Text Similarity Estimation for Targeted Marketing with Outlier Robust Centroids of GloVe Word Embeddings

Tim vor der Brück

School of Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: tim.vorderbrueck@hslu.ch

***Abstract*—Customer segmentation is an important task for marketeers. It is a prerequisite for precise and successful marketing campaigns. The traditional way of conducting it is by clustering based on demographic, geographic and psychographic variables like sex, age, city, or profession. Such an approach has several drawbacks. First, some of these variables might be hard to obtain in practice. Second, deducing from them actual interests for certain products is very hard in practice. In this paper, we present a different approach, in which we use short text snippets provided by users in an online contest to come up with a much more precise user interest profile. In particular, these text snippets are matched to keyword lists representing several marketing target groups like *Freestyle Action Sportsmen*, *Young Performer*, etc. For that, we employed the cosine measure on outlier robust centroids of GloVe word embeddings. These centroids are determined in an iterative fashion that gives most focus on non-outlier vectors and tends to disregard vectors, which are far off from the others. The evaluation showed that we obtained superior results with our method than several baseline approaches including one alternative method of noise reduction based on tf-idf weights.**

Keywords–GloVe; Targeted marketing; Outlier robust centroid

I. INTRODUCTION

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables, and behaviors [1]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner operates a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms. An increasing number of them require the members to write short free-text snippets, e.g., to elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency. Based on the results of a broad survey, the platform provider's marketers assume five different target groups called youth milieus. A sixth milieu called *Special groups* comprises all those who cannot be assigned to one of the upper five milieus. For each milieu (with the exception of *special groups*) a keyword list was manually created to describe its main characteristics. For triggering marketing campaigns, an algorithm shall be developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal.

The semantic similarity of the two documents is then estimated by computing the cosine measure on the two centroids. One typical issue in this approach is that text can contain noise in form of words irrelevant for the actual topic. Such words are often either function words or have a very general meaning and can partly be filtered out using stop word lists. Additionally, one can mitigate this problem by weighting the word vectors according to their associated tf-idf value. We follow in this paper an alternative method to tf-idf word vector weighting, which is the use of an outlier robust centroid. This centroid reduces the influence of outliers by an iterative approach that weights the individual word vectors by their distance to the current centroid.

The remainder of this paper is structured as follows. In Section II, we summarize existing work on semantic similarity estimation. In Section III we describe the process of obtaining the outlier robust centroid in detail. Section IV describes the application to targeted marketing. Section V contains the evaluation results obtained on two manually annotated contests including a discussion. Finally, the paper concludes with Section VI, which summarizes the obtained results and gives an outlook to possible future work.

II. RELATED WORK IN SEMANTIC SIMILARITY ESTIMATION

Semantic similarity estimation is usually based on word or sentence vectors that are first aggregated document-wise to centroid vectors, which are afterwards compared by the cosine similarity. The most popular method to come up with word vectors is Word2Vec [2], which is based on a 3 layer neural network architecture in which the word vectors are obtained as the weights of the hidden layer. Alternatives to Word2Vec are GloVe [3], which is based on aggregated global word co-occurrence statistics and the Explicit Semantic Analysis (or shortly ESA) [4], in which each word is represented by the column vector in the tf-idf matrix over Wikipedia.

The idea of Word2Vec can be transferred to the level of sentences as well. In particular, the so-called Skip Thought Vector model (STV) [5] derives a vector representation of the current sentence by predicting the surrounding sentences.

 ISBN: 978-1-61208-738-2

Sond and Roth [6] propose an alternative approach to applying the cosine measure to the two word vector centroids for ESA word embeddings called Dense-ESA. In particular, they establish a bipartite graph consisting of the best matching vector components by solving a linear optimization problem. The similarity estimate for the documents is then given by the global optimum of the objective function. However, this method is only useful for sparse vector representations. In case of dense vectors, Mijangos et al. [7] suggested to apply the Frobenius kernel to the embedding matrices, which contain the embedding vectors for all document components (usually either sentences or words) (cf. also [8]). However, crucial limitations are that the Frobenius kernel is only applicable if the number of words (sentences respectively) in the compared documents coincide and that a word from the first document is only compared with its counterpart from the second document. Thus, an optimal matching has to be established already beforehand.

Another similarity estimate that employs the entire embedding matrix is the word mover's distance [9], which is a special case of the earth mover's distance, a well studied transportation problem. Basically, this approach determines the minimum effort (with respect to embedding vector changes) to transform the words of one text into the words of another text. The word mover's distance requires a linear optimization problem to be solved. Linear optimization is usually tackled by the simplex method, which has in the worst case, which rarely occurs however, exponential runtime complexity.

In two former papers we proposed for the task of customer segmentation two additional text similarity estimates, one based on an ontology [10] and the other on matrix norms [11] applied on the word similarity matrix over the two texts to compare.

A drawback of most conventional similarity estimates as described above is that slightly related word pairs can have in aggregate a considerable influence on their values, i.e., these estimates are sensitive to noise in the data.

III. Outlier Robust Centroids

The outlier robust centroid is illustrated in Figure 1. We basically use a variant of the Huber centroid but applied on ordinary vectors instead of covariance matrices [12]. The red dot denotes the ordinary centroid of the black dots, the blue dot is the outlier robust centroid. As one can perceive from the figure, the ordinary centroid is much more drawn in direction of the outlier (the black dot on the very bottom) than its outlier robust counterpart.

The procedure to obtain the latter is given in pseudo-code (see algorithm 1). First, the word vector weights are initialized with 1 divided by the number of word vectors. In this way, each word vector is weighted identically at the beginning. Afterward (step 2), our initial centroid is computed as the weighted sum of all word vectors. Now we update the weight, where each vector is weighted by the reciprocal of its distance to the centroid. In bid to avoid weights of infinity, we add a tiny positive amount to the distance prior to building the reciprocal. Using this weighting procedure, very distant vectors, typical outliers, are weighted less than closeby ones. Now we repeat this process returning to step 2 until the coordinates of the centroid have sufficiently converged and remain basically unchanged.

Algorithm 1 Outlier Robust Centroid

```
1:  procedure ROBUST_OUTIER(vec)
2:  numit = size(vec)
3:  w ← [1/numit, ..., 1/numit]
4:  for ever:
5:      C := [0, ..., 0]
6:      w_sum := 0
7:      i := 0
8:      for vec in vecs:
9:          C+ = w[i++] · vec
10:     i := 0
11:     for vec in vecs :
12:         w[i] = 1/(dist(vec, C) + 0.00001)
13:         w_sum+ = w[i++]
14:     for wi in w:
15:         wi := wi/w_sum
16:     if dist(C, last_C) < threshold then
17:         break
18:     last_C := C
19: return C
```

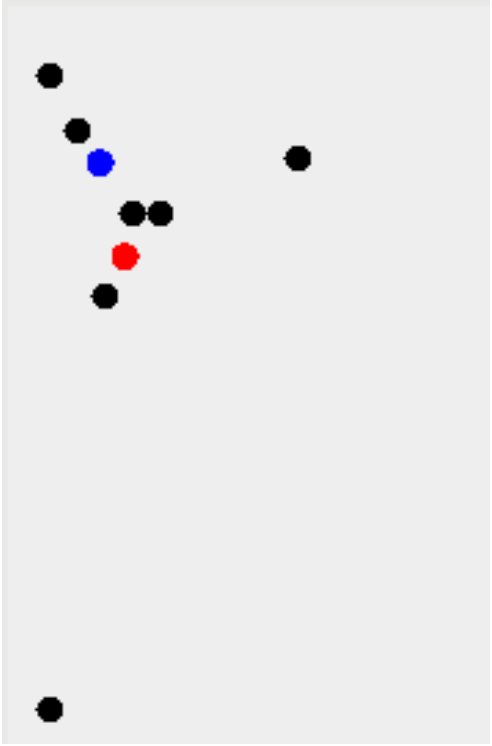

Figure 1. Outlier robust centroid (blue dot) vs ordinary centroid (red dot).

IV. Application to Targeted Marketing

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables and behaviors [1]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner Jaywalker GmbH operates

a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms. Some of them require the participants to write short free-text snippets. For instance, in one of our contests, the participants should specify their three preferred travel destination countries and elaborate on how a perfect holiday there would look like. An example of a participant's answer is given below:

1) Jordanien: Ritt durch die Wüste und Petra im Morgengrauen bestaunen bevor die Touristenbusse kommen
2) Cook Island: Schnorcheln mit Walhaien und die Seele baumeln lassen
3) USA: Eine abgespaceste Woche am Burning Man Festival erleben

English translation:

1) Jordan: Ride through the desert and marveling Petra during sunrise before the arrival of tourist buses
2) Cook Island: Snorkeling with whale sharks and relaxing
3) USA: Experience an awesome week at the Burning Man Festival

Based on the results of a broad survey, the platform provider's marketers assume five different target groups (called *milieus*) being present among the platform members: *Progressive Postmodern Youth* (people primarily interested in culture and arts), *Young Performers* (people striving for a high salary with a strong affinity to luxury goods), *Freestyle Action Sportsmen*, *Hedonists* (rather poorly educated people who enjoy partying and disco music) and *conservative youth* (traditional people with a strong concern for security). A sixth milieu called *Special Groups* comprises all those who cannot be assigned to one of the upper five milieus. For each milieu (with the exception of *Special Groups*) a keyword list was manually created to describe its main characteristics. For triggering marketing campaigns, an algorithm shall be developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal. In case the highest similarity estimate falls below the 10 percent quantile for the distribution of highest estimates, the special groups' milieu is selected. Since the keyword list typically consists of nouns (in the German language capitalized) and the user contest answers might contain a lot of adjectives and verbs as well, which do not match very well to nouns in the Word2Vec vector representation, we actually conduct two comparisons for our Word2Vec based measures, one with the unchanged user contest answers and one by capitalizing every word beforehand. The final similarity estimate is then given as the maximum value of both individual estimates.

Note that we apply the outlier robust centroid only to the word vectors derived from the user snippets since the keyword list is manually defined and should usually be free of noise.

V. Evaluation

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination, speculated about potential experiences with a pair of fancy sneakers (contest 2) and explained why they emotionally prefer a certain product out of four available candidates. We experimented with different keyword list sizes (see Table IV) but obtained the best results with rather few, and therefore precise keywords (see Table V).

TABLE I. Obtained accuracy values for several similarity measures and for several baseline methods. (W)W2VC=(tf-idf-weighted) Word2Vec Embedding Centroids. GloVe,R.=GloVe using Outlier Robust Centroids.

Method	Contest			
	1	2	3	All
Random	0.167	0.167	0.167	0.167
ESA	0.357	0.254	0.288	0.335
ESA2	0.355	0.284	0.227	0.330
W2VC	0.347	**0.328**	0.227	0.330
WW2VC	0.347	0.299	0.197	0.322
GloVe	0.350	0.269	0.258	0.328
GloVe,R.	**0.365**	0.239	0.303	**0.342**
STV	0.157	0.313	0.258	0.189

TABLE II. Minimum and maximum average inter-annotator agreements (Cohen's kappa) / average inter-annotator agreement values for our automated matching method.

Method	Contest		
	1	2	3
Min kap.	0.123	0.295/0.030	0.110/0.101
Max. kap.	0.178	0.345/0.149	0.114/0.209
# Entr.	1544	100	100

TABLE III. Corpus sizes measured by number of words.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

In bid to provide a gold standard, three professional marketers from different youth marketing companies annotated independently the best matching youth milieus for every contest answer. We determined for each annotator individually his/her average inter-annotator agreement with the others (Cohen's kappa). The minimum and maximum of these average agreement values are given in Table II. Since for contest 2 and contest 3, some of the annotators annotated only the first 50 entries (last 50 entries respectively), we specified min/max average kappa values for both parts. We further compared the youth milieus proposed by our unsupervised matching algorithm with the majority votes over the human experts' answers (see Table I).

The Word2Vec and GloVe word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper Corpus and 34 249 articles of the news journal *20 minutes* (see http://www.20min.ch), where the latter is targeted to the Swiss market and freely available at various Swiss train stations (see Table III for a comparison of corpus sizes). By employing articles from *20 minutes*, we want to ensure the reliability of word vectors for certain Switzerland specific expressions like *Velo* or *Glace*, which are underrepresented in the German

 ISBN: 978-1-61208-738-2

TABLE IV. ORIGINAL KEYWORD LISTS (TRANSLATED FROM GERMAN).

Youth Millieu	Keywords
Progressive Postmodern Youth	Trend, Trendsetter, Opinion Maker, Opinion Maker, Opinion, Opinion Leader, Individuality, Individual, Self-realization, Urban, Urbanity, City, Urban, Forward Thinker, Academic, Academic, Conscious, Design, Culture, Cultural, Choice, Freedom, Flexibility, Unbound, Progressive, Ecology, sustainability, new, discover, discovery, postmodern, hip, future, cultural journey, history, buildings, architecture, theatre, language travel, ipster, acting, instrument, musical, vegan, vegetarian, vegetarian, vegetarian, arthouse, independent, beard, criticism, rehearsal, band, books, literature, language, green, secondhand, fair, human right, human rights
Young Performer	Mobility, mobile, flexible, flexibility, performance, performance, performance-oriented, elite, elitist, risk, risk-averse, luxury, luxurious, income, self-realization, self-management, career, spontaneous consumption, education, educated, student, Status, bespoke, Brands, individual, Individuality, excessive, Success, materialistic, materialistic, possession, wealth, enjoyment, Enjoy, Wellness, City break, Gourmet, First class, Business, Opera, Metropole, Money, Account, MBA, CAS, MAS, business, Tailored, Individual
Freestyle Action Sportsman	Apprentice, Music, Sports, Sporty, New, New, Action, Action, Joy, Joy, Experience, Freestyle, Social, Joy, Just, Justice, Adventure, Adventurous, Optimism, Optimist, Extreme, Casual, Sportmania, Improvisation, Improvise, Freestyle, Freedom, Unbound, free, positive, celebrate, party, party, yolo, rap, rhythm, freeride, adventure, snow, mountains, bladen, skating, board, authenticity, interculturality, self-determination, left-liberal, curiosity, sea, nature, natural, video, film, nature, chill, group, homies, style, cool, go-pro
Hedonist	Mainstream, enjoy, enjoyment, intense, casual, unectritical, mass, communicative, entertainment, variety, inconspicuous, carefree, consumption, hedonist, materialistic, materialistic, joy, pleasure, lust, desire, painless, selfish, momentary, present, decadence, decadent, egoism, celebrate, party, party, lazy, lazy, all-inclusive, discount, cheap, last minute, beach, rock, pop, hits, new, current, charts, cinema, stadium, exit, club, drink, event, weed, grass, smoking, street parade, carnival, television, sofa, playstation, xbox
Conservative Youth	conservative, bourgeois, bourgeois, tradition, traditional, modest, modesty, community, common, down-to-earth, down-to-earth, associations, considered, orderly, Switzerland, future, middle class, virtue, virtuous, preserve, Existing, Stable, Stability, Preserve, Protect, Protection, Social, Craft, Democracy, Democratic, People, Hiking, Mountains, History, Homeland, Folklore, Popular, Carnival, Guugen, SVP, Work, Former, Quarter, Closing time, Stammtisch, Beiz

TABLE V. REDUCED KEYWORD LIST (TRANSLATED FROM GERMAN).

Youth milleu	Keywords
Progressive Postmodern Youth	clothing, music, art, freedom, culture, educated
Young Performer	rich, elite, luxury, luxurious
Freestyle Action Sportsmen	Sports, Fitness, Music
Hedonist	poor, communication, self-fulfilment, entertainment, party, music, disco
Conservative Youth	conservation of value, conservativity, citizenship, Switzerland

Wikipedia and the Frankfurter Rundschau corpus. ESA is usually trained on Wikipedia, since the authors of the original ESA paper suggest that the articles of the training corpus should represent disjoint concepts, which is only guaranteed for encyclopedias. However, Stein and Anerka [13] challenged this hypothesis and demonstrated that promising results can be obtained by applying ESA on other types of corpora like the popular Reuters newspaper corpus as well. Unfortunately, the implementation we use (Wikiprep-ESA, URL: https://github.com/faraday/wikiprep-esa) expects its training data to be a Wikipedia Dump. Furthermore, Wikiprep-ESA only indexes words that are connected by hyperlinks, which are usually lacking in ordinary newspaper articles. So we could train Wikiprep-ESA on Wikipedia only but additionally have developed a version of ESA that can be applied on arbitrary corpora (in the following referred to as ESA2) and which was trained on the full corpus (Wikipedia+Frankfurter Rundschau+20 minutes). The STVs were also trained on the same corpus as GloVe and Word2Vec embedding centroids. The actual document similarity estimation is accomplished by the usual centroid approach. An issue we were faced with is that STVs are not bag of word models but actually take the sequence of the words into account and therefore the obtained similarity estimate between milieu keyword list and contest answer would be dependent on the keyword ordering. However, this order could have arbitrarily been chosen by the marketers and might be completely random. A possible solution is to compare the contest answers with all possible permutation of keywords and determine the maximum value over all those comparisons. However, such an approach would be infeasible already for medium keyword list sizes. Therefore, we use a beam search approach instead, which extends the keyword list iteratively and keeps only the n-best performing permutations.

Finally, to verify the general applicability of our approach, we conducted a second experiment, where a novel by Edgar Allen Poe (The purloined letter) was independently translated by two different translators into German. We aim to match a sentence from the first translation to the associated sentence of the second by looking for the assignment with the highest semantic relatedness disregarding the sentence order. The obtained accuracy values based on the first 200 sentences of both translations are given in Table VI. To guarantee an 1:1 sentence mapping, periods were partly replaced by semicolons.

The evaluation showed that the use of outlier robust centroids leads to superior results on our evaluation set in terms of classification accuracy in comparison with its non-robust counterpart on 2 of the three contests and also for the overall comparison, for which all entries of the three contests are merged by concatenation as well as for the second task of translation matching. Furthermore, our method also

 ISBN: 978-1-61208-738-2

TABLE VI. ACCURACY VALUE OTBAINED FOR MATCHING A SENTENCE OF THE FIRST TO THE ASSOCIATED SENTENCE OF THE SECOND TRANSLATION.

Method	Accuracy
ESA	0.672
GloVe	0.706
GloVe,R.	**0.726**
STV	0.716
W2VC	**0.726**

clearly outperforms the use of centroids of tf-idf weighted embeddings, which is an alternative method for noise reduction in the data.

VI. CONCLUSION

We proposed a similarity measure to compare GloVe embeddings from different documents based on outlier robust centroids. This measure was evaluated on the task to assign users to the best matching marketing target groups. We obtained superior results compared to the usual non-robust centroid / cosine measure similarity estimation for contests 1 and 2 as well as overall (just appending the individual contests to form one large contest). As future work, we plan to evaluate additional similarity measures like Dense ESA or novel sentence embedding approaches on our dataset.

ACKNOWLEDGEMENT

Hereby we thank the Jaywalker GmbH as well as the Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus.

REFERENCES

[1] M. Lynn, "Segmenting and targeting your market: Strategies and limitations," Cornell University, Tech. Rep., 2011, online: http://scholorship.sha.cornell.edu/articles/243.

[2] T. Mikolov, I. Sutskever, C. Ilya, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Katar, 2014.

[4] E. Gabrilovic and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," Journal of Artificial Intelligence Research, vol. 34, 2009.

[5] R. Kiros, Y. Zhu, R. Salakhudinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fiedler, "Skip-thought vectors," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2015.

[6] Y. Song and D. Roth, "Unsupervised sparse vector densification for short text similarity," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado, 2015.

[7] V. Mijangos, G. Sierra, and A. Montes, "Sentence level matrix representation for document spectral clustering," Pattern Recognition Letters, vol. 85, 2017.

[8] K.-J. Hong, G.-H. Lee, and H.-J. Kom, "Enhanced document clustering using wikipedia-based document representation," in Proceedings of the 2015 International Conference on Applied System Innovation (ICASI), 2015.

[9] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 957–966.

[10] T. vor der Brück, "Estimating semantic similarity for targeted marketing based on fuzzy sets and the odenet ontology," in Proceedings of SemaPro, Athens, Greece, 2018.

[11] T. vor der Brück and M. Pouly, "Text similarity estimation based on word embeddings and matrix norms for targeted marketing," in Proceedings of NAACL, Minneapolis, USA, 2019.

[12] I. Ilea, H. Hajiri, S. Said, L. Bombrun, C. Germain, and Y. Berthoumieu, "An m-estimator for robust centroid estimation on the manifold of covariance matrices: performance analysis and application to image classification," in Proceedings of the 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 2016.

[13] T. Gottron, M. Anderka, and B. Stein, "Insights into explicit semantic analysis," in Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, UK, 2011, pp. 1961–1964.

 ISBN: 978-1-61208-738-2