Wen et al. chose the traditional method combined with machine learning to extract the aspect terms of e-commerce reviews [5]. Wu et al. proposed a hybrid unsupervised method that combined rules and machine learning to address ATE tasks [21]. Manek et al. proposed a Gini Index that combined a features selection method with Support Vector Machine (SVM) for sentiment classification of a large movie review dataset [22]. A Weichselbraun et al. pursued a flexible and automated strategy by linking input entities to DBpedia to obtain background knowledge of these entities (aspects) [23].

Compared with these related works, the current approach differs from them significantly. First, to solve the problem that word embedding ignores the latent contextual information of text, the authors improve the distributed representation of text to extract additional latent semantic through Multi-Feature Embedding (MFE). Second, clustering classes are obtained by kmeans++ algorithm and then clustering classes are used as additional position features to strengthen the effect of CRF. MFE-CRF can extract useful information from complex semantics and shows effectiveness in ATE by the experimental results.

2.2. Aspect Term Extraction Based on BIO

Aspect term extraction is one of the sub-tasks of aspect-based sentiment analysis (ABSA). Its goal is to identify and extract entities in the text that are endowed with review information.

2.2.1. Sequence Representation of Aspect Term

Sequence labeling algorithms in machine learning are widely used in the sequence prediction, such as the Hidden Markov Model (HMM) [24] and Conditional Random Field (CRF) [25]. Text should be transformed into a corresponding set of observation sequences and state sequences in sequence prediction. The observation sequence is the list of words in the text, and the state sequence is the mark list including B, I and O [26], where B stands for the beginning word of the aspect term, I stands for the continuation word of the aspect term and O stands for the non-aspect term.

2.2.2. Sequence Prediction Based on CRF

Due to the good performance of Conditional Random Field (CRF) in sequence prediction, many models selected CRF as the algorithm to predict labels [9,10,12].

Regarding the observation sequence X and the state sequence Y, the problem of sequence labeling is to find the optimal conditional probability distribution $P(Y|X)$ [9].

It is assumed that the joint probability distribution of state sequence Y constitutes the undirected graph model (Markov Random Field), and each node in the state sequence satisfies the conditional independence. Given the observation sequence (input sequence) X, the conditional probability distribution model $P(Y|X)$ of state sequence (output sequence) Y is called the conditional random field (CRF). Therefore, for each node Y_i in the output sequence:

$$P\left(Y_i|X, Y_i, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n\right) = P\left(Y_i \mid X, Y_{i-1}, Y_{i+1}\right), \quad (1)$$

The undirected graph in CRF is constructed on the joint distribution probability of state sequence Y. The set of two adjacent points on sequence Y is the maximum clique $C_i = \{Y_{i-1}, Y_i\}$. Given the observation sample sequence $x = [x_1, x_2, \cdots, x_n]$ and the corresponding output sequence $y = [y_1, y_2, \cdots, y_n]$, the potential function of maximum clique can be expressed as:

$$\Psi_{C_i} = exp\left(\sum_k \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_l \mu_l S_l\left(y_i, x, i\right)\right), \quad (2)$$

The conditional probability and normalized factor can be expressed as:

$$P\left(Y|X\right) = \frac{exp\left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l} \mu_l s_l\left(y_i, x, i\right)\right)}{Z}, \quad (3)$$

$$Z = \sum_y exp\left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l} \mu_l s_l\left(y_i, x, i\right)\right), \quad (4)$$

Using the formula, t_k and s_l are the transfer and state feature functions respectively. λ_k and μ_l are the corresponding feature coefficients. Concerning the observation sequence x of test data, the sequence labeling is to find the output sequence $Y = y$ that maximizes the conditional probability

$P(Y \mid X = x)$. The key to solve text sequence labeling by CRF is to construct the position features on the text. Extracting the most effective position features of text is an affirmation for the good effect of CRF.

3. CRF Model Based on MFE Clustering Reinforcement

3.1. MFE

Multi-Feature Embedding (MFE) is an improved model based on word embedding. First, this model captures the effective semantic features in the original texts and then converts the texts into feature sequences. Second, the feature sequences are used as the input of the word embedding training model. Finally, the obtained semantic features are mapped to MFE for reinforcement.

3.1.1. Semantic Capture

- **Word-POS:** Traditional word embedding models cannot recognize the polysemy of words. There are two sentences "jack works hard every day" and "his works are amazing". The meaning of "works" is different in the two sentences. The former is "work", but the latter is "production". To eliminate ambiguity, MFE combines Part-of-Speech (POS) with word to get the Word-POS feature. The authors specify a text $D = [w_1, w_2, \cdots, w_l]$ contains l words, and then $P = [p_1, p_2, \cdots, p_l]$ is obtained by part-of-speech tagging. Finally, combine D and P to get $WP = [(w_1, p_1), (w_2, p_2), \cdots, (w_l, p_l)]$. "Works", will be distinguished by verb and noun in Word-POS. Table 1 shows the comparison between traditional word embedding model and the Word-POS model regarding similar elements in IMDB (http://ai.stanford.edu/$\sim$amaas/data/sentiment/) dataset (part of speech tagging is done by NLTK toolkit (http://www.nltk.org/#), word embedding is trained by Skip-Gram). The "nn" means noun and "vb" means verb in the Table 1.

Table 1. The comparison between traditional word embedding model and the Word-Part-of-Speech (POS) model regarding similar elements.

Word/Word-POS	The Closest Word of the Vector
works	work, worked, done, plays, working, crafted, quite
(works, nn)	(work, nn), (films, nn), (works, nn), (art, nn), (kundera, nn), (krzysztof, nn), (masterpiece, nn)
(works, vb)	(worked, vb), (work, vb), (work, vb), (succeed, vb), (working, vb), (plays, vb), (done, vb)

It can be clearly observed that in the Word-POS model, the words close to "works" in the verb form are verbs, such as "play", "succeed" and "worked", but "works" is similar to some nouns in the noun form, such as "films" and "art". Taking the word vector model, the polysemy of this word cannot be captured, and it will only regard "works" as a verb to a large extent.

- **Stemming:** To eliminate semantic redundancy, the stem of each word is extracted in the text. Then, the word stem sequence of the text is set as the input of the word embedding training model. Table 2 shows the comparison between traditional word embedding model and the stem model regarding similar elements.

Table 2. The comparison between traditional word embedding model and the stem model regarding similar elements.

Word/Stem	The Closest Word of the Vector
Run (word)	runs, running, ran, mill, afoul, walk, amok
Running (word)	run, runs, walking, run, minutes, around, screaming
Run (stem)	walk, chase, amok, go, afoul, get, wander

It can be seen from the table that the distance between the corresponding vectors of the same stem is very close. Therefore, there is semantic redundancy in the traditional word vector model. Stemming can eliminate this phenomenon by normalization and the more compact semantics of words in the text.

3.1.2. Training of MFE

- **Step1.** Sequence transformation. Prior to the semantic features being mapped to the features vector, the text needs to be transformed into a features sequence. Table 3 shows an example of sequence transformation. The "nn" means noun, "vb" means verb and "jj" means adjective in the Table 3.
- **Step2.** Vector mapping. Subsequent to the text being transformed into the features sequence, it can be used as the input data of the word embedding training model. Through the training process of word embedding, the semantic features sequence can be mapped to the MFE vector. MFE assumes that the similar semantic text features have the similar structures of context. The window slides on the features sequence can map texts that have similar semantic features to vectors with similar distance. The overall structure of the training model is shown in Figure 1.

Table 3. Example of Sequence Transformation.

	Sequence
Original Sequence	Natural language processing is a field concerned with the interactions between computers and human languages.
Pretreatment	natural language processing field concerned interactions computers human languages
Word-POS	(natural, jj) (language, nn) (processing, nn) (field, nn) (concerned, vb) (interactions, nn) (computers, nn) (human, jj) (languages, nn)
Stemming	natur languag process field concern interact comput human languag

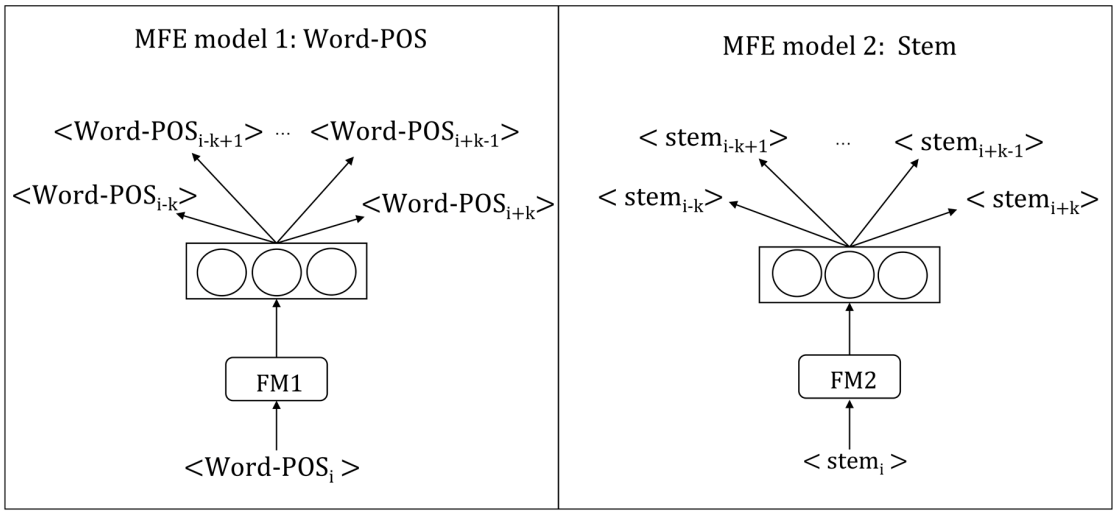

Figure 1. The training of Multi-Feature Embedding (MFE) (Skip-Gram).

3.2. Features of MFE-CRF

The position features of text have a direct influence on the performance of Conditional Random Field (CRF), this section shows the position features used for the aspect term extraction task.

3.2.1. General Features

Considering the aspect term extraction as a BIO sequence labeling problem, the extraction of position features is turned to obtain the features of a word in the text sequence. The following basic position features are extracted as the baseline features:

- **Lower-case word:** Information of uppercase and lowercase needs to be considered separately; make lowercase as the general feature .
- **Suffix of word:** The suffix of a word can determine whether the word is aspect term. The last two and three words are extracted as the general features.
- **Capital:** Proper nouns and special words usually begin with uppercase words; they are more likely to be aspect terms.
- **Part-of-speech:** Aspect terms are usually nouns and other words with specific parts of speech, therefore, part-of-speech is one of the most important general features.
- **Stemming:** To get more compact sentences.
- **Dependent syntactic:**

  1. amod: Whether the word is modified by other words;
  2. nsubj: Whether the word is used as the subject for other words;
  3. dobj: Whether the word is used as a direct object for other words.

3.2.2. Cluster Features

Word vector is trained by CBOW and Skip-Gram [6,7], Multi-Feature Embedding (MFE) is obtained by the method mentioned in Section 3.1.2. Then, all the words in the corpus are mapped to the fixed length vector. Through clustering, all the words in the corpus can be grouped into the clustering class; words in each cluster will have similar semantics due to the close distances of vectors.

This paper selects Kmeans [27] and Kmeans++ [28] as the cluster algorithms. Kmeans++ is an algorithm for choosing the initial center points to avoid the poor clustering found by standard Kmeans algorithms. Using the initialization of Kmeans++, the algorithm is guaranteed to find the optimal result. Since there are two kinds of word vectors, the authors' process of clustering contains two parts: word clustering and MFE clustering.

1. Word Clustering

There are two kinds of word vectors that are trained by CBOW and Skip-Gram respectively, then they are input into Kmeans++ to find the clustering classes respectively. Finally, two kinds of word clustering models can be obtained. The general process of word clustering is as follows:

- **Step1.** Determine the value of k and choose one center uniformly at random among the sample set (all word vectors).
- **Step2.** Taking each word vector w_i, calculate the distance between the vector and the nearest center that has already been chosen by Formula (5). Specify two word vectors of length m $w_1 = [w_{1,1}, w_{1,2}, \cdots, w_{1,m}]$ and $w_2 = [w_{2,1}, w_{2,2}, \cdots, w_{2,m}]$. The distance between vectors can be calculated by Euclidean Distance:

$$\mathrm{d} = \sqrt{(w_{1,1} - w_{2,1})^2 + (w_{1,2} - w_{2,2})^2 + \cdots + (w_{1,m} - w_{2,m})^2}, \tag{5}$$

- **Step3.** Choose one new center. Calculate the selected probability $p(w_i)$ of each word vector w_i using Formula (6), then select the new center by roulette wheel selection:

$$p(w_i) = \frac{d^2(w_i)}{\sum_{w_i \in W} d^2(w_i)}, \tag{6}$$

- **Step4.** Repeat Step2 and Step3 until k centers have been chosen.
- **Step5.** Regarding all the word vectors, calculate the distance between the vector and k centers using Formula (5). Mark the clustering class of the center that is the closest to each point as the clustering class for this point.
- **Step6.** Update the centers of the k classes, specify that the word vectors set belonging to the i^{th} clustering class as D_i. The new centroid C_i of this class is calculated by the mean value of vectors in D_i:

$$C_i = \frac{\sum_{w \epsilon D_i} w}{|D_i|}, \tag{7}$$

Through the iteration of Step4 and Step5, the clustering of word vectors is completed until the centers of mass no longer change obviously or reach the number of iterations.

Following the completion of word clustering, the clustering classes can be extracted as the position features. To increase the richness of clustering features, this paper uses CBOW and Skip-Gram to train two sets of word vectors. Finally, two word clustering models can be extracted.

Concerning the input sequence, the corresponding word vector of each word is found. The closest word clustering class of the word vector is marked as the word clustering class of this word. Since there are two word clustering models, each word in the input sequence extracts two word clustering classes as the position features of CRF.

2. MFE Clustering

According to the description in Section 3.1, text can be transformed into additional semantic sequences. The stem sequence and Word-POS sequence have the same structure as the original word sequence, which means each word of a sentence can be presented by MFE and has a unique feature in the additional semantic sequence.

MFE contains Word-POS embedding and stem embedding. There are four kinds of feature embedding where two kinds of feature embedding are from Word-POS embedding and the others are from stem embedding. Then, they are input into Kmeans++ to get the clustering classes respectively. Finally, four kinds of word clustering models are obtained. The process of MFE clustering based on word clustering is as follows:

- **Step1.** According to the steps described in Section 3.1.2, train Word-POS embedding and stem embedding in corpus.
- **Step2**. Cluster the Word-POS embedding and stem embedding by the method used in Word Clustering.

Following the completion of MFE clustering, the clustering classes can be extracted as the position features. To increase the richness of clustering features, this paper uses CBOW and Skip-Gram to train four sets of MFE. Finally, four MFE clustering models can be extracted.

Concerning the input sequence, the corresponding MFE of each word is found. The closest MFE clustering class of the MFE is marked as the MFE clustering class of this word. Since there are four MFE clustering models, each word in the input sequence extracts four MFE clustering classes as the position features of CRF.

Finally, MFE-CRF extracted six clustering position features where two clustering position features are from the word clustering and the others were from the MFE clustering. Thus, each word of the input sequences can extract six clustering classes as the position features of CRF.

3.3. Process of MFE-CRF

Regarding the observation sequence and state sequence, take a phone comment about phones such as “I think the screen is good but its battery is weak to use”. Specify the “screen” and “battery” as the aspect term, then the state sequence is {‘O’, ‘O’,‘O’, ‘B’, ‘O’, ‘O’, ‘O’, ‘O’, ‘B’, ‘O’, ‘O’, ‘O’, ‘O’} and

the observation sequence is {'I', 'think', 'the', 'screen', 'is', 'good', 'but', 'its', 'battery', 'is', 'weak', 'to', 'use'}. The specific process is as follows:

- **Step1.** Extract the general features. According to the general features shown in Section 3.2.1, it needs to extract the features from each position of the observation sequences. Take the word "screen" for example, the general features dictionary for the word "screen" is $d_1\left('\text{screen}'\right)$ ={lower:'screen', lower[–2]:'en', lower[–3]:'een', isTitle: false, POS:'NN', stem: 'screen', amod: true, nsubj: true, dobj: false}.
- **Step2.** Extract the clustering features. According to the word and MFE clustering methods given in Section 3.2.2, six clustering models are constructed by Skip-Gram and CBOW. Get the clustering classes of each word w_i in the observation sequence and set it as $C_k(w_i)$, the clustering features dictionary of w_i is $d_2(w_i) = \left\{ \begin{array}{c} C_1 : C_1(w_i), C_2 : C_2(w_i), C_3 : C_3(w_i), \\ C_4 : C_4(w_i), C_5 : C_5(w_i), C_6 : C_6(w_i) \end{array} \right\}$.
- **Step3.** Construction of CRF model. Figure 2 shows the overall process of the MFE-CRF model. CRF model is trained by position features of the observation and state sequence. Once the model converges, the position features of the test text are taken as the input and then the BIO sequence of the test text can be predicted.

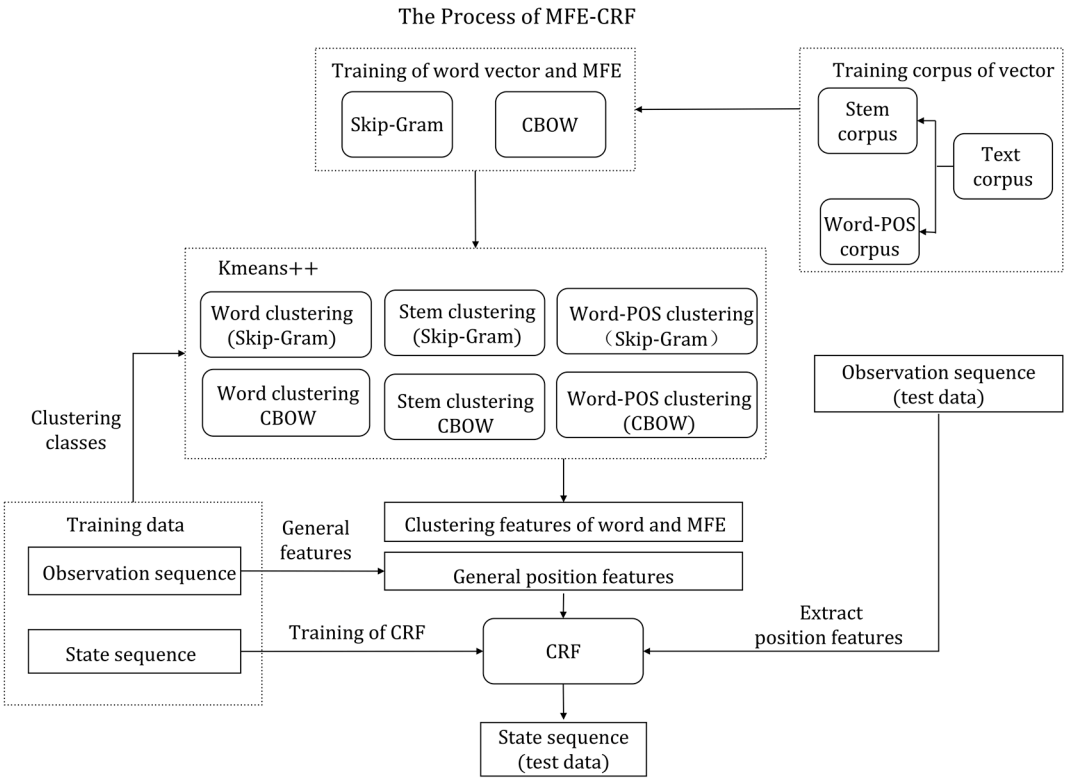

Figure 2. The process of MFE-Conditional Random Field (CRF).

Using the MFE-CRF, the clustering procedure constructs all semantic classes of MFE and word vectors. The word vectors within the same clustering class have similar semantics, so the clustering classes of the observation sequence are the effective position features of CRF. Additionally, there are different results for the clustering class through training vector using different training algorithms (such as Skip-Gram and CBOW), thus, the clustering classes are richness. The MFE-CRF proposed in this paper uses six clustering models to extract the clustering features of each word, which is to obtain a semantic class of words from multiple aspects so as to enrich the position features of CRF and get a better effect of training.

4. Experiments

This experiment used the open datasets and evaluation criteria of SemEval, which showed the different results between MFE-CRF and the models that used the general features. Concurrently, the selection of the best value of clustering was analyzed through experiments. (The code and data are publicly available at https://github.com/xymcsu/MFE-CRF)

4.1. Experiment Description

The research task is to label the positions of the aspect terms in the text and requires submission of a set of start and end positions of the aspect terms in the unlabeled test sets for evaluation.

The datasets used in the experiments are as follows:

- **SemEval2014** (http://alt.qcri.org/semeval2014/task4/index.php?id=important-dates): The data set includes laptop and restaurant trial data. Restaurant trial data contains 3041 pieces of train data and 800 pieces of test data; laptop trial data contains 3045 pieces of train data and 800 pieces of test data. During this experiment, all the train data sets were used for position features extraction and model training, and test data sets were used for the evaluation. L-14 means laptop dataset and R-14 means restaurant dataset in SemEval2014.
- **SemEval2015** (http://alt.qcri.org/semeval2015/task12/)/2016 (http://alt.qcri.org/semeval2016/task5/): The data sets include restaurant trial data that are similar to SemEval2014. R-15 contains 1363 pieces of train data and 685 pieces of test data; R-16 contains 2048 pieces of train data and 676 pieces of test data. During this experiment, all the training data sets were used for model training and test data sets were used for the additional evaluation.
- **Yelp dataset** (https://www.yelp.com/dataset): The training of word embedding and MFE requires a large number of training data, therefore the experiment used an additional yelp dataset for vector training to achieve a better effect. Yelp dataset contains 335,022 samples of restaurant-comment data; 200,000 comments were selected randomly for the expansion of vector training.
- **Amazon product data** (http://jmcauley.ucsd.edu/data/amazon/): The data set contains 1,689,188 electronic product reviews; the experiment selected randomly 200,000 comments containing the words "laptop" or "computer" to expand the training text.

4.2. Experiment Setup and Evaluation Measures

CRF models were trained by CRFsuite (http://www.chokkan.org/software/crfsuite/). The details of each model are as follows:

- **CRF1**: CRF1 was trained by the general features shown in Section 3.2.1. This model that was set as the baseline obtained the basic effect of aspect term extraction.
- **CRF2**: CRF2 model was reinforced by Skip-Gram word clustering on the basis of CRF1. The dimension of word embedding was 300, negative sample size was 5, the width of window was 10 and it ignored words that appeared less than 3 times.
- **CRF3**: CRF3 model was reinforced by CBOW word clustering on the basis of CRF2; the parameters of CBOW were consistent with Skip-Gram.
- **CRF3+Stem**: CRF3+Stem was reinforced by stem MFE clustering including CBOW and Skip-Gram on the basis of CRF3. The parameters of MFE were consistent with word embedding.
- **CRF3+WP**: CRF3+WP was reinforced by Word-POS MFE clustering, including CBOW and Skip-Gram on the basis of CRF3.
- **CRF3+ALL**: CRF3+ALL was reinforced by stem and Word-POS MFE clustering on the basis of CRF3.

The results of the experiments were evaluated with Precision, Recall and F1. Regarding each test text S, the position set of the aspect terms that are predicted is $w1(S)$, the position set of true

aspect terms is $w2(S)$, the number of elements in $w1$ is $pre(S)$ (the number of the aspect terms that are predicted), the number of elements in $w2$ is $true(S)$ (the number of true aspect terms), and the number of intersections between the two sets is $cor(S)$ (the number of the aspect terms that are predicted correctly). The test text set as D. The formulae of Precision Recall and F1 are as follow:

$$\text{Precision} = \frac{\sum_{S \in D} cor(S)}{\sum_{S \in D} pre(S)}, \tag{8}$$

$$\text{Recall} = \frac{\sum_{S \in D} cor(S)}{\sum_{S \in D} true(S)}, \tag{9}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Precision reflects the proportion of the aspect terms that are predicted correctly to the terms that are predicted. Recall reflects the ratio of the aspect terms that are predicted correctly to all the terms, and F1 is the harmonic mean of Precision and Recall.

4.3. Results Comparison and Analysis

4.3.1. Overall Assessment

The results of MFE-CRF and each benchmark model for restaurant and laptop datasets are shown in Table 4. Regarding aspect term extraction, the bottleneck of CRF models using the general features was the value of Recall. Concerning the laptop dataset, the Recall rate of CRF1 was only 56.7%, but the Precision reached 83.09%. The reason for this phenomenon is that the general features had a poor ability to recognize conjunctions but had a high recognition accuracy for aspect terms.

Table 4. Results of the Aspect Term Extraction.

Models	Restaurant			Laptop		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
CRF1	84.91	73.83	78.98	83.09	56.70	67.71
CRF2	86.53	79.52	82.88	84.37	60.71	70.61
CRF3	86.27	80.02	83.03	85.92	64.63	73.77
CRF3+Stem	86.24	81.33	83.71	87.07	67.02	75.74
CRF3+WP	86.19	82.16	83.05	87.64	66.39	75.55
CRF3+ALL	**86.41**	**82.35**	**84.33**	**87.81**	**67.82**	**76.53**

CRF2 and CRF3, which introduced word clustering, achieved a significant improvement compared with CRF1. F1 was improved by 3.9% and 2.9% respectively through CRF2, compared to CRF1 in the restaurant and laptop data set. Recall was increased by 5.69% and 4.01% in two data sets respectively. Word clustering as the position features captured more features of context, improved the accuracy of CRF in identifying aspect term and strengthened its ability to recognize conjunctions. Additionally, F1 was increased by 1.30% and 2.76% respectively through CRF3+ALL in comparison with CRF3. CRF3 combined with CBOW and Skip-Gram word clustering further improved the experimental effect, which proves that the CRF model with multi-clustering classes can obtain a better ability of identification. Compared with CRF1, CRF3+ALL increased Recall rate up to 8.52% and 11.12%, but the precision rate was only 1.5% and 4.72%. The improvement of MFE clustering in the laptop data set was more obvious because MFE clustering improved the recognition accuracy by identifying the semantics of proper nouns in the laptop data. To summarize, MFE clustering greatly improved the Recall rate of the CRF model and it also improved the Precision rate of recognition when the semantics were complex.

4.3.2. Comparison with Other Methods on F1

To further validate the performance of the current model on aspect term extraction, the authors selected some models to compare against this model:

- **HIS_RD, DLIREC(U), EliXa(U) and NLANGP(U):** HIS_RD was the best result of restaurant dataset in SemEval2014 [29], DLIREC(U) was the best result of laptop dataset in SemEval2014 [11]; EliXa(U) was the best result of restaurant dataset in SemEval2015 [30]; NLANGP(U) was the best result of restaurant dataset in SemEval2016 [31]. U means unconstrained and using additional resources without any constraint, such as lexicons or additional training data.
- **LSTM, Bi-LSTM and Deep LSTM:** These deep learning models were provided by Wu et al. [21].
- **MTCA:** MTCA was a multi-task attention model that learned shared information among different tasks [32].

Table 5 shows the results of these models in F1 on SemEval2014, 2015 and 2016 datasets. Compared to the best results in SemEval2014, 2015 and 2016, CRF3+ALL achieved 1.98%, 0.32%, 0.26% and 1.47% F1 score gains over HIS_RD, DLIREC(U), EliXa(U) and NLANGP(U) on L-14, R-14, R-15 and R-16. Compared with the basic models of deep learning, CRF3+ALL showed a significant advantage. F1 was increased by 12.26% and 6.66% respectively compared with the best system, Deep LSTM, in the basic models of deep learning, which indicates the current model has a better effect than the basic deep learning models. Even compared with MTCA, CRF3+ALL still gave the best score for L-14, R-14, R-16 and a competitive score for R-15. Overall, considering all the results, it can be seen that the current model can capture contextual semantic information effectively. The good performance in the laptops dataset reflects the current model has an improvement when the semantics of text is complex. Additionally, compared to other clustering features, CRF3+ALL can capture more semantic information to enrich features and show a big improvement in laptops and restaurants datasets.

Table 5. Comparison in F1 on SemEval2014, 2015 and 2016 datasets.

Models	L-14	R-14	R-15	R-16
HIS_RD	**74.55**	79.62		
DLIREC(U)	73.78	**84.01**	**-**	-
EliXa(U)	-	**-**	**70.05**	-
NLANGP(U)	-	-	67.12	**72.34**
LSTM	62.72	75.78	-	-
Bi-LSTM	61.98	78.63	-	-
Deep LSTM	64.27	77.67	-	-
MTCA	69.14	-	**71.31**	**73.26**
CRF3+Stem	75.74	83.71	68.74	72.73
CRF3+WP	75.55	83.05	69.42	73.49
CRF3+ALL	**76.53**	**84.33**	**70.31**	**73.81**

4.3.3. The Evaluation of K

Using Kmeans++ clustering, the selection of k is of vital importance. When the value of k is too high it leads to low efficiency and over-fitting, while if the value of k is too low it leads to inaccurate clustering results. Thus, the choice of k directly affects the recognition of aspect terms. Selected 10–300 as the value of k to cluster the word vector and MFE. The experiments of aspect term extraction were conducted for each model and F1 values were used for comparison in R-14 and L-14. Figures 3 and 4 show the changing trend of F1 with the increase of k.

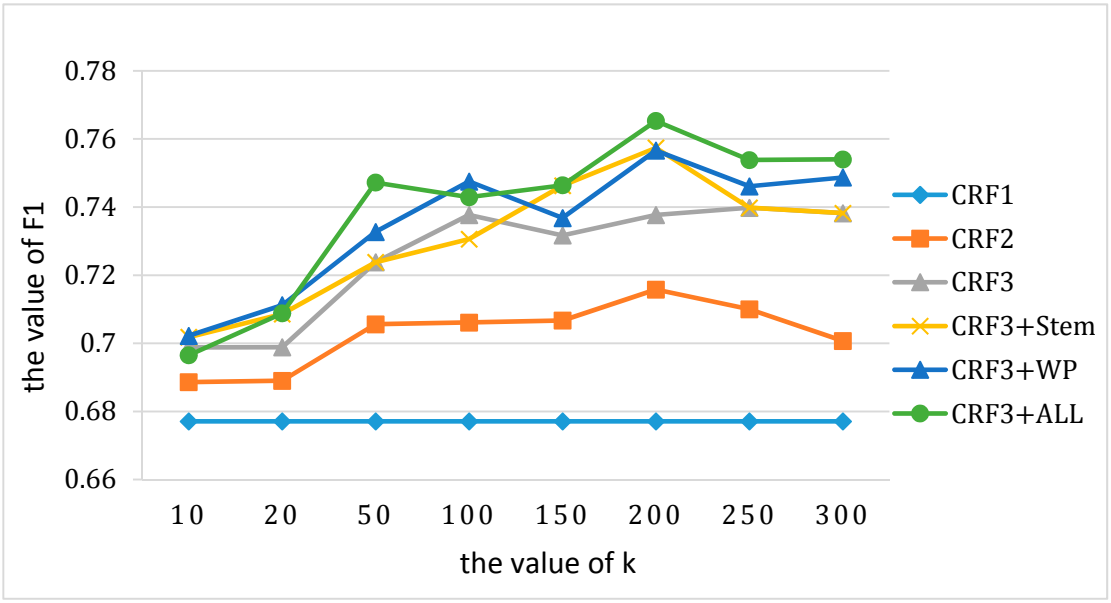

Figure 3. The effect of K on experimental results (R-14).

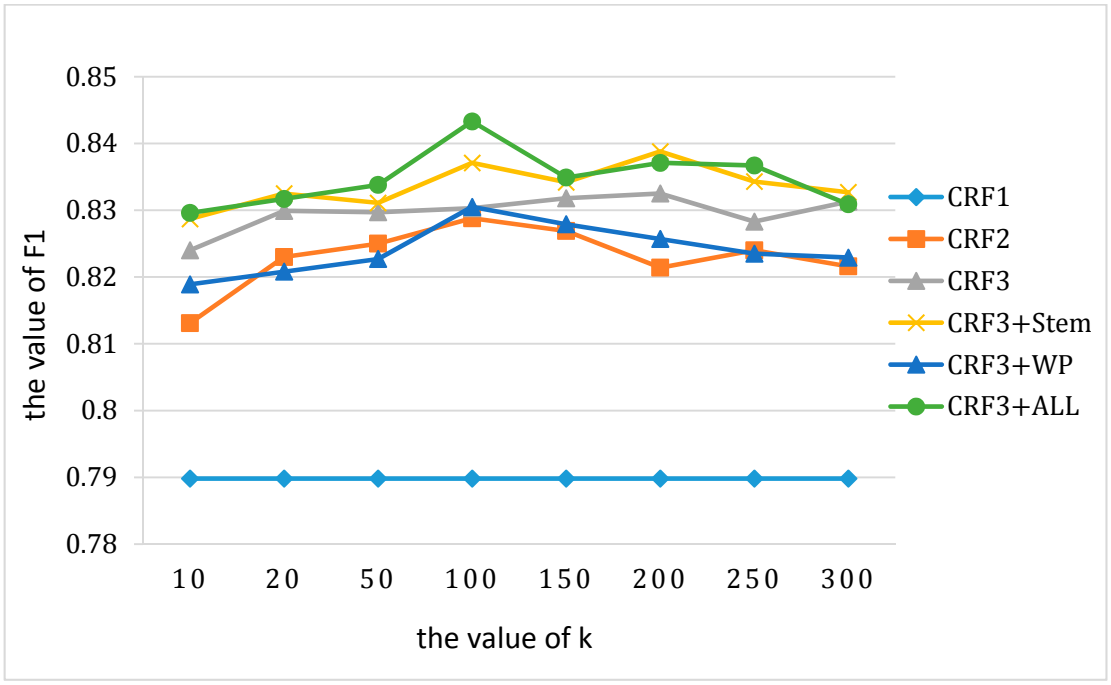

Figure 4. The effect of K on experimental results (L-14).

It can be observed that the word and MFE clustering made the CRF model achieve the best result at a certain point. Due to the complexity of the clustering model, the increase of k will result in over-fitting, which means the F1 will decline. Concerning the R-14, the best choice of k was approximately 100, while it was about 200 in the L-14 that had more complex semantics. The optimal k was also the value used in the previous experiments.

Additionally, a low value of k still brought a visible improvement. When k = 10, F1 of the improved CRF model was significantly higher than CRF1. Concerning the case of k = 10, the F1 of CRF2 exceeded 81% and CRF3+ALL exceeded 82% in the restaurant data set. To summarize, the value of k should not be selected too high, otherwise it will result in a low efficiency of training

and decrease the effect of recognition. Furthermore, the effect of clustering is related to the field of text; when the text contains more proper nouns and more complicated semantics, the value of k should be improved accordingly.

5. Conclusions and Future

MFE clustering reinforcement is introduced on the basis of CRF using the general position features. First, the authors improved the text representation by MFE to capture more semantic information. Second, MFE and word clustering were obtained by kmeans++ algorithm to enrich the position features of CRF and get a better effect of training. The function of MFE clustering to capture implicit contextual information was validated through the experiment of aspect term extraction on SemEval datasets. During the experiments, the effect of MFE-CRF was significantly higher than the CRF models, which only used the general position features and word clustering features.

This paper has achieved visible results in aspect term extraction by improving text representation. However, there are still the following properties that need to be further explored:

- Better MFE vector training strategy: MFE in this paper is a derivative of word vector technology; the vector training algorithm is not thoroughly studied and improved. Additionally, the introduction of more additional semantic features also brings more burdens to training. Thus, more efficient MFE training methods need to be developed.
- Apply and improve deep learning in aspect term extraction: the MFE-CRF presented in this paper is still in the realm of traditional machine learning, so it is still necessary to explore how to introduce deep learning into aspect term extraction.

Author Contributions: Y.X. and H.H. performed the experiment and data collection. J.Z. was involved in theoretical investigation and optimization.

Funding: This work is supported by the Fundamental Research Funds for the Central Universities of Central South University (2018zzts587).

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (61379109). The authors would to thank the reviewers for their valuable suggestions and comments.

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

References

1. Yadollahi, A.; Shahraki, A.G.; Zaiane, O.R. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.* **2017**, *50*, 25. [CrossRef]
2. Giachanou, A.; Crestani, F. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv.* **2016**, *49*, 28. [CrossRef]
3. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *30*, 1–167. [CrossRef]
4. Thet, T.T.; Na, J.C.; Khoo, C.S. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.* **2010**, *36*, 823–848. [CrossRef]
5. Wen, H.; Zhao, J. Aspect term extraction of E-commerce comments based on model ensemble. In Proceedings of the 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 15–17 December 2017; pp. 24–27.
6. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 5–10 December 2013; MIT Press Ltd.: Cambridge, MA, USA, 2013; pp. 3111–3119.
7. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *Comput. Sci.* **2013**.
8. One-Hot. Available online: https://en.wikipedia.org/wiki/One-hot (accessed on 25 July 2018).
9. Choi, Y.; Cardie, C.; Riloff, E.; Patwardhan, S. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; Association for Computational Linguistic: Stroudsburg, PA, USA, 2005; pp. 355–362.

10. Jakob, N.; Gurevych, I. Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011.
11. Miao, Q.; Li, Q.; Zeng, D. Mining Fine Grained Opinions by Using Probabilistic Models and Domain Knowledge. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, Canada, 31 August–3 September 2010; IEEE Computer Society: Washington, DC, USA, 2010; pp. 358–365.
12. Toh, Z.; Wang, W. DLIREC: Aspect Term Extraction and Term Polarity Classification System. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014.
13. Parkhe, V.; Biswas, B. Aspect Based Sentiment Analysis of Movie Reviews: Finding the Polarity Directing Aspects. In Proceedings of the International Conference on Soft Computing and Machine Intelligence, New Delhi, India, 26–27 September 2014; IEEE Computer Society: Washington, DC, USA, 2014; pp. 28–32.
14. Guha, S.; Joshi, A.; Varma, V. SIEL: Aspect Based Sentiment Analysis in Reviews. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, CO, USA, 4–5 June 2015.
15. Román, J.V.; Cámara, E.M.; Morera, J.G.; Zafra, S.M. TASS 2014. The Challenge of Aspect-based Sentiment Analysis. *Proces. Del Leng. Nat.* **2015**, *54*, 61–68.
16. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Mohammad, A.S.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014; pp. 27–35.
17. Poria, S.; Ofek, N.; Gelbukh, A.; Hussain, A.; Rokach, L. Dependency Tree-Based Rules for Concept-Level Aspect-Based Sentiment Analysis. *Commun. Comput. Inf. Sci.* **2014**, *475*, 41–47.
18. Khalid, S.; Khan, M.T.; Durrani, M.; Khan, K.H. Aspect-based Sentiment Analysis on a Large-Scale Data: Topic Models are the Preferred Solution. *Bahria Univ. J. Inf. Commun. Technol.* **2015**, *8*, 22–27.
19. Poria, S.; Chaturvedi, I.; Cambria, E.; Bisio, F. Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4465–4473.
20. Schouten, K.; Baas, F.; Bus, O.; Osinga, A.; van de Ven, N.; van Loenhout, S.; Vrolijk, L.; Frasincar, F. Aspect-Based Sentiment Analysis Using Lexico-Semantic Patterns. In Proceedings of the Web Information Systems Engineering—WISE 2016, Shanghai, China, 7–10 November 2016.
21. Wu, C.; Wu, F.; Wu, S.; Yuan, Z.; Huang, Y. A Hybrid Unsupervised Method for Aspect Term and Opinion Target Extraction. *Knowl. Based Syst.* **2018**, *148*, 66–73. [CrossRef]
22. Manek, A.S.; Shenoy, P.D.; Mohan, M.C.; Venugopal, K.R. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web* **2017**, *20*, 35–154. [CrossRef]
23. Weichselbraun, A.; Gindl, S.; Fischer, F.; Vakulenko, S.; Scharl, A. Aspect-Based Extraction and Analysis of Affective Knowledge from Social Media Streams. *IEEE Intell. Syst.* **2017**, *32*, 80–88. [CrossRef]
24. Schusterböckler, B.; Bateman, A. An Introduction to Hidden Markov Models. *Curr. Protoc. Bioinform.* **2007**, *18*. [CrossRef]
25. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
26. Liu, P.; Joty, S.; Meng, H. Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1433–1443.
27. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100–108. [CrossRef]
28. Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Acm-Siam Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; Louisiana Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007; pp. 1027–1035.
29. Chernyshevich, M. IHS R&D Belarus: Cross-domain extraction of product features using CRF. In Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, 23–24 August 2014.

30. Vicente, I.S.; Saralegi, X.; Agerri, R. EliXa: A modular and flexible ABSA platform. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, CO, USA, 4–5 June 2015.
31. Toh, Z.; Su, J. NLANGP at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In Proceedings of the 10th International Workshop on Semantic Evaluation, San Diego, CA, USA, 16–17 June 2016.
32. Wang, W.; Pan, S.J.; Dahlmeier, D.; Xiao, X. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; AAAI Press: San Francisco, CA, USA, 2017; pp. 3316–3322.