symmetry

MDPI

Article

A Photo Post Recommendation System Based on Topic Model for Improving Facebook Fan Page Engagement

Chia-Hung Liao [1], Li-Xian Chen [2,*], Jhih-Cheng Yang [1] and Shyan-Ming Yuan [1,*]

[1] Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan; aiallen.cs07g@nctu.edu.tw (C.-H.L.); qqoo12342001.pcs05g@nctu.edu.tw (J.-C.Y.)
[2] School of Technology, Fuzhou University of International Studies and Trade, Fuzhou 350202, China
* Correspondence: lixian.cs98g@g2.nctu.edu.tw (L.-X.C.); smyuan@cs.nctu.edu.tw (S.-M.Y.)

Received: 9 June 2020; Accepted: 30 June 2020; Published: 2 July 2020

Abstract: Digital advertising on social media officially surpassed traditional advertising and became the largest marketing media in many countries. However, how to maximize the value of the overall marketing budget is one of the most concerning issues of all enterprises. The content of the Facebook photo post needs to be analyzed effectively so that the social media companies and managers can concentrate on handling their fan pages. This research aimed to use text mining techniques to find the audience accurately. Therefore, we built a topic model recommendation system (TMRS) to analyze Facebook posts by sorting the target posts according to the recommended scores. The TMRS includes six stages, such as data preprocessing, Chinese word segmentation, word refinement, TF-IDF word vector conversion, creating model via Latent Semantic Indexing (LSI), or Latent Dirichlet Allocation (LDA), and calculating the recommendation score. In addition to automatically selecting posts to create advertisements, this model is more effective in using marketing budgets and getting more engagements. Based on the recommendation results, it is verified that the TMRS can increase the engagement rate compared to the traditional engagement rate recommended method (ERRM). Ultimately, advertisers can have the chance to create ads for the post with potentially high engagements under a limited budget.

Keywords: Facebook advertising post; social media marketing; text mining; recommendation system; topic model; post engagement

1. Introduction

Web activity data, as in e-commerce, e-learning, e-government, social networks, and so on, represent diverse information that can provide useful data for particular users. Several studies have proposed a variety of recommendation systems to solve the problem of information retrieval and filtering. General used recommendation methods are content-based (CB), knowledge-based (KB) and collaborative filtering (CF) techniques [1]. However, CB and KB require a lot of domain knowledge and have limited expanded ability problems. CF has data sparseness, synonymous and shilling attacked problems. Many improved methods are proposed to solve these problems, such as social relationship-based recommendation systems [2–5] and context-awareness-based recommendation systems [1,6,7]. Recommendation systems have been widely regarded as an effective mechanism that contributes to social media companies' (i.e., Facebook, Instagram, LinkedIn, and Twitter) digital advertising aims and strategy.

Precise digital advertising brings greater business benefits to enterprises and customers. In 2017, Taiwan Media White Paper pointed out two important interpretations. First, digital advertising has accelerated growth and traditional media encounters are suffering the decline. Second, the growth and decline have changed faster in Taiwanese tradition and digital media [8]. Digital advertising volume

surpassed magazines in 2009 and the newspaper in 2012. In 2016, Taiwan's overall advertising volume reached 60.46 billion, of which digital ads were NT$25.87 billion, surpassing NT$22.53 billion of TV (including $19.16 for cable TV and $3.37 for wireless TV) ads for the first time and then digital media became the largest media. Therefore, how to effectively use the largest media in the advertising market is our goal.

Most social media managers have a heavy workload. In addition to spending a lot of time writing postscripts, adjusting photos, and even making videos, the fan page managers have to squeeze time to manipulate the ads. For example, Taiwan Apple's Daily fan page team needs to process more than 120 posts in a day. It is a difficult and time-consuming task for the fan page manager to pick out high-quality posts to create ads. Furthermore, managers painstakingly operating the fan pages have not received a relative return. The organic reach rate of Facebook posts all over the world declines year by year due to constant changes in the news feed algorithm on Facebook. According to Buzzsum's statistics of 880 million posts, the analysis of the engagement rate dropped by 20% from 2016 to 2017 [9]. This is viewed as Facebook's alternative claim for advertisers to improve the quality of their material or to spend more money on advertising to maintain the discussion of the fan pages.

This research aims to help advertisers or social media managers to concentrate on the content of their fan pages. Therefore, the advertising part is handed over to our topic model recommendation system (TMRS). We use text mining technology to automate the selection of posts with a high engagement rate. Thus, this system can help advertisers to get the most benefit within the same advertising budget. In response to the above issues, (a) we choose posted photo posts to be the training data. (b) Then, input the texts of the target post into the trained topic model, (c) find similar ad posts in the training set, (d) sort these similar ad posts in the order of cosine similarity, and (e) take the appropriate number of ad post samples. (f) Then, use the advertising insight data of these similar ad posts, such as positive feedback filed, to make the weight for the recommendation score. The positive feedback field has three levels, which have been verified to be highly correlated with the cost per post engagement (CPE). i.e., each target post can use the topic model to find their own similar ad posts, and then combine the similarities with positive feedback levels to calculate the recommendation scores to make recommendations for these target posts.

The TMRS includes six stages. First, we preprocess the data. Second, the Chinese word segmentation. Third, we do the word refinement, which means the words that would not be the topic of the post will be removed after the word segmentation. Fourth, the words are converted into TF-IDF vectors. Fifth, we use Latent Semantic Indexing (LSI) or Latent Dirichlet Allocation (LDA) to create a model to identify potential topics or features of the ad post texts. Finally, after feeding the target post texts into the trained topic model, the similarity calculation is performed and the similar post texts are output. Use the positive feedback levels and the similarities of the similar ad posts to calculate the recommendation score for the target post.

The rest of this paper is organized as follows. Section 2 talks about the background knowledge and the related work of recommendation systems, recommendation techniques, and topic modeling. Section 3 presents how we analyze important advertising insight data. Section 4 introduces the procedure of building the model structure. Section 5 describes the experiment scenarios and dataset. Section 6 shows the way we decide the model hyperparameters such as the number of topics and the number of sampling. Section 7 illustrates the idea of how we evaluate the effectiveness of TMRS. Section 8 discusses the summary of the results. Finally, Section 9 gives the conclusion of this study.

2. Related Work

2.1. Recommendation Systems

Recommendation systems (RSs) attempt to recommend the most effective items (advertisements, products, or services) to particular users (individuals, social media managers, or advertising companies). Those use some relevant item information and the interaction between users and items to predict a

user's interest [6,7]. These systems are really critical in specific industries as they can generate a large amount of profit when they are efficient or also be a way to transcend significantly from competitors. RSs methods have been developed by academic researchers and applied in a variety of different social media applications, including marketing, movie box-office, information dissemination, elections, macroeconomic, and many others [10].

Our research mainly focuses on social media marketing, especially on Facebook advertising. The methodologies may include text mining, topic model, document similarity, and recommendation system. Therefore, existing recommendation systems relevant to this study can be roughly categorized into two different groups: content awareness-based and social relationship-based recommendation systems.

2.1.1. Context-Awareness Recommendation Systems

Traditional content-based recommendation systems use a lot of information generated by a large number of user activities to analyze group preferences. Content-based recommendation system refers to the description of the product and the configuration file that matches the active users' interests to suggest products that are similar to those that the active user used to like [11,12]. News content was analyzed by using Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) method to make personalized recommendations [13]. They explored the internal relationships between news articles, and the different characteristics of news items. Effective clustering of newly published news articles, as well as high-level recommendations. Moreover, a new method for Facebook fan page ranking that the ranks of pages estimated by this new method are close to the ranks estimated by an engagement-based method [14]. The traditional ranking methods rely on user engagement including the number of posts, comments, and "likes". However, the polarity of each comment is ignored in these methods, which can be positive, neutral, or negative. It has developed a content-based ranking method that takes into account users' engagement and comment polarity. In addition, the new page ranking method concerning the comment polarity is more accurate regarding users' opinions.

Social media identify individuals shared a connection with others and view their connected information within a public or semi-public profile system. Facebook, a giant social media in the world, always refined their recommendation systems. Content analysis in the official Facebook pages of 70 global brands was used to explore the companies' marketing and advertising strategy in social media [15]. Different fan page types will post in different ways. Interestingly, it also found that a large number of fans on a fan page cannot clearly measure sales figures or purchase intentions. This gives us a perspective on how many fans may not be the focus, but the quality of the post or the product itself. Moreover, Facebook's popularity and effectiveness are largely related to the content and semantic of the posts. The popularity of the post and the engagement rate were used to dynamically adjust the model parameters [16]. The purpose was to increase the exposure of the fan page and apply it to the political fan page. There is also a multifactor model that shows how time, the number of people and their genders, and how the media type contributes to the popularity and effectiveness of the post. It is also found that a fan page with more fans does not necessarily lead to more popular posts or higher engagements. For analyzing the text and photo posts, we need to consider the semantic and context of the posts.

2.1.2. Social Relationship-Based Recommendation Systems

The users' trust relationship, direct trust, and indirect trust are established according to whether or not users directly give trust value [3]. People can receive friends' recommendations through social media such as Facebook or Twitter. The trust relationship between users in a social network can be inferred based on past user interaction records or users' specified items. Users can also use indirect data like rating information to compute and infer their trust relationships. Text mining was an effective method to explore business value from a large amount of available data. The value of social media competitive analysis was demonstrated by analyzing the text content on Facebook and Twitter of the

three largest pizza chains [17]. Its recommendation system provides help to companies to develop their social media strategy. It is also found that references to other competitors' posts with high engagements published in their own social media with the same concept had good engagement with customers.

Many social relationship-based recommendation techniques were analyzed according to implicit trust-based information like user trust relation, interaction, product popularity, and user credibility [3,18]. Users' implicit trust relationship and corresponding degrees can be inferred by their common items with coratings and social networks' role importance. Some studies using machine learning techniques, Connectionist Inductive Learning, to generate recommendations in Web communities or supporting Web navigation [19]. Social relationship-based recommendation systems discriminate against the users' commonalities according to their ratings and generate new recommendations considering the comparisons between different users. Our previous work analyzed the separation degree problem from two aspects: (a) between two normal-persons or famous-persons and (b) two individuals with special characteristics [20]. The six degrees of separation theory was re-evaluated and extended by using the real Facebook tool "We T So Close." The result pointed out that the average acquaintance number was 3.9 regardless of two normal individuals or two persons with rare features. This study aims to select posts with a high interaction rate automatically, so the designed recommendation system needs to consider implicit trust-based information.

2.2. Topic Modeling

In order to investigate the representation of documents, generative topic models, such as Latent Semantic Indexing (LSI) [21,22] and Latent Dirichlet Allocation (LDA) [23,24] are widely adopted methods. Using knowledge representation form established by term frequency-inverse document frequency (TF-IDF) or bag-of-words (BOW) can improve LSI and LDA methods' effects. Topic models refer to a statistical model used to discover abstract topics in a series of documents [25,26]. Intuitively, if an article has a central idea, then some specific words will appear more frequently. For example, if an article is about a dog, the words "dog" and "bone" appear more frequently. If an article is about cats, the words "cat" and "fish" appear more frequently. However, the real situation is that an article usually contains multiple topics, and each topic has a different proportion. Therefore, if an article is 10% related to cats and 90% is related to dogs, the number of keywords associated with dogs will be approximately nine times the number of keywords associated with cats. A topic model uses a mathematical framework to implement the document feature and it decides what topics are included in the current document by analyzing each document, the word counts in the document and other statistical document information.

LSI refers to how to find the relationship between words through massive literature [21,22]. When two words or a group of words appear in the same document in numbers, they can be considered semantically related. LSI uses singular value decomposition (SVD) to decompose the word-document matrix. SVD maps the original data into the semantic space by finding irrelevant index variables so that two dissimilar documents in the word-document matrix may resemble in the semantic space. In addition, LDA is a document-generation model that considers each topic in an article corresponding to different words [23,24]. In the process of constructing an article, it will select a topic with a certain probability, then selects a word with some specific probability under this topic, and finally generates the first word for this article. Repeat this process and generate an entire article. LDA assumes that there is no order between words. At the same time, it is an unsupervised learning algorithm and it does not need to manually label the training set. Using LDA only requires the document set and specifying the number of topics. Moreover, LDA can find some words for each topic to describe it. In order to select a topic model suitable for analyzing social media copywriting, this study uses LSI and LDA topic models for analysis.

Using the topic model toolkit, Gensim, to process corpus data created by LSI and LDA topic models can analyze paradigmatic and syntagmatic relations between lemma within topics [27,28]. The topical similarity can be queried between plain text documents and other documents when the semantic

topics were found. Gensim is a free python module dedicated to working with raw, unstructured text that automatically extracts semantic topics from documents [29]. Modules are developed from three concepts: corpus, vector, and model. According to the topic model toolkit Gensim official document, the latent semantic index (LSI) converts documents from word bag or TF-IDF weighted space to lower-dimensional potential space. The 200–500 topic dimension is recommended as the "gold standard." However, this standard is suitable for long articles. We investigate the social media managers' community copies. The findings show the average number of words is about 50–150 words and the number of concepts to be expressed is about five or so. Therefore, in our research, we will not use the number of topics in the general article, but use the number of topics from 1 to 15 to build a topic model experiment on the TMRS. Moreover, the computer cuts the Chinese characters into units of "meaning" that are important [30]. Without special treatment, the computer will treat each Chinese character separately, but this is meaningless for analyzing semantics and potential topics. To process the Chinese word segmentation correctly, Jieba library was needed to import in this study. Jieba is an open-source project. This Chinese word segmentation program is written by a developer of Baidu in China [30]. Its core is actually Simplified Chinese. However, since it is an open-source project, there are already enthusiastic developers on the Internet plus a traditional Chinese dictionary.

3. Advertising Insight Analysis

3.1. System Overview

Facebook posts can be composed of a variety of attributes, such as texts, images, photos, videos, and call to action [31]. The text is an attribute that can be found on every fan page. If the post texts are well written, it will resonate with the user. The TMRS obtains a weighted score by calculating the cosine similarity [32] of current texts for the target post and past texts for the ad posts. The higher the score, the higher the post engagement that the system predicts will be, and the lower the cost per post engagement (CPE). When each target post is fed into the TMRS, it can form a recommendation order according to the weight score, and provide a priority reference for the social media manager to post advertising.

3.2. Advertising Performance Indicator

In order to evaluate our system, we should compare the results of TMRS with the existing system, engagement rate recommended method (ERRM). Its recommended ranking of the post is based on the level of post engagement rate (PER), and the PER is calculated as follows:

$$PER = \frac{post\_engaged\_users}{post\_impressions\_unique} \quad (1)$$

post_engaged_users is the number of post engagements that users interact with the post, after posting the ad. post_impressions_unique is the number of exposures that the post appears on the users' screen. Both post_engaged_users and post_impressions_unique are collected from Facebook Graph API.

According to Facebook's official document, Facebook's advertising insights API provides a variety of advertising insights for developers [31]. Post engagements refer to all actions taken by the user for the advertisement during the delivery of the advertisement; for example: convey a mood, leave a message or share, request for a discount, view photos or videos, or click a link. In the case of a limited marketing budget, the lower the CPE is, and the more user engagements the ad post gets. The public API can collect lots of the average cost of post engagements data. Our goal is how to effectively use the largest media resources in the advertising market so that we are most concerned about the CPE. Therefore, we hope to find the field of Facebook associated with the post engagement most.

This indicator is calculated by dividing the total cost by the number of post engagements, which is shown as Equation (2):

$$CPE(Cost\ per\ Post\ Engagement) = \frac{total\_ad\_spending}{num\_post\_engage} \quad (2)$$

where total_ad_spending is the total amount of ad post spending, and num_post_engage is the number of post engagements.

In addition, Facebook ads will have different benchmarks for calculating post engagement depending on the type of post. For example, the movie has three seconds, 10 s of views. The photo has photo clicks, etc. In order to avoid the difference of the benchmark, we have chosen the most popular type of post, photo post for experiment, and analysis.

3.3. Ad Insights Select

3.3.1. Relevance Score

The role of the relevance score is to allow the advertiser to evaluate how much the ad resonates with the user he or she wants to reach. The higher the relevance score of an ad post, the better the performance of the ad. This score is based on a comparison between the ad posted by social media manager and other ads that lock the same customer. The factors also include positive feedback (ex: clicks, app installs, video views) and negative feedback (ex: someone clicks "I don't want to see this" on your ad). The relevance score is scored on a scale of 1–10.

3.3.2. Observation

We first analyze the relevance scores in the advertising insights to see if the relevance score can be used to judge the quality of the post texts. In addition, the relevance scores have extended fields, which are positive feedback and negative feedback. The feedback level of the advertisement may be low, medium, or high.

First, we want to find out how the relevance score of the post is related to post engagement, and we will calculate the correlation coefficient between them. In addition, Facebook fan page types are divided into 10 categories, and each category is subdivided into different items. In every experiment, we pick out the same category of related fan pages and calculate the correlation coefficient between their CPEs and relevance scores.

3.3.3. Correlation Coefficient

After the analysis, the correlation coefficient of the relevance score is not very high. Then, we take the extended two fields at the same time, the effect after the analysis is not good too. It may be because the negative feedback does not reflect on the CPE. Finally, we pick positive feedback for analysis, as shown in Table 1. The correlation coefficient between positive feedback and cost per post engagement has come to −0.65, which is strongly correlated. Table 2 shows the degree of correlation strength. It means that the higher the positive feedback level is, the lower the CPE will be. This trend can also be seen in Figure 1.

Table 1. The correlation coefficient between ad insights and the cost per post engagement.

Ad Insights Fields	Correlation Coefficient
Relevance Score Field	−0.41
Negative and Positive Feedback Field	−0.32
Positive Feedback Field	−0.65

Table 2. Pearson product-moment correlation coefficient table.

Degree of Relationship	Negative	Positive
No relationship	−0.09 to 0.0	0.0 to 0.09
Weakly correlated	−0.3 to −0.1	0.1 to 0.3
Moderately correlated	−0.5 to −0.3	0.3 to 0.5
Highly correlated	−1.0 to −0.5	0.5 to 1.0

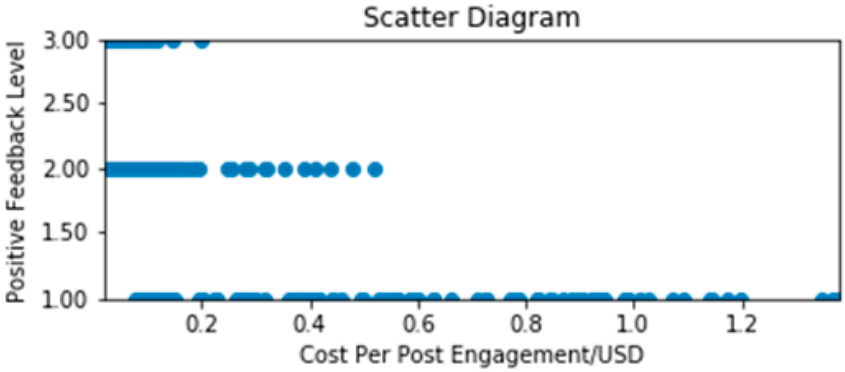

Figure 1. Scatter diagram for positive feedback and cost per post engagement.

Therefore, in our study, we will use a positive feedback level to be the main weighted factor for calculating the recommendation score.

4. System Architecture and Implementation

In order to provide a recommended post list with high engagement potential for social media managers, we design a system for computing a recommendation score by comparing the target post and ad posts. We use Facebook Graph API to get the post data which we need, then input them to model and get the score. Finally, sort the score from high to low. There are six stages for the system: preprocessing, word segmentation, word refinement, TF-IDF vector conversion, creating the LSI/LDA model, and calculating the recommendation score. The system structure is shown in Figure 2.

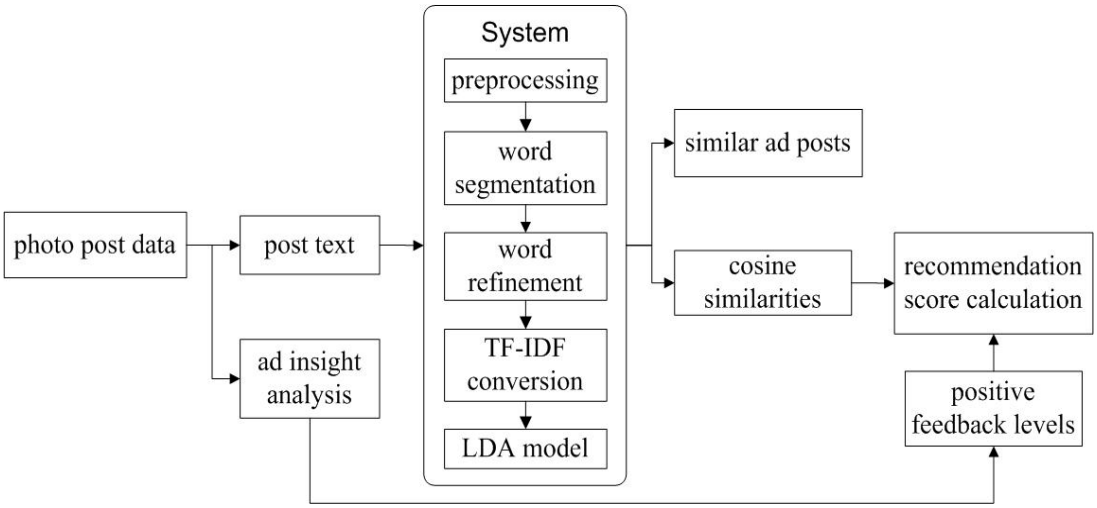

Figure 2. Schematic for system structure.

Procedure

Stage 1: Preprocessing

The actual data will be affected by different factors, so there may exist extreme value. In order to prevent the extreme value from affecting the result accuracy, and avoiding the influences on analyzing posts, we preprocess the data. First, the model removes the extreme value. If there exists some extreme value in the data that has a big difference with others, the credibility of the overall data may be reduced. Therefore, we remove the data in which the CPEs fall outside the two plus and minus standard deviations from the mean. Through this step, we can prevent the overall data from being affected by values that are too large or too small. Then, the model removes special characters. After we got the post from Facebook Graph API, it may contain emoji and special characters, for example, ♡ or line break symbol. It is relatively irrelevant to the quality of the post content. We hope to retain only the story or artistic concept of the post, therefore, we use the program to remove these special characters from the post. Additionally, there are some URLs in the post. These URLs may be an official website or event registration page, but the URL has nothing to do with the quality of the post and will be removed from the post texts here.

Stage 2: Word Segmentation

Chinese word segmentation [33,34] is the most important preprocess in Chinese. If the Chinese word segmentation correctly identifies the words with the smallest unit of meaning, we may have a way to conduct higher-level natural language analysis. This study used Jieba participle (an open-source project) to do Chinese word segmentation. After doing participles, a sequence of words is regrouped into a sequence according to certain specifications. Therefore, the correctness of the Chinese word segmentation has affected the success or failure of many natural language processing applications.

Stage 3: Word Refinement

We remove the words that should not be the topic of the post after the word segmentation. For example, if words such as "it, is, that", are not removed and appear many times in the post, it will be misunderstood for the post topic. Therefore, before training the model, it is necessary to remove such words from the bag of words after the word segmentation. There are three steps to do for word refinement. First, synonym replacement replaces words with the same or similar meaning, such as wine and spirit. In the post texts, it would be better if the words with the same meaning are expressed by the same word, to ensure better performance when calculating the similarity of the posts [35]. Second, removing the brand or product name from the words bag makes this recommended system common to any fan page copywriting. If the brand or product name do not be removed from the post texts, the model will misjudge them to be kinds of topics when doing text analysis. Moreover, the brand and the product name will disturb the similarity and will make the post texts too similar to each other. Finally, removing the hashtag lets irrelevant text be deleted. The hashtag is composed of # with a word or a sentence without spaces. Users can link to the same platform with the same hashtag. The reason for the removal of the hashtag is the same as the brand name.

Stage 4: TF-IDF Conversion

This study uses the Gensim module in the topic model to convert words into vectors and feed them to the TF-IDF model. TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and text mining [36,37]. TF-IDF is a statistical method that is used to evaluate the importance of a word for a document in a group of documents or corpus. The importance of a word is proportional to the number of times it appears in the document, but the word importance also decreases inversely with the frequency it appears in the corpus. After TF-IDF conversion, the meaningful words' weights will be increased.

Stage 5: Create the LSI/LDA Model

After the TF-IDF vector conversion, each word has its own weighted vector. Then, use these weighted vectors and specify the number of topics via Gensim library, the LSI/LDA model is generated separately for the cosine similarity of the subsequent target post.

Cosine similarity is commonly used for file comparison in text mining, and the similarity between them is measured by the cosine of the angle between the two vectors [32]. Cosine similarity is usually used in positive spaces and the value is between 0 and 1. For example, cosine similarity is one when two vectors have the same orientation and the value is 0 when two vectors angle is 90°.

Stage 6: Recommendation Score Calculation

After the LSI and LDA are established, the target post can be fed into these trained topic models to calculate the similarity between the target post and each ad post in the training set. Then, according to the similarity order, output those indices of similar ad posts. The indices here are the index numbers of the ad posts in the training set. Then, return the advertising data of the ad posts and observe its positive feedback levels. Use these levels to calculate the recommendation score of the target post, which is calculated by Equation (3).

$$\hat{R}_i = \frac{\sum \text{sim}\left(a_i, t_j\right) \times p_i}{\sum \text{sim}\left(a_i, t_j\right)} \tag{3}$$

$\hat{R}_i$: The target post recommendation score predicted by the similar ad posts. $\sum \text{sim}\left(a_i, t_j\right)$: Cosine similarity of the target post the ad posts. p_i: Positive feedback level of ad post (high = 3, medium = 2, low = 1).

For example, assuming that the target post texts are fed into the system, the system takes the positive feedback rating of the first 10 most similar posts, like Table 3. We put the "high" level for three points, "medium" for two points, and "low" for one point. The level scores are multiplied by the similarity then added, and finally divided by the total score. This is the final recommendation score.

Example: $(0.99123 \times 3 + 0.97456 \times 2 + 0.96111 \times 2 + \ldots \ldots + 0.86666 \times 1) / (0.99123 + 0.97456 + \ldots \ldots + 0.86666)$

Table 3. Similar ad post index, similarity and its positive feedback level (example).

Ad Post Index	Positive Feedback Level	Cosine Similarity
3	high	0.99123
9	medium	0.97456
10	medium	0.96111
90	high	0.96000
100	low	0.95444
200	low	0.93214
305	medium	0.93001
446	medium	0.88888
555	medium	0.87777
666	low	0.86666

5. Scenarios and Dataset

5.1. Experiment Scenarios

To select a topic model suitable for analyzing social media copywriting, this paper designs three scenarios for experimentation which are shown by Table 4. In the first scenario, we consider the recommendation effectiveness of the post texts of the wine fan page under the LSI and LDA models. The second scenario is to use the ad post screened by marketing experts from the wine fan pages, then re-generate LDA models and check the recommendation effectiveness. The third scenario will be based on the above two experiments, to see which way will win the most. Then, we select the best way to do

experiments on different types of fan pages. In scenario3, we choose makeup/skincare fan pages to test our TMRS.

Table 4. Three kinds of scenarios.

Scenario1	All the advertising post texts of the Wine/Spirits fan pages.
Scenario2	Selected ad post texts of the Wine/Spirits fan pages by three marketing experts.
Scenario3	Apply the better solution from 1 and 2 to the Makeup/Skincare fan pages.

5.2. Marketing Expert Screening

In the experiment of a photo post, which is composed of photos and texts, our experiments focus on analyzing the post texts. Therefore, in order to reduce the variation factor of the photo, we invited three marketing experts to vote for each ad post to see whether the positive feedback of each post was influenced by the texts or the photo or the half. When an ad post is more than 1.5 points after the experts' vote, it will be selected into our data set. The voting score rules are shown in Table 5.

Table 5. The voting score rule for marketing experts.

Score	Voting of Rule
0	The positive feedback level of this ad post is mainly caused by the photo.
0.5	The positive feedback level of this post is mainly caused by the photo and the texts.
1	The positive feedback level of this ad post is mainly caused by the texts.

5.3. Dataset

If one wants to do a text analysis of ad posts, one must use the past post data. This study used Facebook's Graph API to access individuals' information without requiring their passwords. After accessing fan pages' tokens, this API collected the post data including manage_pages and ads_management to do the following analysis. Instead of the common 80:20 rule, we use older data as the training data, and later data as test data. Thus, it can be in line with the actual advertisement created by social media managers. In this case, we can also know whether our TMRS will get more post engagements (PEs). We use the ad post and ad data that implement the promotion of PE ads from Mar. 2015 to Mar. 2018 as training data for scenario1 and scenario2. The training data of scenario3 come from Oct. 2016 to Mar. 2018. Test data were all obtained from Apr. 2018 to Jun. 2018. Table 6 shows the ad post data for different scenarios. Taking scenario1 as an example, in the training data that actually has the post for creating ads, there are 688 posts with delivery data, from 18 wine-related fan pages. Test data has 92 posts for 11 fan pages to create ads. Table 7 shows the important ad data for different scenarios, such as their total advertising spending (AS), post engagements (PE), cost per post engagement (CPE), total post moods, and exposures. Taking the training data from scenario1 as an example, its total AS is 14,253528 New Taiwan dollars (NTD), total PE is 3,039,045 times, and CPE is 4.69 NTD which is calculated by using AS, PE, and Equation (2). It has 2,353,152 post moods and 141,386,675 exposures. Scenario2 and scenario3 are similar, and so on. PE includes all actions taken by the user for the ad during the delivery. PE includes the following actions: conveying moods, messages or sharing, requesting offers, viewing photos or videos, or clicking on a link. Post mood is the amount of mood the ad receives. The mood button of the ad post allows the user to express different moods for the post, such as "like", "big heart", "ha", "wow", "cry" or "angry". Exposures are the number of times an ad appeared on the user's screen. When the ad is first displayed on the user's screen, it is counted as one exposure. (For example, if a user scrolls down after seeing an ad and then scrolls back up to the same ad, it counts as one exposure. If the user sees the same ad two times a day, it counts as two exposures.)

Table 6. Ad post data for different scenarios.

Experiment	Use	Time Interval	Total Fan Pages	Total Posts
Scenario1	Training data	March 2015 to March 2018	18	688
	Test data	April 2018 to June 2018	11	92
Scenario2	Training data	March 2015 to March 2018	18	411
	Test data	April 2018 to June 2018	11	77
Scenario3	Training data	October 2016 to March 2018	20	590
	Test data	April 2018 to June 2018	8	104

Table 7. Ad data for different scenarios.

Experiment	Use	AS (NTD)	PE	CPE (NTD)	Post Moods	Exposures
Scenario1	Training data	14,253,528	3,039,045	4.69	2,353,152	141,386,675
	Test data	688,398	241,564	2.84	155,212	3,225,919
Scenario2	Training data	6,825,854	1,571,306	4.34	1,116,576	60,549,680
	Test data	419,032	164,317	2.55	101,236	2,041,252
Scenario3	Training data	6,926,666	2,365,664	2.93	1,313,777	51,991,448
	Test data	862,344	364,772	2.36	111,879	9,726,080

6. Model Hyperparameter Selection

6.1. Number of Topics

The topic model refers to a set of methods for extracting hidden topics from a document [26]. When training the model, we need to set the number of topics in advance, manually adjust the parameters according to the results of the training, optimize the number of topics, and then optimize the text classification results. The length of the post texts in social advertising is generally not too long, so the experiment will set the number of topics to 1–15 and use the training data to obtain the best number of topics for the TMRS.

6.2. Number of Samples

When calculating the recommendation score, how many most similar post samples are needed to be taken from the training set? Through the experiment, we will test by sampling 1% to 10% of the total number of training data to obtain the most suitable number of samples for the TMRS.

6.3. Settings and Methods

First, after the training data are preprocessed, the ad post texts are sent to the LSI and LDA models respectively. The difference between each model is the number of topics. Then, compare the CPE of each monthly ad post list of each fan page of the training data, and then decide the best number of topics for each scenario. While experimenting with the best number of topics, we also experiment with the optimal number of samples required for TMRS. We took 1% to 10% of the total number of training samples. For example, there are 411 posts in the training data of scenario2. We will take 4, 8, 12, . . . , ad posts to be the similar numbers of samples, and use these numbers of samples to calculate the recommendation score.

Then, we segment the training data of scenario2 according to the fan page and the month, to form a total of 46 cases. Note that we have removed the case where there are only one or two ad posts for the month. We use these segmented cases to compare the recommendation effectiveness of ERRM and TMRS. If the CPE from the TMRS is relatively low, the number of topics and the number of samples of the model are recorded, and the LSI-based TMRS number of wins table is constructed as shown in Table 8. Then, use the table lookup method to find the combination of the number of topics and the number of samples, that this best combination means TMRS has the most wins. If the best

combination has more than one, choose a smaller number of topics and the number of samples as the optimal combination to reduce the time to build the model.

Table 8. Wins count table for the LSI-based topic model recommendation system (TMRS) in scenario2.

Sampling Topics	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
1	22	22	22	22	22	22	22	22	22	22
2	20	21	22	20	20	20	20	20	20	20
3	22	22	22	20	20	20	21	20	20	20
4	21	23	21	21	21	21	20	20	20	21
5	23	23	22	22	22	22	22	22	22	21
6	21	22	23	23	23	22	22	22	21	21
7	23	24	22	23	22	22	22	22	21	21
8	24	24	23	24	23	23	21	22	22	21
9	21	22	21	21	21	21	21	21	22	22
10	24	24	22	21	21	21	21	21	21	20
11	21	23	23	23	21	21	22	21	22	21
12	22	24	23	22	21	21	21	21	21	22
13	23	24	22	21	22	21	20	20	19	19
14	23	23	22	21	20	21	21	21	21	21
15	24	22	23	24	22	22	22	22	21	20

Construct the tables for the LSI and LDA of each scenario in the same way, and take the most wins combination of the number of topics and the number of samples. The obtained (sampling number, topic number) of each scenario is shown in Table 9. Taking the LSI of scenario2 as an example, the training data can be segmented into 46 cases, and the maximum number of wins of TMRS is 24. Therefore, the number of topics suitable for the wine fan page is set to seven for the LSI model, and the words that make up a certain topic are, for example, "Activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends ", and the top 2% of samples for the ad posts are used to evaluate the recommendation scores. Others and so on.

Table 9. The results of model hyperparameter (sampling and topics) after experiments.

Experiment	Model	Cases/Max Wins	Sampling	Topics	Representative Words for One of the Topics (Example)
Scenario1	LSI	48/26	7%	2	Activities, classics, events, tastes, fans, messages, flavors, first time, sharing, original intentions
	LDA	48/28	4%	12	Absolute, travel bag, taste, flavor, aroma, oak barrel, limited edition, one bite, departure, greet
Scenario2	LSI	46/24	2%	7	Activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends
	LDA	46/24	2%	15	Cherry blossoms, appearance, flower season, cans, faces, couples, friends, aftertaste, rogue, lobster
Scenario3	LSI	58/34	1%	13	Official website, consumption, limited gift, reward, purchase, discount, full, forgive, exclusive, gift
	LDA	58/34	8%	13	Official website, exclusive, essence, purchase, repair, moisturizing, activities, skin, discount, reservation

The test data uses LSI- or LDA- based TMRS to compare their CPE with traditional ERRM to see how the recommendation effectiveness works. Test data is also segmented according to the fan page and month. The test data for each scenario form 20 cases. If the fan page has three target posts in the month, take the first one for creating an ad, and take the first two to do so if the fan page has six target posts, and calculate the average CPE of these first posts, and so on.

7. Evaluation Method for Recommendation

After building the model by using training data and setting the model hyperparameters, we use the test data and go through the following steps to evaluate our TMRS. Here, we show an example that

the results are shown in Table 10, to illustrate the idea of the evaluation method. Table 10 shows the effectiveness of LSI-based TMRS by using test data in scenario1. In accordance with the usual habits of the social media managers, we do the sorting of the posts in cases with monthly units. (Let $Apr\_P_i^A$ denotes the i^{th} post of fan page A in April. In this case, i ranges from one to five.)

Step 1 Start from the fan page A in April, and this case has five target posts.

Step 2 Sort these five posts by traditional ERRM and TMRS, respectively.

Step 3 Take the first two posts respectively from ERRM and TMRS, and calculate their average cost per post engagement (ACPE). (Take the same fan page month as the case unit, choose the first one if there are three posts, or choose the first two if there are six posts, etc.)

Step 4 Compare which ACPE is lower and decide whether the ERRM wins or TMRS wins. For the fan page A in April, the ERRM-ACPE is 2.52 NTD, and it is lower than the TMRS-ACPE 2.85 NTD. Therefore, the TMRS loses this round.

Step 5 Calculate the CPE gain (CPEG). Here, the CPEG is −13%, which is calculated by Equation (4)

$$\text{CPEG} = \frac{\text{ERRM\_ACPE} - \text{TMRS\_ACPE}}{\text{ERRM\_ACPE}} \tag{4}$$

Step 6 Recursively implement the above steps to the other cases in test data, until all cases belonging to this test data have been done. (According to the rule for sorting the posts in cases with monthly units, each test data for each scenario are divided into 20 cases.)

Table 10. The effectiveness of LSI-based TMRS by using test data in scenario1.

Fan Page/Post Month	Posts/Selected Posts	ERRM-ACPE (NTD)	TMRS-ACPE (NTD)	Win Lose	CPEG
A/Apr	5/2	2.52	2.85	lose	−13%
A/May	4/1	2.97	2.52	win	15%
B/Jun	5/2	1.47	1.29	win	12%
C/May	5/2	3.63	3.18	win	12%
C/Jun	4/1	2.94	2.97	lose	−1%
D/Apr	7/2	1.95	1.92	win	2%
D/May	8/3	2.37	2.46	lose	−4%
D/Jun	6/2	2.25	2.55	lose	−13%
E/May	2/1	4.8	4.8	tie	0%
F/Apr	3/1	5.97	1.26	win	79%
F/May	2/1	2.61	2.70	lose	−3%
F/Jun	3/1	2.28	2.28	tie	0%
G/May	3/1	3.24	3.33	lose	−3%
H/Apr	8/3	2.34	1.98	win	15%
H/May	7/2	2.49	2.19	win	12%
H/Jun	6/2	2.04	2.16	lose	−6%
I/Apr	4/1	2.88	2.88	tie	0%
I/May	3/1	2.91	2.91	tie	0%
J/Jun	5/2	2.13	3.24	lose	−52%
K/May	2/1	33.09	32.37	win	2%

Note that there are two possible situations in the tie: In situation1, assume the first two posts of ERRM and TMRS in the same case are the same. This means both ERRM and TMRS obtain the same best post list to get the same value of ACPE. In situation2, assume the first two posts of ERRM and TMRS in the same case are different. For example, the traditional ERRM obtains $Apr\_P_1^A$, $Apr\_P_2^A$, and TMRS obtains $Apr\_P_3^A$, $Apr\_P_4^A$. This implies that although it is in the tie, our TMRS can still obtain better ad posts for earning more engagements under the same budget. The reason is that the TMRS is prerecommended and does not need to be publicized first, it will be considered to have won the ERRM.

(ERRM needs to publish the post on the community for a while to calculate the engagement rate.) That is, in situation1, TMRS and ERRM are in a true tie. However, when it comes to situation2, TMRS will be recognized to win over the ERRM.

8. Results

After recursively implementing the steps of the evaluation method, there are effectiveness tables that are similar to Table 10 for LSI- and LDA-based TMRS by using test data in the three scenarios. Then, we count the numbers of win, lose and tie, and calculate the win rate, lose rate and tie rate, which are defined by the following equations:

$$Win\ rate = \frac{number\ of\ win}{total\ number\ of\ win\ and\ lose} \tag{5}$$

$$Lose\ rate = \frac{number\ of\ lose}{total\ number\ of\ win\ and\ lose} \tag{6}$$

$$Tie\ rate = \frac{number\ of\ tie\ in\ situation1}{total\ number\ of\ tie} \tag{7}$$

Then, taking Table 10 as an example, there are a total of 20 cases in scenario1, and we can find that the numbers of win, lose, and tie are eight, eight, and four, respectively. Furthermore, we calculate the average CPEG to see how much gain percentage of the post engagements under the same ad budget. When calculating the average CPE increasing gain (ACPE-IG), we only consider and add the cases where CPEGs are larger than 0%, and take the average. When it comes to the average CPE decreasing gain (ACPE-DG), we only consider the ones lower than 0%. Repeating the above procedure, we can obtain the results of LSI and LDA for the three scenarios, which are summarized in Table 11.

Table 11. Summary of results.

	Scenario1		**Scenario2**		**Scenario3**	
	LSI	**LDA**	**LSI**	**LDA**	**LSI**	**LDA**
Win	8	6	10	7	11	9
Lose	8	9	5	4	3	5
Tie	4	5	5	9	6	6
cases	20	20	20	20	20	20
Win rate	50%	40%	67%	64%	79%	64%
ACPE-IG	18.6%	11.5%	21.9%	15.4%	22.5%	20.3%
Lose rate	50%	60%	33%	36%	21%	36%
ACPE-DG	11.9%	18.9%	13.6%	8%	14%	11.6%
Tie rate	75%	60%	100%	78%	17%	33%

In scenario1, we directly use the photo post texts of the wine fan page and compare the recommendation effectiveness by the LSI- and LDA-based TMRS. LSI-based TMRS achieves a 50% win rate and increases the ACPE-IG by 18.6%, while it reduces the ACPE-DG by 11.9% in the lost part. LDA-based TMRS only achieves a 40% win rate and increases the ACPE-IG by 11.5%, while it reduces the ACPE-DG by 18.9% in the lost part. The tie rates for LSI and LDA are 75% and 60%, respectively. Additionally, we can see the results of scenario2 and scenario3, which are shown in Table 11.

An advertising post example from a wine fan page recommended by LSI-based TMRS was shown in Figure 3. The slogans of figure were “Burn your passion, win your Bud beer.“ ”Login invoices and win the prize,“ and “No drunk driving. Don’t drive after drinking let you safe and secure.” The number of likes was 1750 times, and this recommended post received about 150 comments and 180 shared times. The TMRS analysis results showed that the representative words included prize, share it, limited gifts, invoice, and so on. This result was similar to the result of manual inspection.

Figure 3. A recommended post example of TMRS.

According to the results of the above experiments, the engagement effect of LSI is better than that of LDA. Take scenario2 in Table 9 as an example, the representative words for one of the topics extracted by LSI are "activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends", among which "activities, messages, fans, friends" and "Classics, first time, time" have a certain correlation with each other. Those representative words for one of the topics extracted by LDA in scenario2 are "cherry blossoms, appearance, flower season, cans, faces, couples, friends, aftertaste, rogue, lobster", among them, only "cherry blossoms, flower season" are related to each other, and other words are less relevant. That is to say, the topic formed by the words obtained by LSI is more obvious than the topic of LDA. This is due to the weak correlation between the components of the random vector of the Dirichlet distribution (The reason why there is some "relevance" is that the sum of the weights must be 1), making the potential topics of the LDA hypothesis almost irrelevant. Therefore, from the results of each scenario, it can be inferred that in the fan page posts, if the topics extracted by LSI or LDA are not completely independent, it will affect the recommended effectiveness of TMRS.

Then, from the comparison of the results of scenario2 and scenario1, it indicates that photo post data identified by marketing experts and then used in TMRS is significantly better than ERRM. Therefore, one can gain more post engagements under the same marketing budget. Finally, we apply the best setting and method for TMRS from scenario1 and scenario2 to scenario3 to verify whether the TMRS is still as effective. From Table 11, it can be seen that the experimental results of scenario3 are in line with expectations, and LSI has a 79% win rate, which is higher than the LDA model. ACPE-IG is also as high as 22.5%.

9. Conclusions

In this paper, we successfully propose a Facebook photo post recommendation system based on the topic model that can increase the fan page post engagement rate, and develop an automated method to select posts to create ads to replace the manual selection by social media managers, and reduce the managers' daily workload. The text mining method we proposed here, LSI is more suitable for the TMRS than LDA from the experiment results, and effectively improves the traditional ERRM of the existing system. These results confirm that LSI and LDA techniques are useful in context-awareness-based recommendation systems [13]. In the recommendation results from the

experimental fan page, we have helped more than half of the fan pages to effectively increase the post engagement rate or achieve the effect of saving the budget. TMRS can also provide social media managers with popular keywords referring to the previous Facebook ad posts. The need of considering using implicit trust-based information to select fan page posts with a high interaction rate automatically is also verified [20]. In addition, the photo post datasets of the wine fan page identified by marketing experts are more effective in improving the effectiveness of the TMRS, and we have proved the effectiveness of the TMRS by applying it to other types of fan pages, such as makeup/skincare fan pages. Furthermore, even in the tie situation of TMRS and ERRM, our TMRS is still better than ERRM, since it is not necessary to publish posts or create post ads first to help the social media managers to prerecommend. All the above results prove that the advertising budget can be saved and more engagements can be achieved than the existing recommendation methods.

In the future, there are still several points that can be improved. For example, designing an automatic classifier to replace the experts' identification for improving the winning rate of the recommendation system. This requires many times to communicate with experts to learn and analyze their identification knowledge. Furthermore, how to determine the number of model topics for different fan page types is difficult. Although we can decide a value based on past advertising data, whether this value will cause overfitting or underfitting remains to be evaluated. In addition, Facebook posts have a comment mechanism, so that users can leave their feelings under the related post. Therefore, we can consider the sentiment analysis of the comments under the post, which can be used as another reference indicator to provide a more accurate recommended post order. Finally, TMRS is constructed using the text content of the photo post selected by experts, but the photo is another important factor. In the future, we will also think about how to include the advertising features of photos to the recommendation system, so as to enhance the recommendation effectiveness of the entire model and provide more reference value for the social media managers.

Author Contributions: Conceptualization, S.-M.Y.; methodology, S.-M.Y. and C.-H.L.; software, J.-C.Y.; validation, C.-H.L. and J.-C.Y.; formal analysis, J.-C.Y.; investigation, C.-H.L. and L.-X.C.; resources, S.-M.Y.; data curation, J.-C.Y.; writing—original draft preparation, J.-C.Y.; writing—review and editing, C.-H.L. and L.-X.C.; visualization, C.-H.L. and L.-X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Ministry of Science and Technology of Taiwan under the grant no. 108-2511-H-009 -009 -MY3.

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

References

1. Lu, J.; Wu, D.; Mao, M.; Wang, W.; Zhang, G. Recommender system application developments: A survey. *Decis. Support Syst.* **2015**, *74*, 12–32. [CrossRef]
2. Kumar, P.; Reddy, G.R.M. Friendship recommendation system using topological structure of social networks. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 237–246.
3. Lai, C.-H.; Lee, S.-J.; Huang, H.-L. A social recommendation method based on the integration of social relationship and product popularity. *Int. J. Hum. Comput. Stud.* **2019**, *121*, 42–57. [CrossRef]
4. Lee, D.; Brusilovsky, P. Recommendations based on social links. In *Social Information Access*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 391–440.
5. Ma, X.; Ma, J.; Li, H.; Jiang, Q.; Gao, S. ARMOR: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks. *Future Gener. Comput. Syst.* **2018**, *79*, 82–94. [CrossRef]
6. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl. Based Syst.* **2013**, *46*, 109–132. [CrossRef]
7. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: Context-aware recommender systems. In *Recommender Systems Handbook*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 191–221.
8. MAA. Taipei Media Agency Association (MAA). Available online: https://maataipei.org/ (accessed on 10 May 2018).

9. Rayson, S. Facebook Engagement for Brands and Publishers Falls 20% in 2017. Available online: https://buzzsumo.com/blog/facebook-engagement-brands-publishers-falls-20-2017/ (accessed on 20 February 2018).
10. Yu, S.; Kak, S. A survey of prediction using social media. *arXiv preprint* arXiv:1203.1647 2012.
11. Yera, R.; Martinez, L. Fuzzy tools in recommender systems: A survey. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 776–803. [CrossRef]
12. Rosaci, D. Finding semantic associations in hierarchically structured groups of Web data. *Form. Asp. Comput.* **2015**, *27*, 867–884. [CrossRef]
13. Li, L.; Wang, D.; Li, T.; Knox, D.; Padmanabhan, B. SCENE: A Scalable Two-Stage Personalized News Recommendation System. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 25–29 July 2011; ACM: New York, NY, USA, 2011; pp. 125–134.
14. Ngoc, P.T.; Yoo, M. The Lexicon-based Sentiment Analysis for Fan Page Ranking in Facebook. In Proceedings of the 2014 International Conference on Information Networking (ICOIN), Phuket, Thailand, 10–12 February 2014; IEEE: Piscataway, NJ, USA; pp. 444–448.
15. Parsons, A. Using Social Media to Reach Consumers: A Content Analysis of Official Facebook Pages. *Acad. Mark. Stud. J.* **2013**, *17*, 27.
16. Goncalves, J.; Liu, Y.; Xiao, B.; Chaudhry, S.; Hosio, S.; Kostakos, V. Increasing the Reach of Government Social Media: A Case Study in Modeling Government–Citizen Interaction on Facebook. *Policy Internet* **2015**, *7*, 80–102. [CrossRef]
17. He, W.; Zha, S.; Li, L. Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. *Int. J. Inf. Manag.* **2013**, *33*, 464–472. [CrossRef]
18. Poongodi, M.; Vijayakumar, V.; Rawal, B.; Bhardwaj, V.; Agarwal, T.; Jain, A.; Ramanathan, L.; Sriram, V. Recommendation model based on trust relations & user credibility. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4057–4064.
19. Rosaci, D. CILIOS: Connectionist inductive learning and inter-ontology similarities for recommending information agents. *Inf. Syst.* **2007**, *32*, 793–825. [CrossRef]
20. Daraghmi, E.Y.; Yuan, S.-M. We are so close, less than 4 degrees separating you and me! *Comput. Hum. Behav.* **2014**, *30*, 273–285. [CrossRef]
21. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
22. Hofmann, T. Probabilistic Latent Semantic Indexing. *Acm Sigir Forum* **2017**, *51*, 211–218. [CrossRef]
23. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
24. Hoffman, M.; Bach, F.R.; Blei, D.M. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Vancouver, BC, Canada, 2010; pp. 856–864.
25. Blei, D.M.; Lafferty, J.D. A Correlated Topic Model of Science. *Ann. Appl. Stat.* **2007**, *1*, 17–35. [CrossRef]
26. Steyvers, M.; Griffiths, T. Probabilistic topic models. *Handb. Latent Semant. Anal.* **2007**, *427*, 424–440.
27. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608. [CrossRef]
28. Mitrofanova, O. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In Proceedings of the International Workshop on Language, Music, and Computing, St. Petersburg, Russia, 20–22 April 2015; pp. 69–76.
29. Řehůřek, R. Gensim Tutorial. Available online: https://radimrehurek.com/gensim/tut2.html#id6 (accessed on 20 March 2018).
30. Fxsjy. Jieba. Available online: https://github.com/fxsjy/jieba (accessed on 25 April 2018).
31. Facebook. Facebook Ad Insights. Available online: https://developers.facebook.com/docs/marketing-api/insights/?locale=en_US (accessed on 28 February 2018).
32. Nguyen, H.V.; Bai, L. *Cosine Similarity Metric Learning for Face Verification. Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 709–720.
33. Xue, N. Chinese Word Segmentation as Character Tagging. *Comput. Linguist. Chin. Lang. Process.* **2003**, *8*, 29–48.
34. Sproat, R.; Emerson, T. The First International Chinese Word Segmentation Bakeoff. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing—Volume 17; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 133–143.

35. Keskisärkkä, R. Automatic Text Simplification via Synonym Replacement. Master's Thesis. Available online: http://www.diva-portal.org/smash/get/diva2:560901/FULLTEXT01.pdf (accessed on 10 April 2018).
36. Ramos, J. Using TF-IDF to Determine Word Relevance In Document Queries. *Proc. First Instr. Conf. Mach. Learn.* **2003**, *242*, 133–142.
37. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]