applied sciences

MDPI

Article

Incorporating Word Significance into Aspect-Level Sentiment Analysis

Refuoe Mokhosi [1], ZhiGuang Qin [1], Qiao Liu [1,*] and Casper Shikali [2]

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, Sichuan, China
[2] Department of Information Technology, South Eastern Kenya University, 170, Kitui 90200, Kenya
* Correspondence: qliu@uestc.edu.cn

Received: 28 July 2019; Accepted: 20 August 2019; Published: 27 August 2019

Featured Application: This study demonstrates the relevance of word significance in aspect-level sentiment analysis, which may be applicable in guiding attention to the most significant items in attention based machine learning tasks. Novelty which can be independent or a component of word significance is applicable in scenarios where there is a possibility of time based novelty decay with repeated exposure to an item, with the decay factor varying from 0 to 1 based on the application area. Ultimately, this study closely represents the human attention mechanism in machine learning tasks.

Abstract: Aspect-level sentiment analysis has drawn growing attention in recent years, with higher performance achieved through the attention mechanism. Despite this, previous research does not consider some human psychological evidence relating to language interpretation. This results in attention being paid to less significant words especially when the aspect word is far from the relevant context word or when an important context word is found at the end of a long sentence. We design a novel model using word significance to direct attention towards the most significant words, with novelty decay and incremental interpretation factors working together as an alternative for position based models. The interpretation factor represents the maximization of the degree each new encountered word contributes to the sentiment polarity and a counter balancing stretched exponential novelty decay factor represents decaying human reaction as a sentence gets longer. Our findings support the hypothesis that the attention mechanism needs to be applied to the most significant words for sentiment interpretation and that novelty decay is applicable in aspect-level sentiment analysis with a decay factor $\beta = 0.7$.

Keywords: aspect-level sentiment analysis; attention mechanism; novelty decay; incremental interpretation; stretched exponential law

1. Introduction

Aspect-level sentiment analysis is a prominent task in Natural Language Processing (NLP) and has attracted growing attention in research and business community as it identifies sentiment as a key driver of human behavior [1,2]. Given a sentence with context and aspect words, this task aims to infer a positive, negative or neutral sentiment polarity of a context word towards an aspect word. For example, in the sentence "Well, it happened because of a graceless manager and a rude bartender who had us waiting 20 min for drinks and then tells us to chill out", the aspect "drinks" would be assigned a neutral polarity and aspect words "waiting," "manager" and "bartender" would have a negative polarity.

Recent advances in aspect-level sentiment analysis have gravitated towards neural network based models [3,4]. Such models learn text representations from high dimensional data without careful feature engineering and capture semantic relations between context and aspect words in a more scalable manner [5,6]. Researchers have further used the attention mechanism in conjunction with Recurrent Neural Networks (RNNs) as effective sequence modeling techniques to achieve higher model accuracy [7,8]. Continuing from the sentence example provided earlier, one can see that the context word "graceless" is more relevant to the aspect word "manager," the word "rude" is more applicable to the word "bartender", while the words "20 min" are more pertinent to the word "waiting." In this example, the attention mechanism is used to highlight the most influential context word with respect to an aspect word. In existing approaches, the magnitude of attention on a context word correlates to its relevance to the aspect word in predicting the output at each position of a sentence [9,10]. These existing approaches model the attention weight based on explicit positional encodings [11,12] or generate target specific sentence representations [5,7] but there are still some outstanding limitations in these models as they overlook the notion that humans use attention filters to direct attention towards items deemed as important.

The significance of attention filters is highlighted by psychologists, who define attention as a basic component of human cognitive biology which is limited in terms of capacity and duration [13,14]. An everyday example showing that there is a limit to items humans can pay attention to is, one can not effectively listen to a telephone conversation if someone else in the room is simultaneously giving them complex instructions. Consequently selective towards stimuli [15] and this is supported by theories behind language reading comprehension. These theories illustrate that text elements are first processed at a minimal level and graded for significance, then extra attention is paid to elements in proportion to their significance [16,17].

In this backdrop, the first problem with current aspect-level sentiment analysis models is paying attention to less significant words as they allocate attention weights to words without grading for significance. For example, it is common for a writer to use italics or bold to draw attention to important words that may otherwise be missed in a sentence. Taking the sentence "Great food, REASONABLE prices, makes for an evening that can't be beat" as an example, human attention is drawn to the word "REASONABLE" with respect to the aspect word "prices" because of the capitalization. Current methods bypass this step resulting in most methods encompassing of embedding, contextual, attention and output layers [2,8,12], resulting in attention not being directed to the most significant words.

The second problem in current models is overlooking important words that may be at the end of a sentence. Such models ignore the notion that humans make incremental interpretations of a sentence by maximizing the degree each new encountered word affects sentence interpretation [18,19]. This implies that the ultimate sentiment polarity keeps being reviewed as more words are accessed. For example in this sentence "My day off after a wedding consist of wii zelda and chinese food. I feel like im in jr high. This rocks!!", the polarity towards the aspect word "wii" is neutral if only "My day off after a wedding consist of wii zelda and chinese food" is considered, the polarity is still neutral if the part sentence "My day off after a wedding consist of wii zelda and chinese food. I feel like im in jr high" is considered. The ultimate sentiment polarity is established when the ending words "This rocks!!" are considered, then the polarity towards the aspect word "wii" is positive. We therefore argue that incorporating incremental interpretation into aspect-level sentiment analysis may tackle sentences where there the relevant aspect and context words are far apart.

The last problem current models have is determining sentiment polarity when an aspect word and the relevant context word are far apart from each other. Current models overlook the concept of decaying human reaction to repeated stimuli [20,21] and therefore word importance decreases as sentence length increases until there is a complex stimuli triggered by an encountered context word [22,23]. Current methods do not take this quality into consideration but instead use global attention based models which do not take explicit positions of a word [5,6,24,25], while position based

attention assumes that a context word has higher importance if it is closer to an aspect word [11,12]. We argue that computational methods could benefit from taking language processing psychological evidence into consideration in aspect-level sentiment analysis.

To mitigate the first problem we propose a unified module dubbed as word significance as an attention filter to guide our model to pay more attention to the most significant words. The word significance module also acts as a propagation mechanism that allows a high significance context word that may be far from an aspect word to still have influence on the sentiment polarity. The word significance factor is made up of two aspects being an increasing incremental interpretation factor as a solution to the second problem, which is counter balanced by a novelty decay factor as a solution to the third problem. We solve incremental interpretation using a growth function that increases as more words are accessed in a sentence, as previously used in determining research paper significance growth [21,26]. We propose to represent novelty decay using the stretched exponential law which has proven successful in representing phenomena relating to human nature dynamics with a natural time span decay. To the best of our knowledge the use of the stretched exponential has not been used in NLP, particularly in aspect-level sentiment analysis. Experimental results show that our model performs competitively with current models.

The aim of this paper is to demonstrate the importance of word significance in directing attention to the most attention worthy words in aspect-level sentiment analysis. The contributions of this paper are as follows:

- We use the word significance factor to model attention worthiness in aspect-level sentiment analysis using the Significant Attention Network (SAN) model.
- We introduce two novel factors in aspect-level sentiment analysis being incremental interpretation and novelty decay as an alternative to position based models.
- We conduct qualitative research on three real world datasets to prove the universality of stretched exponential novelty decay in aspect-level sentiment analysis.

The rest of the paper is organized as follows: We first review related works in aspect level sentiment analysis, novelty decay and incremental interpretation in Section 2. Then the proposed model is presented in Section 3, while Section 4 is an extensive comparison of experiments conducted to test the effectiveness of the proposed model. Finally, Section 5 summarizes this work and provides future directions.

2. Related Works

Aspect-level sentiment analysis is a fundamental task in NLP, aimed at determining the polarity with respect to a specific aspect term. Early studies typically depend on the quality of handcrafted intensive lexicons, ngram and parse trees for feature engineering [27–29] to train sentiment classifiers such as Support Vector Machines (SVM) [30,31]. Since effective features are based on expensive, complicated domain expert knowledge, recent methods prefer neural networks which offer efficient generation of low dimensional sentence representations [10,32,33]. The attention mechanism as a human capability has been used to achieve higher performance in aspect-level sentiment analysis [5,6,25], and since we propose that word significance can be used to guide models to apply attention to the most significant words, we discuss prevailing attention, incremental interpretation and novelty decay research in the following section.

2.1. Attention Mechanism in Aspect-Level Sentiment Analysis

Inspired by visual attention, researchers in various domains such as machine translation [34], question answering [35,36] and machine comprehension [37] have successfully used attention to achieve performance gains. When used in conjunction with neural networks, the attention mechanism is used to do a soft selection over hidden representations and automatically learn the most important context works for a specific target word in a sentence [8,9]. Due to numerous influential papers in

attention research [34,36,38], there is a wide array of Long Short Term Memory(LSTM) attention based architectures which use aggregated attention scores to aggregate contextual features for prediction. Tang et al. [10] model the aspect word dependent left and right context and concatenate both as the final representation before prediction. Wang et al. [7] then introduce the concept of modeling aspect words with context words by concatenating aspect word embeddings to each context word, before using an LSTM and attention to generate the final representation. However these models only consider word level attention without taking the overall sentence meaning into consideration, making it difficult to make correct predictions when related words are far apart in long sentences.

Ma et al. [5] bridge this gap by separately modeling aspect word attention to extract mutual features of aspect and context words at sentence level though the IAN model. They use two attention based LSTMs to represent aspect and context words, then use the hidden context states to calculate attentions towards aspect words using a pooling operation and vice versa, thus capturing important parts in both the context and aspect words interaction. Huang et al. [6] make modifications by forgoing the pooling operation as it ignores interactions between word-pairs through the AOA-LSTM model. These models generate mutual attentions to concentrate on important aspect and context words at sentence level, however they do not explicitly take the distance between aspect and context words into consideration. Fan et al. [25] therefore enhance LSTM hidden states with position encodings before applying interactive attention, while Li et al. [11] introduce position weightings to transformation LSTM outputs before applying a single convolution to extract important features. Liu et al. [24] also use a position weighted memory module in conjunction with sentence level attention to capture aspect global attention and context attention to simultaneously capture the order and correlations of words. However, these methods assume that a context word closer to an aspect word has higher importance, which is not always the case. Our model presents novelty decay and incremental interpretation as an alternatives to position weighted models, with attention being guided by word significance.

2.2. Novelty Decay

It has been observed that repeated exposure to a stimuli has diminishing effects, this phenomenon is described as novelty decay [20]. This concept has been widely accepted in modeling of attention dynamics in research journals [26,39] and social media [15,40,41] but there is no consensus on whether the universal distribution shape of novelty decay is exponential [26,42] or logarithmic [39,43]. The main disadvantage of the exponentially shaped novelty decay is that it makes wrong predictions in short decay periods [21,42,44], resulting in stretched exponential shaped novelty decay which describes data over many orders of magnitude [45,46]. The stretched exponential law provides a simple, economical mechanism for novelty decay as it has only one extra parameter β as a novelty decay factor [47,48]. Therefore, Wu et al. [21] use a stretched exponential based model to simulate novelty decay in collective attention in online stories, while Laherrere et al. [47] find the stretched exponential based models to perform better in paper citation dynamics. Asur et al. [49] also use stretched exponential in modeling novelty decay in Twitter trends and Feng et al. [50] go further by modeling influence maximization and novelty decay in Twitter. Our model examines the relevance of novelty in aspect-level sentiment, to determine whether novelty decay can be characterized by the stretched exponential law.

2.3. Incremental Interpretation

One of the basic assumptions of language interpretation is that it processes incrementally by seeking to maximize the degree of interpretation with each new word encountered [19]. Researchers have consequently adopted a divide and conquer approach to guide the interpretation process where the ultimate contribution of a context word to the sentiment polarity prediction is not determined by an individual word but is instead an accumulation of its linguistic relations with the rest of the sentence [51,52]. The application of this found in psychology linguist research areas in sentence and dialogue comprehension [18,53] but not in neural networks based models of sentiment analysis.

3. Proposed Model

In this section we provide a detailed description of the proposed model, with a high-level illustration of the model in Figure 1. We first provide a task definition of aspect-level sentiment analysis, then look at the word embedding and LSTM conceptual layers. We then examine the word significance and interactive attention layers and lastly present the output layer and the training details of our model.

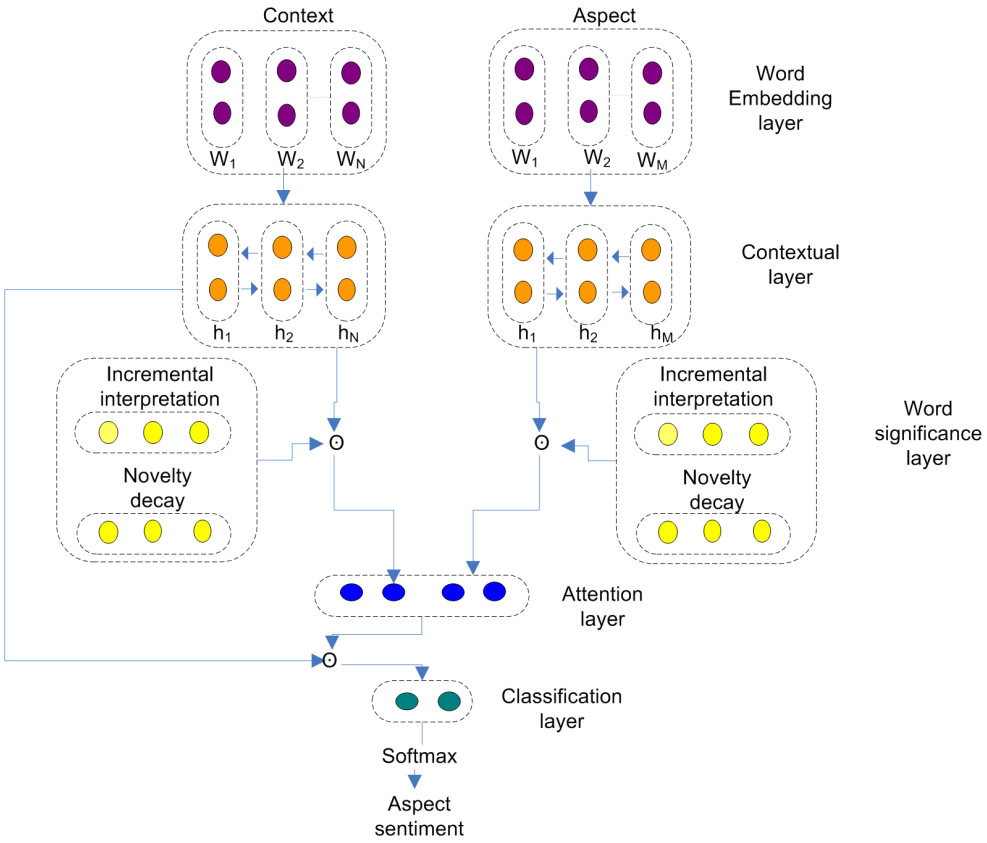

Figure 1. The architecture of the proposed ISAN model.

3.1. Task Definition

In this aspect level sentiment classification task, we have a sentence $s = [w_1, w_2, ...w_i, ..., w_n]$ consisting of n words and an aspect list $a = [w_i, w_{i+1}..., w_{i+m-1}]$ consisting of m words. Taking the polarity and the sentence-aspect word pair as input, the goal is to predict the polarity of sentence s towards aspect a.

3.2. Word Embedding Layer

The word embedding layer maps each word to a high dimensional vector using pretrained Glove embeddings [54] to get fixed low dimensional word embeddings. Each word w_i is represented by vector $v_i \in R^{d_w}$ from lookup matrix $R^{V \times d_w}$, where V is vocabulary size and d_w is vocabulary dimension. After the embedding lookup we get two vectors $[v_1; v_2; ...v_n] \in R^{n \times d_w}$ and $[v_i; v_{i+1}; ...v_{i+m-1}] \in R^{m \times d_w}$.

3.3. Contextual Layer

We feed the vectors from the previous layer into two bidirectional LSTMs to capture temporal interactions in the context and aspects words respectively. Each Bi-LSTM is created using two stacked LSTMs to learn long term dependencies and avoid vanishing gradient to produce hidden states

$\overrightarrow{h}_s \in R^{n \times d}$ where d is the hidden states dimension. Given embedding x at each time step t, the update process of the forward LSTM can be formalized as follows:

$$i_t = \sigma(\overrightarrow{W_i} \cdot [\overrightarrow{h}_{t-1}, \overrightarrow{x}_t] + \overrightarrow{b_i}) \quad (1)$$

$$f_t = \sigma(\overrightarrow{W}_f \cdot [\overrightarrow{h}_{t-1}, \overrightarrow{x}_t] + \overrightarrow{b}_f) \quad (2)$$

$$o_t = \sigma(\overrightarrow{W_o} \cdot [\overrightarrow{h}_{t-1}, \overrightarrow{x_t}] + \overrightarrow{b_o}) \quad (3)$$

$$g_t = tanh(\overrightarrow{W_g} \cdot [\overrightarrow{h}_{t-1}, \overrightarrow{x}_t] + \overrightarrow{b_g}) \quad (4)$$

$$\overrightarrow{c}_t = f_t * \overrightarrow{c}_{t-1} + i_t * g_t \quad (5)$$

$$\overrightarrow{h}_t = o_t * \tanh(\overrightarrow{c}_t) \quad (6)$$

where σ is the sigmoid activation function to control the outflow of irrelevant information and i_t, f_t and o_t are the input forget and output gates respectively to control the inflow of information. $\overrightarrow{W_i}, \overrightarrow{W_f}, \overrightarrow{W_o}, \overrightarrow{W_g} \in R^{d \times 2d}$ are weights $\overrightarrow{b_i}, \overrightarrow{b_f}, \overrightarrow{b_o}, \overrightarrow{b_g} \in R^d$ are biases, where d is the hidden dimension size. The symbol "*" is element-wise multiplication, the symbol "·" is matrix multiplication. The backward LSTM performs a similar process to produce $\overleftarrow{h_s}$, resulting in $h_s = [\overrightarrow{h_s}, \overleftarrow{h_s}] \in R^{n \times 2d}$ as the output of the BiLSTM. Given sentence s and aspect word a, we separately perform two BiLSTMs for the sentence output as $h_s \in R^{n \times 2d}$ and aspect output as $h_a \in R^{m \times 2d}$.

3.4. Word Significance Layer

We introduce the word significance layer which guides the attention layer to the most significant words for sentiment prediction. The word significance factor is determined using a growth model whereby incremental interpretation is counter balanced by a novelty decay factor. We first represent word significance a_t without novelty decay in a sentence at time t with $a_t = (1 + z_t)\, a_{t-1}$, where $z_t \in R^n$ as the incremental interpretation variable and a_0 as the initial word significance are learned variables. In order to incorporate novelty decay into word significance, the arrangement of word significance is adjusted as follows:

$$a_t = (1 + r_t * z_t)\, a_{t-1} \quad (7)$$

where $r_t \in R^n$ is the novelty decay variable.

As previously discussed, incremental interpretation is counterbalanced by novelty decay which is parameterized as r_t where $r_1 = 1$ and $r_t \downarrow 0$ as $t \uparrow \infty$.

$$r_t \sim e^{-\beta t^{\beta}} \quad (8)$$

where $0 < \beta < 1$ is decay factor. The same process is repeated for context words. The final significant word vector for context words is represented as $a_s = [a_1; a_2 ... a_n]$ and the final significant word vector for aspect words is represented as $a_a = [a_1; a_2 ... a_m]$. The significant context representation $\hat{h_s}$ and significant aspect representation $\hat{h_a}$ are derived as following using element-wise multiplication:

$$\hat{h}_s = h_s * a_s \quad (9)$$

$$\hat{h}_a = h_a * a_a \quad (10)$$

where $a_s \in R^n$, $a_a \in R^m$, $\hat{h}_s \in R^{n \times 2d}$ and $\hat{h}_a \in R^{m \times 2d}$.

3.5. Interactive Attention Layer

Given representations $\hat{h}_s \in R^n$ and $\hat{h}_a R^m$, we calculate our attention to exploit mutual information between the two representations following a methodology similar to the AOA-LSTM model [6] as

shown in Figure 2. Using the significant aspect and context representations, we calculate the interaction matrix **K** as $\hat{h_s} \cdot \hat{h_a}^T \in R^{n \times m}$ using a dot product with each value representing a context-aspect pair. Using a column-wise softmax we get aspect to context attention $\alpha \in R^{n \times m}$, while a row-wise softmax gets a context to aspect attention $\delta \in R^{n \times m}$. We then use column-wise averaging to get aspect specific attention $\overline{\delta} \in R^m$, to finally have the final sentence level attention $\gamma \in R^n$ as the weighted sum of each aspect to context attention as given by Equation (14).

$$\alpha_{ij} = \frac{exp(K_{ij})}{\sum_i exp(K_{ij})} \tag{11}$$

$$\delta_{ij} = \frac{expK_{ij}}{\sum_j exp(K_{ij})} \tag{12}$$

$$\overline{\delta}_j = \frac{1}{n} \sum_i \delta_{ij} \tag{13}$$

$$\gamma = \alpha \cdot \overline{\delta}^T \tag{14}$$

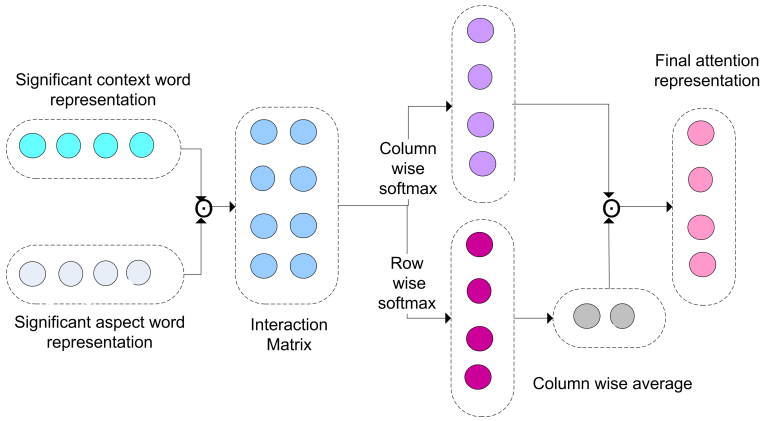

Figure 2. The interactive attention layer.

3.6. Output Layer

The final sentence representation is a weighted sum of sentence hidden semantic states $h_s \in R^{n \times 2d}$ using the attention score $\gamma \in R^n$.

$$m = h_s^T \cdot \gamma \tag{15}$$

where $m \in R^{2d}$ This final representation is then feed into a linear layer to project m into C aspect classes:

$$q = W_q \cdot m + b_q \tag{16}$$

where W_q and b_q are weight and bias respectively, $q \in R^C$ where sentiment polarity $c \in C$ and C is set to 3 as the number of aspect classification classes. After the linear layer we use a softmax layer to determine polarity.

$$P(y = c) = \frac{exp(q_c)}{\sum_{i \in C} exp(q_i)} \tag{17}$$

3.7. Model Training

The model is trained in an end-to-end manner through back-propagation, using cross-entropy loss with L_2 regularization as the objective function:

$$loss = -\sum_{i=1}^{C} y_i log(\widehat{y}_i) + \lambda_r (\sum_{\theta \in \Theta} \theta^2) \tag{18}$$

where y is the target sentiment distribution and $\hat{y}$ predicted sentiment distribution, λ_r is L_2 regularization coefficient and Θ is the parameter set.

4. Results And Discussion

4.1. Datasets and Parameter Setting

We conduct the experiments on three datasets being ACL 14 Twitter dataset [4] and SemEval 2014 Task 4 composed of Restaurant and Laptop reviews [55]. Experienced annotators tagged the aspects terms and polarities as positive, negative and neutral. Table 1 illustrates the distribution of sentiment polarity.

Table 1. Dataset polarity distribution.

Dataset	Positive		Neutral		Negative	
	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	637	196	807	196
Laptop	994	341	464	169	807	196
Twitter	1561	173	3127	346	1560	173

We initialize word embeddings with 300-dimensional Glove vectors [54] with vocabulary size 1.9 M. The dimension of LSTM hidden states is set to 300, the L2 regularization coefficient is set to 10^5 and dropout rate is 0.1. The Adam optimizer with learning rate 0.01 is used to update parameters and we adopt Accuracy and Macro-F1 as metrics.

4.2. Baseline Comparison

In order to adequately evaluate and analyze the performance of our model, we use the following models as baselines:

- Majority is a basic baseline method which assigns the sentiment polarities in the test set according to the largest polarities in the training set.
- Feature-SVM [29] uses an SVM classifier based on lexicon, parse and ngram features to achieve state-of-the-art performance.
- LSTM uses an LSTM network to learn hidden states without considering aspect words and uses the averaged vector as the sentence representation to predict polarity.
- AE-LSTM [7] models context word representations using an LSTM, then combines the hidden states with aspect embeddings to generate attention weights for classification.
- ATAE-LSTM [7] improves on the AE-LSTM by appending the aspect embedding to each word embedding in representing the context.
- IAN [5] individually models context and aspect words in attention based LSTMs using interactive attention. The two representations are then concatenated to predict polarity.
- AOA-LSTM [6] based on the individual hidden states of aspect and context words, the model uses an interaction mechanism to focus on the important context words.

We also list the ISAN model variants to analyze the effects of stretched exponential attention:

- ISAN-I: models word significance based on incremental interpretation, the final prediction is based on the an interactive attention module and word significance.
- ISAN-D: models word significance based on novelty decay, includes an interactive attention module module.
- ISAN: the complete interactive significant attention model.

4.3. Overall Performance Comparison

Table 2 illustrates the model comparison results of baseline methods. We observe that the Majority method achieves the worst performance as it only uses data distribution information, while Feature+SVM achieves improved performance on all datasets because of well designed features. Both models are outperformed by our ISAN model and other LSTM based models which automatically learn high quality representations for predictions. The LSTM method has the lowest performance in the neural network based methods because it equally treats aspect and target words. This highlights the importance of aspect words as illustrated by Jiang et al. [56]. Consequently, the ATAE-LSTM achieves higher performance as it introduces aspect embeddings together with the attention mechanism. Our proposed model outperforms these methods as it exploits mutual information between the aspect and context representations through attention interaction.

Table 2. Performance comparison of different models on the three datasets. The baseline models results are accessed from published papers except for the ones marked with "*". The values in bold represent the highest performance values.

Model	Laptop		Restaurant		Twitter	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Majority	0.535	0.333	0.650	0.333	0.500	0.333
Feature-SVM	0.705	-	0.802	-	0.634	0.633
AE-LSTM	0.689	-	0.762	-	-	-
ATAE-LSTM	0.687	-	0.772	-	-	-
IAN	0.721	-	0.786	-	0.716 *	0.705 *
AOA-LSTM	0.745	-	0.812	-	0.725 *	0.706 *
ISAN	**0.749**	**0.721**	**0.824**	**0.744**	**0.734**	**0.716**

The IAN model achieves improved performance in comparison to previously mentioned LSTM based methods as it interactively learns the aspect and context attention weights, our model and the AOA-LSTM are able to outperform it because it uses a pooling operation to get the final representation. The AOA-LSTM model achieves higher performance as it replaces the pooling operation in the IAN with an interaction matrix, with each value representing a context-aspect pair to produce mutual attentions to concentrate on important aspect and context words. Our complete ISAN model performs competitively with state of the art as it incorporates word significance.

4.4. Analysis of Isan Model

Table 3 illustrates the performance comparison of the ISAN model versions. The efficacy of incremental interpretation in aspect-level sentiment analysis is illustrated by the ISAN-I model, showing the importance of considering relevant aspect words that may be at the end of a sentence. This model is outperformed by the novelty decay based ISAN-D model, meaning that when these two factors are considered individually the impact of decaying novelty throughout a sentence is stronger. The influence of novelty decay implies that irrespective the length of a sentence, novelty is still an important factor as the stretched exponential law does not allow novelty to become zero even when the sentence length reaches large lengths. The performance of our model improves when novelty decay acts as a discounting factor of incremental interpretation as shown by the complete ISAN results, leading to 3.5%, 1.8% and 1% performance increase in the Laptop, Restaurant and Twitter datasets respectively. We also collected the statistics of the aspect word position in sentences from the Restaurant, Twitter and Laptop dataset as shown by Table 4. When each sentence length is divided into three segments, the majority of aspect words are located in the last section of the sentence for all datasets. This shows the importance of simultaneously considering words at the beginning and at the end of a sentence when determining sentiment polarity. The results also illustrate the

effectiveness of word significance in propagating the significance grading of a context word that may be far from an aspect word. Our results support our intuition that word significance can be used as a representation of how much a word contributes to the sentence interpretation and ultimately draws our model to words that are worthy of attention.

Table 3. Performance comparison of ISAN models on the three datasets.

Model	Laptop		Restaurant		Twitter	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**	**Acc**	**Macro-F1**
ISAN-I	0.719	0.665	0.801	0.705	0.711	0.674
ISAN-D	0.723	0.673	0.806	0.713	0.727	0.710
ISAN	0.749	0.721	0.824	0.744	0.734	0.716

Table 4. Aspect word location in sentence. Sentences are divided into three fragments, being the beginning, middle and end to determine the fragment each aspect word is found in.

Sentence Fragment	**Beginning** (%)	**Middle** (%)	**End** (%)
Laptop	20.30	22.56	55.02
Restaurant	22.84	22.56	54.59
Twitter	27.98	21.87	50.13

This study demonstrates the relevance of word significance in aspect-level sentiment analysis and can be extended to other machine learning tasks based on the attention mechanism. Word significance may be used to incorporate attention filters which can guide the attention mechanism to items more worthy of attention, more especially in other NLP tasks where there is a possibility of time based novelty decay with repeated exposure. Depending on the field, the stretched exponential decay factor ranges between 0 and 1 as shown by Figure 3 [48], putting emphasis on initial occurrences of an item without nullifying its later occurrences. Our results are also in line with previous studies which illustrate the universality of novelty decay in fields investigating social media and economics [21,26,42,57], which may be further investigated using neural network based models. Ultimately, our model strives to closely represent human attention mechanisms in machine learning tasks.

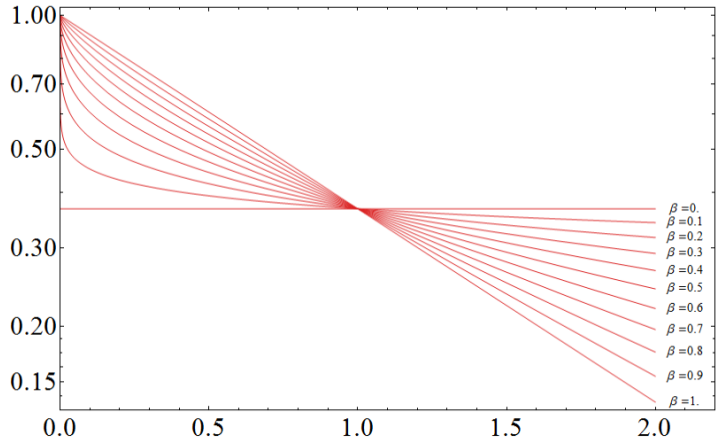

Figure 3. The stretched exponential on a log plot of novelty versus time for varying β values.

4.5. The Effects of Novelty Decay

In order to validate the effects of attention decay, we vary the novelty decay exponent beta β as shown in Figure 4. From the results we can see that the ISAN model reaches its peak Macro-F1 when

the novelty decay factor $\beta = 0.7$, afterwards there is a general decline in performance. The novelty factor value of 0.7 is larger than the one found by Huberman et al. [21] of 0.4 for collective attention decay in websites, implying faster novelty decay in aspect-level sentiment analysis. Asur et al. [49] find the novelty decay factor in new story detection in Twitter to decrease from 0.8 to 0.4 as the number of stories increase, while Feng et al. [50] find novelty decay to range from 0.8 to 0.3 as the number of people in social networks increases. The novelty decay factor of our model is higher up the range demonstrated in previous studies, maybe due to lower number of words in sentences in comparison to the number of stories or people on social media.

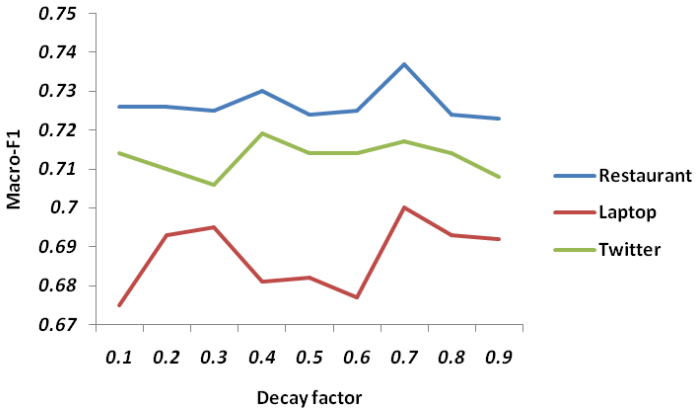

Figure 4. The effect of varying the novelty decay factor on Macro-F1.

4.6. Case Study

In this section we perform a case study on the ISAN model using two sentences as shown in Figure 5, with a deeper color intensity representing higher weight allocation to a word towards an aspect word. We first use the review sentence "as much i like the food there i cant bring myself to go back", with "food" as the aspect word from the Restaurant dataset. We apply the ISAN model to get the correct sentiment polarity prediction as negative. Although the sentiment towards food is positive, the model puts higher weight to the words "ca nt bring" despite being a few words away from the aspect word "food". This shows that our model is able to counterbalance novelty decay with incremental interpretation and allocates a word that is further down the sentence with a higher attention weights to detect negation and make the correct prediction. The attention given to common words such like "as" and "the" is also relatively low, verifying the intuition that such words do not contribute much to the sentiment polarity. The second sentence for our case study is "highly recommend this as great value for excellent sushi and service", where "service" is the aspect word as shown by Figure 6. Our model correctly predicts a positive sentiment polarity with higher attention given to words "great value" instead of word "excellent" which according to our intuition is more relevant to the word "sushi".

Figure 5. Attention weights where food is the aspect word.

Figure 6. Attention weights where service is the aspect word.

4.7. Error Analysis

The first type of error our model encounters is related to non-compositional sentiment expression which appears in numerous previous words [6,10]. For example, for the sentence "we requested they re-slice the sushi and it was returned to us in small cheese-like cubes" as there is no direct sentiment towards the aspect "sushi". This problem is further highlighted in the sentence "entrees include classics

like lasagna, fettuccine alfredo and chicken parmigiana" where aspects words "lasagna", "fettucine" and "alfredo". Another type of error our model encounters is caused by the use of complex sentences like "Anywhere else, the price would be 3x as high!" where our models misses to give the word "3x" a high attention weight.

5. Conclusions

In this paper, we propose the ISAN model which is made of word significance to guide the attention mechanism towards words most worthy of attention. We first use a BiLSTM to separately model context and aspect hidden states, then apply a word significance module based on incremental interpretation and novelty decay to determine sentiment polarity. Our results highlight the importance of simultaneously considering incremental interpretation and novelty decay in determining polarity, showing that our model can be an alternative for position based models in aspect-level sentiment analysis. We perform experiments on three public datasets to prove the applicability of stretched exponential novelty decay in aspect-level sentiment analysis and establish a novelty decay factor of 0.7. Although results illustrate the effectiveness of word significance, there is still space for improvement as illustrated by the error analysis.

Author Contributions: R.M., metodology, investigation and writing-original draft preparation, Z.Q., supervision, Q.L. supervision and funding acquisition, C.S., writing-review and editing.

Funding: This research has been supported in part by the National Natural Science Foundation of China (61772117), the General Equipment Department Foundation (6140312010203), the Science and Technology Foundation of the Military Commission (1816321TS00105301), the National Engineering Laboratory of Big Data Application Technology for Improving Goverment Management Efficiency Open foundation Program (10-2018039), and the 54th Research Institute of China Electronics Technology Group Corporation (185686).

Acknowledgments: We thank the anonymous reviewers for their helpful and insightful advice.

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory

References

1. Deng, L.; Liu, Y. *Deep Learning in Natural Language Processing*; Springer: Berlin/Heidelberg, Germany, 2018.
2. 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 Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
3. Socher, R.; Pennington, J.; Huang, E.H.; Ng, A.Y.; Manning, C.D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 151–161.
4. Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; Xu, K. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; Volume 2, pp. 49–54.
5. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive Attention Networks for Aspect-Level Sentiment Classification. *arXiv* **2017**, pp. 4068–4074, arXiv:1709.00893.

6. Huang, B.; Ou, Y.; Carley, K.M. Aspect level sentiment classification with attention-over-attention neural networks. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington, DC, USA, 10–13 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 197–206.
7. 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.
8. 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.
9. Chen, P.; Sun, Z.; Bing, L.; Yang, W. Recurrent attention network on memory for aspect sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 452–461.
10. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for target-dependent sentiment classification. *arXiv* **2015**, arXiv:1512.01100.
11. Li, X.; Bing, L.; Lam, W.; Shi, B. Transformation networks for target-oriented sentiment classification. *arXiv* **2018**, arXiv:1805.01086.
12. Zeng, J.; Ma, X.; Zhou, K. Enhancing Attention-Based LSTM With Position Context for Aspect-Level Sentiment Classification. *IEEE Access* **2019**, *7*, 20462–20471. [CrossRef]
13. Kahneman, D. *Attention and Effort*; Citeseer: State College, PA, USA, 1973; Volume 1063.
14. Styles, E. *The Psychology of Attention*; Psychology Press: Hove, UK, 2006.
15. Weng, L.; Flammini, A.; Vespignani, A.; Menczer, F. Competition among memes in a world with limited attention. *Sci. Rep.* **2012**, *2*, 335. [CrossRef] [PubMed]
16. Anderson, R.C. Allocation of attention during reading. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1982; Volume 8, pp. 292–305.
17. Shirey, L.L.; Reynolds, R.E. Effect of interest on attention and learning. *J. Educ. Psychol.* **1988**, *80*, 159. [CrossRef]
18. Schlesewsky, M.; Bornkessel, I. On incremental interpretation: Degrees of meaning accessed during sentence comprehension. *Lingua* **2004**, *114*, 1213–1234. [CrossRef]
19. Sedivy, J.C.; Tanenhaus, M.K.; Chambers, C.G.; Carlson, G.N. Achieving incremental semantic interpretation through contextual representation. *Cognition* **1999**, *71*, 109–147. [CrossRef]
20. Wu, F.; Huberman, B.A. Popularity, novelty and attention. In Proceedings of the 9th ACM Conference on Electronic Commerce, Chicago, IL, USA, 8–12 July 2008; ACM: New York, NY, USA; pp. 240–245.
21. Wu, F.; Huberman, B.A. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 17599–17601. [CrossRef] [PubMed]
22. Atchley, P.; Lane, S. Cognition in the attention economy. In *Psychology of Learning and Motivation*; Elsevier: Amsterdam, The Netherlands, 2014; Volume 61, pp. 133–177.
23. Falkinger, J. Limited attention as a scarce resource in information-rich economies. *Econ. J.* **2008**, *118*, 1596–1620. [CrossRef]
24. Liu, Q.; Zhang, H.; Zeng, Y.; Huang, Z.; Wu, Z. Content attention model for aspect based sentiment analysis. In Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Lyon , France, 23–27 April 2018; pp. 1023–1032.
25. 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.
26. Wang, D.; Song, C.; Barabási, A.L. Quantifying long-term scientific impact. *Science* **2013**, *342*, 127–132. [CrossRef] [PubMed]
27. Kaji, N.; Kitsuregawa, M. Building lexicon for sentiment analysis from massive collection of HTML documents. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 1075–1083.
28. Kiritchenko, S.; Zhu, X.; Mohammad, S.M. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **2014**, *50*, 723–762. [CrossRef]

29. Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 437–442.
30. Qu, L.; Ifrim, G.; Weikum, G. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; pp. 913–921.
31. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
32. Vo, D.T.; Zhang, Y. Target-dependent twitter sentiment classification with rich automatic features. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
33. Zhang, M.; Zhang, Y.; Vo, D.T. Neural networks for open domain targeted sentiment. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 612–621.
34. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
35. Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; Socher, R. Ask me anything: Dynamic memory networks for natural language processing. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1378–1387.
36. Sukhbaatar, S.; Weston, J.; Fergus, R. End-to-end memory networks. In Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 2440–2448.
37. Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. *arXiv* **2016**, arXiv:1611.01603.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
39. Radicchi, F.; Fortunato, S.; Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17268–17272. [CrossRef] [PubMed]
40. Candia, C.; Jara-Figueroa, C.; Rodriguez-Sickert, C.; Barabási, A.L.; Hidalgo, C.A. The universal decay of collective memory and attention. *Nat. Hum. Behav.* **2019**, *3*, 82. [CrossRef] [PubMed]
41. Lehmann, J.; Gonçalves, B.; Ramasco, J.J.; Cattuto, C. Dynamical classes of collective attention in twitter. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; ACM: New York, NY, USA, 2012; pp. 251–260.
42. Higham, K.W.; Governale, M.; Jaffe, A.; Zülicke, U. Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. *Phys. Rev. E* **2017**, *95*, 042309. [CrossRef] [PubMed]
43. Stringer, M.J.; Sales-Pardo, M.; Amaral, L.A.N. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE* **2008**, *3*, e1683. [CrossRef]
44. Higham, K.W.; Governale, M.; Jaffe, A.; Zülicke, U. Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields. *J. Inf.* **2017**, *11*, 1190–1200. [CrossRef]
45. Krapivsky, P.L.; Redner, S.; Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **2000**, *85*, 4629. [CrossRef]
46. Krapivsky, P.L.; Redner, S. Organization of growing random networks. *Phys. Rev. E* **2001**, *63*, 066123. [CrossRef]
47. Laherrere, J.; Sornette, D. Stretched exponential distributions in nature and economy:"Fat tails" with characteristic scales. *Eur. Phys. J. B Condens. Matter Complex Syst.* **1998**, *2*, 525–539. [CrossRef]
48. Elton, D.C. Stretched exponential relaxation. *arXiv* **2018**, arXiv:1808.00881.
49. Asur, S.; Huberman, B.A.; Szabo, G.; Wang, C. Trends in social media: Persistence and decay. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.

50. Feng, S.; Chen, X.; Cong, G.; Zeng, Y.; Chee, Y.M.; Xiang, Y. Influence maximization with novelty decay in social networks. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014.
51. Pereira, F.C.; Pollack, M.E. Incremental interpretation. *Artif. Intell.* **1991**, *50*, 37–82. [CrossRef]
52. Altmann, G.T.; Kamide, Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* **1999**, *73*, 247–264. [CrossRef]
53. DeVault, D.; Sagae, K.; Traum, D. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue Discourse* **2011**, *2*, 143–170. [CrossRef]
54. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
55. 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-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 19–30.
56. Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; Zhao, T. Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 151–160.
57. Lorenz-Spreen, P.; Mønsted, B.M.; Hövel, P.; Lehmann, S. Accelerating dynamics of collective attention. *Nat. Commun.* **2019**, *10*, 1759. [CrossRef]