From Hotel Reviews to City Similarities: A Unified Latent-Space Model

electronics

MDPI

Article

From Hotel Reviews to City Similarities: A Unified Latent-Space Model

Luca Cagliero*, Moreno La Quatra and Daniele Apiletti

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi, 24-10129 Torino, Italy; moreno.laquatra@polito.it (M.L.Q.); daniele.apiletti@polito.it (D.A.)

* Correspondence: luca.cagliero@polito.it; Tel.: +39-011-090-7179

Received: 13 December 2019; Accepted: 15 January 2020; Published: 20 January 2020

Abstract: A large portion of user-generated content published on the Web consists of opinions and reviews on products, services, and places in textual form. Many travellers and tourists routinely rely on such content to drive their choices, shaping trips and visits to any place on earth, and specifically to select hotels in large cities. In the context of hospitality management, a challenging research problem is to identify effective strategies to explain hotel reviews and ratings and their correlation with the urban context. Under this umbrella, the paper investigates the use of sentence-based embedding models to deeply explore the similarities and dissimilarities between cities in terms of the corresponding hotel reviews and the surrounding points of interests. Reviews and point of interest (POI) descriptions are jointly modelled in a unified latent space, allowing us to deeply investigate the dependencies between guest feedbacks and the hotel neighborhood at different aggregation levels. The experiments performed on public TripAdvisor hotel-review datasets confirm the applicability and effectiveness of the proposed approach.

Keywords: text mining; deep natural language processing; open data; frequent itemset mining

1. Introduction

Online hotel bookings have radically changed the hospitality industry. Thanks to the increasing availability of user-generated data, hotel reputation is nowadays strongly influenced by guest-provided ratings [1]. In fact, guests are strongly encouraged to rate hotels and comment on various hotel aspects, including location, service, cleanliness, and price. Online reviews are known to have a major impact on hotel revenues and on customer behaviours [2]. For these reasons, in the last decade the academic and industrial communities have devoted an increasing effort to analyzing hotel reviews and ratings.

In this paper we address the problem of explaining hotel ratings by analyzing the textual content of the hotel reviews. An overview of the related literature is given in Section 2. Since guest reviews are often related to the context in which the hotel is located, studying the correlations between hotel reviews the surrounding context is particularly appealing. For example, hotel appeal strongly depends on the accessibility to various types of points of interest (POIs) such as touristic places (e.g., museums, monuments, squares), public transports (e.g., bus stops, underground stations), and food services (e.g., restaurants, pubs). Understanding which POIs are influencing the hotel reviews and to what extent they relate guest opinions with the hotel ratings is a challenging task due to the following reasons:

- Hotel reviews are short pieces of text, which often miss a clear description of the related context.
- POIs are not explicitly tagged in most of the hotel reviews.
- Although for each of the most important cities a list of geo-referenced POIs is easy to retrieve, POI descriptions are often unstructured.

- A list of the most popular geo-referenced POIs is available for the most important cities, the detailed POI descriptions are often unstructured. Hence, their preparation and processing is prone to errors and time consuming.

In light of the above-mentioned issues, correlating review text and hotel ratings with POI descriptions can be challenging. Therefore, there is a need for integrated and flexible solutions allowing domain experts to easily explore the semantic correlation between hotel reviews, ratings, and surrounding POIs with limited effort and acceptable reliability.

The present work investigates the use of a unified deep natural language processing (NLP) model to jointly explain hotel reviews and ratings and their correlation with the surrounding POIs. Deep NLP entails applying deep learning techniques to accomplish natural language processing task [3,4]. Embedding models produce a latent vector space, where the learned vectors explicitly encode many linguistic regularities and patterns. The key idea behind their use is to effectively combine in a unified space all the information needed to characterize hotel reviews, ratings, and surrounding POIs. POI descriptions, acquired from open Web sources, are unstructured text snippets that are not required to be previously annotated (except for POI geo-localization).

In this work, we exploit BERT embeddings [5] to transform the raw text data into a unified review-POI latent space. The unified model encodes all the semantic relationships between hotels as well as between hotels and POIs. Thus, it allows a deep exploration of the causes influencing hotel rating. More specifically, the following aspects can be analyzed by exploring the unified model: (i) the similarity between different cities in terms of guest feedbacks; (ii) the correlation between the review text and the POI descriptions; (iii) the correlation between hotel ratings and POIs; (iv) the characteristics of the analyzed cities (independently of the hotels).

We have conducted an experimental campaign on public TripAdvisor hotel-review datasets. The results provide valuable insights into the viewpoints of anglophone hotel guests spread all over the world. Despite guest opinions significantly change according to the continent, the density of POIs in the urban area, the characteristics of the POI neighborhood, the hotel review descriptions also highlight interesting regularities in specific city groups.

The proposed approach can be useful at different levels, from local authorities and policy makers aiming at improving the tourism ecosystem, to hotel chains targeting new openings, to provide better insights to both hotel owners and users besides the specific rating score.

The structure of the paper is defined as follows. Section 2 overviews the previous research studies. Section 3 thoroughly described the methodology used in the experiments. Section 4 reports a selection of the most relevant experimental results. Finally, Section 5 draws the conclusions of the present work and discusses the future research directions.

2. Literature Review

2.1. Explaining Hotel Reviews

Explaining hotel ratings and reviews is a popular research problem. The main challenge is to understand the key factors that contribute to customer satisfaction or dissatisfaction by means of data-driven approaches [6]. As shown by [2], the presence of positive rating has a strong influence on customer actions to the extent that it could be deemed as more influential than advertising strategies. Therefore, analyzing review data and exploring the extracted knowledge is particularly appealing. In [7] the authors try to find out what aspects of hotel experience are taken into consideration by customers while rating a hotel. They analyze hotel reviews from Booking.com to by means of latent semantic analysis to identify the main features referred to by customers and to determine their correlation with customer ratings. A supervised approach to predicting review polarity is given by [8]. In [9] a similar approach, based on word co-occurrences, has been applied to hotel reviews from TripAdvisor.com. Since hotel ratings can be specialized to different semantic categories (e.g., location, cleanliness), customer judgements can be explained under various review aspects.

2.2. Predict Hotel Rates and Recommend the Most Suitable Hotels

Predicting review ratings is an another remarkable research task. For example, one of the most recently proposed approaches [10] have tried to capture user preferences comprehensively by jointly analyzing hotel ratings and customer reviews. Specifically, the authors have proposed a neural network-based approach to consider users' opinions on specific aspects. The idea behind is that different rating predictors share the same review encoder model in order to exploit the inner dependencies, but tailor networks to different facets. A practical application of hotel rating predictors is their tight integration into hotel recommender systems (e.g., [1,11–14]). Existing solutions rely on text sentiment analysis [12,13], contextual information [14], reviewer reputation from the social community [11], and opinions from friends external to the community [1].

2.3. Correlate Hotel Location With the Surrounding POIs

Parallel studies have deepened the analysis of the correlation between hotel location and Points Of Interests (POIs). For example, in [15] the authors have investigated the correlation between guest satisfaction (expressed by the hotel location rating) and three main location-related features, i.e., accessibility to points of interest, transport convenience, surrounding environment. The results confirm that the presence of attractions, airports, universities, public transportation, and green spaces are significant determinants. The reasons why hotel choose their locations are discussed in [16], whereas an overview of the most popular hotel location models is given in [17].

Position of the present study in the related literature: This paper falls into category hotel review explanation. Rating prediction and hotel recommendation are out of the scope of this work. The aim of the present study is to explore the applicability of a unified latent vector representation of Wikipedia POI descriptions and hotel reviews to analyze guest reviews at a city-wise aggregation level. To this aim, we first enrich hotel reviews with general-purpose POI descriptions, model the underlying text relationships in the latent space of sentence-based text embeddings, and then explain hotel reviews according to both guest ratings and surrounding POIs. To the best of our knowledge, the use of Deep NLP techniques to combine hotel reviews and ratings with POI descriptions is new.

3. Proposed Methodology

In this section we propose a method to explore reviews $\mathbb{R}$ and ratings $\mathbb{W}$ of hotels $\mathbb{H}$ in a sentence-based embedding space. $r_{ij} \in \mathbb{R}$ denotes the i-th review of hotel $h_j \in \mathbb{H}$, whereas $w_j \in \mathbb{W}$ denotes the hotel rating. The geographical context of a hotel h_j is described by the set of POIs P_j located in its neighborhood. Each POI $p_k \in P_j$ is described by a set $D_{k,j}$ of textual descriptors.

Figure 1 depicts the data analytics process, which entails the following steps: (i) Data enrichment: review data were enriched with textual descriptions of the surrounding POIs, which were acquired from open data repositories. (ii) Processing: reviews and POIs were transformed into a unified latent space, which was produced by a state-of-the-art sentence embedding model [5]. The latent model preserved the semantic relationships among textual words and phrases, thus allowing to correlate the review content with the POIs located in the hotel neighborhood. (iii) Analysis: review and POI embeddings were explored in order to gain insights into their possible semantic relationships.

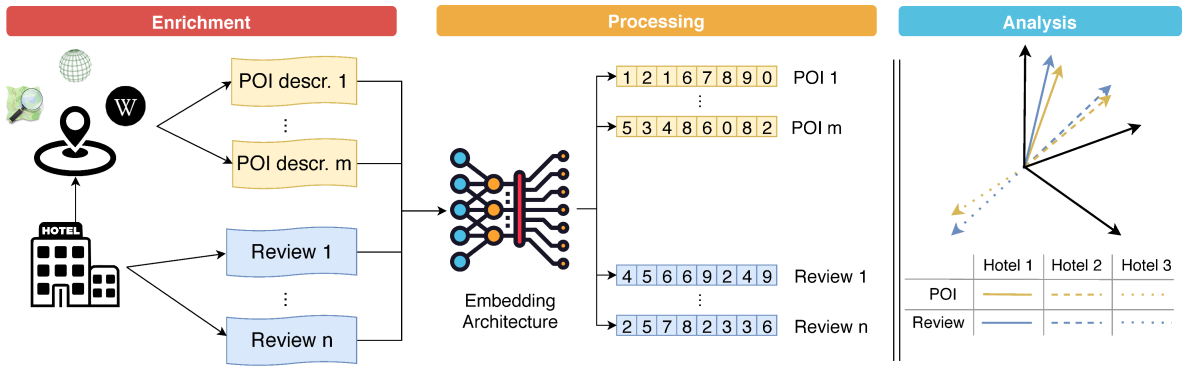


Figure 1. Data enrichment, data processing, and analysis steps.

3.1. Dataset Description

We considered the set of English-written reviews and ratings available in OpinRank dataset [18]. They were crawled from the TripAdvisor website (www.tripadvisor.com). It collected the reviews of 3096 hotels located in 10 popular cities spread all over the world, i.e., Bejing, Shangai, New Delhi, Dubai, London, Montreal, New York City, Chicago, Las Vegas, and San Francisco. For each hotel, a discrete rating (from 1 to 5) was provided, separately evaluating different aspects: value, location, cleanliness, room, service, and the overall experience. Notice that TripAdvisor is a U.S.-based crowd-sourced travel review service. Hence, it mainly collects travel experiences and hotel reviews submitted by anglophone users.

Table 1 summarizes the main data characteristics. Hotels and reviews were grouped by city. Cities were sorted by geographical location, according to their longitude, from East (Shangai and Bejing) to West (Las Vegas and San Francisco), with respect to the prime meridian in Greenwich. Cities with less than 100 hotels (i.e., Montreal and New Delhi) were discarded to reduce the bias due to the presence of possible outliers.

Table 1. Description of OpinRank Dataset

City	Hotels (All)	Hotels (Geo-located)	Avg. Num. of Reviews per Hotel
Shanghai	280	114	34.6
Beijing	260	118	42.4
New Delhi	277	97	40.9
Dubai	275	204	49.5
London	988	644	103.4
Montreal	142	100	81.8
New York City	260	229	218.3
Chicago	157	110	151.8
Las Vegas	230	121	148.7
San Francisco	227	179	151.6
Total	**3096**	**1916**	**103.1**

3.2. Data Enrichment

To enrich review data with contextual information, we geo-localized the hotels using the OpenStreetMap (OSM) (www.openstreetmap.org) and GeoNames (www.geonames.org) services. Specifically, out of 3096 hotels, we detected the geographical coordinates of 1916 of them (around 62%). Hereafter, we exclusively analyzed the reviews of geo-localized hotels (197,473 reviews overall, corresponding to approximately 103 reviews per hotel on average).

Cities located in different areas showed rather peculiar characteristics. Specifically, for Eastern cities (i) the number of reviews per hotel was on average lower (35–50 vs. 100–200), and (ii) the number of geo-located hotels was proportionally lower (less than 50% of hotels are geo-located in Eastern cities vs. more than 60% in Western cities). The above trends are mainly due to the provenance of the

TripAdvisor service (U.S.). Notice also that TripAdvisor was founded in February 2000 and has started to operate abroad later, e.g., it launched its official site in China in April 2009. Overall, the enriched dataset consisted of 1719 geo-located hotels located in eight cities and three different continents, along with their textual reviews. The coverage of the hotels with respect to the city extension was higher for Western countries.

Reviews were enriched also with a description of the context surrounding the hotel. Specifically, the names and characteristics of the nearby POIs were given. POIs are locations that were deemed as worthy by most of the citizens. They may represent tourist attractions (e.g., museums, city landmarks, famous spots), key places for transports (bus stops, underground stations), or food services (e.g., restaurants, pubs). POIs are geo-localized and mapped to hotels using the API offered by GeoNames. Specifically, for each hotel we got the 100 nearest POIs no farther than 1 km.

We retrieved a textual description of each POI from open Web sources. Specifically, from the GeoNames geographical database we retrieved the description of each POI by following the corresponding Wikipedia link.

Notice that the GeoNames geographical database covers all countries and contains over 11 million free place names, but the country coverage is not uniform. For instance, it provided over 2 million place names in the United States, with a density of 0.23 names/km^2, while only 768,000 place names were available for China, with a density of 0.08 names/km^2, at the time of writing (www.geonames.org/statistics)lastaccess:December2019.

3.3. Data Processing

We explored the correlation between hotel reviews, ratings, and the nearest POIs by analyzing review text and POI textual descriptions.

An embedding function $f: \mathbb{R} \rightarrow \mathbb{V}'$ mapped reviews to a high-dimensional vector space V', where vector $v_{ij} \in \mathbb{V}'$ corresponds to an arbitrary review r_{ij} (hereafter denoted as review embedding). Similarly, an embedding function g: $\mathbb{D} \rightarrow \mathbb{V}''$ mapped POI descriptions to a high-dimensional vector space V'' (hereafter denoted as POI embedding). Embedding functions were obtained using DistilBERT [19], which is a recently proposed sentence-based embedding model [5].

Thanks to the notable properties of the recently proposed vector representations of text [4], vectors in the latent space could be conveniently combined by performing basic vector operations. Specifically, review embeddings could be averaged by computing a pointwise average [20] over the vector dimensions in order to generate the vector representations of hotels $h_j^r = \text{avg}_i v'_{ij}$. Notice that averaging the vector components separately for each dimension allows us to generate an aggregated embedding vector with the same dimension.

Hotel-review embeddings h_j^r described hotel h_j by reflecting the corresponding review content. Hereafter we will denote as $\mathbb{H}'_r$ the resulting hotel-review embedding space. Likewise, we generated a complementary vector representation $p_{d,j} = \text{avg}_i v''_{kj}$ of each hotel h_j reflecting the surrounding POI descriptions. The resulting hotel–POI embedding space will be denoted as $\mathbb{H}'_p$.

Since hotel reviews were annotated with ratings describing different hotel aspects, we first mapped ratings to discrete bins (i.e., Low, Medium, High) and then generated the corresponding vector hotel–review and hotel–POI sub-spaces (respectively denoted as $\mathbb{H}_r^L$, $\mathbb{H}_r^M$, $\mathbb{H}_r^H$, and $\mathbb{H}_p^L$, $\mathbb{H}_p^M$, $\mathbb{H}_p^H$).

We propose to map reviews $\mathbb{R}$ and POI descriptions $\mathbb{D}$ to a unified latent space, i.e., $f = g$, $V' = V''$, to investigate textual data correlations between hotel reviews and POIs descriptions. Thanks to the vector operations made available in the latent space, data could be analyzed at different aggregation levels: review, hotel, city, or larger geographical aggregations.

Figure 2 graphically shows an example of the process used to embed each review (or POI description) as well as an example of multiple vector aggregation using the point-wise average.

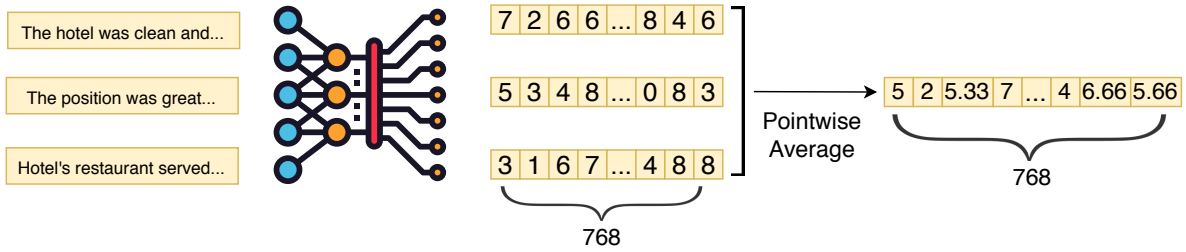


Figure 2. Illustration of the embedding and aggregation steps.

The unified model enables the following analyses.

Review–POI similarity. The presence of POIs in the hotel neighborhood is likely to influence hotel reviews. To investigate to what extent review-POIs relationships hold, for each hotel $h_i \in \mathbb{H}$ we identified the POIs whose description was mostly related to the hotel reviews, i.e., $\arg_{p_j \in \mathbb{P}} \max(sim(v'_{ij}, v''_{ij})$. Embedding similarity was computed using the established cosine similarity measure [20].

The review–POI similarity could be also explored at an aggregated view in the hotel–review embedding space. The hotels whose reviews were most correlated to the surrounding POIs were those identified by the following objective function: $\arg_{h_j \in \mathbb{H}} \max sim(h_{r,j}, p_{d,j})$. Further aggregated views could be conveniently generated by aggregating hotels, based on their geographical location, into districts, cities, countries, and continents.

Rating-POI similarity. The presence of POIs in the hotel neighborhood is likely to influence hotel ratings as well. To investigate the correlation between review-POI relationships and ratings, we drilled down the review-POI similarity analyses to the hotel–review and hotel–POI sub-spaces. Specifically, for each hotel we identified the POIs that positively influenced hotel reviews by optimizing $\arg_{p_j \in \mathbb{P}} \max(sim(v'_{ij}, v''_{ij}) \; | v'_{ij} \in \mathbb{H}^H_r, v'_{kj} \in \mathbb{H}^H_p$. Similarly, we derived negative influences by jointly exploring sub-spaces $\mathbb{H}^L_r$ and $\mathbb{H}^L_p$. The hotels that were most influenced by the surrounding POIs in a positive way could be identified by specializing the hotel embeddings to the sub-spaces related to high ratings i.e., $\arg_{h_j \in \mathbb{H}} \max sim(h^H_{r,j}, p^H_{d,j})$.

Similarities among geographical areas: hotels located in different areas could be compared with each other in terms of the similarity between the corresponding reviews and surrounding POIs. Without any loss of generality, hereafter we present a method for comparing hotels in different cities. However, the same approach can be trivially extended to other types of geographical areas (e.g., districts, regions, continents).

First, we generated city-review and city-POI embeddings (denoted as $c_{r,j}$ and $c_{p,j}$, respectively) by combining the embeddings of the corresponding hotels. More specifically, let c_q be an arbitrary city and let h_j be a hotel located in c_q. Hotel-review embeddings h^r_j are computed as the pointwise average of all the review embeddings of h_j. Likewise, hotel-POI embeddings h^p_j are defined as the pointwise average of all the POI embeddings in the neighborhood of h_j.

Next, given a list of cities and the corresponding hotels per city, we estimated pairwise city similarities. To this aim, we computed a city similarity matrix $\mathcal{M}$, where each cell m_{yz} denotes the similarity between a given pair of cities in terms of hotel reviews/POIs. Given an arbitrary cell m_{yz} in the matrix (corresponding to the city pair $\langle c_y, c_z \rangle$), its value depends on the similarity (in the embedding space) between the hotels in c_y and those in c_z. Specifically, for each hotel h^*_i in c_y we find the neighbor-weighted k-nearest neighbor hotels h^*_j in terms of cosine similarity in the embedding space (We set k to 100 in the experiments reported in the paper.). Each neighbor hotel (either located in c_z or not) is weighted by its inverse rank $\frac{1}{R_{ij}}$ (i.e., the nearest hotel takes rank 1, the second $\frac{1}{2}$, and so on). Hence, m_{yz} takes as value $\sum_{h*_i, h*_j in c_z} \frac{1}{R_{ij}}$. The rationale was to smooth the contribution of farther hotels while giving more importance to the nearest ones.

3.4. Descriptive Analysis of The Unified Model

To explore and interpret the results of the review-POI and rating-POI similarity analyses, we applied two established data mining techniques on top of processing results: clustering and frequent itemset mining [20].

Clustering

Hotel-review embeddings were clustered in order to identify groups of similar hotels. Clustering aims at generating groups of data samples such as samples within the same group are highly similar (i.e., high intra-cluster similarity), whereas samples in different groups are rather different (i.e., low inter-cluster similarity).

The goal of the clustering analysis is manifold. For example, clustering review embeddings is aimed at grouping similar reviews together (indipendently of the hotel and location), whereas clustering hotel embeddings aims at identifying groups of similar hotels in terms of text reviews (independently of the location). In the experiments reported in Section 4, we have applied the well-known K-means clustering algorithm [20].

Frequent Itemset Mining

Frequent itemset mining is an established unsupervised data mining technique to analyze co-occurrences among multiple data items [21]. It has been successfully applied in many contexts to extract recurrent association rules, even from high-dimensional [22], multiple-level [23], and weighted [24] transactional data.

In our context, itemsets represented recurrent co-occurrences among multiple words in the hotel review. To our purposes, words represented items, whereas each review was modeled as a transaction consisting of a set of words. During this phase, we removed domain-specific stopwords, the ones occurring in more than 10% of reviews, as long as the least frequent words, i.e., those occurring in less than 0.01% of the reviews. Frequent itemsets were extracted by using the popular FP-growth algorithm [25]. To reduce the number of potentially redundant patterns, a representative subset of the frequent itemsets, i.e., the closed itemsets, is extracted. Closed itemsets are frequent itemsets such that there exists no superset that has the same support count as the original itemset [20]. Itemsets provide an interpretable description of the most recurrent trends in hotel reviews. Hence, they can be used, in combination with the aforesaid techniques, to support domain experts in the process of extracting interpretable models.

4. Results and Discussion

In this section we present the most salient results achieved on the real dataset. The main goal of the empirical analysis was to identify relevant patterns, correlations and insights on hotel reviews, their contribution to the overall city-wide experience, and the influence of the geographical information carried by the POIs.

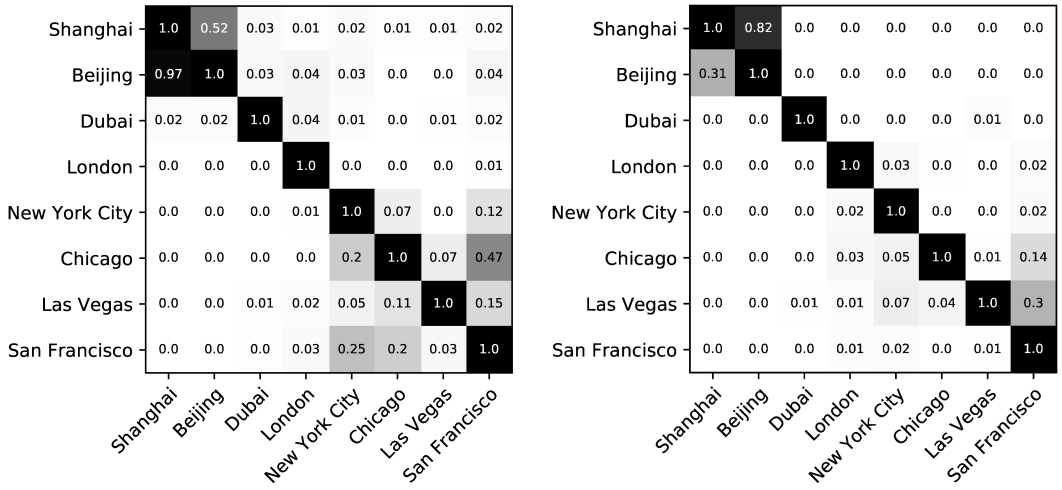
The experiments were run on a machine equipped with an Intel® Xeon® X5650, 32 GB of RAM and running Ubuntu 18.04 LTS.

For the sake of readability, we separately present the results related to different analytics goals. Specifically, Section 4.1 addresses the study of the similarities among different cities. Section 4.2 and 4.3 separately discuss the outcomes of the POI-review and POI-rating similarity analyses. Finally, Section 4.4 and 4.5 report the results of a city characterization based on single words and word sets, respectively.

4.1. City Similarity

In this section, we evaluate the similarity among different cities according to both the review- and POI embeddings.

The heat-map charts depicted in Figure 3 show the city-to-city similarities computed on (i) the review embeddings (see Figure 3a) and (ii) the POI embeddings (see Figure 3b).



(**a**) Review-based natural language processing (NLP) similarity

(**b**) Point of interest (POI)-based NLP similarity

(**c**) Clustering of review vectors

(**d**) Geographical proximity

Figure 3. City similarity heat maps.

Both maps follow a similar trend, with a higher similarity among cities nearer in longitude. For instance, Beijing and Shanghai, two Chinese cities, have very high similarity scores between them, and they are also more similar to Dubai (still in Asia) than any other Western cities. A second group of similar cities are New York, Chicago, Las Vegas, and San Francisco, all from US. The trends between the two maps, even if with different intensity, are similar in quality. If the POI-based embedding similarity was expected due to their physical presence in the same place as the hotels of the same city, the similarity in reviews triggers the question of the influence of the POIs on the experience of the hotel customers: how much is the presence of nearby points of interests correlated with the reviews and their ratings? We further investigate this issue in the following sections.

To confirm the city-to-city review similarity, we ran an unsupervised clustering of the review embeddings. Being the clustering unaware of the actual city of the hotels, the goal was to measure the match between each resulting cluster and each city. We run the K-means algorithm on the deep-NLP

representation of each hotel review, using the cosine distance, and setting the number of clusters equal to 8, since this was the number of cities. In Figure 3c we report the cluster assignments (in percentage), normalizing the number of items on each cluster per row, i.e., for each city we reported the portion of hotel reviews of that city assigned to each cluster. For instance, the first cell indicates that 89% of the hotel review vectors of Shanghai have been assigned to Cluster 0.

Results confirm that Shanghai and Beijing were very similar, as their hotel reviews were assigned predominantly (89%) to the same Cluster 0. In Cluster 0 no other city had assignments, apart from 6% of Dubai hotel reviews. Hence, Cluster 0 confirmed that both Chinese cities were very similar, they shared a very limited similarity with Dubai, another Asian city, and they were dissimilar with respect to all other cities. Dubai was on a Cluster of its own, i.e., Cluster 2, where almost 80% of its reviews were assigned. The majority of London reviews were assigned to Cluster 3 (58%), and a minority to Cluster 1 (27%): London was probably a multi-faceted city from a review perspective. All US cities had reviews assigned to Clusters 4, 5, and 6, with New York City predominantly on Cluster 4 (64%), Chicago was split over Clusters 4 and 5 (40% each), Las Vegas predominant on Cluster 5, and finally San Francisco predominant on Cluster 5. To sum up, we can describe the clusters as follows: Cluster 0 for Chinese cities (Beijing and Shanghai), Cluster 2 for Dubai, Cluster 3 and 1 for London, Cluster 4 for New York City, Cluster 5 for San Francisco, Cluster 6 for Las Vegas. Chicago was similar to both New York and San Francisco Clusters. Cluster 7 was an outsider, collecting outlier reviews from all cities.

For the city clustering results, we computed the Rand index [26] with respect to the true city assignment, obtaining 0.44. We note that cities belonging to the same continent are typically similar. This trend was confirmed from the continent-based clustering (Asia, Europe, America), whose Rand Index was higher, and equal to 0.88.

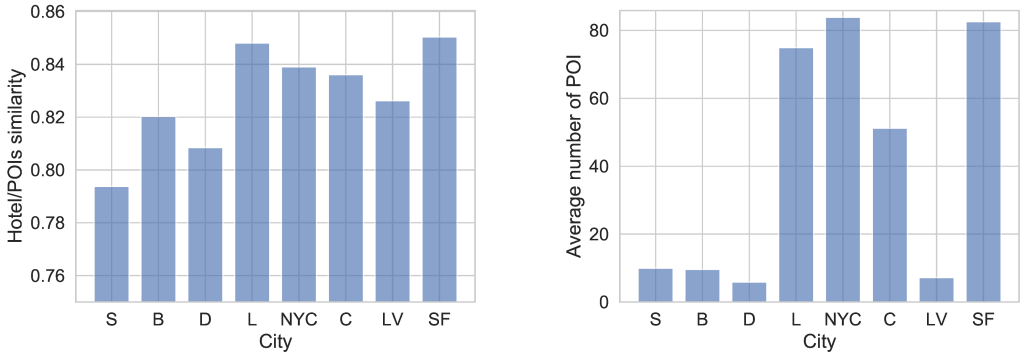
We also note that, taking into account only the geographical-distance city-to-city matrix, we obtained a completely different grouping from a magnitude perspective, as reported in Figure 3d, even if qualitatively the continent trend was still evident.

4.2. Review–POI Similarity

Given the latent-space representation and the location coordinates for both hotels and POIs, we computed the hotel–review and hotel–POI embeddings, respectively representing, on the one hand, the hotel from the reviewers' perspective and, on the other hand, hotel from the viewpoint of the nearest POI descriptions. Both embeddings were separately averaged for each city. Finally, we computed the cosine similarity measure between the average city-wide review vector and the corresponding POI vector, thanks to the unified latent space. Such a metric aims at capturing the possible influence of the POIs in the review evaluation, i.e., if the reviews take into consideration aspects linked to the surrounding POIs.

The average similarity value for each city is reported in Figure 4a. For the sake of the analysis, we also evaluated the average number of POIs nearby each hotel, reported in Figure 4a.

Even if the two figures were different in magnitude, they shared some qualitative patterns. We note that the top three cities with the highest average number of POIs nearby their hotels, i.e., San Francisco, New York City, and London, were also the top three cities for review–POI similarity: the more POIs were available, the more the experience was related to those POIs. The difference in magnitude, instead, is probably due to the missing POI coverage of the Eastern cities with respect to the Western ones. To address the POI coverage imbalance, in Section 4.3 we correlated the review ratings of each city with respect to the POIs of the same city. Notice also that the similarity between hotel reviews and POI descriptions not only depended on the number of POIs, but also on the in-text POI references contained in the text of the reviews. This motivated the presence of exceptions, such as Las Vegas and Denver, where the number of POIs was rather low but the similarity was high.



(**a**) Average similarity with respect to nearest POIs

(**b**) Number of POIs within 1 km

Figure 4. Average distances analysis.

4.3. Rating–POI Correlations

We now analyze the correlation between the latent-space similarity values (described in Section 4.2) and the rating scores of the reviewers, with the aim of explaining the influence of the POIs not only on the review text, but also on the overall rating and specifically on the location rating.

For each hotel we computed (i) the similarity between its POI-based representation (i.e., the hotel-POI embedding) and the review-based one (i.e., the hotel–review embedding), (ii) the average overall and location score for the hotel itself, across all of its reviews.

On top of the evaluation of the above-mentioned scores, for each city we can compute the Spearman correlation coefficient (ρ_s) [27] among the metrics (i) and (ii), as reported in Table 2. We chose the Spearman correlation since it evaluates the monotonic relationship between two ordinal variables. In our case we want to correlate whether an increasing similarity value in the latent space matches an increasing overall/location score by the reviewers (or vice versa).

Table 2. Spearman correlation analysis between latent-space similarity and overall/location ratings.

City	ρ_s (Overall)	ρ_s (Location)	Area
Shanghai	0.01	0.25	6.340 km^2
Beijing	−0.02	0.36	16.808 km^2
Dubai	0.05	0.11	4.114 km^2
London	0.33	0.52	1.572 km^2
New York City	0.15	−0.01	7838 km^2
Chicago	0.24	0.36	6061 km^2
Las Vegas	0.07	0.25	352 km^2
San Francisco	0.13	0.05	1214 km^2

From Table 2 we note that the correlation was stronger with the location score, as expected, than with the overall score, since the latter also included indoor services of the hotel. Both Eastern and Western cities shared the same correlation values on the location score (e.g., Shanghai, Bejing, Las Vegas, and Chicago). London and New York City (NYC) were exceptions, with the former being highly correlated and the latter not correlated at all. The stronger correlation of London with the location score (0.52) is probably due to the large extension of the city with respect to the other Western cities (it has the widest area): this means that being near a POI is a key feature of hotels. New York City hotel reviews appeared to be not correlated with the location ratings. Such exception might be motivated by the following facts: (i) Iconic places in NYC (e.g., the Empire State Building) were cited also when they were relatively far away (low similarity with nearby POIs but high location scores). (ii) NYC has

a good transportation network so geographical distances were not so meaningful in that particular context. (iii) NYC has a relatively high average number of reviews per hotel, hence more sparse terms and citations to different, possibly far, POIs emerged. (iv) NYC has the highest number of hotels and the largest extension among the U.S. cities in our dataset.

4.4. City Characterization

To confirm trends and extract human-readable insights on the reviews, at a city-wide aggregation level, we considered two aspects: (i) the word clouds of the hotel reviews of each city, and (ii) the k nearest POIs in the embedding space for all the hotels of each city. We expected to be able to capture the mention of POIs in some cities and to characterize the key features of the cities as reviewed by their hotel customers.

In Figure 5 we report four representative word-clouds. Those illustrations are created taking into account the most frequent unigrams and bigrams in the hotel reviews. Such visual representations provide an at-a-glance overview of the main characteristics that visitors (or better, English-speaking reviewers) consider relevant of their experience.

We note that Eastern cities (Figure 5a,b) were predominantly characterized by emerging terms related to the hotel itself and its services, meaning that a comfortable experience provided by the hotel was deemed worthy to be shared as it might not be so likely to happen. On the contrary, Western cities Figure 5c,d) also included important city-specific features such as landmarks and special attractions, such as “casino” and “tube station”. For this reason, we confirmed the experience in Western cities to be more related with the surrounding POIs, as reported in Figure 4a.

The second analysis on the POIs was performed as follows. For each hotel, in the embedding space the 50 nearest POIs were selected. A majority voting approach was applied to identify the most similar POIs for the whole city across all its hotels. Finally, the top three POIs for each city were selected, as those related the most, in the embedding space, with the hotel reviews of the given city. Results are reported in Table 3. We note that most POIs were peculiar landmarks, tourist attractions, or public services that actually characterize each city. For instance, Shanghai was characterized by (i) Xinzhuang, a town located in Minhang District, (ii) the Rockbund Art Museum in central Shanghai, and (iii) the Waibaidu Bridge, called the Garden Bridge in English, being the first all-steel bridge, and the only surviving example of a camelback truss bridge in China. Beijing was characterized by POIs located in the Dongcheng District, such as (i) the Jiaodoukou Subdistrict, (ii) Nanluoguxiang, a narrow alley, that gives its name to an old part of the Beijing city centre, with traditional architecture, and (iii) the Great Leap Brewing, which operates three popular brewpubs in Beijing, two of which are located in the Dongcheng District. Dubai’s top landmarks were (i) Al Muraqqabat, a locality in the heart of eastern Dubai in Deira, (ii) the Union station of Dubai Metro on both the Green and Red Lines, known as the busiest metro station of Dubai, and (iii) Dubai Municipality. Western cities highlighted avenues, churches, casinos, and theatres as special POIs. This approach allowed us to automatically extract peculiar attractions in the cities from the point of view of hotel customers, which is of paramount importance for the tourism industry.

(**a**) Beijing Word Cloud (**b**) Shanghai Word Cloud

(**c**) Las Vegas Word Cloud (**d**) London Word Cloud

Figure 5. Word Clouds.

Table 3. Most influencing POIs for each city.

City	Most influencing POIs	% of hotels
Shanghai	Xinzhuang	6.52%
	Rockbund Art Museum	3.11%
	Waibaidu Bridge	2.8%
Beijing	Jiaodoukou Subdistrict	4.69%
	Nanluoguxiang	4.43%
	Great Leap Brewing	4.43%
Dubai	Al Muraqqabat	12.6%
	Union (Dubai Metro)	11.02%
	Dubai Municipality	9.45%
London	Holy Trinity, Paddington	1.91%
	Cleveland Square	1.91%
	Leinster Gardens	1.86%
New York City	450 Lexington Avenue	1.18%
	Black Chamber	1.04%
	St. Agnes Church	1.0%
Chicago	Medinah Temple	5.75%
	Millennium Centre	5.41%
	55 East Erie Street	5.08%
Las Vegas	Main Street Station Hotel and Casino and Brewery	4.17%
	Mob Museum	4.17%
	Las Vegas Fire & Rescue Department	4.17%
San Francisco	Curran Theatre	11.6%
	Ruby Skye	11.45%
	City of Paris Dry Goods Co.	11.45%

4.5. Frequent Itemsets

Besides the straightforward word cloud approach, we applied also the exploratory itemset mining technique to gain more insights into the main review topics separately for each country. In Table 4 we report the top three word sets per city as well as their relative frequency count (i.e., the support index [21]) in the set of review related to that city.

It is worth noting that in our analysis we set a minimum support threshold equal to 1% in order to extract all the closed frequent word combinations (independently of the size). Therefore, the mined itemsets were potentially more expressive than simple word clouds as they represented not only single words but also larger word combinations.

The results are reported in Table 4. They show that the itemset-based descriptions of the most relevant topics differed in Western and Eastern cities. Specifically, the reviews in Beijing and Shanghai clearly shows that the topic of spoken English language was a frequent issue for English-speaking visitors because of the low percentage of local English-speaking people [28] in hotels. Not surprisingly, in the English-spoken cities, the reviews focus more on the nearby points of interest. Remarkably, in Dubai, the emerging concepts were also related to the nearby POIs. This pattern could be due to the high percentage of English-speaking people [29] in this country.

Table 4. Top three closed word-sets for each city.

City	Frequent Words Set	Support
Shanghai	nanjing, road	8.1%
	bund, nanjing	5.9%
	chinese, english	5.52%
Beijing	chinese, english	10.7%
	chinese, taxi	8%
	forbidden, shopping	7.8%
Dubai	beach, taxi	7.13%
	beach, jumeirah	6.76%
	beach, fantastic	6.64%
London	road, tube	4.01%
	minute, tube	3.99%
	easy, tube	3.82%
New York City	empire, state	7.62%
	building, empire	5.94%
	building, state	5.93%
Chicago	avenue, michigan	7.99%
	navy, pier	6.93%
	magnificent, mile	6.87%
Las Vegas	casino, strip	19%
	buffet, strip	8.24%
	shuttle, strip	7.79%
San Francisco	cable, car	9.85%
	fisherman, wharf	6.96%
	car, parking	6.62%

5. Conclusions and Future Research Agenda

This work has addressed the research issue of identifying effective strategies to correlate hotel reviews and ratings with the surrounding geo-referenced open data. Specifically, this paper has explored the use of sentence-based embedding models on public sources of points of interest, reviews and ratings, including thousands of hotels in eight popular cities in three continents, with a total of almost 200 thousands reviews and ratings. The proposed methodology has allowed us to project all

textual data from both reviews and POI descriptions into a unified latent space, thus enabling the investigation of the influence of surrounding POIs on customer reviews at a city-wise aggregation level.

Experimental results addressed many facets of such correlations, from city similarities, to review-POI influence, to insights on the specific words characterizing each city. To explore the content of the unified latent space, unsupervised data mining techniques such as clustering and frequent itemset mining have been applied with the goal of automatically extract correlations among words and data-driven hotel similarities. Different trends have been identified, such as similarities in the latent-space for cities in the same continent, correlation between reviews and POIs, and explanations thanks to frequent word occurrence extraction.

The proposed approach can be useful for different stakeholders, from local authorities to policy makers aiming at improving the tourism ecosystem, to hotel chains targeting new openings, to provide better insights to both hotel owners and users besides the specific rating score. Being able to explain the correlation of POIs with hotel reviews opens up various exploitation possibilities related to the importance of the POI distance witto the hotels, the local mobility patterns, the urban transport vh respect ectors, and the interest of the visitors towards nearby POIs or far away landmarks. As an example, cities with iconic landmarks and an efficient and pervasive public transport system tend to present reviews referring to far away POIs (i.e., the iconic landmarks) with high location ratings, hence (i) new hotels do not need to be opened in downtown to reach high location scores and (ii) customers can save money by booking hotels far from the city centre, without affecting their experience.

The preliminary results achieved on real hotel review data leave room for further extensions of the current research. In the future research agenda, we plan to extend the current method to cope with multi-lingual reviews and POI descriptions. In fact, since hotel reviewers come from all over the world, the review text is, necessarily, multi-lingual. Likewise, POI descriptions can be retrieved from Web sources written in different languages. However, analyzing cross-lingual text is potentially challenging. This poses the question of how to integrate hotel reviews and POI description in a a unified cross-lingual vector space. We plan to overcome this issue by applying ad hoc embedding alignment strategy (e.g., [30,31]).

Author Contributions: Conceptualization, L.C., M.L.Q., and D.A.; Investigation, L.C., M.L.Q., and D.A.; Methodology, L.C., M.L.Q., and D.A.; Software, M.L.Q.; Supervision, L.C., and D.A.; Writing—Original Draft, L.C., M.L.Q., and D.A.; Writing—Review and Editing, L.C., M.L.Q., and D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The research leading to these results has been partly funded by the Smart-Data@PoliTO center for Big Data and Machine Learning technologies.

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

References

1. Gellerstedt, M.; Arvemo, T. The impact of word of mouth when booking a hotel: Could a good friend's opinion outweigh the online majority? *Inf. Technol. Tour.* **2019**, *21*, 289–311. doi:10.1007/s40558-019-00143-4. [CrossRef]
2. Hollenbeck, B.; Moorthy, S.; Proserpio, D. Advertising Strategy in the Presence of Reviews: An Empirical Analysis. In *Proceedings of the 2018 ACM Conference on Economics and Computation*; ACM: New York, NY, USA, 2018; p. 7.
3. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
4. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.

5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. doi:10.18653/v1/N19-1423. [CrossRef]
6. Nicholas, C.K.W.; Lee, A.S.H. Voice of Customers: Text Analysis of Hotel Customer Reviews (Cleanliness, Overall Environment & Value for Money). In Proceedings of the 2017 International Conference on Big Data Research, New York, NY, USA, 22–24 October 2017; pp. 104–111.
7. Chanwisitkul, P.; Shahgholian, A.; Mehandjiev, N. The Reason Behind the Rating: Text Mining of Online Hotel Reviews. In Proceedings of the 2018 IEEE 20th Conference on Business Informatics (CBI), Vienna, Austria, 11–13 July 2018, pp. 149–157.
8. Hirokawa, S.; Hashimoto, K. Simplicity of Positive Reviews and Diversity of Negative Reviews in Hotel Reputation. In Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, Thailand, 15–17 November 2018; pp. 1–6.
9. Berezina, K.; Bilgihan, A.; Cobanoglu, C.; Okumus, F. Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *J. Hosp. Mark. Manag.* **2016**, *25*, 1–24. [CrossRef]
10. Wu, C.; Wu, F.; Liu, J.; Huang, Y.; Xie, X. ARP: Aspect-aware Neural Review Rating Prediction. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 3–7 November 2019; pp. 2169–2172.
11. O'Mahony, M.P.; Smyth, B. Learning to Recommend Helpful Hotel Reviews. In Proceedings of the Third ACM Conference on Recommender Systems, New York, NY, USA, 23–25 October 2009; pp. 305–308.
12. Takuma, K.; Yamamoto, J.; Kamei, S.; Fujita, S. A Hotel Recommendation System Based on Reviews: What Do You Attach Importance To? In Proceedings of the 2016 Fourth International Symposium on Computing and Networking (CANDAR), Hiroshima, Japan, 22–25 November 2016; pp. 710–712. doi:10.1109/CANDAR.2016.0129. [CrossRef]
13. Sharma, Y.; Bhatt, J.; Magon, R. A Multi-criteria Review-Based Hotel Recommendation System. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 26–28 October 2015; pp. 687–691.
14. Jalan, K.; Gawande, K. Context-aware hotel recommendation system based on hybrid approach to mitigate cold-start-problem. In Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 1–2 August 2017, pp. 2364–2370. doi:10.1109/ICECDS.2017.8389875. [CrossRef]
15. Yang, Y.; Mao, Z.; Tang, J. Understanding Guest Satisfaction with Urban Hotel Location. *J. Travel Res.* **2018**, *57*, 243–259. doi:10.1177/0047287517691153. [CrossRef]
16. Yang, Y.; Wong, K.K.; Wang, T. How do hotels choose their location? Evidence from hotels in Beijing. *Int. J. Hosp. Manag.* **2012**, *31*, 675 – 685. doi:10.1016/j.ijhm.2011.09.003. [CrossRef]
17. Yang, Y.; Luo, H.; Law, R. Theoretical, empirical, and operational models in hotel location research. *Int. J. Hosp. Manag.* **2014**, *36*, 209–220. doi:10.1016/j.ijhm.2013.09.004. [CrossRef]
18. Ganesan, K.; Zhai, C. Opinion-based entity ranking. *Inf. Retr.* **2012**, *15*, 116–150. [CrossRef]
19. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
20. Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*, 2nd ed.; Pearson: New York, NY, USA, 2018.
21. Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.* **1993**, *22*, 207–216. doi:10.1145/170036.170072. [CrossRef]
22. Apiletti, D.; Baralis, E.; Cerquitelli, T.; Garza, P.; Pulvirenti, F.; Michiardi, P. A Parallel MapReduce Algorithm to Efficiently Support Itemset Mining on High Dimensional Data. *Big Data Res.* **2017**, *10*, 53–69. doi:10.1016/j.bdr.2017.10.004. [CrossRef]
23. Cagliero, L. Discovering Temporal Change Patterns in the Presence of Taxonomies. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 541–555. doi:10.1109/TKDE.2011.233. [CrossRef]

24. Baralis, E.; Cagliero, L.; Garza, P.; Grimaudo, L. PaWI: Parallel Weighted Itemset Mining by Means of MapReduce. In Proceedings of the 2015 IEEE International Congress on Big Data, New York City, NY, USA, 27 June–2 July 2015; pp. 25–32. doi:10.1109/BigDataCongress.2015.14. [CrossRef]
25. Han, J.; Pei, J.; Yin, Y. Mining Frequent Patterns without Candidate Generation. *SIGMOD Rec.* **2000**, *29*, 1–12. doi:10.1145/335191.335372. [CrossRef]
26. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
27. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef]
28. Wei, R.; Su, J. The statistics of English in China: An analysis of the best available data from government sources. *English Today* **2012**, *28*, 10–14. [CrossRef]
29. Dorsey, C. The Role of English in the United Arab Emirates and Resulting Implications for English Teaching. *Preprint* **2018**. [CrossRef]
30. Schuster, T.; Ram, O.; Barzilay, R.; Globerson, A. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 1599–1613. doi:10.18653/v1/N19-1162. [CrossRef]
31. Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; Jégou, H. Word Translation without Parallel Data. *arXiv* **2017**, arXiv:1710.04087.