Guiding Creative Design in Online Advertising

Shaunak Mishra
Yahoo Research
shaunakm@verizonmedia.com

Manisha Verma
Yahoo Research
manishav@verizonmedia.com

Jelena Gligorijevic
Yahoo Research
jelenas@verizonmedia.com

ABSTRACT

Ad creatives (text and images) for a brand play an influential role in online advertising. To design impactful ads, creative strategists employed by the brands (advertisers) typically go through a time consuming process of market research and ideation. Such a process may involve knowing more about the brand, and drawing inspiration from prior successful creatives for the brand, and its competitors in the same product category. To assist strategists towards faster creative development, we introduce a recommender system which provides a list of desirable keywords for a given brand. Such keywords can serve as underlying themes, and guide the strategist in finalizing the image and text for the brand's ad creative. We explore the potential of distributed representations of Wikipedia pages along with a labeled dataset of keywords for 900 brands by using deep relevance matching for recommending a list of keywords for a given brand. Our experiments demonstrate the efficacy of the proposed recommender system over several baselines for relevance matching; although end-to-end automation of ad creative development still remains an open problem in the advertising industry, the proposed recommender system is a stepping stone by providing valuable insights to creative strategists and advertisers.

CCS CONCEPTS

• **Information systems → Online advertising**.

KEYWORDS

Online advertising; relevance matching; ad creative; creative design

ACM Reference Format:
Shaunak Mishra, Manisha Verma, and Jelena Gligorijevic. 2019. Guiding Creative Design in Online Advertising. In *Thirteenth ACM Conference on Recommender Systems (RecSys '19), September 16–20, 2019, Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3298689.3347022

1 INTRODUCTION

Online advertising is a huge industry with a worldwide revenue of ~100,000 million USD projected for 2019 [3], and is crucial for increasing brand awareness as well as influencing online users towards purchases (conversions) [10, 23]. The ad creatives, *i.e.*, the text and images used for ad campaigns, are typically designed by creative strategists employed by the brands (advertisers), or by third party advertising agencies. In general, the creative design process may involve multiple steps. Creative strategists usually start with conducting market research about the brand, its products, prior ad creatives or creatives of other competitor brands. Following such market research, with target themes in mind, the strategist may choose the ad image along with the accompanying text, and send it to the brand's legal team for feedback and approval. The entire design process, as described above, is very time consuming, ranging from a few days to weeks, depending on the experience of the strategist, and the number of creatives requested.

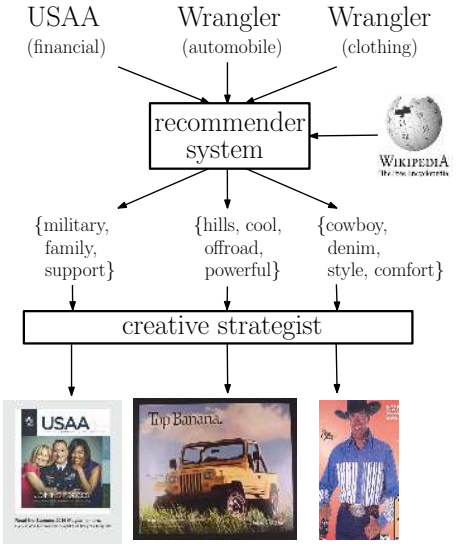

Figure 1: Illustrative example of a recommender system providing brand specific keywords to assist the creative strategist. The images in this example were sourced from the ad creatives dataset [2], and the recommended keywords were sourced from the brand Wikipedia pages respectively.

Vis-a-vis the challenges described above, in this paper, we propose a recommender system for guiding creative strategists. Specifically, the proposed recommender system can ingest features (*e.g.*, brand name, category, Wikipedia page) associated with a brand (input), and recommend a list of keywords (output) which can be consumed by a creative strategist as underlying themes to come up with creatives as illustrated in Figure 1.

The strategist can use this list of brand specific keywords to query a stock image library [1] or design a custom image and text around this theme. As shown in Figure 1, the category also represents important piece of information (in the case of *Wrangler*) as it is helpful towards disambiguation and output relevance (*automobile* versus *clothing*). Therefore, we design a deep neural network based recommender system *Creative-Assist* with the goal of inferring

RecSys '19, September 16–20, 2019, Copenhagen, Denmark

ACM ISBN 978-1-4503-6243-6/19/09…$15.00
https://doi.org/10.1145/3298689.3347022

brand specific keywords as outputs for a given *brand-category* pair. The proposed system exploits two datasets (details in Section 3.2):

(1) *creatives dataset*: publicly available dataset [2] of 64,000 ad creatives over 900 brands in 39 categories (1,579 brand-category pairs), along with questions and answers (~ 3 per creative) about the brand and its product (*e.g.*, the reasons for buying from the brand depicted in the creative; we extract target keywords from these answers).
(2) *brand wiki pages*: Wikipedia pages associated with the 1,579 brand-category pairs inferred in the creatives dataset as well as Wikipedia pages of other ~ 69,000 companies identified by Wikipedia's *Template:Infobox company* [8].

In particular, we treat a brand-category pair's *wiki page* as a proxy *query*, and *target keywords* as relevant *documents* in the standard query-document relevance ranking framework [15]. In addition, we employ the state-of-the-art deep relevance ranking model (DRMM [15]) with doc2vec embeddings [20] for the Wikipedia pages and target keywords. The proposed approach shows a significant lift above several competitive baselines in terms of precision and recall; in particular, DRMM achieves a lift of 11% over the second best (baseline multilayer perceptron, details in Section 4). To the best of our knowledge, we are the first to study the problem of recommending keywords for a brand's creative from a relevance ranking perspective, and demonstrate that the *creatives dataset* [2] and *wiki pages* can provide valuable creative insights.

2 RELATED WORK

In this section, we first cover relevant background on online advertising, followed by related work in automatic understanding of ad images, relevance matching, and sentiment analysis.

Online advertising and ads exploration: In a typical online advertising scenario, advertisers work with ad platforms [10, 14, 23, 31], and launch campaigns to show ads on online publishers served by the ad platform. Advertisers create one or more creatives with the help of creative strategists to target relevant online users. Typical performance metrics that advertisers care about include click-through-rate (CTR), and conversion rate (CVR) [10]. It is common for advertisers to *rotate* the choice of creatives, and study which ones have better performance (*e.g.*, CTR, CVR). Such *exploration and exploitation* based on performance feedback can be enhanced via reinforcement learning approaches as studied in [30] and [21]. In this context, the proposed recommender system in this paper can enable the faster development of a large initial set of creatives, which can be further pruned via feedback based strategies described above (*i.e.*, [21, 30]).

Automatic understanding of ad images: The *creatives dataset* [2] is one of the key enablers of the proposed recommender system. This dataset was introduced in [17], where the authors focused on automatically understanding the content in ad images and videos from a computer vision perspective. The dataset has ad creatives with annotations including topic (category), questions and answers (*e.g.*, reasoning behind the ad, expected user response due the ad).

Relevance matching and collaborative filtering: Since our goal is to recommend a set of highly relevant creative keywords for a given brand-category pair, our task can be modeled both as a collaborative filtering problem where *user-item* latent representations can be used for recommendations [16, 19, 27], as well as query-document relevance ranking [15]. In the context of collaborative filtering, approaches based on matrix factorization [19], low rank feature interactions (factorization machines [27]), and recent neural network based approaches [16] have been effective in *user-item* recommendation scenarios. Since we can only suggest a handful of keywords to creative strategists, we rely on approaches that can order keywords relative to their importance with respect to a given brand-category pair. It is worth noting that there may be cases where two keywords are *related* (*e.g.*, obtained via matrix factorization) but are *non-relevant* for a given brand or its creative. Given the restriction on number of keywords and their relevance to the brand under consideration, we argue that employing ranking models is better suited to our problem. Hence, we pose this as a query-document relevance matching problem in this paper. Specifically, we leverage the state-of-the-art deep relevance matching model (DRMM) proposed in [15], and use it to solve the keyword ranking problem for a given brand-category pair (details in Section 3.3). Our experiments in Section 4 also show that the ranking model performs better than a collaborative filtering baseline.

Sentiment analysis: Data sources like Wikipedia, tweets, and product reviews are all potential sources of keywords which can guide creative design for a brand. In this context, extracting keywords, and understanding the sentiment associated with the keywords is another candidate method for recommending creative keywords. Many state-of-the-art sentiment analysis methods are readily available as tools (*e.g.*, NLTK-SentiWordNet [5] and VADER [13]). Prior work on sentiment analysis includes analyzing review data sets [25, 32], tweets [12, 29], and online news articles [24].

3 CREATIVE-ASSIST

3.1 Problem setup

Given a list of keywords in the vocabulary ($\mathcal{V} = \{w_1, w_2, \ldots\}$), our objective is to recommend keywords that are relevant for a given brand-category pair. We formalize this as a ranking problem, wherein the input query is a brand-category pair and a ranked list of keywords is the output of the proposed system. As in ranking problems, each word corresponds to a document, where a document can be either relevant or non-relevant for a brand-category pair. Formally, $query\ (q_i) = brand-category\ pair$, $document\ (d_i) = word \in \mathcal{V}$, $label\ (y_i) = relevant/not\ relevant$ for creative design. The details on data sources, label generation (*i.e.*, relevant/not relevant), representation of query and document (*i.e.*, embeddings for brand-category pair and words), and the relevance ranking model are covered in the following subsections.

3.2 Data sources

As previously mentioned in Section 1, we use two data sources in this paper: (i) creatives dataset, and (ii) brand wiki pages. While we focus on the Wikipedia dataset in this work, it is worth noting that our methods can be easily extended to use any dataset rich in brand-related keywords such as tweets or product reviews.

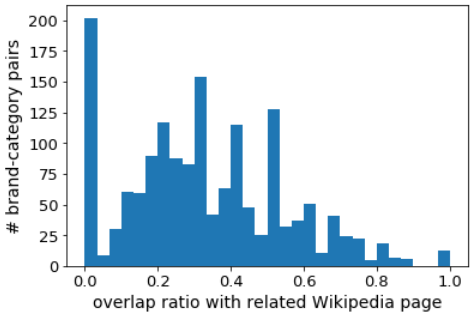

Figure 2: Histogram showing the Wikipedia-keyword overlap ratio across $1,579$ brand-category pairs. For a brand-category, the overlap ratio is the count of target keywords (inferred via [2]) present in the related Wikipedia page divided by the total count of target keywords. Over 50% of the brand-category pairs have overlap ratio $\geq 30\%$.

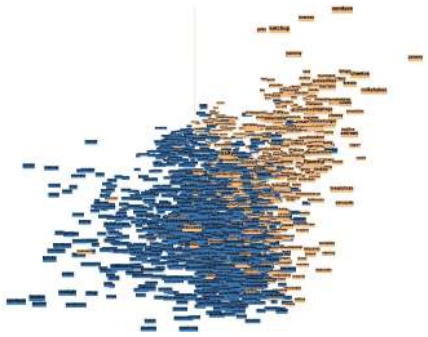

Figure 3: Separability of target keywords for category *food* (in yellow) and category *cars* (in blue) in doc2vec [20] space.

Creatives dataset: This is a publicly available data set [2, 17] with about $64,000$ advertisement creatives, spanning over about 900 brands across 39 categories. The dataset provides the following annotations for each creative: (i) topics (39 types), (ii) questions and answers as reasons for buying from the brand depicted in the creative (~3 per creative). For example, in the case of *Wrangler (clothing)* input as shown in Figure 1, a sample question is "*why do you want to buy Wrangler jeans?*", and a sample answer is "*because it has style and is very comfortable*". Using such a sample, the (target) keywords for a *Wrangler (clothing)* creative can include *style* and *comfortable* as shown in Figure 1. In addition to the existing annotations, we add the following annotations: (i) brand present in a creative, (ii) Wikipedia page relevant to the brand-category pair in a creative, and (iii) the set of target keywords associated with each brand-category pair (as described in Section 3.3). This led to a derived dataset from [2] with 1579 unique brand-category pairs with over 900 unique brands spanning 39 categories.

Brand wiki pages. We also consider the Wikipedia pages linked to the 1579 brand-category pairs, as well as Wikipedia pages of ~69,000 brands identified by *Infobox company* template [8].

Our preliminary data analysis on these two datasets revealed two insights which shaped our proposed methods:

(1) There was reasonable overlap between the target keywords for a brand-category pair, and the words in the Wikipedia page related to the brand-category (as shown in Figure 2).
(2) Distributed word embeddings [20] trained using *brand wiki pages*, and the answers documents in the *creatives dataset* (details in Section 3.3) showed reasonable separability of target keywords across different categories of brands; Figure 3 shows an example for two categories (*food* and *cars*).

3.3 Methodology

Label generation. For each brand-category pair, the answers for brand preference provided by annotators in [2] was used to extract the ground truth set of relevant keywords for each brand-category pair. We used state-of-the-art NLTK part-of-speech (POS) tagger to filter proper nouns, pronouns and stop words. These filters were used to ensure selection of keywords that can be eloquently used for creative design. We randomly sampled words from the Wikipedia page of brand-category pair to construct the set of non-relevant keywords (with relevant:non-relevant words ratio $1:15$). These set of relevant and non-relevant words were used for training our ranking model, and for evaluation. For cross validation, we used a $80\% - 20\%$ train-test split across the 1579 brand-category pairs.

Query-document representations. For the vocabulary derived from the answers documents of ~64,000 creatives, and ~69,000 *brand wiki pages*, we obtained distributed word embeddings using doc2vec [20]. The documents for doc2vec were: (i) Wikipedia pages, and (ii) documents formed by concatenating all answers for a creative.

DRMM based ranker. We further build on recent advances in deep learning, especially in relevance ranking, to order target keywords with respect to a brand-category input. We use the state-of-the-art deep relevance matching model (DRMM) model proposed in [15] whose architecture is shown in Figure 4. DRMM is designed to capture variable length local interactions between query and document terms. It is initialized with fixed length matching histograms which are precomputed similarities between query and document terms. Each query term matching histogram is passed through a matching multilayer perceptron (MLP), and the overall score is aggregated with a query term gate which is a softmax function over all terms in that query. For gating, we use the softmax function as follows:

$$g_i = \frac{\exp(w_g {x_i}^{(q)})}{\sum_{j=1}^{M} \exp(w_g {x_j}^{(q)})}, \tag{1}$$

where w_g denotes the weight vector of the term gating network and ${x_i}^{(q)}$ denotes the i-th term in the Wikipedia page.

DRMM is provided with two inputs: (i) the embedding of the brand-category's Wikipedia page as a query, and (ii) the embedding of the word. It then learns to predict the relevance of the given word with respect to the query brand-category pair. In addition, the corresponding doc2vec embeddings are used for initializing the query and document embeddings for DRMM. Given that our input documents (*i.e.*, keywords) are short, we employ a variant of DRMM which only selects top k interactions between a Wikipedia document embedding, and the embedding of keywords. We consider a pairwise ranking hinge loss to train DRMM as explained below.

For ease of explanation, we denote the brand-category by just the brand below. Given a triple (brand, w^+, w^-) where w^+ is ranked higher than w^- with respect to query brand, the loss function is defined as:

$$\mathcal{L}(brand, w^+, w^-; \theta) = max(0, 1 - s(brand, w^+) + s(brand, w^-))$$

where $s(brand, w)$ denotes the predicted matching score for word w and the query $brand$; θ includes the parameters for the feed forward matching network and those for the term gating network. We tuned three hyperparameters: the embedding size of the last layer (we found that 10 works the best), batch size (set to 300) and iterations (we found the model to converge on 10 iterations).

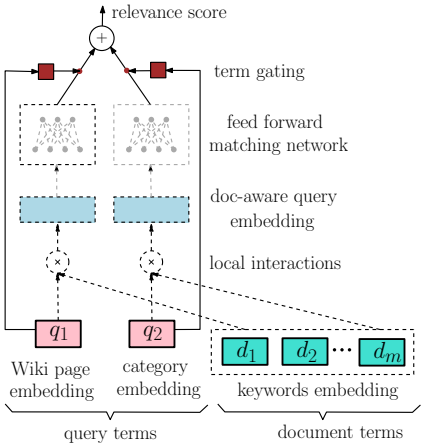

Figure 4: DRMM architecture.

4 EXPERIMENTS

Here, we review the methods used to compare the performance of our model, and the metrics used to evaluate each baseline approach.

4.1 Baseline approaches

We consider several baseline methods (listed below) to compare with the performance of our proposed model.

K-nn on doc2vec space: We train doc2vec model on the Wikipedia documents and documents formed by concatenating all answers for a creative, as described in Section 3.3. To suggest target keywords for a hold-out brand-category, we select the closest K keywords vectors to the vectors of training set creatives in the given category.

TF-IDF: Term frequency based statistics are frequently used to order documents in information retrieval. We use the well known TF-IDF metric to determine relevance of each keyword with respect to a brand-category pair's Wikipedia page. The inverse document frequency (IDF) of each keyword is computed using all the brand Wikipedia articles in the training set.

Logistic regression (LR): We train an LR model that determines binary relevance of keywords for a given brand-category pair. We use one-hot encoding of the tokens in brand name, categories of the brand, and the keyword as features to train the model (using Vowpal Wabbit [7]).

Logistic regression + sentiment (LR-sent): Given that one may want to recommend only positive keywords for ad creative design, we also train a model that exploits the keyword sentiment as features to determine its relevance for a given brand-category pair. In particular, we concatenate a three dimensional vector of sentiment scores obtained via VADER [6] (corresponding to *pos, neg, neu*) to the feature vector described for LR above.

Factorization machine (FM): The FM model learns lower order projections of two sets of interactions: (i) interaction between tokens in the brand name and keywords, and (ii) interaction between brand categories and keywords. We used Vowpal Wabbit [7] for our experiment.

Multilayer perceptron (MLP): We use an MLP that takes brand-category and keyword pair as input (where we initialize their representations with doc2vec embeddings) and predicts a binary relevance of the keyword. We use cross-entropy loss for training. We experimented with both 2-layer and 3 layer hidden units and found that models with 2 layers of hidden units performs the best. The MLP model was implemented using MatchZoo library.

In addition to the above baselines, for our DRMM experiments, we used the implementation in MatchZoo [4] with 2 hidden layers for the feed forward matching network. The 10 dimensional doc2vec embeddings were used for initialization.

4.2 Results

We evaluate the performance of our approach (and baselines) using two metrics: precision at K (P@K), and recall at K (R@K). We calculate these metrics for each brand-category pair, and obtain the aggregate metrics (over all the brand-category pairs in the test set) as reported in Table 1. We observe that DRMM significantly outperforms all baselines (with more than 17% lift over LR in terms of precision). K-nn and TF-IDF show performance significantly lower than the other supervised approaches. The sentiment feature, and FM provide no improvement over the LR baseline. The MLP baseline is better than LR, but still inferior compared to DRMM.

Model	P@5	P@10	P@20	R@5	R@10	R@20
K-nn	0.05	0.05	0.03	0.00	0.00	0.00
TF-IDF	0.02	0.02	0.01	0.00	0.01	0.01
LR	0.32	0.31	0.29	0.01	0.02	0.03
LR-sent	0.32	0.31	0.26	0.01	0.02	0.03
FM	0.32	0.28	0.27	0.01	0.02	0.03
MLP	0.44	0.33	0.23	0.10	0.13	0.17
DRMM	0.51	0.42	0.34	0.24	0.27	0.32

Table 1: Precision and recall for baselines and DRMM.

5 CONCLUSION

In this paper, we make progress towards the goal of end-to-end creative automation by proposing a keyword recommendation system for creative design. The DRMM based *Creative-Assist* can be extended to tweets and product reviews about brands. As an extension of our work, one can generate a *seed* set of initial creatives (via a stock image library) using combinations of recommended keywords, and then perform online A/B tests to identify the creatives leading to superior CTR and CVR performance.

REFERENCES

[1] Adobe creative cloud stock photos. https://www.adobe.com/creativecloud/stock.html.

[2] Automatic understanding of image and video advertisements. http://people.cs.pitt.edu/~kovashka/ads.

[3] Digital advertising. https://www.statista.com/outlook/216/100/digital-advertising/worldwide.

[4] Match zoo. https://github.com/NTMC-Community/MatchZoo.

[5] NLTK SentiWordNet. http://www.nltk.org/howto/sentiwordnet.html.

[6] Vader sentiment. https://github.com/cjhutto/vaderSentiment.

[7] Vowpal wabbit. https://github.com/JohnLangford/vowpal_wabbit/wiki.

[8] Wikipedia infobox company template. https://en.wikipedia.org/wiki/Template:Infobox_company.

[9] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[10] N. Bhamidipati, R. Kant, S. Mishra, and M. Zhu. A large scale prediction engine for app install clicks and conversions. In *CIKM 2017*.

[11] E. Colleoni, A. Arvidsson, L. K. Hansen, and A. Marchesini. Measuring corporate reputation using sentiment analysis. In *Proceedings of the 15th International Conference on Corporate Reputation: Navigating the Reputation Economy, New Orleans, USA*, 2011.

[12] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[13] C. H. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[14] J. Gligorijevic, D. Gligorijevic, I. Stojkovic, X. Bai, A. Goyal, and Z. Obradovic. Deeply supervised model for click-through rate prediction in sponsored search. *Data Mining and Knowledge Discovery*, 2019.

[15] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016.

[16] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *WWW*, 2017.

[17] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. In *CVPR*, 2017.

[18] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[19] Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *KDD*, 2008.

[20] Q. Le and T. Mikolov. Distributed representations of sentences and documents. ICML'14, 2014.

[21] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, and R. Jin. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[22] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[23] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: a view from the trenches. KDD 2013.

[24] S. Mishra, A. Pappu, and N. Bhamidipati. Inferring advertiser sentiment in online articles using wikipedia footnotes. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, 2019.

[25] T. Munkhdalai and H. Yu. Neural semantic encoders. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 397. NIH Public Access, 2017.

[26] A. Radford, R. Jozefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

[27] S. Rendle. Factorization machines. In *IEEE International Conference on Data Mining*, ICDM 2010.

[28] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[29] N. A. Vidya, M. I. Fanany, and I. Budi. Twitter sentiment to analyze net brand reputation of mobile phone providers. *Procedia Computer Science*, 72:519–526, 2015.

[30] J. Zhao, G. Qiu, Z. Guan, W. Zhao, and X. He. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[31] Y. Zhou, S. Mishra, J. Gligorijevic, T. Bhatia, and N. Bhamidipati. Understanding consumer journey using attention based recurrent neural networks. *KDD*, 2019.

[32] Z.-H. Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. *arXiv preprint arXiv:1702.08835*, 2017.