Comparison of Sentiment Analysis and User Ratings in Venue Recommendation

Xi Wang[1], Iadh Ounis[2], Craig Macdonald[2]

Unversity of Glasgow, Glasgow, UK
[1]x.wang.6@research.gla.ac.uk
[2]{firstname.secondname}@glasgow.gla.ac.uk

Abstract. Venue recommendation aims to provide users with venues to visit, taking into account historical visits to venues. Many venue recommendation approaches make use of the provided users' ratings to elicit the users' preferences on the venues when making recommendations. In fact, many also consider the users' ratings as the ground truth for assessing their recommendation performance. However, users are often reported to exhibit inconsistent rating behaviour, leading to less accurate preferences information being collected for the recommendation task. To alleviate this problem, we consider instead the use of the sentiment information collected from comments posted by the users on the venues as a surrogate to the users' ratings. We experiment with various sentiment analysis classifiers, including the recent neural networks-based sentiment analysers, to examine the effectiveness of replacing users' ratings with sentiment information. We integrate the sentiment information into the widely used matrix factorization and GeoSoCa multi feature-based venue recommendation models, thereby replacing the users' ratings with the obtained sentiment scores. Our results, using three Yelp Challenge-based datasets, show that it is indeed possible to effectively replace users' ratings with sentiment scores when state-of-the-art sentiment classifiers are used. Our findings show that the sentiment scores can provide accurate user preferences information, thereby increasing the prediction accuracy. In addition, our results suggest that a simple binary rating with 'like' and 'dislike' is a sufficient substitute of the current used multi-rating scales for venue recommendation in location-based social networks.

1 Introduction

Location-Based Social Networks (LBSNs), such as Yelp, are increasingly used by users to discover new venues and share information about such venues. These networks are nowadays collecting a large volume of user information such as ratings, check-ins, tips, user comments, and so on. This large volume of interaction data makes it more difficult for users to select venues to visit without the help of a recommendation engine. Indeed, many systems [1–3] have been proposed to address the data overload problem on LBSNs by automatically suggesting venues for users to visit based on their profile and visiting history. In particular, the

explicit ratings of venues by users are widely used in various recommendation systems to elicit users' preferences, including in collaborative filtering systems [4, 5], matrix factorization (MF) approaches [6] and more recent advanced venue recommendation approaches [7–9].

However, in practice, user ratings are not always effective in representing the users' preferences. For example, it has been reported that users have distinct and inconsistent rating behaviour [10] and find it difficult to provide accurate feedback on the venues when faced with selecting among multi-rating values [11]. Several previous studies aimed to assess the impact of users' location [10], their personal and situational characteristics [12], and the nature of rating scales available to users on the quality of the users' ratings [11].

On the other hand, sentiment analysis is a widely used technique to gauge users' opinions and attitudes, for instance towards products and venues, from textual user reviews [13]. Sentiment analysis not only predicts the polarity of user opinions (e.g. positive vs. negative) but can also provide a summary of the users' opinions of a product or a venue from their reviews [13, 14]. In fact, sentiment analysis has been adopted by many studies to adjust user ratings or provide extra features to enhance the performance of recommendation systems [15–17].

The integration of sentiment analysis into recommendation systems is still limited to adjusting users' ratings to overcome their inconsistency. Instead, Lak et al. [18] substituted user' ratings with sentiment analysis, but concluded that sentiment analysis was insufficient to replace ratings in their experiments. Since then, sentiment analysis has seen a lot of attention in the literature cumulating in the development of advanced effective neural networks-based sentiment analysis of long and short texts [19, 20]. Indeed, previous sentiment analysis approaches mainly relied on human-crafted sentiment dictionaries [21, 22], which are not necessarily sufficiently effective on the wide variety of words used in LBSNs [20].

Therefore, in this paper, we hypothesise that it is possible to replace the users' explicit ratings by leveraging state-of-the-art sentiment analysers on the users' comments, thereby increasing the consistency of the user's preferences when making venue recommendations. We integrate the obtained users' preference scores through sentiment analysis into the widely used MF and GeoSoCa multi feature-based venue recommendation models [8]. Our results, using three different Yelp Challenge-based datasets, show that it is indeed possible to effectively replace users' ratings with sentiment scores when state-of-the-art sentiment analysers are used and still produce accurate venue recommendations. Our findings also suggest that it is possible to alleviate the users' ratings inconsistency by substituting the popular five-star rating scale used by LBSNs with a binary rating scale (i.e. 'like' or 'dislike') based on the sentiment analysis of their comments. In particular, the main contributions of our study are as follows:

– We explore the sentiment polarity classification accuracy of various recent sentiment analysis approaches based on neural-networks such as convolutional neural networks (CNN) and long short-term memory (LSTM) along the more conventional SentiWordNet and support vector machine (SVM)-based approaches.

– We replace the user ratings with sentiment scores in the MF model, as well as within the popular multi-feature GeoSoCa venue recommendation model [8]. To the best of our knowledge, this is the first study to examine the performance of solely using sentiment scores to substitute explicit user ratings in venue recommendation.
– We conduct thorough sentiment classification and venue recommendation experiments on datasets from the Yelp Dataset Challenge[1]. First, we use part of the dataset to conduct sentiment classification experiments and then conduct venue recommendation experiments on two other different types of datasets, which are extracted from Yelp. These two types of datasets include two city-based datasets (i.e. Phoenix and Las Vegas) and one cross-city dataset (multiple cities are covered).

The rest of our paper is structured as follows. We review the literature on the effectiveness of rating, sentiment analysis development and the application of sentiment analysis to venue recommendation in Section 2. In Section 3, we describe the GeoSoCa model and the rating substitution strategy. Section 4 describes the sentiment analysis techniques that we deploy. After that, we detail the setup for our experiments (Section 5). Then, in Section 6, we present our obtained results for evaluating the effectiveness of substituting ratings with sentiment scores. Finally, Section 7 provides concluding remarks.

2 Related Work

Ratings and Rating Scale. Many venue recommendation systems use explicit user ratings as the ground truth, both when learning user preferences and to evaluate the performance [6, 23, 24]. Ratings are a simple way for users to express opinions. However, the effectiveness of rating and rating scale have been well studied in the literature – for instance, Cosley et al. [11] argued that ratings are not sufficient for recommendation systems to effectively model the complex user preferences and opinions, while also biasing recommendation evaluation with inaccurate user opinion information. Moreover, as argued by Amoo et al. [25], the users' rating distributions are affected by different rating scales. Pennock et al. [26] recognised that the ratings of a user for the same item may not be consistent at different times. Therefore, in this paper, we similarly argue that using explicit ratings may not well represent users' opinions or attitudes. In particular, with the development of further refined techniques for sentiment analysis (discussed next), we propose to use sentiment scores to replace ratings for representing users opinions.

Sentiment Analysis. Sentiment analysis has been developed over many years, to automatically estimate users' opinions and attitudes from review texts. In the early period, sentiment analysis mainly relied on manually collected sentiment words. Ohana et al [27] used SentiWordNet[2] [28] to identify word features, and

[1] Yelp Dataset Challenge: https://www.yelp.co.uk/dataset/challenge. [2] SentiWordNet is an opinion lexicon, where the sentiment and polarity of each term is quantified.

constructed a learned model using SVM on such features. Mohammad et al. [21] leveraged tweet-specific sentiment lexicons to construct features and also included a collection of negated words. However, with the increasing applications of deep neural networks (NN), NN-based sentiment analysis techniques have achieved excellent accuracies [20]. We note the work of Kim [19], who exploited a convolutional NN (CNN) to run on a pre-trained word embedding vector and obtained much improvement in the sentiment analysis performance. Moreover, Baziotis et al. [20] leveraged long short-term memory (LSTM) to capture word order information when conducting sentiment analysis on tweets. Their approach outperformed other approaches in the 2017 SemEval competition [29]. Therefore, in this paper, with these comparably recent improvements in sentiment analysis performances, we aim to measure their usefulness in leveraging sentiment scores expressed in user reviews for the purposes of venue recommendation.

Venue recommendation with Sentiment Analysis Various studies [15, 30, 31] have been concerned with integrating sentiment analysis in venue recommendation. However, to the best of our knowledge, the used sentiment analysis in venue recommendation models have not encompassed the most recent state-of-the-art sentiment analysis approaches. For instance, Yang et al. adopted SentiWordNet3.0 [28] and the NTLK toolkit [32] for sentiment analysis. Gao et al. [30] used unsupervised sentiment classification on the sentiment polarity of words to generate a user sentiment indication matrix. Zhao et al. [31] constructed a probabilistic inference model to predict the user sentiment based on a limited number of sentiment seed words. Wang et al. [17] extracted latent semantic topics from the user reviews using a Latent Dirichlet allocation (LDA) model and inferred a user preference distribution. According to the recent sentiment analysis competition in SemEval [29], these sentiment analysis approaches are not competitive with the current state-of-the-art approaches for sentiment analysis. Therefore, we postulate that applying the state-of-the-art sentiment analysis approaches in venue recommendation could improve the performance of the sentiment-based venue recommendation approaches. Moreover, we argue that the sentiment scores could effectively replace the users' ratings to represent users' preferences.

3 Venue Recommendation Model and Rating Substitution Strategy

In this section, we first state the venue recommendation problem and the notations used in this paper (Section 3.1). Next, we introduce the two models (i.e. MF and GeoSoCa) that we use to learn the ability of sentiment analysis in capturing user preferences instead of user ratings. Finally, in Section 3.3, we describe the rating substitution strategy which we apply to MF and GeoSoCa.

3.1 Problem Statement

The venue recommendation task aims to rank highly venues that users would like to visit, based on users' previous venue visits and other sources of informa-

tion. For instance, considering sets of U users and V venues (of size m and n, respectively), the previous ratings of users can be encoded in $R \in \mathbb{R}^{m \times n}$, where entries $r_{u,v} \in R$ can represent the previous venue ratings (1..5) or the checkins (0..1) of user $u \in U$ to venue $v \in V$. Venue recommendation can then be described to accurately estimate the value $r_{u,v}$ for a venue that the user has not previously visited, or to rank highly venues that they would highly likely visit.

3.2 Venue Recommendation Approaches

In this work, we examine the behaviour of two venue recommendation approaches, namely MF and GeoSoCa, and how they perform when we change the definition of $r_{u,v}$.

Matrix Factorization (MF). MF is a classic recommendation approach, which has been adopted by many recommendation model studies as a baseline [33–35]. MF adopts singular value decomposition to learn latent semantic vectors q_v and p_u for user u and item v, respectively on known ratings $r_{u,v} \in R$.

GeoSoCa. GeoSoCa [8] is a popular venue recommendation approach, proposed by Zhang et al. in 2015. Compared to MF, it encompasses three additional important sources of information in making improved venue recommendation, namely geography, social and category information [23, 30]. Since then, it has been frequently used and discussed in various studies [1, 36, 37]. GeoSoCa estimates the probability of users visiting an unvisited venue according to the influence of three additional sources of information, namely the geography, social and category features. The geographical and social influence features use the geographical distance and users' social connections to measure the influence of different venues on users, respectively. The categorical influence estimates users' preferences distribution over categories of venues (restaurants, bars, etc.). In particular, $p_{c,v}$ indicates the popularity of venue $v \in V$, which belongs to category $c \in C$, where C denotes all venue categories[3]. In computing all these three additional features, GeoSoCa makes use of the users' ratings to estimate the probability of user u visiting venue v [8]. Note that, following Zhang et al., in our experiments, we also deploy both GeoSoCa and its components individually, i.e. Geo, So and Ca.

3.3 Rating Substitution Strategy

As argued earlier, the advent of accurate sentiment analysis approaches offers new opportunities for more refined venue recommendation. In particular, since the resulting sentiment classifiers can be formulated in a probabilistic manner, we assume that the users' preferences are indicated by the classifier's confidence, denoted as sentiment score $s_{u,v}$. This score captures the classifier confidence in

[3] A venue might belong to more than one category in the Yelp dataset. For such venues, we use the category that is uppermost in the hierarchy.

user's u comment on venue v is positive. Indeed, our work examines if the sentiment score, $s_{u,v}$, can effectively replace the rating $r_{u,v}$ as an indicator of users' preferences. We now describe our adaptations of MF and GeoSoCa.

In MF, the sentiment-based MF approach replaces user ratings on venues, $r_{u,v} \in R$, with sentiment scores $s_{u,v}$. In contrast, for GeoSoCa, we consider the substitution strategy on each component: In the geographical and social influence features, we replace users' ratings $r_{u,v} \in R$ with $s_{u,v} \in R$. Moreover, different from the previous two features, in the categorical influence feature, we not only replace users' ratings $r_{u,v} \in R$ with $s_{u,v}$, but we also modify the venue category popularity, $p_{c,v}$, as follows:

$$p_{c,v} = \sum_{u \in U} s_{u,v} \tag{1}$$

Therefore, we evaluate the ability of the sentiment scores to accurately capture the overall venue popularity. In the next section, we discuss the sentiment classification approaches that we apply to calculate $s_{u,v}$.

4 Sentiment Classification Approaches

As discussed in Section 2, sentiment analysis approaches can be broadly classified into dictionary-based, learned, and deep-learned. We apply four approaches that represent all of the categories, as well as a **Random** (Rand) classifier that matches the class distribution in its predictions, as a weak baseline.

1. **SentiWordNet-based Classifier (SWN).** The SentiWordNet-based classification approach is constructed following the approach proposed by [38], which used the updated SentiWordNet3.0 dictionary [28]. In addition, we use the 'geometric' weighting strategy that considers the word frequency to compute the prior polarity of each sentiment lexicon. The sentiment score is obtained by averaging the sentiment score of words in each user's comments.
2. **SVM-based Classifier (SVM).** Following the experimental setup of Pang et al. [39], we implement an SVM-based classifier, using the labelled word frequency vector for each review, trained using a linear kernel.
3. **CNN-based Classifier (CNN).** We use a CNN-based classifier [19] for sentiment classification. In addition, we also follow the 'CNN-Static' model setup in [19], which reported a good performance without the need for tuning the word embedding vectors.
4. **LSTM-based Classifier (LSTM).** We deploy an LSTM-based classifier [20], which obtained the top performance in the sentiment classification competition in SemEval 2017. We follow the experimental model construction process and configuration described by Baziotis et al. [20].

5 Experimental Setup

In the following, we evaluate the sentiment classification approaches compared to the corresponding users' ratings of these venues. Thereafter, we examine the

difference in venue recommendation effectiveness between models that leverage user ratings and those that use sentiment scores instead. Therefore, our experiments aim at answering the following research questions:

- **RQ1** Which sentiment analysis approaches exhibit the highest performance for user review classification?
- **RQ2** Can sentiment scores sufficiently capture the users' preferences so as to replace ratings for the purposes of effective venue recommendation? Does increased sentiment classification accuracy results in improved venue recommendation effectiveness?

To address these research questions, we perform two experiments using the Yelp Challenge dataset Round 11. We use the Yelp dataset as it is the only available public dataset that fulfils our experimental requirements, i.e. to include geographical, social and category information, as well as user reviews.

Sentiment Classification: The statistics of the dataset extracted from Yelp for the sentiment classification experiments are shown in Table 1. In Yelp, all ratings are given in a 5-star rating scale (1 is poor, 5 is great). Following Koppel et al. [40], we label the polarity of each review according to the user's rating of the venue, which we regard positive if the rating ≥ 4, and negative if rating ≤ 2. Then we randomly select equal numbers of positive and negative reviews to construct the training and testing datasets, which also avoids the class bias phenomena. Moreover, as the CNN and LSTM approaches rely on trained word embedding models, we use the remaining reviews (minus the reviews found in the Phoenix and Las Vegas city-based Yelp datasets, discussed below) in the Yelp dataset to train a word embedding model using the GenSim tool. For out-of-vocabulary words, we randomly initialise the embedding vectors, as suggested by Yang et al. [41].

We vary the size of the training dataset, from 10,000 to 600,000 reviews, to examine the stability and accuracy of the sentiment classification approaches. We use a 5-fold cross-validation setup on the training dataset before reporting the accuracy on the test datasets.

Venue Recommendation: We use three subsets of the Yelp dataset to evaluate the performance differences between the rating-based and sentiment-based venue recommendation models. Unless otherwise stated, the sentiment scores are generated after the classifiers have been trained on 600,000 comments [4]. Table 2 shows the statistics of the three Yelp-based datasets we use to evaluate the venue recommendation effectiveness. In particular, for generalisation purposes, we include two city-based datasets (namely Phoenix and Las Vegas) following other recent works [42–44], and one cross-city dataset. Indeed, we use these different Yelp subsets to obtain an overall understanding of the venue recommendation models' performances in different settings. To alleviate extreme sparseness, following Yuan et al. [43], for each dataset, we remove users with less than 20

[4] As will be shown in Section 6, this is the best training setup in terms of sentiment classification accuracy.

Dataset Name	Number of Reviews
Training	600,000
Testing	200,000

Table 1: Dataset Summary for Sentiment Classification Usage

Dataset	#Users	#Venues	#Reviews	Density
Phoenix	2,781	9,678	124,425	0.46%
Las Vegas	8,315	17,791	386,486	0.26%
Cross city	11,536	54,922	564,216	0.089%

Table 2: Datasets Summaries for the Venue Recommendation Task

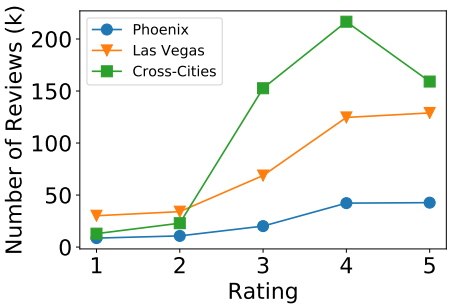
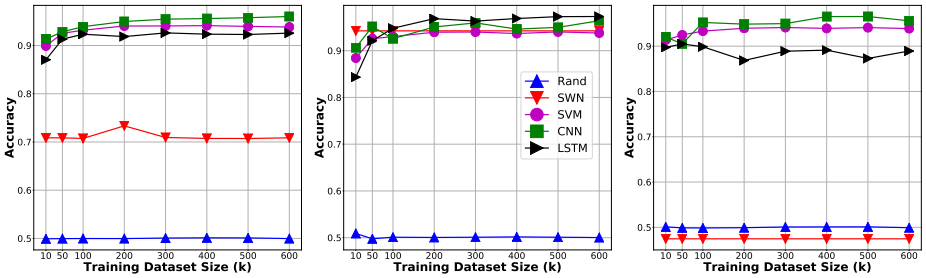


Fig. 1: Ratings Distribution within the Datasets

reviews and venues with less than 5 visits. Figure 1 shows the ratings distribution of the three datasets. It is of note that for all three datasets, the number of positive reviews (ratings 4 & 5) outweighs the number of negative reviews (ratings 1 & 2) by quite a margin. Finally, experiments are conducted using a 5-fold cross-validation on each dataset, and evaluated for recommendation quality using Precision@5 & @10 and mean average precision (MAP)[5].

6 Results Analysis

We now present the experiments that address our two research questions, concerning the sentiment classification accuracy (Section 6.1) and the usefulness of sentiment classification as an effective proxy for ratings in venue recommendation (Section 6.2).



(a) Overall Accuracy (b) Accuracy on Positive Comments (c) Accuracy on Negative Comments

Fig. 2: Sentiment classification accuracy of different approaches.

[5] The NDCG metric is not used since not all users will consistently use the rating scale (1-5), as discussed in this paper.

6.1 RQ1: Opinion Classification

Figure 2 presents the classification accuracy of our sentiment classification approaches described in Section 4, while varying the amount of training review data. In particular, we show the overall accuracy (Figure 2(a)) as well as the accuracy on the positive and negative comments alone ((b) & (c), respectively).

Figure 2(a) shows that our selected sentiment analysis approaches are divided into three groups with different classification performances – SVM, CNN and LSTM all exhibit similar top performances, followed by SWN with medium accuracy, and Rand with (expected) low accuracy. Among the highest performing group (SVM, CNN and LSTM), CNN is the highest performer.

Next, we consider the accuracy of the classifiers separately on the positive and negative classes. From Figures 2(b) & (c), we find that SVM, CNN and LSTM still provide high accuracy (≥ 0.9) for both classes, while LSTM varies in accuracy across the classes. Indeed, LSTM surpasses CNN on the positive comments yet underperforms on the negative comments (indicating a higher false positive rate). Finally, since SWN exhibits a high accuracy on the positive comments but a low accuracy on the negative comments, it is mostly identifying comments as having a positive polarity.

Overall, in answer to research question RQ1, we find that SVM, CNN and LSTM exhibit high sentiment classification accuracy, with CNN outperforming all other techniques in terms of overall accuracy. In particular, LSTM performs better than SVM and CNN for positive comments and CNN is more accurate than the other classifiers for negative comments.

6.2 RQ2: Sentiment Classification in the MF and GeoSoCa Models

We now consider if the sentiment scores generated from the classifiers evaluated in Section 6.1 can be used for effective venue recommendation by MF and GeoSoCa. All results are presented in Table 3. Each column denotes the evaluation metric on the corresponding datasets. Each group of rows defines a particular venue recommendation approach: MF, GeoSoCa, or the latter's respective components (Geo, So, Ca). Each row in a group specifies the rating-based performance or the sentiment scores-based performances from the corresponding applied sentiment classification approaches. Finally, the rightmost column indicates the number of significant increases and decreases compared to the rating-based (baseline) model in that group of rows.

On analysing the general trends between MF, Geo|So|Ca and GeoSoCa, we find that the MF approach exhibits a weak effectiveness for this ranking-based recommendation task. Indeed, the observed performances for the combined GeoSoCa approach are markedly higher (0.0029 vs. 0.0223 MAP)[6]. Overall, the lower performance of MF is expected, as MF is intended as a rating prediction approach, rather than a ranking approach, where the objective is to rank highly the actual venues that the user visited. Using sentiment information shows some

[6] The low absolute MAP values on this dataset are inline with other papers, e.g. [45].

		Phoenix (*100)			Las Vegas (*100)			Cross-City (*100)			Signf. #
		P@5	P@10	MAP	P@5	P@10	MAP	P@5	P@10	MAP	(↑ / ↓)
MF	Rating	0.12	**0.11**	**0.29**	0.01	0.02	0.12	0.02	0.01	0.05	—
	Rand	0.06	0.06	0.23	0.01	0.01	0.10	**0.05**	0.02	0.02	0 / 0
	SWN	**0.13**	**0.11**	0.28	**0.05**	**0.05**	**0.17**	0.02	0.02	0.05	0 / 0
	SVM	0.06	0.06	**0.29**	0.01	0.01	0.12	0.02	0.01	0.04	0 / 0
	CNN	0.08	0.08	0.27	0.01	0.01	0.11	0.02	**0.03**	**0.06**	0 / 0
	LSTM	0.04	0.03	0.25	0.03	0.02	0.11	0.04	0.02	**0.06**	0 / 0
Geo	Rating	0.67	0.73	0.83	**0.47**	0.43	**0.55**	0.78	0.74	1.02	—
	Rand	0.64	0.61	0.78	0.35	0.39	0.51	0.73	0.75	1.01	0 / 0
	SWN	**0.69**	0.69	0.86	0.44	**0.45**	**0.55**	0.73	0.77	1.03	0 / 0
	SVM	0.61	0.73	0.81	**0.47**	0.44	0.53	**0.76**	**0.79**	1.03	0 / 0
	CNN	0.66	0.73	0.82	0.45	0.44	**0.55**	0.75	**0.79**	**1.04**	0 / 0
	LSTM	0.67	**0.75**	0.84	0.45	**0.45**	**0.55**	0.75	0.78	**1.04**	0 / 0
So	Rating	3.38	2.88	1.95	**2.81**	**2.36**	1.76	2.97	**2.61**	1.53	—
	Rand	2.52↓	2.10↓	1.14↓	1.98↓	1.76↓	0.98↓	2.07↓	1.80↓	0.77↓	0 / 9
	SWN	2.73↓	2.36↓	1.74↓	2.15↓	1.82↓	1.36↓	2.15↓	1.89↓	1.28↓	0 / 9
	SVM	3.34	2.27↓	2.02	2.69	2.33	1.67	2.56↓	2.15↓	1.49	0/ 3
	CNN	3.39	2.87	2.06	2.70	2.33	1.70	2.84	2.49	1.57	0/ 0
	LSTM	**3.43**	**2.89**	**2.16↑**	2.71	2.34	**1.75**	**2.98**	2.54	**1.71**	1/ 0
Ca	Rating	3.51	**3.17**	**2.79**	2.54	2.27	**2.11**	**0.79**	**0.72**	0.54	—
	Rand	3.03↓	2.83	2.57	2.35	2.16	1.98↓	0.72	0.68	0.50	0/ 1
	SWN	1.88↓	2.14↓	2.72	0.72↓	0.85↓	2.01↓	0.49↓	0.55↓	0.54↓	0/ 9
	SVM	3.35	3.15	2.69	2.53	2.25	2.07	0.74	0.70	0.53	0/ 0
	CNN	**3.52**	**3.17**	2.77	2.51	2.26	2.07	0.78	0.71	0.54	0/ 0
	LSTM	3.50	3.15	2.78	**2.56**	2.26	2.10	0.78	**0.72**	**0.55**	0/ 0
GeoSoCa	Rating	3.68	**3.04**	2.23	2.64	2.09	1.51	3.92	3.17	**2.22**	—
	Rand	3.08↓	2.57↓	1.79↓	2.44	2.03	1.47	3.59	2.97	2.06	0/ 3
	SWN	1.32↓	1.35↓	1.39↓	0.52↓	0.58↓	1.06↓	1.83↓	1.75↓	1.44↓	0/ 9
	SVM	3.52	2.93	2.17	2.84	2.21	1.78↑	3.68↓	2.98↓	2.06	1/ 2
	CNN	3.62	2.90	2.18	2.86↑	2.29	1.79↑	3.71↓	3.02↓	2.09	2/ 2
	LSTM	**3.73**	2.96	**2.28**	**2.97↑**	**2.38↑**	**1.87↑**	**3.96**	**3.21**	2.15	3/ 0

Table 3: Recommendation performances of rating and sentiment-based approaches on three datasets (reported evaluation measures are *100). Using the t-test, statistically significant increases (resp. decreases) with respect to the corresponding rating-based baseline are indicated by ↑ (resp. ↓).

minor improvements, but none of the sentiment classifiers causes significant enhancements to this weaker rating-based MF baseline.

Next, we consider GeoSoCa and its components Geo|So|Ca for each dataset. For the geographical information, the rating and sentiment-based models provide statistically indistinguishable results (according to a paired t-test; p-value < 0.05), regardless of the sentiment classification approach used. Next, for the social influence model (i.e. So), the distinction among the approaches is clear: SWN and Rand significantly degrade effectiveness compared to the rating-based baseline in 9 cases; the learned approach (SVM) significantly degrades effectiveness in 3 cases (P@10 for Phoenix and P@5 & P@10 for Cross-City); on the other hand, the deep-learned sentiment approaches (CNN and LSTM) are at least statistically indistinguishable from the corresponding rating-based baseline (only one significant increase: $1.95 \rightarrow 2.16$). Indeed, it is promising that the latest approaches (CNN and LSTM), which were shown to be the most effective sentiment classifiers in Section 6.1, also result in the recommendation models

with the highest effectiveness, suggesting that they could be a suitable proxy for user ratings.

For the categorical information (i.e. Ca), recall that our substitution strategy replaces not only the users' preferences but also the aggregated popularity of the category for that user, as per Equation (1). On examining Table 3, the learned and deep learning sentiment approaches are able to provide comparable performances to the corresponding rating-based baseline. The same observation also holds with the social information-based model. Moreover, similar to the social information-based model, the recommendation effectiveness also aligns with the performances of sentiment classifications, with CNN and LSTM providing the most effective results.

Finally, we consider the combined GeoSoCa model - where we observe that the product of the geographical, social and category influence scores, when using the sentiment scores from CNN or LSTM, could still provide performances that cannot be statistically distinguished from those based on ratings (only 1 significant decrease). Moreover, in 5 cases there were actually significant increases in effectiveness by deploying CNN or LSTM. Therefore, in answer to research question RQ2, we find that only the sentiment-based user preference scores from the state-of-the-art deep-learning-based sentiment classification approaches (i.e. CNN and LSTM) can provide similar effectiveness to the rating-based models. It is also of note that in these experiments, given that we regard a user rating ≥ 4 as positive, and a user rating ≤ 2 as negative, our results in Table 3 suggest that the sentiment scores can simply be binary (i.e. 'like' and 'dislike'). Such a binary rating scale (as might be determined by a sentiment polarity classifier) is a sufficient substitute for the currently used multi-rating scales to effectively capture the users' preferences in venue recommendation.

7 Conclusions

In this paper, we explored the performances of various sentiment analysis approaches at identifying the polarity of comments about venues, while also considering their use as a replacement for the users' explicit ratings in venue recommendation. For the sentiment classification approaches, we found that CNN outperforms other approaches in terms of overall accuracy, while LSTM performs better in classifying positively labelled reviews. Next, when substituting users' ratings with sentiment scores from state-of-the-art sentiment classification approaches (i.e. CNN and LSTM), we found that the resulting GeoSoCa-models were rarely significantly degraded in effectiveness, and were actually seen to be significantly enhanced in several cases. Overall, our results suggest that, for venue recommendation, a simple binary rating with 'like' and 'dislike' (as might be determined by a sentiment polarity classifier) is an effective substitute for the currently used multi-rating scales in location-based social networks. As future work, we plan to apply our rating substitution strategy in additional venue recommendation approaches. We will also investigate how to improve the performances of venue recommendation models by exploiting user reviews posted on other platforms (e.g. Twitter or Facebook) where no multi-scaling rating is used.

References

1. Manotumruksa, J., Macdonald, C., Ounis, I.: Modelling user preferences using word embeddings for context-aware venue recommendation. In: Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval. (2016)
2. Noulas, A., Scellato, S., Lathia, N., Mascolo, C.: A random walk around the city: New venue recommendation in location-based social networks. In: Proc. of SocialCom-PASSAT. (2012)
3. Lian, D., Ge, Y., Zhang, F., Yuan, N.J., Xie, X., Zhou, T., Rui, Y.: Scalable content-aware collaborative filtering for location recommendation. IEEE Transactions on Knowledge and Data Engineering **30**(6) (2018) 1122–1135
4. Frankowski, D., Herlocker, J., Sen, S., et al.: Collaborative filtering recommender systems. The Adaptive Web **4321** (2007) 291–324
5. Hu, L., Sun, A., Liu, Y.: Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In: Proc. of SIGIR. (2014)
6. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8) (2009) 30–37
7. Manotumruksa, J., Macdonald, C., Ounis, I.: A contextual attention recurrent architecture for context-aware venue recommendation. In: Proc. of SIGIR. (2018)
8. Zhang, J.D., Chow, C.Y.: Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In: Proc. of SIGIR. (2015)
9. Zhu, Q., Wang, S., Cheng, B., Sun, Q., Yang, F., Chang, R.N.: Context-aware group recommendation for point-of-interests. IEEE Access **6** (2018) 12129–12144
10. Hu, R., Pu, P.: Exploring relations between personality and user rating behaviors. In: Proc. of Workshop on Emotions and Personality in Personalized Services (EMPIRE at UMAP). (2013)
11. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., Riedl, J.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: Proc. of SIGCHI. (2003)
12. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction **22**(4-5) (2012) 441–504
13. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
14. López Barbosa, R.R., Sánchez-Alonso, S., Sicilia-Urban, M.A.: Evaluating hotels rating prediction based on sentiment analysis services. Aslib Journal of Information Management **67**(4) (2015) 392–407
15. Yang, D., Zhang, D., Yu, Z., Wang, Z.: A sentiment-enhanced personalized location recommendation system. In: Proc. of ACM Conference on Hypertext and Social Media. (2013)
16. Gurini, D.F., Gasparetti, F., Micarelli, A., Sansonetti, G.: Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. Future Generation Computer Systems **78**(P1) (2018) 430–439
17. Wang, H., Fu, Y., Wang, Q., Yin, H., Du, C., Xiong, H.: A location-sentiment-aware recommender system for both home-town and out-of-town users. In: Proc of SIGKDD. (2017)
18. Lak, P., Turetken, O.: Star ratings versus sentiment analysis–a comparison of explicit and implicit measures of opinions. In: Proc. of HICSS. (2014)
19. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)

20. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proc. of SemEval. (2017)
21. Mohammad, S., Kiritchenko, S., Zhu, X.D.: NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proc. of SemEval. (2013)
22. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research **50** (2014) 723–762
23. Cheng, C., Yang, H., King, I., Lyu, M.R.: Fused matrix factorization with geographical and social influence in location-based social networks. In: Proc. of AAAI. (2012)
24. Manotumruksa, J., Macdonald, C., Ounis, I.: Regularising factorised models for venue recommendation using friends and their comments. In: Proc. of CIKM. (2016)
25. Amoo, T., Friedman, H.: Do numeric values influence subjects' responses to rating scales. International Marketing & Marketing Research **26**(1) (2001) 41–46
26. Pennock, D.M., Horvitz, E., Lawrence, S., Giles, C.L.: Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach. In: Proc. of UAI. (2000)
27. Ohana, B., Tierney, B.: Sentiment classification of reviews using SentiWordNet. In: Proc. of Information Technology & Telecommunication. (2009)
28. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proc. of LREC. (2010)
29. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proc. of SemEval. (2017)
30. Gao, H., Tang, J., Hu, X., Liu, H.: Content-aware point of interest recommendation on location-based social networks. In: Proc. of AAAI. (2015)
31. Zhao, K., Cong, G., Yuan, Q., Zhu, K.Q.: Sar: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In: Proc. of ICDE. (2015)
32. Bird, S., Loper, E.: NLTK: the natural language toolkit. In: Proc. of ACL. (2004)
33. Lian, D., Zhao, C., Xie, X., Sun, G., Chen, E., Rui, Y.: GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: Proc. of SIGKDD. (2014)
34. Zhao, G., Qian, X., Xie, X.: User-service rating prediction by exploring social users' rating behaviors. Transactions on multimedia **18**(3) (2016) 496–506
35. He, J., Li, X., Liao, L., Song, D., Cheung, W.K.: Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. In: Proc. of AAAI. (2016)
36. Zhao, S., Zhao, T., King, I., Lyu, M.R.: Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In: Proc. of WWW. (2017)
37. Zhao, S., Zhao, T., Yang, H., Lyu, M.R., King, I.: Stellar: Spatial-temporal latent ranking for successive point-of-interest recommendation. In: Proc. of AAAI. (2016)
38. Guerini, M., Gatti, L., Turchi, M.: Sentiment analysis: How to derive prior polarities from sentiwordnet. In: Proc. of EMNLP. (2013)
39. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proc. of ACL. (2002)
40. Koppel, M., Schler, J.: The importance of neutral examples for learning sentiment. Computational Intelligence **22**(2) (2006) 100–109
41. Yang, X., Macdonald, C., Ounis, I.: Using word embeddings in Twitter election classification. Information Retrieval Journal **21**(2-3) (2018) 183–207

42. Zimba, B., Chibuta, S., Chisanga, D., Banda, F., Phiri, J.: Point of interest recommendation methods in location based social networks: Traveling to a new geographical region. arXiv preprint arXiv:1711.09471 (2017)
43. Yuan, F., Jose, J.M., Guo, G., Chen, L., Yu, H., Alkhawaldeh, R.S.: Joint geospatial preference and pairwise ranking for point-of-interest recommendation. In: Proc. of ICTAI. (2016)
44. Guo, Q., Sun, Z., Zhang, J., Chen, Q., Theng, Y.L.: Aspect-aware point-of-interest recommendation with geo-social influence. In: Proc. of UMAP. (2017)
45. Liu, Y., Pham, T.A.N., Cong, G., Yuan, Q.: An experimental evaluation of point-of-interest recommendation in location-based social networks. In: Proc. of VLDB. (2017)