Classifying Travel-related Intents in Textual Data

Zae Myung Kim[1], Young-Seob Jeong[1], Jonghwan Hyeon[1], Hyungrai Oh[2], and Ho-Jin Choi[1]

***Abstract*—**Intent classification refers to the process of identifying a set of intents of interest that appear in a given document. This work considers the task of annotating travel-related reviews with travel intents that best represent the reviewer's reason for visiting the place of interest (POI). A domain-tailored word embedding model is learned to construct intent-specific feature vectors, thereby improving classification accuracy. The feasibility of multiclass intent classification is explored using an intent corpus, consisting of 6,560 labelled reviews.

***Keywords*—**intent classification, text mining, travel and tourism, word embedding.

I. Introduction

THERE has been increasing interest in mining useful information from the vast amount of user-generated web content. One of the tasks that has not been well explored is intent classification, which classifies a set of intents of interest appearing in a given document. Identifying intentions from texts is especially useful for building a personalized recommendation system (targeted marketing) [12] and for enhancing a context-aware module in a smartphone-based intelligent personal agent [1].

Recent studies [3], [7] consider the task of identifying purchasing intents, e.g. "I'm planning to buy this smartphone," from product reviews or posts in online social forums by utilizing n-grams and dependency structure of sentences as features. These works target identifying somewhat explicit intents in the sense that the intents are represented by key phrases that appear in a subset of document. However, for some domains such as traveling, users do not always explicitly state the purpose of visit to a certain place of interest (POI) in the text.

TABLE 1
Reviews with explicit and implicit intents

A travel review with an explicit intent
"We spent more than 3 hours *shopping at the mall*! The place had a lot of our fav brand! …"
A travel review with an implicit intent
"What I like most about Disneyland is that it has a wide variety of events! Our kids loved …"

Therefore, these implicit travel-related intents must be inferred from the whole document (review). Classifying such implicit intents is a difficult task, mainly because it is not clear how to devise a set of features that effectively captures these latent intents.

In this work, we present a data-driven approach that classifies eight travel intents that travellers can exhibit when visiting a certain POI. A travel intent represents the main reason the traveller had visited the POI. The main task is to identify it from the review that he/she wrote after the journey. The contributions of this work can be summarized as follows:

- To the best of our knowledge, there are no existing studies on multiclass intent classification in texts; we begin the first investigation into the feasibility of such task.
- We develop an approach to construct intent-specific features that do not require manual effort.

We propose eight intents in the traveling domain, and create an intent corpus, consisting of 6,560 labelled reviews. This can be utilized as a benchmarking dataset for further research.

II. Background

Unlike many other text mining fields, studies on intent classification from texts have just begun and yet to gain much prominence. Recent works [3], [7] consider the task of identifying purchase intents from product reviews or posts in online social forums. Chen et al. in [3] propose a transfer learning method that constructs a classifier for a single product domain, and subsequently applies it to a different domain. Their supervised classification features include n-grams and terms with high information gain values. In [7], Gupta et al. follow an approach that defines domain-specific features, such as purchase action words, using the dependency structure of sentences. In web search, Strohmaier and Kröll [15] develop a method that learns a classifier from syntactic structure of (explicit) intent phrases and constructs a knowledge base for those intents with the search results obtained from the intent phrases as queries. The knowledge base is used to mark intents in a given document according to similarity measurements.

The notion of understanding users' intents is also prevalent in the domain of spoken language understanding (SLU). The task includes classifying a user's series of utterances into corresponding words (speech recognition), assigning an appropriate semantic label related to the domain of concern to each word (semantic slot filling), and eventually classifying what user has meant from the series of utterances, i.e. a user's intent within the target domain (intent classification) [8],

[1]School of Computing, KAIST, Daejeon, South Korea.
[2]Samsung Electronics Inc., Suwon, South Korea.

[13]. As the purpose of SLU systems is to aid the user with a particular task, e.g., viewing a flight schedule from A to B, it is assumed that users express their intent explicitly in their utterances.

Our work shares similar goals with the above studies, in that we also annotate a given document using the travel-related intents it contains. However, the above works target explicit (and often a single class of) intent that appears in a subset of a document. As shown above in Section 1, for some domains such as traveling, a document may not always contain explicit intent phrases. Furthermore, it can encompass multiple intent classes; these factors make the feature selection process more difficult.

To effectively classify non-explicit intents that are hidden in text, the entire context of the document should be considered. Utilizing word embedding models is one such strategy; these models have proven their effectiveness in many text mining fields such as information retrieval [5] and sentiment analysis [16], [17]. The intuition is to discover explanatory factors from the data and relieve classification algorithms from heavy feature engineering [2]. In essence, a word embedding algorithm assigns each word a dense, low-dimensional and real-valued vector that captures the word's syntactic and (occasionally) semantic properties.

III. Intent Corpus

A. The Eight Travel Intents

As intent classification in textual data is at an early stage of research, there is no publicly available corpus for the research. To the extent of our knowledge, this is the first study on intent classification that takes diverse intents into account. We chose to explore the task in traveling domain because of the abundance of data, i.e., the user-generated travel reviews. The annual statistical report on world tourism [18] published by the United Nations World Tourism Organization presents nine categories for the purpose of traveling. After some minor modifications—we removed transit and other categories, and added eating out—we established the eight intents shown in Table 2 for more fine-grained intent classification. The reason for removing transit category is that, according to our corpus (refer to Section 3.2), although travellers do visit airports of different cities for such purpose, they do not seem to write a dedicated review about it. Regarding eating out category, we noticed that it is not included in the UN report, presumably because the document concerns the purpose of travelling between the countries, and travellers generally do not visit a different country with the sole purpose of eating food. Nevertheless, in this work, we are looking at the purpose of visiting a certain POI, therefore eating out category is a sensible class to include.

TABLE 2
THE EIGHT TRAVEL INTENTS

0	1	2	3	4	5	6	7
Business	Eating out	Education	Health	Holidays	Religion	Shopping	Socializing

For clarity, note that a travel intent does not mean whether or not a traveller would visit a POI, but the reason he/she made the trip to the POI in the first place. In addition, it is important to notice that these intents are not necessarily mutually exclusive to each other; for instance, a traveller may visit a restaurant to have dinner with friends, in which case, both eating out and socializing intents may be present in the review.

B. Intent Corpus Construction

As with any classification task, a set of labelled training instances is required to train a classifier. We crawled a large number of reviews from three travel websites: TripAdvisor[1], Yelp[2], and Booking.com[3]. Alongside the review content, POI meta-data such as POI name, location, and overall score were also collected. We targeted all 13 states in the Western United States. The dataset is 7 GB in size, and contains nearly 6.8 million reviews for approximately 83,000 POIs.

From this dataset, we sampled 820 reviews for each intent by selecting reviews that have at least two keywords in them. These keywords for each intent were empirically chosen. In total, 6,560 reviews were labelled by nine annotators, all majoring in computer science. The annotators were instructed to select only the most prevalent intent for a given review. Each review was labelled by two annotators separately, and a supervisor was available to mediate any labelling disagreements.

TABLE 3
AN EXAMPLE OF ANNOTATED REVIEW

Review: "What I loved about Disneyland is that it has a variety of events! Our kids loved …"
Intent Index: 4 (Holidays)

The average value of Cohen's kappa coefficient for each intent was 0.61. Table 4 shows kappa values for each intent. Annotators agreed well on most of the intent classes (kappa ≥ 0.61) except business and eating out intents. The kappa value for business intent is significantly lower possibly, due to the fact that there were not many reviews that actively talk about the journey being a business trip. Most reviewers wrote about the quality of the venue that the business meeting took place, rather than the nature of the trip itself. Regarding eating out class, most reviewers went to places to eat with companions, and for many such cases, the annotators tend to label them as socializing. This illustrates the point that the travel intents are not mutually exclusive. This property, albeit reduces the top-1 classification performance, is not a bad news in the long run; such characteristics of travel intents provide us with more insights into POIs (refer to Section 5.3) and can be utilized in recommender systems. We maintained the skewed distribution of intents as this reflects travellers' tendency to visit POIs with certain intents.

[1] https://www.tripadvisor.com
[2] https://www.yelp.com/
[3] https://www.booking.com/

TABLE 4
COHEN'S KAPPA COEFFICIENT AND DISTRIBUTION OF EACH INTENT

	Bus.	Eating	Edu.	Health	Hol.	Rel.	Shop.	Soc.
Kappa	0.3	0.5	0.7	0.7	0.6	0.6	0.7	0.8
Dist. (%)	8	25	5	7	33	5	3	14

IV. APPROACH

A. Travel-related Word Embedding Model

In order to manage the implicit nature of intents contained in reviews, we consider the general context of the review. Rather than attempting to devise a set of complicated yet limited features based on grammatical structures of sentences by hand, a word embedding model is trained using the review data and employed as the building block of features. Among many approaches to constructing word embedding models [6], [11], [14], the Word2Vec algorithm [10] is selected to train the embedding model for its efficiency on large dataset. We used the skip-gram model with a window size of five words to generate vectors of 200 real numbers in range between -1 and 1. The model is learned using the crawled review dataset, and consists of more than 270,000 words, each mapped to a vector of 200 dimensions.

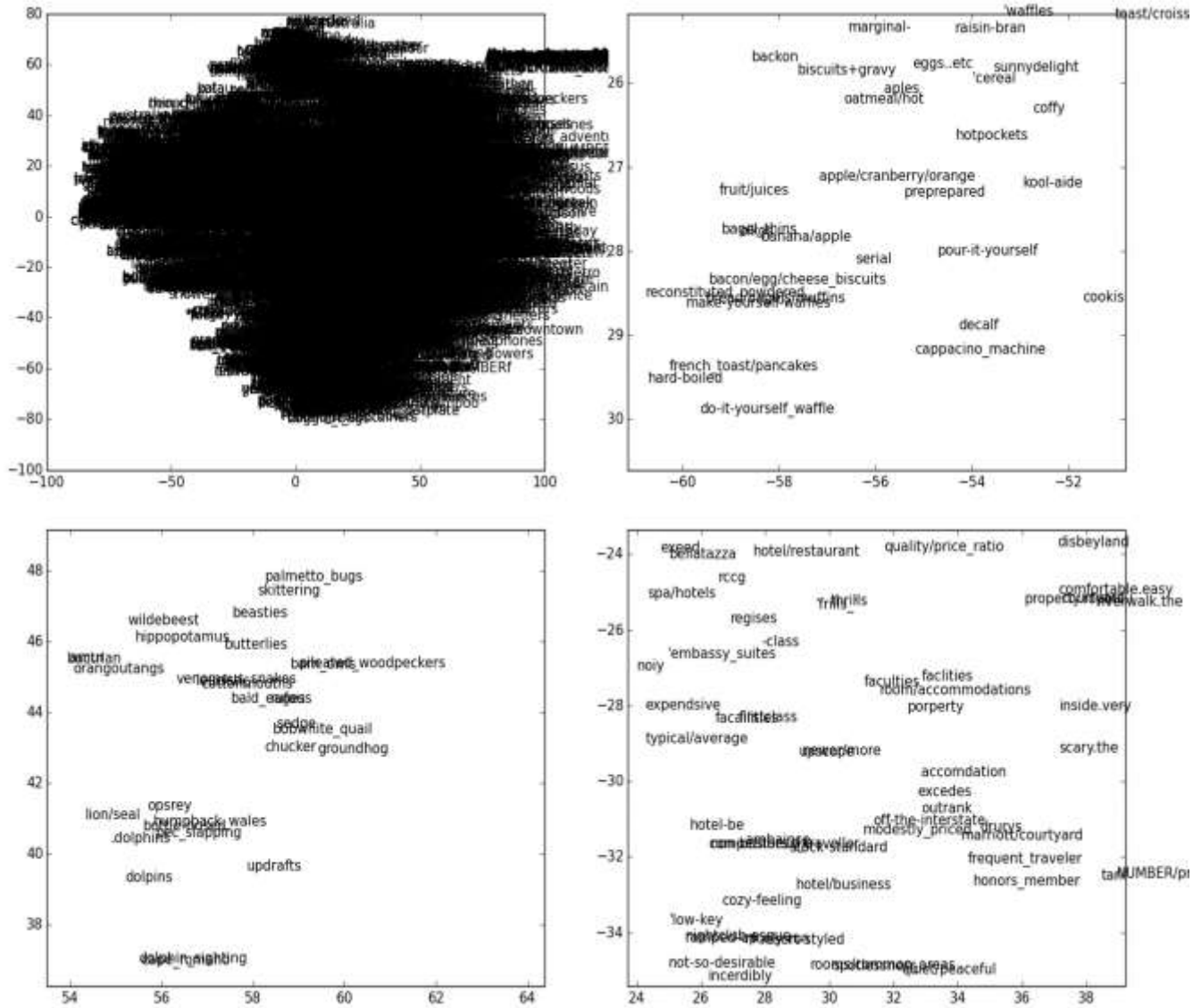

Fig. 1: 2-D embedding of the learned word representation. The top left one illustrates 5000 randomly selected words. The other three figures show zoomed-in view of specific regions.

Fig. **1** presents the 2-D embedding of the words using the learned word embedding matrix. The projection was done by Barnes-Hut-SNE (t-SNE) [9] – a dimensionality reduction technique suitable for visualizing high-dimensional datasets. We can observe that semantically similar words are clustered together (see the zoomed-in plots in Fig. **1**).

B. Intent-specific Feature Generation

The learned word embedding model is used to create three types of intent feature vectors: *a review vector* (**r**), *an intent vector*, and *a visit intent vector* (**i**).

Review vector (**r**): a review undergoes a series of pre-processing steps: tokenization, removal of infrequent words, stopwords filtering, and lemmatization. As described in Section **Error! Bookmark not defined.**, the word embedding model is trained using the entire dataset. For all the words in the review, corresponding embedding vectors are retrieved and summed together. The summed vector is divided by the number of words in the review, i.e., an average embedding vector for the review is computed. This averaged vector is named as the *review vector* for the given review. Essentially, a review of an arbitrary length is represented by a *review vector* with 200 dimensions.

Intent vector: utilizing our labelled intent corpus, a set of intent vectors is constructed using the term frequency-inverse document frequency (TF-IDF) method. The top-50 TF-IDF words for each intent are selected and their embedding vectors are averaged to produce an *intent vector*. As the corpus consists of eight intent categories (classes), we obtain eight intent vectors of 200 dimensions that best represent each category.

Visit intent vector (**i**): given a *review vector* and the eight *intent vectors*, a *visit intent vector* is computed by taking cosine similarities between the *review vector* and each *intent vector*, storing the eight similarity (scalar) values. In essence, the *visit intent vector* can be regarded as a higher-level feature representation of the *review vector* with respect to the eight intents.

Finally, the *review vector* is concatenated with its corresponding *visit intent vector*, yielding a vector (**r+i**) of 208 dimensions. This concatenated vector becomes an input to train a classifier. We experiment with four classification algorithms: naive Bayes (NB), random forest (RF), support vector machine (SVM), and deep neural network (DNN). Fig. 2 below summarizes the process.

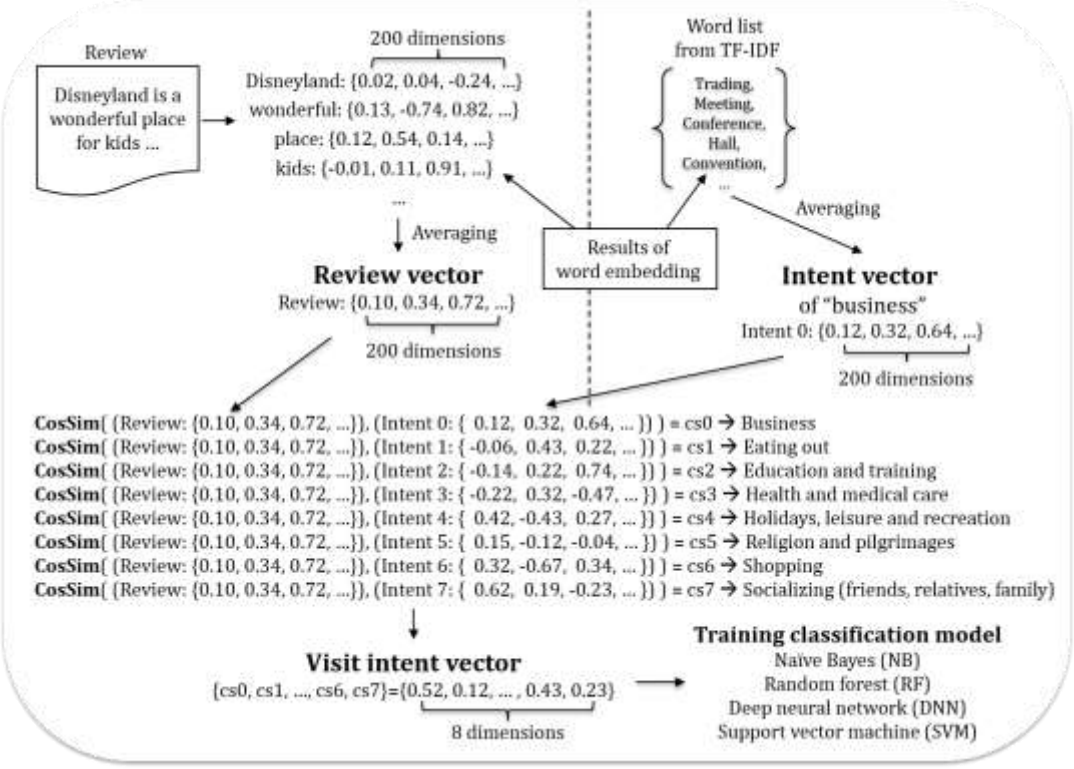

Fig. 2: The process of feature generation

V. EXPERIMENTAL RESULTS

A. Effectiveness of Intent Vectors

Several experiments are conducted to investigate the feasibility of multiclass intent classification, and the effectiveness of intent feature vectors. The performance of each classifier using different *intent vector* combinations is investigated under 10-fold cross-validation—*StratifiedShuffleSplit* [4] is used to sample from the skewed intent distribution. The baseline method is naive Bayes classifiers with token-level bigram model.

The hyperparameters of each classifier are chosen by grid searching:

- Naive Bayes: Gaussian
- Random Forest: 100 trees, Gini impurity

- Deep Neural Network: Tanh(150 units)→ReLU(100 units)→ReLU(50 units)→Softmax, learning_rate=0.01
- Support Vector Machine: Liblinear, l2 penalty, hinge loss, tolerance=0.0001

TABLE 5 shows the average weighted precision, recall, and F-measures of each classifier under 10-fold cross-validation. Compared to the lexical baseline, the use of word embedding features boosts the performance in general. SVM outperforms DNN possibly due to the small size of the intent corpus; in fact, the precision score of DNN is ~3% greater than SVM, but SVM beats DNN in recall score by ~6%, resulting in higher F-measure. We believe that this is probably due to the small size of intent corpus; SVMs are generally known to work well in small dataset, while DNNs require greater volume of data.

The addition of *intent vector* improves the results except for DNN. This is due to the fact that DNN learns higher-level features from lower layers naturally, and may have already learned *intent-vector-like* features with the review vectors alone.

TABLE 5
THE AVG. WEIGHTED PRECISION, RECALL AND F1 SCORES OF CLASSIFIERS UNDER 10-FOLD CV

Classifier	r (%)			i (%)			r+i (%)		
SVM	73	73	71	69	68	64	74	74	72
DNN	76	68	70	74	57	59	76	68	70
RF	73	68	65	69	69	68	73	70	67
NB	63	61	61	59	62	58	63	62	62
NB(bigram)	precision=57, recall=54, F-measure=56								

B. Classifying each Intent

We now investigate how well each travel intent is classified by SVM(r+i) model.

TABLE **6** below shows the precision, recall, and F-measures of each intent.

TABLE 6
THE AVG. WEIGHTED PRECISION, RECALL AND F1 SCORES OF EACH INTENT UNDER 10-FOLD CV

	Bus.	Eating	Edu.	Health	Hol.	Rel.	Shop.	Soc.
Pr.	85	85	88	55	66	88	100	78
Rc.	53	97	66	52	87	85	76	32
F1.	65	90	75	54	75	86	87	45

The precision score for each intent is quite high apart from *health* and *holidays* intents. This is because the *health* reviews often talk about receiving a spa massage or having a facial treatment at resorts or boutique hotels where the *holidays* intent prevails. This is likely to reduce the precision scores for both intents.

Business and *socializing* intents suffer from low recall possibly due to the fact that there are not enough distinct features to learn for these classes. A regular business review often focuses on the quality of venue that the businessperson went to, but less on the nature of the business trip itself.

In the case of *socializing* intent which often takes place in a restaurant-like setting, the reviewer mainly talks about the food she ate or activities that she did, and the fact that she was with her companions is not mentioned enough in the review or buried deeply under the context.

Nevertheless, in general, the top-1 classification results seem promising considering the fact that this is a fine-grained classification. Remember that some POIs such as spa resorts can be annotated as *health* or *holidays* as these intents are not necessarily mutually exclusive. Due to such characteristics of travel intents, it may be more meaningful to look at the top-k classification results.

C. Looking at Intents for POIs

An interesting application of the work is to look at the top-k intents of POIs; intents of a POI effectively represent the major reasons that travellers visit the POI for.

TABLE **7** shows five examples. We observe that *Waikiki Beach* is mostly for *holiday* purpose, but one can also visit there to go *shopping*, as there are shopping centers around the beach. Also for the *Stanford Shopping Center*, we can notice that the mall offers places to *eat* as well as to *shop*. The top-k

We believe that this kind of refined information from mass data can be useful in the field of recommender system as we can additionally utilize the intent similarities of candidate places in the collaborative filtering process.

TABLE 7
THE TOP-K INTENTS OF POIS

POI	Top-3 Intents (%)		
Waikiki Beach	Holidays (79)	Shopping (6)	Education (5)
Affinity Massage and Wellness Center	Health (85)	Socializing (7)	Holidays (5)
Pima Air Space Museum	Education (62)	Holidays (32)	Socializing (3)
Stanford Shopping Center	Shopping (83)	Eating out (11)	Holidays (3)
Canyonland National Park	Holidays (94)	Socializing (2)	Business (1)

VI. Conclusion

We investigated the feasibility of multiclass intent classification in travel review data. We observed that the simple feature generation method using word embedding model yields promising results. The ambiguity of travel intents is explored by creating an intent corpus and through experiments. We noticed that these inherent characteristics of travel intents could represent various reasons people visit a certain POI.

As future works, various methods of computing the review vector can be considered. For example, a review can be represented by a bag-of-centroid using clusters of word embedding vectors. Alternatively, keeping up with the recent trends, a recurrent neural network can be learned to model sentences of arbitrary lengths, and in turn, utilized in the classification. We can also apply the notion of travel intent similarity to recommender systems, and investigate if such additional information makes a difference in performance.

Acknowledgment

This work was supported by Samsung Electronics Co. Ltd.

References

[1] N. Banerjee et al., "RU-In?-exploiting rich presence and converged communications for next-generation activity-oriented social networking," in *Proc. 10th Int. Conf. on Mobile Data Management: Systems, Services and Middleware*, 2009, pp. 222-231.

[2] Y. Bengio, "Deep learning of representations: Looking forward," in *Statistical Language and Speech Processing*, Springer Berlin Heidelberg, 2013, pp. 1-37. http://dx.doi.org/10.1007/978-3-642-39593-2_1

[3] Z. Chen et al., "Identifying Intention Posts in Discussion Forums," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 1041-1050.

[4] G. Churchill and R. Doerge, "Permutation tests for multiple loci affecting a quantitative character," *Genetics*, 1996, Vol: 142, pp. 285-294.

[5] S. Clinchant and F. Perronnin, "Aggregating continuous word embeddings for information retrieval," in *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 100-109.

[6] R. Collobert et al., "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, 2011, Vol: 12, pp. 2493-2537.

[7] V. Gupta et al., "Identifying purchase intent from social posts," in *Proc. Conf. on Weblogs and Social Media*, 2014, pp. 180-186.

[8] P. Haffner et al., "Optimizing SVMs for complex call classification," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2003, pp. 632-635. http://dx.doi.org/10.1109/icassp.2003.1198860

[9] L. van der Maaten, "Accelerating t-SNE using Tree-Based Algorithms," in *Journal of Machine Learning Research*, 2014, Vol: 15 pp. 3221-3245.

[10] T. Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," in *Proc. Conf. on Neural Information Processing Systems*, 2013, pp. 3111-3119.

[11] T. Mikolov et al., "Extensions of recurrent neural network based language model," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 5528–5531.

[12] J. Sang et al., "Contextual and Personalized Mobile Recommendation Systems," *Tools for Mobile Multimedia Programming and Development*, 2013, pp. 82-97. http://dx.doi.org/10.4018/978-1-4666-4054-2.ch005

[13] R. Sarikaya et al., "Deep belief nets for natural language call-routing," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 5680–5683. http://dx.doi.org/10.1109/icassp.2011.5947649

[14] H. Schwenk and J. Gauvain, "Training neural network language models on very large corpora," in Proc. *Joint Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 201-108. http://dx.doi.org/10.3115/1220575.1220601

[15] M. Strohmaier and M. Kröll, "Acquiring knowledge about human goals from search query logs," *Information Processing and Management*, 2012, Vol: 48, No. 1, pp. 63-82.

[16] D. Tang et al., "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 1555-1565. http://dx.doi.org/10.3115/v1/p14-1146

[17] B. Xue et al., "A study on sentiment computing and classification of Sina Weibo with Word2vec," in *IEEE Int. Cong on Big Data*, 2014, pp. 358-363. http://dx.doi.org/10.1109/bigdata.congress.2014.59

[18] UN World Tourism Organization, "Methodological Notes to the Tourism Statistics Database," *UN World Tourism Organization*, 2015.