TERM PROJECT PROGRESS REPORT
for ITU BLG 614E WEB MINING COURSE

.

A Weakly Supervised Graph-based System for Customer Review Categorization

Eray Yıldız
Istanbul Technical University
Computer Engineering Department

I. Project Description

As Web has become one of the most important sources for customers to evaluate and compare products and services, the providers want to monitor the opinion of their costumers with respect to a set of aspects such as price, service and quality. Hence, automated ways to handle and classify customer reviews have been widely studied in the area of Opinion Mining. Most of the studies use supervised classification methods on a manually annotated dataset. Although supervised systems obtain good results for the domain they are trained on, they are almost useless when the domain or language has changed. Annotating data for all domains and languages is not applicable. Because of that, recent studies have focused on unsupervised or semi-supervised methods for customer review categorization task.

In this project, the goal is to develop a system which is able to automatically categorize customer reviews into aspect categories defined by the users. The proposed system is almost unsupervised and requires only a couple of reviews labeled with an aspect category. With a little effort, users are able to explore and analyze customer reviews in any domain.

The architecture of proposed system is illustrated in Figure 1. The system inputs an unlabeled review corpus with a few labeled examples (at least one for each aspect category) and then the system will categorize each review into an aspect category.

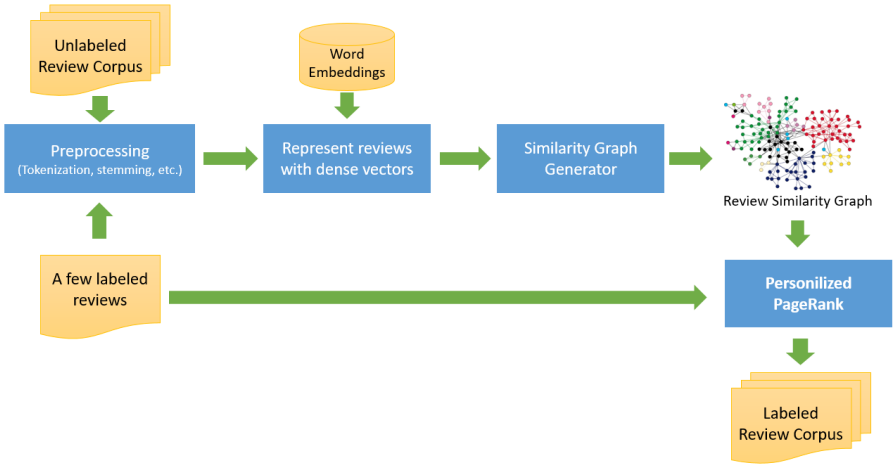


Fig. 1: The architecture of proposed method

In the first step of proposed system, the reviews will be tokenized and lemmatized, the stopwords will be removed to avoid data sparsity problem. Then each review will be represented with a dense vector by leveraging pre-trained word vectors (a.k.a word embeddings) [1] and paragraph vectors (a.k.a doc2vec) [2]. A similarity score will be calculated for each review pair using cosine distance between vector representations of reviews. Using these similarity scores the reviews will be represented with a graph where nodes are reviews and edges are similarity scores. (A threshold parameter will be used to cut off the edges which have similarity scores less than the threshold value). For each user-defined aspect category, Personalized PageRank algorithm [3] will be performed on similarity graph using labeled reviews as topic specific nodes. By this way, each review node will be scored for each aspect category.

II. Project Plan

The schedule of the project is given in Table 2.

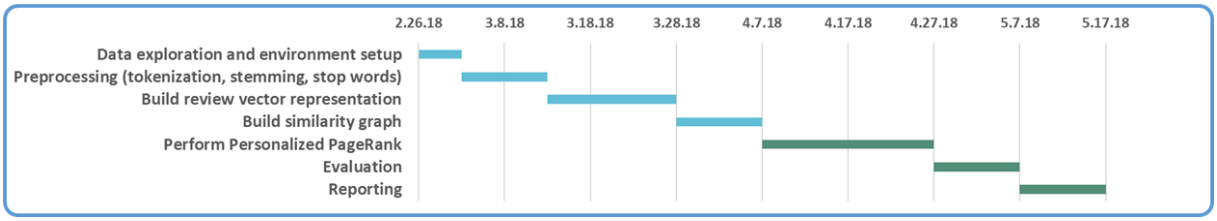


Fig. 2: Project schedule

TABLE I: Example reviews and assigned categories in Turkish restaurant dataset

Review	**Aspect Categories**
bu servise bu fiyatlar ise cok fazla.	RESTAURANT#PRICES
1 kere denedigim fiyatlarinin yüksek yemeklerinin lezzetsiz oldugu bir mekan önermiyorum.	RESTAURANT#PRICES FOOD#QUALITY
Duyduğum kadarıyla suda değil sütte bekletiyorlarmış.	FOOD#STYLE_OPTIONS
Sevmediğimden yiyemedim kenarlarını, ama ortası pek lezizdi.	FOOD#QUALITY FOOD#STYLE_OPTIONS

III. Work Packages

A. Data Exploration

In order to evaluate the performance of proposed method the experiments are performed on a Turkish restaurant review dataset which is published in SemEval 2016 workshop. The dataset contains 350 restaurant reviews which are split into approximately 1.2 sentences. The reviews are labeled with aspect categories in sentence level. The annotation is performed in mutilabel classification setting. Each sentence is assigned wit one or more categories and the categories are organized in two level hierarchal structure. A couple of examples are given in Table I. The frequencies of both first level and second level categories in the dataset are illustrated in Figure 3 and Figure 4.

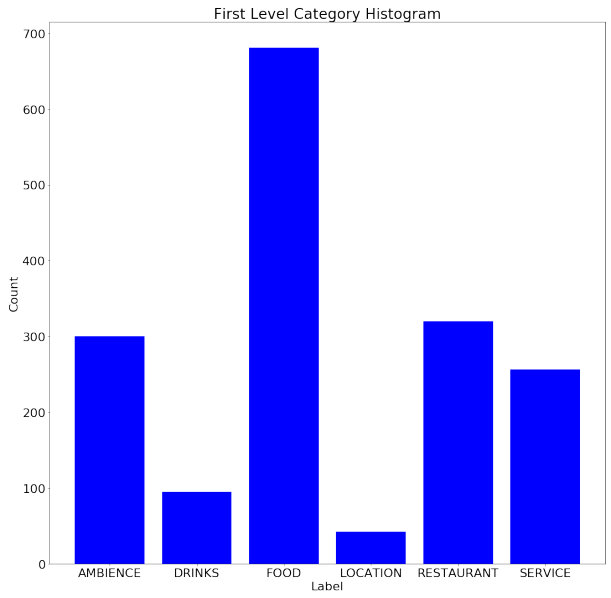


Fig. 3: The frequencies of first level aspect categories in Turkish restaurant dataset

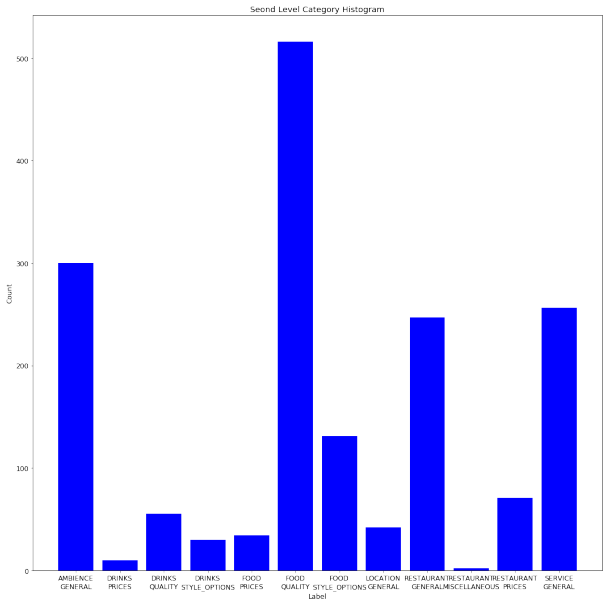


Fig. 4: The frequencies of second level aspect categories in Turkish restaurant dataset

B. Preprocessing

Each sentence in the dataset is lowercased and tokenized using nltk library.[1] Stop words are removed using a Turkish stop words list.[2] A novel context-aware neural stemmer[3] is used to find the root form of each word in the dataset.

C. Vector representation of sentences

The method for obtaining vector representation of reviews has an crucial role in proposed system since the meaning of the reviews will be carried by the vector representations.. TF-IDF vectorization method is chosen as baseline method because it has been widely used in various language processing task for a long time with acceptable performance. TF-IDF vectors of each review sentence in the dataset are obtained using scikit-learn library.[4]

More sophisticated methods to obtain vector representations such as word2vec [1] and paragraph vectors (a.k.a doc2vec) [2] will be further evaluated.

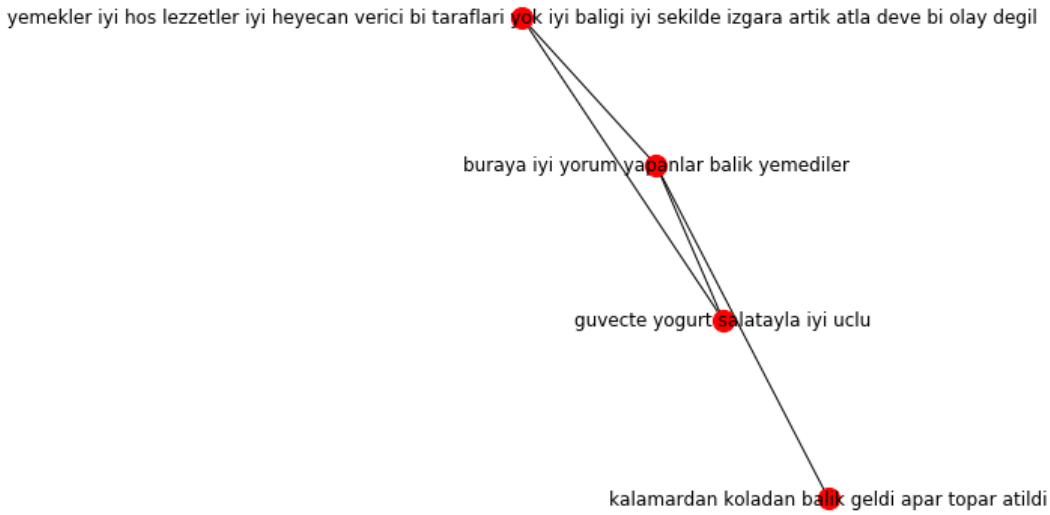


Fig. 5: An example subgraph of similarity graph which is obtained using similarities of TF-IDF vectors

D. Data collection for unsupervised learning of vector representation

Neural network based methods to build semantic vector representations of words and sentences require huge amount of raw text data for capturing high level features of texts. A large text dataset is collected from three different web resources which contain customer reviews to train word2vec and paragraph vectors methods. sikayetvar.com is a heavily used web site where the customers write their complaints about the services they have bought. Hepsiburada is a e-commerce web site where users can buy products from a wide range of sellers and write their reviews about the products. Eksisozluk is a collection of entries submitted by its registered users. The entries are organized in titles. The well known brand names and products are searched in titles and only the entries appear in these titles are fetched. The statistics of the collected dataset which contains approximately 5.6M reviews is given in Table II. This dataset will be used to train neural network based language modeling tools such as word2vec and doc2vec in order to represent each review in Turkish restaurant dataset with dense vectors which carry semantic information about the reviews. In order to evaluate the effect of semantic vector representations of words and sentences, the system will be built using both TF-IDF vectors and semantic vector representations and the performances will be compared.

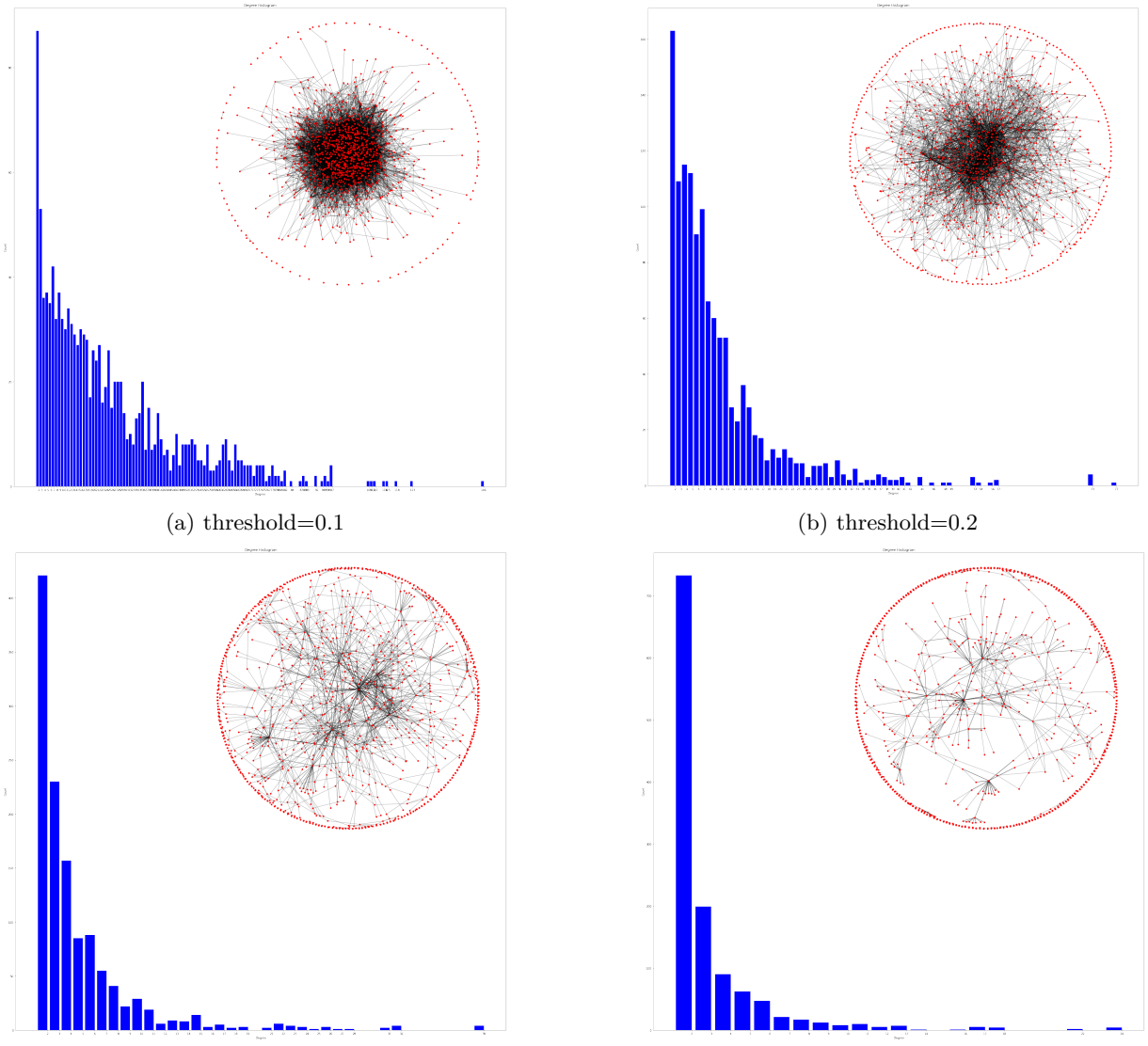
[1]https://www.nltk.org/
[2]https://github.com/xiamx/node-nltk-stopwords/blob/master/data/stopwords/turkish
[3]https://github.com/erayyildiz/Neural-Morphological-Disambiguation-for-Turkish
[4]http://scikit-learn.org

TABLE II: The web resources used to collect customer review text data for unsupervised training

Resource	**# of collected reviews**
sikayetvar.com	570K
hepsiburada.com	853K
eksisozluk	4.2M
total	5.6M

E. Building similarity graph of reviews

A similarity graph is built using the similarities of TF-IDF vectors of reviews. The nodes represent the reviews in the graph and if the similarity of TF-IDF vectors of the reviews is greater than a threshold value an edge is added between the nodes. Cosine similarity method is used to calculate similarity between two TF-IDF vectors. A subgraph of the similarity graph is given in Figure 5.



(a) threshold=0.1

(b) threshold=0.2

(c) threshold=0.3

(d) threshold=0.4

Fig. 6: The degree histograms and overviews of the graphs which are built using different threshold values

Several graphs are built with different threshold values in order to observe how graph structures are effected by chosen threshold values. The degree histograms and overviews of the graphs which are built using different threshold values are shown in Figure 6. All the degree distributions follow power law distribution. The threshold is chosen as 0.1 in order to avoid producing disconnected nodes.

TABLE III: Performance of the system with TF-IDF vectorization on first level categories

	# of Seed Reviews	Threshold	Average Precision	Average Recall	Average F1 Score
Random baseline	-	-	0.24	0.03	0.03
Proposed method	1	0.0005	0.24	**0.40**	**0.32**
	1	0.001	0.29	0.22	0.22
	1	0.002	**0.37**	0.11	0.13
	5	0.0005	0.25	**0.46**	**0.36**
	5	0.001	0.30	0.27	0.26
	5	0.002	**0.43**	0.13	0.15
	10	0.0005	0.26	**0.50**	**0.37**
	10	0.001	0.32	0.32	0.29
	10	0.002	**0.44**	0.15	0.17

TABLE IV: Performance of the system with TF-IDF vectorization on second level categories

	# of Seed Reviews	Threshold	Average Precision	Average Recall	Average F1 Score
Random baseline	-	-	0.13	0.02	0.03
Proposed method	1	0.0005	0.13	**0.51**	**0.26**
	1	0.001	0.15	0.32	0.21
	1	0.002	**0.21**	0.17	0.16
	5	0.0005	0.14	**0.63**	**0.29**
	5	0.001	0.17	0.43	0.25
	5	0.002	**0.27**	0.31	0.23
	10	0.0005	0.14	**0.67**	**0.31**
	10	0.001	0.19	0.49	0.29
	10	0.002	**0.33**	0.38	0.29

F. Personalized PageRank for multi-label classification

The proposed system inputs a customer review collection and represent the reviews in a graph where the nodes are reviews and there is an edge between two nodes if the semantic similarity of the reviews is greater than a threshold value. Then personalized pagerank algorithm [3] is mostly used for finding nodes in a graph that are most relevant to given nodes. In this study, personalized pagerank algorithm is used to find reviews which are related to a given reviews which so that the labels of seed reviews can propagate to other nodes. By this way the proposed system is able to assign labels to most of the nodes using only a few labeled nodes.

Personized pagerank algorithm is applied to the similarity graph which is obtained through TF-IDF vector similarities of the customer reviews. Personlized pagerank scores each nodes in terms of relatedness of the reviews to seed reviews. A threshold value is then used to select related review nodes in the graph. The selected reviews are labeled with the same category of seed reviews.

1) Experimental results: The performance of the system is evaluated on Turkish restaurant dataset with varying number of seed reviews. Various threshold values to determine which nodes are related to given seed reviews are also evaluated. The results of the experiments for both first level and second level categories are reported in Table III and Table IV. The random baseline results are obtained applying standard pagerank algorithm without using seed reviews. As expected the performance improves as the number of seed reviews increase. The best F1 scores are obtained when threshold is set to 0.005.

IV. Conclusion and Future Work

In this study personalized pagerank algorithm is used for multi-label classification. The performance of the system is evaluated on Turkish restaurant dataset where each sentence in the reviews is assigned to multiple aspect categories. TF-IDF vectors are obtained from the review corpus and a similarity graph is built where the nodes are reviews and there are edges between similar reviews. Personalized pagerank algorithm is then applied to the graph in order to propagate the labels of a few seed reviews to the others. The system does not require training data and inputs only a few labeled examples. Although the task is much more harder than a standard supervised classification task, the results are very promising on both first level and second level categories.

In the remaining part of the study, semantic vector representations of words and sentences are used to represent reviews instead of old fashioned TF-IDF vectorization. A large unlabeled customer review data is collected from several web resources and will be used to train neural language models such as word2vec and paragraph vectors. The effects of neural vector representations to the performance of the system will be evaluated.

References

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[2] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[3] Shayan A Tabrizi, Azadeh Shakery, Masoud Asadpour, Maziar Abbasi, and Mohammad Ali Tavallaie. Personalized pagerank clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications*, 392(22):5772–5785, 2013.