Automatically Mapping Ad Targeting Criteria between Online Ad Platforms

Joni Salminen
Qatar Computing Research Institute, Doha, Qatar; and Turku School of Economics, Turku, Finland
jsalminen@hbku.edu.qa

Soon-gyo Jung
Qatar Computing Research Institute, Doha, Qatar
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, Doha, Qatar
bjansen@hbku.edu.qa

Abstract

Targeting criteria in online advertising differ across platforms and frequently change. Because advertisers are increasingly taking a multi-channel approach to online marketing, there is a need to automatically map the targeting criteria between ad platforms. In this research, we test two algorithmic approaches – Word2Vec and WordNet – for mapping ad targeting criteria between Google Ads and Facebook Ads. The results show that Word2Vec outperforms WordNet in finding matches (97.5% vs. 63.6%), covering different criteria (20.0% vs. 13.5%), and having higher similarity scores. However, WordNet outperforms Word2Vec in expert evaluation (Mean Opinion Score = 3.05 vs. 2.46), implying that algorithmic performance metrics may not correlate with expert ratings. Overall, due to specific requirements for mapping ad targeting criteria, automatic means do not (at least yet) offer a satisfactory solution for replacing human judgment.

1. Introduction

Online marketers use various targeting criteria to reach their desired audiences. These criteria include, for example, demographics, search intent, lifestyles, interests, psychographics, brand affinities, and purchase behavior [1, 2, 32]. With the ever-growing amount of customer data, the number of available targeting criteria has exploded to thousands. Facebook has reportedly more than 5,000 targeting criteria [33], while other programmatic advertising platforms also provide similar numbers of targeting criteria.

This space vast space of targeting criteria results in at least four challenges for online marketers:

- First, the management of different possible ways to target customers is difficult due to the large number of targeting criteria and the cognitive limitations of human marketers for managing the criteria.
- Second, the targeting criteria differ by ad platform. There are no standards for online ad targeting criteria, essentially requiring advertisers to manually find similar criteria across the platforms.
- Third, the available targeting criteria frequently change, as new ones are added, and old ones are removed or merged.

Overall, these challenges make it more difficult for advertisers to choose the optimal targeting criteria on one platform and locate them on another platform. This task is essential in **multi-channel marketing** [25].

Suppose, for example, that *Criterion A* is performing well in *Platform 1*, and the advertiser would like to use the same criterion in *Platform 2*. If *Platform 2* is missing the same criterion, the targeting criterion does not directly translate. Some root causes for these challenges relate to (a) the naming conventions differ between the platforms (e.g., one platform can call age "age" while another calls it "age group"), and (b) there are no unique IDs to map the topic between platforms. Thus, one cannot apply the standard practice of ID-based mapping.

However, there is a conceptual overlap between the topic sets of the platforms, meaning that the topics capture the same (or highly similar) customer attribute but are using different terms (as in the above example of age). So, an automated mapping approach would be beneficial for advertisers.

Here, we introduce this problem as the **disambiguation problem of online ad targeting criteria** ('OADP' for short), which is defined as the automatic mapping of targeting criteria between two or more online ad platforms.

There is a limited amount of publicly available research on this problem, which motivates the present study. Here, we test the ability of a word embedding approach called *Word2Vec* [15]to solve the OADP. Word embeddings are a standard approach in NLP (natural language processing), where they have been a robust approach for many language-related problems [22]. Conceptually, the OADP is about matching a pair of the two most concepts that most likely refer to the same idea. The similarity function in Word2Vec allows us to quantify the semantic likeness of two concepts – in this case, two targeting criteria.

We compare the performance of Word2Vec against *WordNet*, which is a human-curated taxonomy for the discovery of related concepts. In a certain sense, this

URI: https://hdl.handle.net/10125/70727
978-0-9981331-4-0

comparison provides implications on marketing automation ("man vs. machine") in the context of the OADP, which we discuss in the last section of the manuscript. Our research questions (RQs) are:

- **RQ1:** *Do word embeddings (Word2Vec) outperform a human-curated baseline (WordNet) to automatically map online ad targeting criteria?*
- **RQ2:** *Is either of the tested approaches good enough for the needs of online ad professionals?*

This study contributes to computational advertising research by (a) introducing the OADP, (b) showing promising results of algorithms and automation to solve this challenge, and (c) sharing resources – ad targeting lists and replication code in the Python scripting language – for further development. For online advertisers, our research contributes by examining how automation can facilitate the work of online ad professionals engaged in multi-channel marketing, burdened with managing multiple targeting criteria lists.

2. Related Literature

The dominant online ad platforms, such as Google Ads and Facebook Ads, command considerable power over online transactions and the evolution of practices and technologies in the advertising industry [28]. Their economic importance is also tremendous, as these companies employ high volumes of data scientists, researchers, marketers, and other professionals. Also, the dominant ad platforms generate substantial incomes from their mother companies – over 90% of the revenue of Google and Facebook is generated via advertising sales [16].

The success of online ad platforms is mostly arising from the performance gains obtained by marketers relative to traditional channels [7]. This success has also contributed to cross-disciplinary research around online marketing. Particularly, computer science is focused on aspects such as algorithmic solutions [27]. Economics research often deals with optimal ad markets [14], [34]. Marketing, Human-Computer Interaction, and Information System Science scholars tend to focus on organizational and human aspects, such as banner blindness [4], ad blocking [8], and design of effective online ads [3].

The common thread of these studies is combining advertising and technology. To this end, Yang et al. have suggested an umbrella phrase of "computational advertising" [36] that refers to the use of computational techniques to facilitate advertising functions, usually through the means of online advertising platforms [30]. It is to this cross-sectional field of computational advertising that our research contributes.

However, despite the broad interest in computational advertising, we could locate no previous study focused on the OADP. Nonetheless, similar problems have been studied in the field of NLP. Word embeddings, in particular, have performed well in many use cases where automated text processing is required. Examples include automatic translation [37], generation of text for chatbot dialogue [24], topic classification [29], and analysis of algorithmic bias [5].

In this research, we test if they perform well for the use case of mapping online ad targeting criteria between different ad platforms. Because such overlaps are not predictable and because the criteria may change over time, there is a need for automated approaches that could flexibly find the matching targeting criteria between the ad platforms.

One approach that relies on finding similarities between words is word embeddings, also known as word vectors [20, 22]. These embeddings represent words, sentences, phrases, or documents in a mathematical space. Each token has a position (coordinates) relative to the other tokens that enables arithmetic calculations, such as determining the distance of two pairs. The distance is often considered a proxy for semantic similarity, as in the classic example of "King is Man what Queen is to Woman" [5]. Thus, if we have two or more lists of targeting criteria, it is possible to convert the items of those lists into vectors and compare, one by one, their similarities in the vector space. Another approach is to use a lexical database of words, such as the publicly available WordNet [9, 23] taxonomy.

3. Research Objective

Our research objective is to evaluate these two approaches for matching targeting criteria between two major online advertising platforms. The benefits of such research include highlighting possible algorithmic concepts as to why one approach performed better than the other, along with significant practical impact for online advertising criteria mapping in support of multi-channel targeting.

We retrieve targeting criteria from two platforms, Google AdWords and Facebook Ads, including 1,074 targeting criteria. We experiment with WordNet and Word2Vec to automatically map the targeting criteria lists. We evaluate the results using quantitative metrics and ratings from three online marketing professionals.

Table 1: Examples targeting categories and criteria. In total, 1074 targeting criteria were used.

Category	Platform	Definition	Example criterion	# of criteria (% of total)
Affinity Audiences	Google	Affinity audiences were created for businesses currently running offline ads and expanding their reach with an online presence [13].	News Junkies & Avid Readers	106 (9.9%)
In-Market Audiences	Google	In-market audiences are in the market, which means that they are researching products and are actively considering buying a service or product. In-market audiences can help reach consumers close to completing a purchase [13].	Consumer Electronics/Game Consoles	467 (43.5%)
Interests	Facebook	Interests are inferred from Facebook's information about the users. For example, a company selling fashionable jewelry can target customers in the category "Shopping and Fashion" [10].	Entertainment / Reading / Magazines	290 (27.0%)
Behaviors	Facebook	Behaviors are inferred from Facebook's information about a user's purchasing behaviors, device usage, and other activities [11].	Digital Activities / Canvas Gaming / Played game in the last 7 days	211 (19.6%)

4. Methodology

4.1. Data Collection

We manually retrieve 1,074 targeting criteria, 573 from Google Display Network, and 501 from Facebook Advertising Manager using advertiser credentials. The targeting criteria fall under four categories designated by the platforms: **Affinity Audiences (Google)**, **In-Market Audiences (Google)**, **Interests (Facebook)**, and **Behaviors (Facebook)**. Table 1 defines the categories and shows examples of the criteria within them. We note that the composition of the criteria lists may have changed at the publication time (the data was collected in February 2018), even though for the study at hand, this does not matter as we are specifically interested in testing word embeddings for the OADP in general.

4.2. Data Processing and Analysis

To process the data, we delete non-alphabetic symbols and transform the words in categories to lower-case format. We experiment with Word2Vec and WordNet because these models are well established, widely available for real-world implementation, and have shown high performance in many tasks in the natural language processing domain [20, 22].

Word2Vec. Word2Vec refers to a group of models that output word embeddings, i.e., distributed numerical representations of words [35]. Word2Vec models are shallow, two-layer neural networks that represent linguistic contexts of words [20]. In practice, word embeddings enable the use of arithmetic operations, such as calculating the distance between words based on their location in a vector space. Word2Vec embeddings can be used for many purposes, e.g., determining the semantic similarity of words, finding analogies, carrying out machine translation, and modeling topics [22, 31].

In this research, we use the *Gensim* library with a Word2Vec model that is trained on the Wikipedia text corpus[1] containing 1.9B words from 4.3M text articles.

In brief, our algorithm for mapping the category lists using Word2Vec works as follows (an example of mapping Affinity category with Interests):

- **Step 1:** For items in the *Affinity category*, create a set of words that are also present in the Wikipedia model.
- **Step 2:** Repeat for items in *Interests*. In case no word is present in the model vocabulary, add an empty set.
- **Step 3:** Create a list of Word2Vec semantic similarities between each item in the *Affinity category* and in *Interests*.
- **Step 4:** Find the highest similarity for each *Affinity category–Interests* pair.
- **Step 5:** Append to a list a tuple of the *Affinity category* under consideration, the original *Interests* item that has the highest similarity, and the value of similarity.
- **Step 6:** Repeat the procedure for *"Interests" and "Affinity categories"*, *"In-market audiences" and "Behaviors"*, and *"Behaviors" and "In-market audiences"*.

[1] https://corpus.byu.edu/wiki/

WordNet. WordNet is a lexical database of manually curated English words. It groups words (nouns, verbs, adjectives, and adverbs) by their conceptual and lexical relations (e.g., synonymy, hyponymy, meronymy, etc.). The goal of WordNet is to provide an interlinked network of meaningfully related words and concepts. The database is publicly available[2]. At the time of the study, the database includes 117,000 synsets, i.e., sets of synonyms with information on the relations among these synonym sets or their members. In this research, we utilize two Python functions in the WordNet library: *WordNetLemmatizer* and *WordNetSimilarity*. The former retrieves the roots of the words in the targeting criteria lists, and the latter finds similar concepts in the WordNet database.

4.3. Experimental Set-Up

The underlying problem that motivates this research is that targeting categories are different across ad platforms. Facebook (FB), Google (GO), and other programmatic ad platforms have thousands of categories that are not readily compatible. However, the underlying concepts, like interests and preferences, that these targeting criteria represent tend to overlap. For example, "dog lovers" and "dog owners," although using different words, express an affinity with dogs. Because the concept of vector similarity in word embeddings assumes that semantic associations are captured by words being close in the vector space, there is a possibility that the mapping could be automated, saving online advertisers much time and effort from a manual mapping of the criteria.

To conduct the experiments, we use Word2Vec and WordNet to map the categories shown in Table 2. We do the mapping both ways: from GO to FB and from FB to GO. Interests and affinities are conceptually referring to the same type of information, whereas interests and behaviors are conceptually apart from one another. Therefore, the pre-mapping of category types is conducted to improve the possibility of producing meaningful matches. We map these four combinations, each using WordNet and Word2Vec, resulting, in total, 4 x 2 = 8 mappings.

Table 2: Experimental set-up. Both algorithms – Word2Vec and WordNet – will be applied to FB to GO and GO to FB mappings.

Facebook to Google	Google to Facebook
Interest → Affinity audiences	Affinity audiences → Interest
Behaviors → In-market audiences	In-market audiences → Behaviors

[2] https://wordnet.princeton.edu/

5. Evaluation

We evaluate each mapping using (1) quantitative metrics and (2) expert evaluation.

5.1. Quantitative Evaluation

For this part of the evaluation, we use three metrics:

- The **success rate** measures how many criteria the algorithm was able to return a match. This is calculated based on a threshold value (for Word2Vec) and based on a direct match (for WordNet).
- The **similarity score** expresses how close the target criterion is to the source criterion. For Word2Vec, this estimate is based on cosine similarity [22]; for WordNet, it is based on word similarity [19].
- **Criteria coverage** measures how many criteria each model addressed. Criteria coverage tells how diverse the mappings were. Ideally, the algorithm can utilize the diversity of the available candidate criteria to find a suitable match.

The WordNet implementation outputs "No similar" when a match is not found, but the Word2Vec implementation always outputs a value. For this reason, we define similarity scores lower than threshold $t = 0.30$ as unsuccessful for Word2Vec. In addition, we normalize WordNet's similarity scores between 0 and 1 using Min-Max normalization [12] to make them comparable with the Word2Vec scores.

Table 3: Quantitative evaluation. Word2Vec obtains higher scores for each metric.

	Success Rate	Similarity Score	Criteria Coverage
Word2Vec (FB to GO)	95.6%	0.675	20.8%
Word2Vec (GO to FB)	99.3%	0.627	19.2%
WordNet (FB to GO)	83.0%	0.505	14.0%
WordNet (GO to FB)	44.2%	0.281	12.9%
Word2Vec (Average)	**97.5%**	**0.651**	**20.0%**
WordNet (Average)	63.6%	0.393	13.5%

Results (see Table 3) show that Word2Vec was able to successfully match 97.5% of the topics. WordNet mapped 63.6% of the topics. Thus, the success rate of Word2Vec was 53% better than that of WordNet.

The average similarity score of the mapped pairs was 0.651 for Word2Vec and 0.393 for WordNet. Thus, the similarity score of Word2Vec was 66% better than that of WordNet. Finally, Word2Vec used 20.0% of the available criteria, whereas WordNet used 13.5%. Thus, Word2Vec made use of 48% more criteria than WordNet.

The limitation of success rate, similarity score, and criteria coverage is that they are all technical metrics that do not tell how useful online advertisers would find the criteria mapping in practice. For this purpose, we conduct a manual evaluation using feedback from professional online marketers.

5.2 Expert Evaluation

The results were manually evaluated with the help of three online advertising experts. Two of the experts were recruited from an online marketing company. Both had more than five years of experience in online advertising, including FB and GO Ads.

The third expert was recruited using UpWork, a freelancer platform with a wide range of specialists from different fields, including online marketing. We ensured the qualifications of the hired person by reviewing her work history, paying attention to:

- the types of jobs completed in UpWork (corresponded to digital marketing)
- the ratings received for those jobs (avg. had to be > 4/5), as well as
- how many jobs they completed successfully in Upwork out of the ones that they started (had to be > 90%).

The offered compensation was USD 70$ for the Upwork expert, whereas the two company representatives were not financially compensated but participated out of general curiosity for the study topic. Two were men, one woman, with an average of 5.3 years of experience in the online marketing industry. All the experts were given the following guidance:

> *"You are shown two targeting criteria from different online ad platforms: Google Ads and Facebook Ads. Your task is to evaluate how well Criterion A from Platform 1 matches Criterion B from Platform 2. Your options:* ***1 = Does not match at all, 2 = Matches poorly, 3 = Matches not particularly poorly or well, 4 = Matches well, 5 = Matches perfectly****."*

Due to the experts' busy schedules, they could not rate each pair. Thus, we opted for a random sample, asking the professionals to rate 836 pairs (38.9% of the total pairs). In total, 2,508 ratings were given. The agreement between the raters was calculated using *Intraclass correlation* (ICC). The obtained result is ICC = 0.623, which indicates moderate reliability [17].

We use the *Mean Opinion Score* (MOS) to measure the overall quality of the matches. This metric captures a user's opinion of a system's output quality [18]. Typically, the scale is in the range of 1-5, where 1 represents the lowest perceived quality, and 5 represents the highest perceived quality. The calculation for MOS is simply the arithmetic mean of all individual ratings provided by the human subjects, such that

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N},$$

where R denotes the individual ratings for an item by N raters. The results are shown in Table 4. Figure 1 shows the distribution of the scores.

Table 4: Mean scores from three raters. The differences between scores of WordNet and Word2Vec, and GO to FB and FB to GO are statistically significant at p < 0.001. The calculation was done using Welch's t-test, transforming categorical variables into a corresponding integer.

	WordNet	Word2Vec	GO to FB	FB to GO
MOS	**3.05****	2.46	**3.03****	2.52
***significantly higher than the reference group at p < 0.001.*				

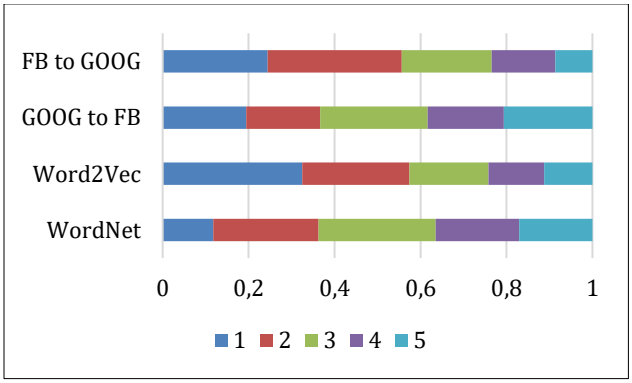

Figure 1: The MSO distribution in rated groups. WordNet has relatively more ratings in 3-5 (higher end), where Word2Vec has more on the lower end (1-2). Similarly, conversions from GO to FB appear to be working better than from FB to GO.

A closer examination (Figure 2) reveals that the higher propensities of WordNet to perform well, especially on the Google Affinities to Facebook Interests that has 47.9% of ratings of 5 ("perfect match"). This corroborates the finding of category specificity. Overall, conversion from Google to Facebook performs better than *vice versa*, and WordNet is better, according to the experts. Surprisingly, converting from Facebook Interests to Google Affinities does not work equally well, achieving only 18.5% of 5s from the total ratings, even though this number is still high relative to most other mappings (see Figure 2).

	Score=1	Score=2	Score=3	Score=4	Score=5
WordNet: B to IM	13.0 %	45.0 %	31.7 %	9.0 %	1.3 %
WordNet: IM to B	17.8 %	17.4 %	44.3 %	19.3 %	1.1 %
WordNet: I to A	15.2 %	21.5 %	19.5 %	25.3 %	18.5 %
WordNet: A to I	2.2 %	8.6 %	14.9 %	26.3 %	47.9 %
Word2Vec: B to IM	49.7 %	32.5 %	13.0 %	4.5 %	0.3 %
Word2Vec: IM to B	46.6 %	26.5 %	24.6 %	1.9 %	0.4 %
Word2Vec: I to A	16.2 %	21.5 %	18.9 %	24.9 %	18.5 %
Word2Vec: A to I	15.2 %	17.8 %	19.4 %	21.0 %	26.7 %

Figure 2: Proportion of scores experts gave to each mapping. Green indicates more, red indicates less. B = Behaviors; IM = In-market audiences; I = Interests; A = Affinities. WordNet's performance on A-to-I is excellent (47.9% of the ratings in the highest category), whereas Word2Vec's performance on B-to-IM and IM-to-B is particularly weak, with 49.7% and 46.6% of the ratings in the lowest category, respectively.

We also asked the experts if the matches were good enough (*"Would you use this criterion (in Column B) to replace the other criterion (in Column A)?"*). The responses to this question support the ones shown in Table 4, with a significantly higher proportion of yes answers to WordNet (M=0.46) than to Word2Vec (M=0.32), $t(417) = 7.52$, $p < 0.001$. Similarly, there were significantly more 'yes' answers for GO to FB mapping (M=0.49) than for FB to GO mapping (M=0.30), $t(789) = 5.75$, $p < 0.001$.

Since we have scores from three experts, we can compute a majority vote where at least two raters agree on the match being good enough for practical use. Computing this reveals a 'yes' rate of 32.1% (N=134) for Word2Vec and 45.7% (N=191) for WordNet.

Table 5: "Would you use this criterion to replace the other criterion?". The expert ratings reveal a large variation among the source and target criteria. Mappings from Affinities (A) to Interests (I) were generally successful (more than 50% of responses indicate a positive answer), whereas the mappings from In-marketing audiences (IM) to Behaviors (B) were considerably less successful.

	A to I	I to A	IM to B	B to IM
yes	167	111	23	24
no	43	87	153	228
total	210	198	176	252
yes rate	79.5 %	56.1 %	13.1 %	9.5 %

In other words, there are 43% more majority 'yes' votes for WordNet than for Word2Vec, and the difference between the groups is statistically significant, $X2(1, N = 418) = 16.36$, $p < 0.001$. As previously, clear differences were observed among the categories (see Table 5). The average yes rate among all categories was 38.9%.

9. Discussion

9.1. Main Implications

This research is one of the rare attempts to investigate the possibilities of automation to facilitate the work of online advertisers. The results indicate that:

- Quantitative metrics and expert assessment disagree. The former indicate Word2Vec performs better than WordNet, and the latter indicates the opposite.
- Automatic mapping performance varies according to the source platform. Mapping from Google to Facebook obtains higher expert ratings than vice versa.
- Experts gave significantly better scores to WordNet than to Word2Vec. This implies that word embeddings (at least those implemented in Word2Vec) are not mature enough to solve the OADP.
- The experts only moderately agree on their ratings. This may arise from the fact that the quality of the match is a subjective measure, with personal preferences and imagined contexts affecting it.
- The performance of automatic mapping varies by category. This suggests the pre-selection of categories to be mapped can improve (or worsen) the performance. Humans can improve the results by selecting conceptually similar categories (e.g., affinities and interests).

9.2. Why Do the Results Differ?

In practice, both tested approaches have their strengths and weaknesses. Whereas Word2Vec treats words as numerical distributions, WordNet treats words as specific associations that are manually coded into the network. Therefore, WordNet aims at capturing the human logic of processing words into an ontological representation, whereas Word2Vec is the purely computational approach that learns the associations from the data.

The connections prescribed in the WordNet database seem to be more conceptually sound (as they have been hand-crafted by humans). In contrast, the meaning of the close associations in words found by Word2Vec is more ambiguous than a specified taxonomy. On the other hand, due to its unsupervised nature, Word2Vec can scale up to billions of words, whereas WordNet is handicapped by its manual curation. Another advantage is that WordNet is only available in English, whereas the Word2Vec (or word embeddings, in general) can be trained for any language.

The results highlight some of the general limitations of algorithms to solve problems that require expert knowledge. Algorithms "always give an answer," but this answer may not always be what a domain expert would accept. In turn, human-curated taxonomies have trouble

covering all possible decision-making situations – in this case, all concepts expressed in the targeting criteria lists.

9.3. Implications for online advertisers

Word2Vec is better when a rough approximation for suitable candidate criteria in another ad platform is sufficient (as it has a higher success rate)

WordNet is better when there is no possibility to manually review the suitable candidate criteria (as it has higher expert ratings)

A suggested approach, given the overall quality of the matches, uses Word2Vec (or another word embedding based approach) to generate a list of candidate matches for professional advertisers. Using this pre-filtered list of matches, advertisers can then make the final choice of criteria they wish to adopt.

9.4. Limitations and Future Work

The research has limitations. First, the mapped lists differ in their range (i.e., they have an unequal number of categories). Therefore, the results can needlessly penalize the algorithms – if there is no match in the target list, the algorithm obviously will not find a match. We used the largest publicly available lists that we could locate for this study, but we are aware that there are much larger targeting criteria lists used in the industry. The algorithms should be tested using those lists in future work.

Second, the similarity threshold for Word2Vec was set based on the intuition that the values lower than that would be close to the minimum of the natural range of the metric; changing the threshold value would affect the obtained success rate. A perhaps more suitable way to determine the threshold value would be to perform sensitivity analyses that would correlate the similarity values with MOS values. We leave this for future work.

Third, even when an algorithm finds a *technically* close match, that match could be judged as poor by the practitioners. This was hinted by one of the raters who stated: "*Because I'm personally interested in camping, 'outdoor enthusiasts' is not sufficient to replace 'camping/hiking' alternative*." In real use cases, the marketing goals also affect the required granularity. For example, sometimes a close match ("lifestyle match," e.g., "mobile phone users") can be adequate. In contrast, other times, a more exact match is needed ("Vodafone users" when mobile operator Orange is campaigning). Thus, the applicability of the algorithms is case dependent, and advertisers should consider automatic mappings as *suggestions* that can potentially save time navigating the targeting criteria of online ad platforms.

Fourth, as the number of items in each category and their abstraction levels differ between the categories, it is unclear whether using only the highest similarity score would work well. Not only one-to-one matching but multiple matching (i.e., 1-to-n) could be considered for better matching results. To this end, future work could investigate how online advertisers make use of top-n recommendations given by the matching algorithm, and if using these recommendations improves the advertisers' task performance of finding matching categories.

Fifth, future work should allocate efforts on the possible creation of standard ontology of online ad targeting criteria, as this would provide guidelines for online ad platforms to harmonize their ad targeting criteria and thus help resolve the disambiguation problem.

Finally, future research could inspect other embedding-based approaches like Glove [26], fastText [21], or Universal Sentence Encoder [6], as well as including more platforms and more targeting criteria in the comparison. We believe that using other embedding methods could possibly *improve* the candidate criteria generation, but we are skeptical that any of the current models can *replace* human judgment in the process. For future research and development in this area, we make the source code of our algorithms available on GitHub[3].

9.5. Conclusion and Future Work

We investigated if automatically computed word embeddings can outperform a human-curated list for the OADP. The results show that Word2Vec gives better technical scores than WordNet, but WordNet gives better expert evaluation scores than Word2Vec. Results suggest that the tested algorithms are not reliable enough for replacing human judgment.

10. Acknowledgments

We thank the anonymous reviewers for their many helpful comments towards improving the quality of this manuscript. We also thank the three online advertising experts that evaluated the quality of the mappings.

11. References

[1] An, J., H. Kwak, J. Salminen, S. Jung, and B.J. Jansen, "Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data", *ACM Transactions on the Web (TWEB) 12*(3), 2018.

[3] https://github.com/joolsa/Automatically-Mapping-Ad-Targeting-Criteria-between-Online-Ad-Platforms

[2] Ashish Kathuria, Bernard J. Jansen, Carolyn Hafernik, and Amanda Spink, "Classifying the user intent of web queries using k-means clustering", *Internet Research 20*(5), 2010, pp. 563–581.

[3] Bayles, M.E., "Designing online banner advertisements: Should we animate?", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2002), 363–366.

[4] Benway, J.P., "Banner Blindness: The Irony of Attention Grabbing on the World Wide Web", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 42*(5), 1998, pp. 463–467.

[5] Bolukbasi, T., K.-W. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings", *Advances in Neural Information Processing Systems*, (2016), 4349–4357.

[6] Cer, D., Y. Yang, S. Kong, et al., "Universal sentence encoder", *arXiv preprint arXiv:1803.11175*, 2018.

[7] Chan, T.Y., C. Wu, and Y. Xie, "Measuring the lifetime value of customers acquired from Google search advertising", *Marketing Science 30*(5), 2011, pp. 837–850.

[8] Despotakis, S., R. Ravi, and K. Srinivasan, "The beneficial effects of ad blockers", *Management Science*, 2020.

[9] Ercan, G., and F. Haziyev, "Synset expansion on translation graph for automatic wordnet construction", *Information Processing & Management 56*(1), 2019, pp. 130–150.

[10] Facebook, "How to reach the right people on Facebook", *Facebook Business*, 2018. https://en-gb.facebook.com/business/m/interest-targeting

[11] Facebook, "Choose your audience", *Facebook Business*, 2018. https://en-gb.facebook.com/business/products/ads/ad-targeting

[12] Furlan, B., V. Batanović, and B. Nikolić, "Semantic similarity of short texts in languages with a deficient natural language processing support", *Decision Support Systems 55*(3), 2013, pp. 710–719.

[13] Google, "About audience targeting - Google Ads Help", 2018. https://support.google.com/google-ads/answer/2497941?hl=en

[14] Graepel, T., J.Q. Candela, T. Borchert, and R. Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine", *Proceedings of the 27th international conference on machine learning (ICML-10)*, (2010), 13–20.

[15] Hu, K., Q. Luo, K. Qi, et al., "Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis", *Information Processing & Management 56*(4), 2019, pp. 1185–1203.

[16] Koetsier, J., "Report: Google Captures Nearly 80% Of All Retail Search Ad Spend", *Forbes*, 2018. https://www.forbes.com/sites/johnkoetsier/2018/03/09/report-google-captures-nearly-80-of-all-retail-search-ad-spend/

[17] Koo, T.K., and M.Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research", *Journal of Chiropractic Medicine 15*(2), 2016, pp. 155–163.

[18] Kumar, B., S.P. Singh, A. Mohan, and A. Anand, "Performance of quality metrics for compressed medical images through mean opinion score prediction", *Journal of Medical Imaging and Health Informatics 2*(2), 2012, pp. 188–194.

[19] Li, Y., Z.A. Bandar, and D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources", *IEEE Transactions on Knowledge and Data Engineering 15*(4), 2003, pp. 871–882.

[20] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[21] Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations", *arXiv preprint arXiv:1712.09405*, 2017.

[22] Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, 3111–3119.

[23] Miller, G.A., "WordNet: a lexical database for English", *Communications of the ACM 38*(11), 1995, pp. 39–41.

[24] Oh, K.-J., D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation", *Mobile Data Management (MDM), 2017 18th IEEE International Conference on*, IEEE (2017), 371–375.

[25] Parise, S., and P.J. Guinan, "Marketing using web 2.0", *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, IEEE (2008), 281–281.

[26] Pennington, J., R. Socher, and C. Manning, "Glove: Global vectors for word representation", *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (2014), 1532–1543.

[27] Qin, J., W. Qi, and B. Zhou, "Research on Optimal Selection Strategy of Search Engine Keywords Based on Multi-armed

Bandit", *2016 49th Hawaii International Conference on System Sciences (HICSS)*, IEEE (2016), 726–734.

[28] Salminen, J., *Power of Google: A study on online advertising exchange*, Master's thesis. Turku School of Economics, Turku, 2009.

[29] Salminen, J., H. Almerekhi, M. Milenković, et al., "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media", *Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018)*, (2018).

[30] Salminen, J., N. Gach, and V. Kaartemo, "Platform as a Social Contract: An Analytical Framework for Studying Social Dynamics in Online Platforms", In A. Smedlund, ed., *Collaborative Value Co-creation in the Platform Economy*. Springer, 2018, 41–64.

[31] Salminen, J., J. Luotolahti, H. Almerekhi, B.J. Jansen, and S. Jung, "Neural Network Hate Deletion: Developing a Machine Learning Model to Eliminate Hate from Online Comments", *Lecture Notes in Computer Science (LNCS 11193)*, Springer (2018).

[32] Salminen, J., S. Seitz, B.J. Jansen, and T. Salenius, "Gender Effect on E-Commerce Sales of Experience Gifts: Preliminary Empirical Findings", *Proceedings of International Conference on Electronic Business (ICEB 2017)*, (2017).

[33] TechCrunch, "Facebook is removing over 5,000 ad targeting options to prevent discriminatory ads", *TechCrunch*, 2018. http://social.techcrunch.com/2018/08/21/facebook-is-removing-over-5000-ad-targeting-options-to-prevent-discriminatory-ads/

[34] Weber, T.A., "Dynamic Learning in Markets: Pricing, Advertising, and Information Acquisition", *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences (HICSS)*, (2019).

[35] Wikipedia, "Word2vec", *Wikipedia*, 2018. https://en.wikipedia.org/w/index.php?title=Word2vec&oldid=869116438

[36] Yang, Y., Y.C. Yang, B.J. Jansen, and M. Lalmas, "Computational Advertising: A Paradigm Shift for Advertising and Marketing?", *IEEE Intelligent Systems 32*(3), 2017, pp. 3–6.

[37] Zou, W.Y., R. Socher, D. Cer, and C.D. Manning, "Bilingual word embeddings for phrase-based machine translation", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (2013), 1393–1398.