future internet

MDPI

Article

Enhancing Short-Term Sales Prediction with Microblogs: A Case Study of the Movie Box Office

Jie Zhao [1,*], Fangwei Xiong [1] and Peiquan Jin [2,*]

[1] School of Business, Anhui University, Hefei 230601, China; m20201021@stu.ahu.edu.cn
[2] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China
* Correspondence: zhaojie@ahu.edu.cn (J.Z.); jpq@ustc.edu.cn (P.J.)

Abstract: Microblogs are one of the major social networks in people's daily life. The increasing amount of timely microblog data brings new opportunities for enterprises to predict short-term product sales based on microblogs because the daily microblogs posted by various users can express people's sentiments on specific products, such as movies and books. Additionally, the social influence of microblogging platforms enables the rapid spread of product information, implemented by users' forwarding and commenting behavior. To verify the usefulness of microblogs in enhancing the prediction of short-term product sales, in this paper, we first present a new framework that adopts the sentiment and influence features of microblogs. Then, we describe the detailed feature computation methods for sentiment polarity detection and influence measurement. We also implement the Linear Regression (LR) model and the Support Vector Regression (SVR) model, selected as the representatives of linear and nonlinear regression models, to predict short-term product sales. Finally, we take movie box office predictions as an example and conduct experiments to evaluate the performance of the proposed features and models. The results show that the proposed sentiment feature and influence feature of microblogs play a positive role in improving the prediction precision. In addition, both the LR model and the SVR model can lower the MAPE metric of the prediction effectively.

Keywords: microblog; sales prediction; sentiment analysis; social influence; regression model; short-term prediction

Citation: Zhao, J.; Xiong, F.; Jin, P. Enhancing Short-Term Sales Prediction with Microblogs: A Case Study of the Movie Box Office. *Future Internet* **2022**, *14*, 141. https://doi.org/10.3390/fi14050141

Academic Editors: Vijayakumar Varadarajan, Rajanikanth Aluvalu and Ketan Kotecha

Received: 13 March 2022
Accepted: 3 May 2022
Published: 4 May 2022

1. Introduction

Microblogs have emerged as primary social-network media for years due to their special properties, including convenient information release, real-time updates, rapid dissemination, and free interaction [1,2]. Microblogs can deliver timely information compared with other media-like web pages and blogs. With the increasing amount of microblog data, utilizing microblog data in decision-making tasks is a critical issue that receives attention from both academia and industries [3].

In particular, the rapid spread of microblogs on social networks can quickly form a hot spot of public opinion. Therefore, by evaluating the influence of microblogs, we can quickly find the news sources that spread fast on social networks and take corresponding actions, e.g., preventing the spread of illegal news and rumors, which is of great significance for public opinion analysis and emergency management [4]. Furthermore, the measurement of microblog influence can also assist the evaluation of microblog users' influence on social networks. Therefore, accurate evaluation of microblog influence is also of great value for business decision-making [5].

On the other hand, the sentiment polarity of microblogs can often accurately reflect users' actual views. Because of the huge number of microblogging users, the opinion of microblogging users towards a specific event can reflect the social sentiment polarity of

the event [6], which can be utilized to analyze and predict users' behavior in social life. Furthermore, microblog data can be obtained freely on the Internet, offering a more economical way for enterprises to conduct data-driven analysis and prediction than traditional investigation-based approaches.

Due to the lack of large-scale datasets for product sales prediction, this paper studies the feasibility of predicting short-term product sales with the support of microblogs based on a case-study-based perspective. We first develop a framework to enhance the prediction of short-term sales. Then, we take movie box office prediction as a case study to demonstrate the effectiveness of our proposal. Briefly, we make the following contributions in this paper:

1. Unlike previous works that focused on event detection or prediction from microblogging platforms, this paper concentrates on short-term product sales based on microblogs and presents a new framework based on sentiment analysis and social influence.
2. We propose a new feature called social influence to reflect the impact of social-network information diffusion on short-term product sales. A new algorithm is presented to measure the social influence in the paper.
3. We conduct experiments on a real dataset to evaluate the performance of the proposed framework. We take movie box office prediction as a case study and analyze the prediction performance of two regression models. The results show that the proposed sentiment feature and influence feature of microblogs play a positive role in improving the prediction precision.

The remainder of the paper is organized as follows. Section 2 summarizes the related work. Section 3 presents the framework of microblog-based sales prediction. Section 4 reports the experimental results, and finally, Section 5 concludes the paper.

2. Related Work

2.1. Microblog Influence Analysis

The social network is the primary platform for communication between people in the Internet era, including e-mail, BBS, and microblogs. Influence is reflected in the interaction and communication between social network users on the social network platform by publishing news, updating statuses, instant comments, paying attention to others, and other behaviors. Users play an essential role in social network relationships. Various user behaviors will directly affect the dissemination of information. Therefore, user influence is the main object of influence research.

Driven by the word-of-mouth effect in the market, researchers have investigated an essential issue in social network analysis, that is, "which customers should be selected to start marketing activities and make the effect of marketing activities cover the largest community." To study the influence maximization problem, we need to find an appropriate influence propagation model. The basic influence propagation process refers to the spreading process of the active nodes in social networks. The active nodes will affect other nodes following a certain propagation mechanism. The commonly used influence propagation models [7] are similar to graph traversing algorithms. The main difference between the models lies in the probability and the mode of influence propagation between nodes. For example, in the linear threshold model [7], the possibility of a node to be traversed depends on whether all active neighbor nodes' probabilities are more significant than a specified threshold.

Based on a given influence propagation model, the influence maximization algorithm mainly focuses on heuristic and greedy algorithms. A heuristic algorithm generally selects an initial node for influence propagation based on the node degree and the network centrality. However, it does not consider the diffusion process of influence and cannot guarantee the optimal influence range. Kempe et al. [7] proposed a greedy algorithm based on mountain climbing, which starts from an empty initial active-node set and

selects the most influential node to join the active-node set through influence evaluation at each step until the target number is reached. The disadvantage of this algorithm is that it has high time complexity in the case of large data scales. In the literature [8], researchers selected the potentially influential nodes in the inspiration stage and the most influential nodes in the greedy stage. The two-step improved algorithm effectively reduced the running time and improved the algorithm's efficiency.

In the research of influence maximization algorithms, the evaluation of user influence is a vital link. The assessment of user influence in social networks is based on network topology and user interaction behavior, mainly the evaluation of user communication ability. The stronger the communication ability of users, the greater the breadth and depth of communication scope and influence. The number of fans usually reflects the influence of microblog users. If a user has more fans, the microblogs published by him or her will be read by more users, meaning that the user has a greater influence on the microblogging platform. In addition to the number of fans, researchers also introduced other features to measure the influence of users, e.g., the number of comments or forwards. Xiao et al. [9] listed various characteristics of users' behavior and studied the impact of these characteristics on the influence of microblogs as well as users. They demonstrated that the number of microblogs and the number of forwards were significant indicators affecting the influence of users. Some other researchers introduced the PageRank algorithm to measure the importance of web pages [10].

So far, most research on the influence of microblogs is aimed at the influence measurement of microblog users, and few were toward the microblog contents. A simple view is to measure the influence of microblogs by calculating the sum of the number of forwards and comments of microblogs in a certain period. For example, the hot posts on Sina Weibo, which is the biggest microblogging platform in China, adopt this calculation method [6]. In order to measure and calculate the influence of microblog contents, Xu et al. [11] defined the influence of each user's microblog according to the number of fans, the number of forwards, and the number of comments. They believed that the number of fans was the main factor in evaluating the influence of the microblog content, and both forwards and comments reflected the topic and influence caused by the microblogs. However, this method might be affected by robot users on microblogging platforms.

2.2. Social Network-Based Information Prediction

The existing research on information prediction based on a social network involves many aspects. Researchers have carried out prediction research from different angles. Some researchers focus on the prediction of the attributes of social networks. For example, Kong et al. [12] predict the life cycle of microblogs based on the information characteristics of publishers and the forwarding characteristics in the first hour of publishing. Other researchers use network data to predict the sales volume of products. For example, Choi et al. [13] used the Google trend index to predict the sales of cars in the short-term future. As a result, they obtained an earlier and more accurate sales forecast than the official data.

In recent years, network users' emotions for information prediction have attracted people's attention and have gradually become an essential direction of sentiment analysis and research. Sentiment is a subjective factor in different social activities. In some areas, sentiment analysis can be predicted independently. However, it can only be an additional factor in some applications. Based on the actual work needs, the sentiment analysis methods used by different researchers are also different. At present, the prediction research based on the social network has made beneficial attempts and explorations in many fields such as political elections, film box office, and stock price [14]. Researchers are committed to studying the relationship between blog, search volume, microblog and book sales, film box office, stock closing price, and other real-world fields. Gruhl et al. [15] proposed using sentiment analysis to obtain the state changes of book reviews and predict book sales. Du et al. [16] combined the quantity information, such as microblog forwarding and microblog content information, and used machine learning methods such

as neural networks to predict the film box office, which obtained more accurate results than the traditional linear prediction. Liu et al. [17] proposed to find the potential emotional factors in blogs and established an autoregressive emotional perception model to predict commodity sales. However, some other researchers claimed that compared with traditional methods, the information on social platforms did not apply to some scenarios. For example, Skoric et al. [18] tried to use Twitter to predict the election in Singapore and found that Twitter's prediction has poor performance.

Microblogging platforms have been one of the major social networks in the Web 2.0 era. The increasing amount of microblog information provides a new idea for information prediction from microblogs. Some recent studies have revealed that the sentiment of microblog messages can highly impact information diffusion and its spread on microblogging platforms [19–21]. Additionally, there are many studies aiming to predict different information from Twitter or other microblogging media, such as stock return prediction [21], election prediction [22], crime prediction [23], and Bitcoin price prediction [24]. Some researchers were concerned about the sales prediction problem on microblogging platforms [25]. However, they relied on other information sources rather than microblogs, which was different from this study. Instead, we only needed microblog data. In addition, we focused on short-term product sales prediction rather than long-term prediction.

3. Framework of the Microblog-Based Short-Term Sales Prediction

3.1. Architecture

Unlike the previous prediction work focusing on microblogs' quantitative characteristics and text characteristics, this paper proposes a new prediction framework based on the topic characteristics of microblogs, the sentiment polarity, and the influence of microblogs (as shown in Figure 1). The original microblog was crawled from Sina Weibo based on keywords, including the original microblogs, comments, and the forwarding of these microblogs. The fields extracted from the microblog data included microblog content, the number of likes, comments, forwards, and user information [26]. After preprocessing the microblog content, the sorted word segmentation results were used to obtain the topic distribution. According to the topic distribution and sentiment dictionary of microblogs, the sentiment polarity of a microblog was calculated. Through some measurement methods of microblog influence, we can obtain the influence of these original microblogs. Then, using the sentiment polarity and microblog influence as the input, we trained a prediction model for a short-term product sales prediction. Below, we detail the critical modules of the architecture, shown in Figure 1.

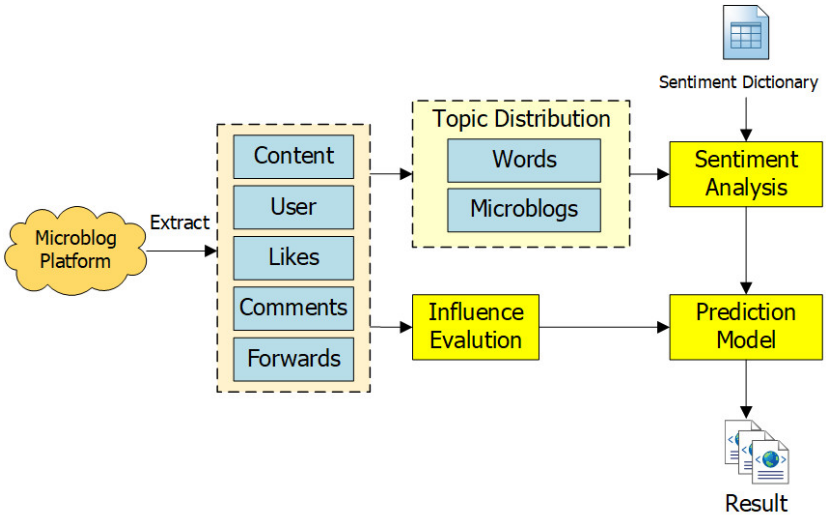

Figure 1. The architecture of microblog-based short-term sales prediction.

3.2. Sentiment Analysis

Sentiment can be expressed through text, and text is usually a sentence or a combination of multiple sentences. Sentences are composed of words and phrases. Some words are emotional, e.g., "happy" means happy, "roar" means angry. Additionally, sentences composed of words may be emotional. For example, "the woman is so beautiful" means the user praises the woman. The sentiment of the text reflects the user's viewpoint and attitude towards things, the sentiment of positive tendency reflects the user's positive attitude, and the sentiment of negative tendency reflects the user's negative attitude. We can analyze the sentiment of the text to obtain the emotion held by the user. Sentiment analysis refers to analyzing the emotional state implied by the user when transmitting information and judging or evaluating the user's attitude and opinions [27]. Text can be divided into subjective text and objective text. Objective text is a text without any emotional factor, and subjective text can be divided into positive tendency text and negative tendency text. Whether it is a single word or the whole text, the sentiment polarity of text can be obtained through sentiment analysis. According to the different text granularity, sentiment analysis mainly has four research directions: word-level sentiment analysis, sentence-level sentiment analysis, text-level sentiment analysis, and topic-level sentiment analysis. The word-level sentiment analysis is the basis of the text sentiment analysis.

Compared with the sentiment analysis of conventional textual documents, microblog text is short, and the microblog content often includes irregular words. In addition, microblog users often invent new symbols or words, complicating microblog text analysis. For example, there is a common phenomenon of polysemy of one word, and the meaning of the same word in different contexts is likely to be the opposite. Notably, the Chinese microblog's short and casual nature brings great challenges to text sentiment analysis. A sentiment dictionary is useful to improve the efficiency of sentiment analysis because we can directly judge the polarity of the corresponding words based on the sentiment dictionary. So far, the sentiment polarity of each microblog can be calculated by defining a list of negative, positive, or neutral words.

Generally, there are two kinds of text sentiment classification methods: sentiment calculation based on a sentiment dictionary and sentiment classification based on statistical learning. The latter method needs to label the training data manually, and there are certain human subjective factors. Therefore, this paper uses the former method to classify the text objectively. Sentiment calculation based on the sentiment dictionary is to judge the sentiment of words through the sentiment dictionary and then judge the sentiment polarity of text. The basic idea of this method is to regard the emotional polarity of words with a positive tendency as 1, the sentiment polarity of words with a negative tendency as -1, and the sentiment polarity of other words as 0. Each word contributes the same to the analysis of text tendency. By accumulating the sentiment polarity of each word in the text, we can determine the sentiment polarity of the text.

Although a microblog only contains simple sentences, it will have a clear topic and may involve multiple topics expressing different views or opinions. However, the sentiment polarities on these sub-topics may be different. Thus, the sentiment polarity of the sub-topic may affect the emotional calculation of the whole text. Therefore, in this paper, we define the weight of the words in microblog sentiment analysis through the topic similarity between words and the microblog to obtain the sentiment polarity of the microblog.

In particular, we use the LDA (Latent Dirichlet Allocation) model for the topic analysis of microblogs [1]. According to the trained topic model, we can get the topic distribution and the word distribution of topics, which are denoted as $p(z_t|m)$ and $p(n|z_t)$, respectively. Given a document m, if we select a topic z_t with the probability $p(z_t|m)$, then the topic with the maximal probability is most likely to be the topic that the document is associated with. Let the document topic be $z_{t_{max},m}$, which can be represented

as $z_{t_{max},m} = \underset{t}{argmax}\, p(z_t|m)$. To make the description clearer, we summarize all notations used in our framework in Table 1.

Table 1. Notations used in our method.

Notation	Description
m	A document, which refers to a microblog in this paper.
n	A word in a microblog
z_t	A topic in a microblog
N	The word count of a document
$similarity_{z_{t_1},z_{t_2}}$	The similarity between two topics: z_{t_1} and z_{t_2}.
$senti_n$	The sentiment polarity of the word *n*.
$senti_m$	The sentiment polarity of the microblog *m*.
d	A day
pd_d	The proportional difference between positive and negative sentiment within the day *d*.
$influ_m$	The influence of the microblog *m*.
$norm\_influ_m$	The normalized influence of the microblog *m*.
sif_m	The social influence of the microblog *m*.
inf_d	The social influence of the microblog within the day *d*.

For each word *n* in a document *m*, the probability of this word belonging to a topic z_t is $p(z_t|m,n)$. Then, the main topic with the maximal probability can be regarded as the most suitable topic that the word *n* belongs to. Thus, we define the main topic of a word *n* within a document *m*: $z_{t_{max},m,n} = \underset{t}{argmax}\, p(z_t|m,n)$.

As all the words belonging to the same topic of a document have the same word distribution, we can know that $p(n|z_t,m)$ equals $p(n|z_t)$. Therefore, according to the Bayesian conditional probability formula, we have the following equation: $(z_t|m,n) = \frac{p(n|z_t)*p(z_t|m)}{p(m,n)}$, based on which we can transform $z_{t_{max},m,n}$ into the following equation: $z_{t_{max},m,n} = \underset{t}{argmax}\, p(n|z_t) * p(z_t|m)$.

Basically, if the main topic of a word is similar to the main topic, the word is closer to the document's main topic. Therefore, the sentiment polarity of the word contributes more to the sentiment polarity of the document. Based on such analysis, we propose to use the Cosine similarity to measure the similarity between two topics, e.g., z_{t_1} and z_{t_2}, which is defined by Formula (1). Here, $\sum_{n=1}^{N} p(n|z_t) = 1, t = t_1, t_2$.

$$similarity_{z_{t_1},z_{t_2}} = \frac{\sum_{n=1}^{N} p\left(n\middle|z_{t_1}\right)p(n|z_{t_2})}{\sqrt{\sum_{n=1}^{n} p_{(n|z_{t_1})}^2}\sqrt{\sum_{n=1}^{n} p_{(n|z_{t_2})}^2}} \quad (1)$$

As each word *n* has different impacts within a document *m*, we define the weight of a word *n* within a document *m* by Formula (2), where $\sum weight_{m,n} = 1$.

$$weight_{m,n} = \frac{similarity_{z_{t_{max},m}, z_{t_{max},m,n}}}{\sum_{n=1}^{N} similarity_{z_{t_{max},m}, z_{t_{max},m,n}}} \quad (2)$$

The sentiment polarity of a word in a sentiment dictionary is usually classified into three states, namely positive, negative, and neutral. Formula (3) shows the definition of sentiment polarity.

$$senti_n = \begin{cases} 1 \ n \text{ is a positive word.} \\ 0 \ n \text{ is a neutral word.} \\ -1 \ n \text{ is a negative word.} \end{cases} \quad (3)$$

Based on the sentiment polarity of each word in a document, we can define the sentiment polarity of the document. Formula (4) shows the definition of the sentiment polarity of a document, which is represented as the weighted sum of the sentiment polarity of each word in the document.

$$senti_m = \sum_{n=1}^{N} weight_{m,n} * senti_n \quad (4)$$

Therefore, we can say that a document has positive sentiment if its $senti_m$ is positive. If $senti_m$ is below zero, the document has negative sentiment. Finally, if $senti_m$ is zero, the document contains neutral sentiment.

If the microblog content contains positive emotions, the user's evaluation of the topic content described by the microblog is positive, which can actively promote real-world information such as commodities and stock prices. On the contrary, if the microblog content contains negative emotions, the evaluation held by users is negative, meaning that it has a negative impact on the information. Therefore, this paper considers the positive and negative sentiment of microblogs simultaneously and calculates the proportional difference between the positive and negative sentiment scores within one day. Such a proportional difference is then used to predict future information. Formula (5) shows the definition of the proportional difference.

$$pd_d = \frac{|Count\_Positive_d - Count\_Negative_d|}{Count_d} \quad (5)$$

Consequently, we present the algorithm for extracting sentiment analysis features in Algorithm 1.

Algorithm 1 *Sentiment Feature Extraction*

Input: *The set of all the related microblogs in one day,* M;
Output: *Sentiment Feature,* pd
Preliminary: M_{pre} is the set of the preprocessed microblogs
T is the set of all the topics in the LDA Model

/* *Preprocessing microblogs* */

1: $M_{pre} \leftarrow M \ after \ preprocessing$

/* *Modeling LDA* */

2: $T, \ p(z_t|m), \ p(n|z_t) \leftarrow LDA \ Model \ in \ M_{pre}$

/* *Calculating topic similarity for each pair* <t_1, t_2> *in T* */

3: $similarity_{z_{t_1}, z_{t_2}} \leftarrow$ computed by Formula (1)

/ Extracting the sentiment feature */*

4: $poscnt,\ negcnt,\ totalcnt \leftarrow 0$
5: **for** each $prem_m$ in M_{pre} **do**
6: $senti_m \leftarrow 0$
7: **for** each $word_n$ in $prem_m$ **do**
8: $senti_n \leftarrow$ computed by Formula (3)
9: $weight_{m,n} \leftarrow$ computed by Formula (2)
10: $senti_m \leftarrow senti_m + weight_{m,n} * senti_n$
11: **end for**
12: **if** $senti_m > 0$ **then**
13: $poscnt \leftarrow poscnt + 1$
14: **else if** $senti_m < 0$ **then**
15: $negcnt \leftarrow negcnt + 1$
16: **end if**
17: $totalcnt \leftarrow totalcnt + 1$
18: **end for**
19: $pd \leftarrow \frac{poscnt - negcnt}{totalcnt}$
20: **return** pd

3.3. Social Influence

The influence of microblogs is reflected in the user's response behavior to microblogs, including reading behavior, such as behavior, comment behavior, and forwarding behavior. However, the influence of microblogs may be different. When the user has many fans, the microblogs posted by the user will be read by more people, which can enhance the influence of the microblogs. When there are many likes to a microblog, we can infer that the microblog is likely to be influential on the microblogging platform.

In this paper, we use Formula (6) to measure the influence of a microblog. Here, *FC* means the number of fans, *PC* is the count of praises, *CC* is the comment count, and *RC* is the count of re-posts.

$$influ_m = \ln (FC_m + 1) + PC_m + CC_m + RC_m \tag{6}$$

In order to reduce the over-impact of a large count in Formula (6), we use $norminflu_m$ to normalize $influ_m$ into a value range [0,1]. The normalization method is shown in Formula (7).

$$norm\_influ_m = \frac{1}{1 + e^{-\tan^{-1}(influ_m)}} \tag{7}$$

Note that the sentiment polarity of a microblog also contributes to the social influence of the microblog. For example, a piece of bad news always spreads fast on social networks. Therefore, this paper proposes combining microblog influence and sentiment polarity to evaluate the social influence of a microblog, which is shown in Formula (8).

$$sif_m = \begin{cases} (1 + senti_m) * (1 + norm\_influ_m) & senti_m > 0 \\ (1 - senti_m) * (1 + norm\_influ_m) & senti_m < 0 \\ 1 * (1 + norm\_influ_m) & senti_m = 0 \end{cases} \tag{8}$$

This paper considers the influence scores of microblogs with three sentiment polarities and calculates the influence of one day as the influence feature for information prediction. The influence combines positive scores, negative scores, and neutral scores, which is defined by Formula (9). Here, pos_sif_d is the sum of sif_m ($senti_m > 0$) within the day d, neg_sif_d is the sum of sif_m ($senti_m < 0$) within the day d, neu_sif_d is the sum of sif_m ($senti_m = 0$) within the day d, and $total\_sif_d$ is the sum of sif_m within the day d.

$$inf_d = \frac{2*pos\_sif_d - neg\_sif_d + neu\_sif_d}{total\_sif_d} \quad (9)$$

To sum up, the extraction algorithm of the microblog influence feature is shown in Algorithm 2.

Algorithm 2 *Influence Feature Extraction*

Input: *The set of all the related microblogs in one day,* M;
The sentiment of each microblog, $senti_m$
Influence of each microblog, $norminflu_m$

Output: *Influence Feature,* inf

$possis, negsis, neusis, totalsis \leftarrow 0$
for each *microblog*$_m$ in M **do**
  $norminflu_m \leftarrow$ computed by Formula (7)
  if $senti_m > 0$ **then**
    $sif_m \leftarrow (1 + senti_m) * (1 + norm\_influ_m)$
    $pos\_sif \leftarrow pos\_sif + sif_m$
  else if $senti_m < 0$ **then**
    $sif_m \leftarrow (1 - senti_m) * (1 + norm\_influ_m)$
    $neg\_sif \leftarrow neg\_sif + sif_m$
  else
    $sif_m \leftarrow 1 * (1 + norm\_influ_m)$
    $neu\_sif \leftarrow neu\_sif + sif_m$
  end if
  $total\_sif \leftarrow total\_sif + sif_m$
end for
$inf \leftarrow \frac{2*pos\_sif - neg\_sif + neu\_sif}{total\_sif}$
return *inf*

3.4. Prediction Model

Considering that product sales are typically counted in days, we make information predictions with days as granularity and assume $x_{t,i}$ is the ith feature extracted from the microblog on day t, y_t is the predicted value of day t, and y_1 is the predicted value for the first day. Therefore, the prediction model proposed in this paper can be represented by: $y_{t+1} = f(x_{t,1}, x_{t,2}, y_t)$.

Here, $x_{t,1}$ is the sentiment polarity feature extracted from the microblog on day t, $x_{t,2}$ is the influence feature of microblog extracted from the microblog on day t, y_t and y_{t+1} are the predicted values of day t and day $t + 1$, respectively. In other words, this paper uses the characteristics of sentiment polarity, microblog influence, and actual value of the previous day to predict the day's results.

After extracting the prediction features, we need to find a suitable model to describe the relationship between input (features) and output (real-world information). Therefore,

this paper selects the following two regression models to test the prediction ability of microblogs.

(1) Linear Regression Model (LR). The existing research on microblog-based information prediction usually assumes a linear relationship between variables. Therefore, this paper selects the linear regression model as the benchmark for comparison. In this paper, the LR model as a prediction model is represented as: $y_{t+1} = \alpha x_{t,1} + \beta x_{t,2} + \gamma y_t + \epsilon$, where α,β, and γ and the error ϵ are obtained from the training using the normal equation method.

(2) Support Vector Regression (SVR) [28]. In real life, the relationship between variables as a prediction may be very complex and not necessarily limited to a linear relationship. Therefore, some nonlinear models need to be used for prediction. Therefore, this paper selects the support vector regression model SVR (support vector regression) to describe the relationship between input and output. The SVR model is derived from the support vector machine (SVM) [29]. SVM is a binary classification model that aims to find a hyperplane to distinguish different categories. It has shown better performance in machine learning algorithms than many existing methods. In this paper, the SVR model as a prediction model can be described by: $y_{t+1} = w^T x + b$. Here, $w = (w_1, w_2, w_3)^T, x = (x_{t,1}, x_{t,2}, y_t)^T$. w^T is the normal vector determining the direction of the hyperplane. b is the offset, which determines the distance between the hyperplane and the origin point.

Note that there are also some other prediction models [30,31] that can be used in product sales prediction. In general, the LR model and the SVR model can be regarded as the two basic models. As we aim to demonstrate the feasibility of the proposed prediction framework based on sentiment polarity and social influence, we do not include more models in this study and save this issue as one to address in our future works.

4. Experiments

4.1. Settings

This paper selects two films as the experimental data, namely the domestic romantic comedy *Breakup Buddies* and the American science fiction film *Interstellar*. The screenings of the two films are shown in Table 2 and were screened for 34 days and 31 days, respectively.

Table 2. Movie information.

Movie ID	Movie Title	Start Time	End Time	Days on Show
#1	*Breakup Buddies*	2014-09-30	2014-11-02	34
#2	*Interstellar*	2014-11-12	2014-12-12	31

The dataset of this paper is the original microblog data within the corresponding performing period and the forwarding and comment data of these microblogs, which are crawled on Sina Weibo with the keywords of "Breakup Buddies" and "Interstellar," respectively. The box office statistics of films come from the daily box office situation released by the official WeChat "China box office" of the China film box office data center (http://www.cbooo.cn, accessed on 12 February 2022). The daily box office records are generally counted at about 16:00 the next day.

After data extraction, the microblog data of the two films are shown in Table 3. The *Breakup Buddies* forwarding rate was only 7.46%, 2273 of 30,484 microblogs were forwarded, and the comment rate was 30.82%. Although the film performed well in the comedy genre, its plot expression was not recognized by the audience. With the participation of comedians, it did not have a box office of CNY 1.267 billion in 2012, but it achieved a good overall result of CNY 1.169 billion. The forwarding rates and comment rates of *Interstellar* are slightly higher than those of *Breakup Buddies*, with 11.23% and 37.61%, respectively. The film also achieved a box office result of CNY 748 million in China.

Table 3. Statistics of the microblogs about the movies.

	Breakup Buddies	***Interstellar***
#Original microblogs	30,484	30,459
#Likes	57,323	73,453
#Comments	59,956	78,167
#Forwards	48,483	91,039
#Micorblogs liked	10,540	12,159
#Microblogs commented	9395	11,456
#Microblogs forwarded	2273	3420
#Most likes	6142	6336
#Most comments	3411	3782
#Most forwards	16,264	17,282

The user statistics of the two films are shown in Table 4. Since the maximum number of followers of Sina Weibo is 3000 (Sina Weibo restricts users' behavior of increasing the number of fans by paying attention to others), only the number of users with the maximum number of followers are counted.

Table 4. Statistics of users related to the movies.

	Breakup Buddies	***Interstellar***
The count of users	87,787	144,221
The highest fans number of users	13,640,601	33,332,635
The users with the most followers	28	47

The sentiment dictionary used in the experiment was independently sorted and labeled by the Information Retrieval Research Office of the Dalian University of Technology. The labeling of emotional words is graded. This paper does not consider the intensity of sentiment but only the positive and negative polarity. The number of topics selected in the LDA model is 100, the value of α is 0.5, β is 0.1, and the Gibbs sampling iteration is 2000 times. Both prediction models are implemented in the R language. The LR model uses the default function in the R language, and the SVR model uses the SVM function in the LibSVM package (LibSVM. https://www.csie.ntu.edu.tw/~cjlin/libsvm/, accessed on 12 February 2022.). The SVR model selects the linear kernel, and the penalty parameter C used in the kernel is 10.

To measure the prediction effect, we use mean absolute percent error (MAPE) [16], the mean value of the percentage error between the predicted value and the actual value, to compare the prediction effect of different features under different prediction models. The smaller the MAPE, the higher the prediction accuracy, and the better the model's performance. The calculation method of MAPE is shown in Formula (10). Here, n is the number of prediction rounds, y is the real value, and $\hat{y}$ is the predicted value.

$$MAPE = \frac{1}{n}\sum \frac{|\hat{y}-y|}{y} \quad (10)$$

4.2. Results

The film screening days are generally about 30 days. Therefore, for the datasets of the two films, this paper uses the data of two weeks or 14 days after screening to train a prediction model to predict the box office of the remaining screening days.

We first explore the impact of emotional propensity on the prediction effect. Then, the feature of sentiment polarity is introduced, and the model based on sentiment polarity can be represented by: $y_{t+1} = f(x_{t,1}, y_t)$.

We proposed to use the topic similarity between words and microblogs to calculate the sentiment polarity of the microblogs. Since the accuracy of subjective labeling microblogs' sentiment polarity to measure the sentiment polarity highly depends on human factors, we chose to compare the advantages and disadvantages of the method based on the sentiment polarity characteristics extracted by these methods for the predicted performance. Without considering the weight of words, we directly sum up the sentiment polarity value of words. The calculation method is shown in Formula (11). When $senti_m > 0$, the microblog expresses a positive sentiment. On the other hand, when $senti_m < 0$, the microblog expresses a negative sentiment. When $senti_m = 0$, the microblog expresses neutral sentiment.

$$senti_m = \sum_{n=1}^{N} senti_n \tag{11}$$

Table 5 shows the MAPE value of the LR model and the SVR model in predicting movie box office when considering or not considering the sentiment feature.

Table 5. MAPE with the sentiment feature.

Prediction Model	Features	*Breakup Buddies*	*Interstellar*
LR	sentiment (weighted)	0.471416	2.636955
	sentiment (not weighted)	0.543178	2.827116
SVR	sentiment (weighted)	0.347621	0.817183
	sentiment (not weighted)	0.363807	0.837459

To prove that the introduction of the microblog influence feature can improve prediction accuracy, we take the sentiment polarity feature alone as the comparison basis and observe the prediction effect of the proposed prediction framework on the film box office by comparing the two models. The box office prediction results of different influence features in different prediction models are shown in Table 6.

Table 6. MAPE with the influence feature.

Prediction Model	Features	*Breakup Buddies*	*Interstellar*
LR	sentiment (weighted)	0.471416	2.636955
	sentiment (weighted) + influence	0.365470	1.128859
SVR	sentiment (weighted)	0.347621	0.817183
	sentiment (weighted) + influence	0.335603	0.694272

As shown in Table 6, in terms of prediction accuracy, whether in the LR model or the SVR model, when the influence feature of a microblog is introduced, the prediction effect is improved, with higher accuracy and less error. Moreover, the error of the SVR model is

generally smaller than that of the LR model, which shows that compared with the linear relationship, the variables used for prediction are more likely to have a more complex nonlinear relationship.

We tested the effect of the microblog influence measurement method on the prediction performance and used the influence based on forwarded comments and the influence based on forwarded comments and user credibility as the comparison method. Based on the influence calculated by these two methods, we used the extraction method of microblog influence characteristics proposed in Section 3.3 to extract the influence feature of microblogs to test the prediction effectiveness. The comparative experiments on the prediction performance of different features show that the SVR model based on the weighted sentiment and influence features achieves the best performance, which achieves the minimum MAPE value on the two models and two datasets, indicating that the influence obtained by our proposed method can improve the accuracy of prediction.

Since the information prediction framework proposed in this paper is based on microblogs, we used the data without introducing any microblog characteristics (sentiment polarity characteristics and microblog influence characteristics). That is, only the box office of the previous day is used as the input for predicting the box office.

Figures 2 and 3 summarize the average prediction results of the LR and SVR models with different features. Here, we use the first movie *Breakup Buddies* as a case study. The data before 2014-10-15 is used as the training data. Then, we plot the prediction results of different models in the figures. To visualize the prediction precision clearly, we also show the real sales of the movie in the figures, which are taken from the China film box office data center. We can see that using the sentiment feature and the influence feature is helpful to improve the prediction precision.

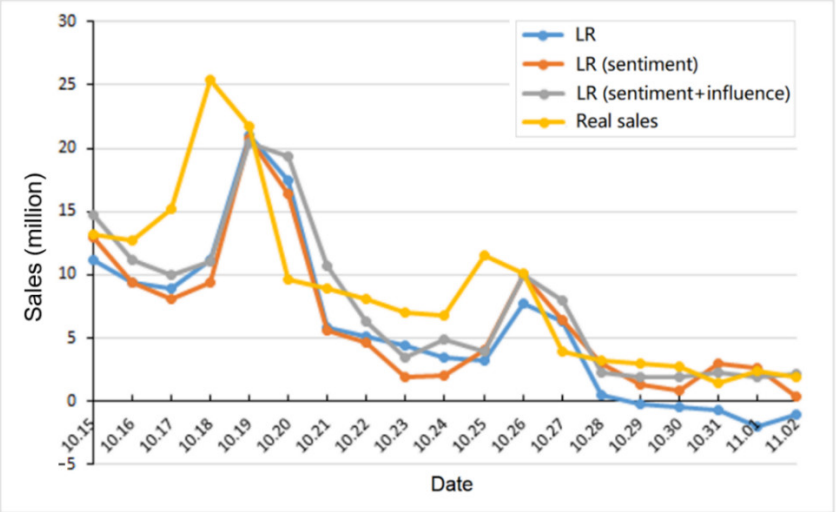

Figure 2. Prediction results of the LR model with different features.

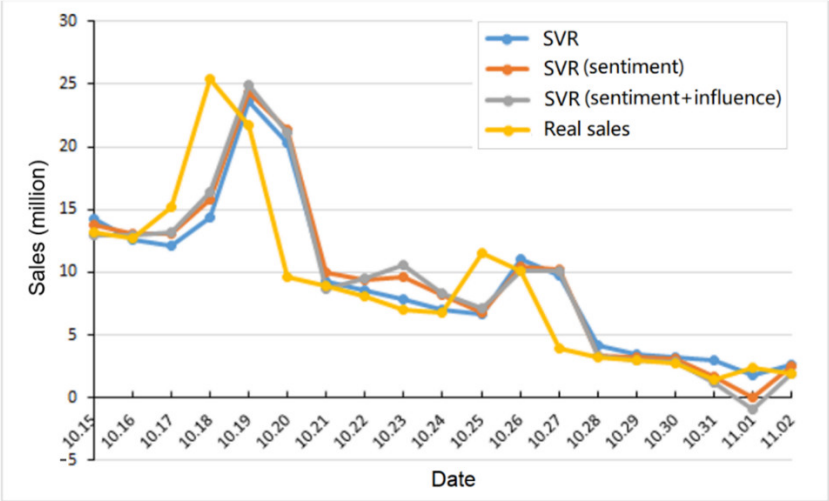

Figure 3. Prediction results of the SVR model with different features.

Figure 4 compares the real sales, the LR model with the two features, and the SVR model with the two features. Compared with the LR model, the predicted value of the SVR model is quite close to the real value on some days, especially from 28 October to 31 October, which almost overlaps with the real sales. Although these two prediction models still have some errors in the prediction of film box office, we can see that both can provide a precise short-term prediction for the box office. In real applications, enterprises can make decisions based on the predicted trend, which can benefit the short-term marketing strategy of enterprises.

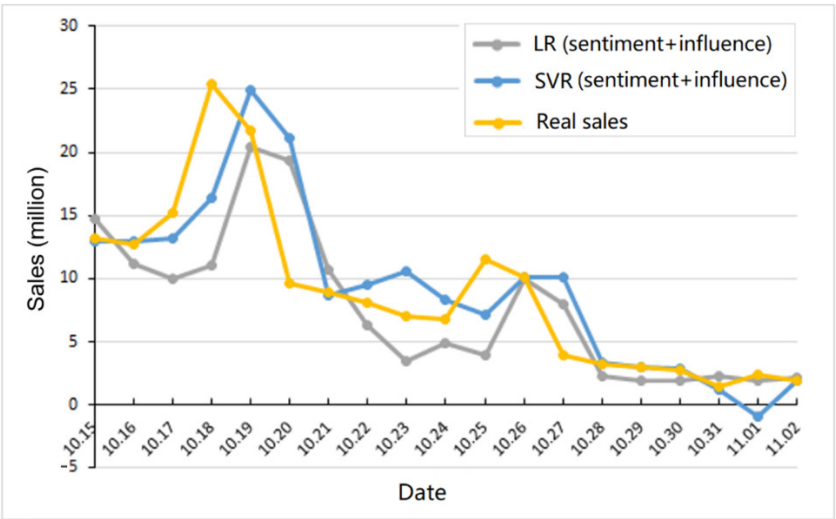

Figure 4. Comparison of the prediction models with real sales.

5. Conclusions and Future Work

With the explosive development of microblogs, predicting future information based on microblogs is a recent research focus. In this paper, considering the timely updates and interaction of microblogs, we presented a new framework to use microblogs to predict short-term product sales. The novel idea of the proposed framework is to integrate the sentiment polarity and the microblog influence into a prediction model, such as the LR model and the SVR model. We conducted experiments on a real dataset concerning the early box office prediction, which was crawled from Sina Weibo, the biggest microblogging platform in China. The results demonstrated that the proposed sentiment polarity and influence feature of microblogs play a positive role in improving the

prediction precision. In addition, both the LR model and the SVR model can lower the MAPE metric of the prediction effectively.

This study only presents an initial step towards short-term product sales prediction, and several limitations exist. First, real applications call for real-time processing and predicting frameworks, which are not tackled in this paper. Real-time processing of microblogs needs to devise some efficient stream-processing methods. Second, this paper only conducted a case study focusing on two small datasets due to the lack of public datasets.

In the future, we will investigate several issues. First, we will collect data from other fields and conduct experimental verification, such as real estate prices and car sales. Second, we will consider other prediction models, e.g., deep learning models [31,32], in our future work. Third, the sentiment polarity detection in this paper adopted a straightforward method. We will study more effective approaches, such as multi-dimensional sentiment analysis [33] and domain-specific sentiment analysis [34], to enhance the effectiveness of the sentiment analysis. Finally, we did not measure the credibility of microblog messages [35,36]. However, this issue has been recognized as a challenge for years. In the future, we will adopt existing approaches in information credibility evaluation to enhance the precision of the sales prediction.

Author Contributions: J.Z., conceptualization, funding acquisition, project administration, supervision, and writing—review and editing; F.X., data curation, methodology, validation, and writing—original draft preparation; P.J., project administration, supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Anhui Philosophy and Social Science Foundation (grant number AHSKY2021D15) and the Humanities and Social Sciences Research Project of the Anhui Provincial Department of Education (grant number SK2020A0036).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank the editors and anonymous reviewers for their suggestions and comments to improve the quality of the paper.

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

References

1. Mu, L.; Jin, P.; Zhao, J.; Chen, E. Detecting evolutionary stages of events on social media: A graph-kernel-based approach. *Future Gener. Comput. Syst.* **2021**, *123*, 219–232.
2. Jin, P.; Mu, L.; Zheng, L.; Zhao, J.; Yue, L. News feature extraction for events on social network platforms. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 69–78.
3. Asur, S.; Huberman, B. Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, Canada, 31 August–3 September 2010; pp. 492–499.
4. Bi, B.; Tian, Y.; Sismanis, Y.; Balmin, A.; Cho, J. Scalable topic-specific influence analysis on microblogs. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, NY, USA, 24–28 February 2014; pp. 513–522.
5. Afyouni, I.; Aghbari, Z.; Razack, R. Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey. *Inf. Fusion* **2022**, *79*, 279–308.
6. Mamo, N.; Azzopardi, J.; Layfield, C. An automatic participant detection framework for event tracking on Twitter. *Algorithms* **2021**, *14*, 92.
7. Kempe, D.; Kleinberg, J.; Tardos, É. Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003; pp. 137–146.
8. Liu, X.; Wu, S.; Liu, C.; Zhang, Y. Social network node influence maximization method combined with degree discount and local node optimization. *Soc. Netw. Anal. Min.* **2021**, *11*, 31.
9. Xiao, Y.; Li, J.; Zhu, Y.; Li, Q. User behavior prediction of social hotspots based on multimessage interaction and neural network. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 536–545.
10. Bo, H.; McConville, R.; Hong, J.; Liu, W. Social network influence ranking via embedding network interactions for user recommendation. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 379–384.

11. Xu, Y.; Liu, Y.; Zhang, X. Analysis of social network user behaviour and its influence. *J. Intell. Fuzzy Syst.* **2020**, *38*, 1159–1171.
12. Kong, S.; Feng, L.; Sun, G.; Luo, K. Predicting lifespans of popular tweets in microblog. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, USA, 12–16 August 2012; pp. 1129–1130.
13. Choi, H.; Varian, H. Predicting the present with google trends. *Econ. Rec.* **2012**, *88*, 2–9.
14. Jiang, W.; Wang, Y.; Xiong, Z.; Song, X.; Long, Y.; Cao, W. Detecting urban events by considering long temporal dependency of sentiment strength in geotagged social media data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 322.
15. Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; Tomkins, A. The predictive power of online chatter. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; pp. 78–87.
16. Du, J.; Xu, H.; Huang, X. Box office prediction based on microblog. *Expert Syst. Appl.* **2014**, *41*, 1680–1689.
17. Liu, Y.; Huang, X.; An, A.; Yu, X. ARSA: A sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, 23–27 July 2007; pp. 607–614.
18. Skoric, M.; Poor, N.; Achananuparp, P.; Lim, E.-P.; Jiang, J. Tweets and votes: A study of the 2011 Singapore general election. In Proceedings of the 45th Hawaii International Conference on System Sciences, Maui, HI, USA, 4–7 January 2012; pp. 2583–2591.
19. Salehan, M.; Kim, D. An investigation of predictors of information diffusion in social media: Evidence from sentiment mining of Twitter messages. In Proceedings of the 53rd Hawaii International Conference on System Sciences, Wailea, HI, USA, 7–10 January 2020; pp. 1–10.
20. Song, G.; Huang, D. A sentiment-aware contextual model for real-time disaster prediction using Twitter data. *Future Internet* **2021**, *13*, 163.
21. Sun, T.; Wang, J.; Zhang, P.; Cao, Y.; Liu, B.; Wang, D. Predicting stock price returns using microblog sentiment for Chinese stock market. In Proceedings of the 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), Chengdu, China, 10–11 August 2017; pp. 87–96.
22. Okimoto, Y.; Hosokawa, Y.; Zhang, J.; Li, L. Japanese election prediction based on sentiment analysis of Twitter replies to candidates. In Proceedings of the 2021 International Conference on Asian Language Processing (IALP), Singapore, 11–13 December 2021; pp. 322–327.
23. Jane, G.L.; Hari, S. Crime Prediction Using Twitter Data. *Int. J. e-Collab.* **2021**, *17*, 62–74.
24. Shahzad, M.; Bukhari, L.; Khan, T.; Islam, S.; Hossain, M.; Kwak, K. BPTE: Bitcoin price prediction and trend examination using Twitter sentiment analysis. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 20–22 October 2021; pp. 119–122.
25. Pai, P.; Liu, C. Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. *IEEE Access* **2018**, *6*, 57655–57662.
26. Zheng, L.; Jin, P.; Zhao, J.; Yue, L. A fine-grained approach for extracting events on microblogs. In Proceedings of the International Conference on Database and Expert Systems Applications, Munich, Germany, 9–14 September 2014; pp. 275–283.
27. Zou, X.; Yang, J.; Zhang, W.; Han, H. Collaborative community-specific microblog sentiment analysis via multi-task learning. *Expert Syst. Appl.* **2021**, *169*, 114322.
28. Yu, H.; Lu, J.; Zhang, G. An online robust support vector regression for data streams. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 150–163.
29. Sun, Q.; Tan, Z.; Zhou, X. Workload prediction of cloud computing based on SVM and BP neural networks. *J. Intell. Fuzzy Syst.* **2020**, *39*, 2861–2867.
30. Türkbayragí, M.; Dogu, E.; Albayrak, Y. Artificial intelligence based prediction models: Sales forecasting application in automotive aftermarket. *J. Intell. Fuzzy Syst.* **2022**, *42*, 213–225.
31. Khodabakhsh, M.; Kahani, M.; Bagheri, E. Predicting future personal life events on twitter via recurrent neural networks. *J. Intell. Inf. Syst.* **2020**, *54*, 101–127.
32. Chang, Y.; Ku, C.; Nguyen, D. Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry. *Inf. Manag.* **2022**, *59*, 103587.
33. Zheng, L.; Jin, P.; Zhao, J.; Yue, L. Multi-dimensional sentiment analysis for large-scale e-commerce reviews. In Proceedings of the International Conference on Database and Expert Systems Applications, Munich, Germany, 9–14 September 2014; pp. 449–463.
34. Fiok, K.; Karwowski, W.; Gutiérrez, E.; Wilamowski, M. Analysis of sentiment in tweets addressed to a single domain-specific Twitter account: Comparison of model performance and explainability of predictions. *Expert Syst. Appl.* **2021**, *186*, 115771.
35. AlRubaian, M.; Al-Qurishi, M.; Hassan, M.; Alamri, A. A credibility analysis system for assessing information on Twitter. *IEEE Trans. Dependable Secur. Comput.* **2018**, *15*, 661–674.
36. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on twitter. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 675–684.