The Automated Copywriter: Algorithmic Rephrasing of Health-Related Advertisements to Improve their Performance

Brit Youngmann
Microsoft Research
Herzliya, Israel
t-bryoun@microsoft.com

Elad Yom-Tov
Microsoft Research
Herzliya, Israel
eladyt@microsoft.com

Ran Gilad-Bachrach
Microsoft Research
Herzliya, Israel
rani.gb@gmail.com

Danny Karmon
Microsoft Healthcare NExT
Herzliya, Israel
Danny.Karmon@microsoft.com

ABSTRACT

Search advertising is one of the most commonly-used methods of advertising. Past work has shown that search advertising can be employed to improve health by eliciting positive behavioral change. However, writing effective advertisements requires expertise and (possible expensive) experimentation, both of which may not be available to public health authorities wishing to elicit such behavioral changes, especially when dealing with a public health crises such as epidemic outbreaks.

Here we develop an algorithm which builds on past advertising data to train a sequence-to-sequence Deep Neural Network which "translates" advertisements into optimized ads that are more likely to be clicked. The network is trained using more than 114 thousands ads shown on Microsoft Advertising. We apply this translator to two health related domains: *Medical Symptoms* (MS) and *Preventative Healthcare* (PH) and measure the improvements in click-through rates (CTR).

Our experiments show that the generated ads are predicted to have higher CTR in 81% of MS ads and 76% of PH ads. To understand the differences between the generated ads and the original ones we develop estimators for the affective attributes of the ads. We show that the generated ads contain more calls-to-action and that they reflect higher valence (36% increase) and higher arousal (87%) on a sample of 1000 ads. Finally, we run an advertising campaign where 10 random ads and their rephrased versions from each of the domains are run in parallel. We show an average improvement in CTR of 68% for the generated ads compared to the original ads.

Our results demonstrate the ability to automatically optimize advertisement for the health domain. We believe that our work offers health authorities an improved ability to help nudge people towards healthier behaviors while saving the time and cost needed to optimize advertising campaigns.

ACM Reference Format:
Brit Youngmann, Elad Yom-Tov, Ran Gilad-Bachrach, and Danny Karmon. 2020. The Automated Copywriter: Algorithmic Rephrasing of Health-Related Advertisements to Improve their Performance. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380211

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380211

1 INTRODUCTION

Spending on search advertising (also known as sponsored ads) in 2019 in the U.S. was valued at US$36.5 billion [6] and US$109.9 worldwide [7]. The justification for these enormous amounts is the high efficiency of such ads, mostly due to the ability to tune the advertisements to explicit user intent, as specified in their queries, rather than on implicit information about their preferences [26]. This, naturally, increases the likelihood of clicks and conversions (product purchases).

In search advertising, ads are shown on a Search Engine Results Page (SERP) whenever a user performs a search. Advertisers bid for specific keywords which they perceive as indicating an interest in their product. When these keywords are searched, their ads can be presented. As these ads are shown only for specific keywords, the displayed advertisement better matches the user's needs. The actual mechanism for matching ads to keywords are similar for most of the popular search engines (e.g., Google, Bing). Generally, search ads are targeted to match key search terms in the user's search query (namely the keywords), which are provided by the advertiser, and additionally, advertisers may express preference for demographics (age, gender), location, and other user parameters. Advertisers compete with other advertisers who chose the same keywords, by submitting bids, representing monetary values for acquiring the ad slots associated with these keywords. The ads to be displayed are chosen by the advertising system using the bids, allowing advertisers to increase their exposure by bidding higher.

Various ad performance measures can be tracked [26], including the number of times an ad was shown (the number of impressions), the percentage of impressions which led to a click (the click-through rate, CTR), and the percentage of clicks which led to a purchase (the conversion rate). These performance measures are reported to the advertiser. The advertiser can optimize their campaign by modifying the ads, bids, or other related campaign parameters. Additionally, the advertiser can usually request that the advertising system optimize the campaign to one of the above-mentioned performance measures.

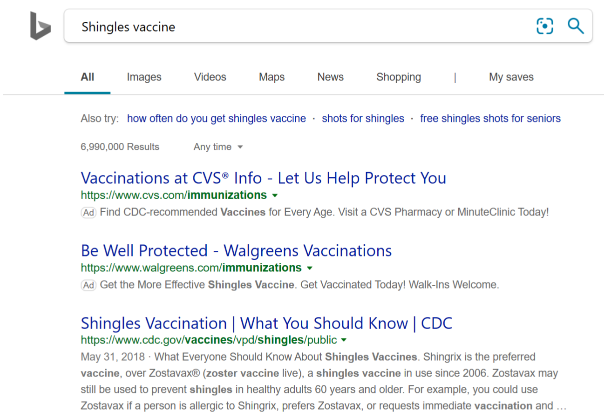


Figure 1: Example Bing SERP with two search ads. The search query is "Shingles vaccine" and the search ads appear as the two top results.

Health authorities, pharmaceutical companies, and other stakeholders in the health domain have recognized that search advertising can assist not only in selling products but (more importantly) in steering people towards healthier behaviors. Ads were shown to be effective in nudging people towards less harmful pro-anorexia websites [57], to quit smoking [58] and towards increased physical activity and better food choices [59]. The conversion optimization mechanism of advertising engines was utilized (in conjunction with advertising campaigns) to improve HPV vaccination rates [44] and to screen for cancer [56].

However, creating effective advertisements, as measured either using the above-mentioned performance measures or in eliciting behavioral change (the two are not synonymous [59]) requires expertise, experience, and testing. The first two are usually provided by advertising agencies, but these are generally proficient in selling products, not in promoting health outcomes. Testing ads is expensive and time-consuming. Thus, a health agency without the expertise or experience required to create an effective advertising campaign may not be able to do so, especially when the campaign needs to be quickly fielded in response to a public health crisis.

Here we offer a possible solution to this problem, based on recent advances in Natural Language Processing and on the vast repositories of ads and their performance, available to Internet advertising systems. We propose an algorithm that can take an (health-related) advertisement, for example, one created by a health agency, and rephrase it so as to maximize a required performance measure. After discussing relevant prior work we describe the building blocks of this algorithm, including ad normalization, preprocessing, and the training of a deep neural network (DNN) sequence-to-sequence model. We then show the performance of the algorithm and our insights into what the network learned in order to optimize the ads. Our experiments show that the proposed method increases the CTR on ads by 68%. Our analysis of the generated ads show that they reflect higher values of psychological attributes associated with user action, including higher valance and arousal, more calls to action, and the amplification of user desires.

2 RELATED WORK

To the best of our knowledge, the proposed algorithm has no directly comparable past work. Instead, our work draws on past literature in several areas, including Machine Translation, advertisement performance prediction, and work in psychology on the effectiveness of emotions in creating compelling advertisements. Here we review relevant work in these three areas.

2.1 Machine Translation

Machine translation (MT) is a sub-field of computational linguistics that investigates the use of a machine to translate text from a source language to a target language, while keeping the meaning and sense of the original text the same.

The MT process can be simplified into three stages: the analysis of source-language text, the transformation from source-language text to target-language text and the target-language generation. Work on MT can be divided into three main approaches: Rule-based MT [45, 46], Statistical MT [15, 38, 53] and Neural MT [21, 47].

In rule-based MT systems (e.g., [24]), a translation knowledge base consists of dictionaries and grammar rules. A main drawback of this approach is the requirement of a very significant human effort to prepare the rules and linguistic resources, such as morphological analyzers, part-of-speech taggers and syntactic parsers.

In statistical MT systems, the translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora [19, 39]. Generally, the more human-translated text is available in a given language, the better the translation quality. A main drawback in statistical MT is that it depends upon huge amounts of parallel texts, and its inability to correct singleton errors made in the source language.

State-of-the-art MT systems use neural networks to predict the likelihood of a sequence of words [20, 35, 51]. As opposed to previous approaches, in neural MT, all parts of the translation model are trained jointly (end-to-end) to maximize performance, requiring minimal domain knowledge. Such systems often use encoder-decoder architectures, encoding a source sentence into a fixed-length vector from which a decoder generates the translation.

In the current work we adopted a simple neural MT model to translate advertisements to ones attracting more users to click on the advertisements. We note that recent work has proposed an attentional mechanism to improve translation performance, by selectively focusing on parts of the source sentence during translation. [10, 43]. However, as we show, even our simple neural MT model achieves significant improvement in terms of click-through rates (See Section Results). We note that a further improvement may be achieved by using an attentional-based model. Nonetheless, as the translation task is only one of the black-box components of our framework and not a part of our contributions, we leave this direction for future research.

2.2 Click-through rate prediction

As mentioned in the Introduction, popular search engines (such as Google and Bing) use keyword auctions to select advertisements to be shown in the display space allocated alongside search results. Auctions are most commonly based on a pay-per-click model where advertisers are charged only if their advertisements are clicked by users. For such a mechanism to function efficiently, it is necessary for the search engine to estimate the click-through rate (CTR) of ads for a given search query, to determine the optimal allocation of display space and payments [28]. As a consequence, the task of CTR prediction has been extensively studied [30, 34, 55, 60], since it impacts user experience, profitability of advertising, and search engine revenue.

CTR prediction is based on a combination of campaign and advertiser attributes, temporal information, and, especially, the keywords used for advertising. While the former are dense, the latter are sparse, and hence need care in their representation.

Bengio et al. [12] suggested learning a model based on a distributed representation for possible keywords, aiming to avoid the curse of dimensionality in language modeling. More recently, the authors of [25, 30] proposed networks with one hidden layer, which first employ an embedding layer, then impose custom transformation functions for target fitting, aiming to capture the combination relations among features. Other works (e.g., [18, 23, 49]) suggested replacing the transformation functions with a complex Multilayer Perceptron (MLP) network, which greatly enhances the model capability. Generally, these methods follow a similar model structure with combination of embedding layer (for learning the dense representation of sparse features) and MLP (for learning the combination relations of features automatically).

Predicting exact CTR values of ads is beyond the scope of the current work. As we explain in Section 3.4, here we focus on generating ads with higher CTR values than those of their original ads, hence we use a simple ranking model for this task. However, our model is based on insights from the above-mentioned prior work.

2.3 The effect of emotion in advertising

A widely-accepted framework proposes that affective (emotional) experiences can be characterized by two main axes: arousal and valence [13, 22, 31, 36, 41] . The dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from calming or soothing on one end of the spectrum to exciting or agitating on the other. Figure 2 shows an illustration of this, where different affective states are located in the space spanned by these axes [14].

Professionals of advertising aim to create ads which capture consumers' attention with the aim of increasing advertising effectiveness. It has been shown that the inclusion of high arousal and valence sequences in ads increases user attention and interest [11, 41, 42].

Arousal is related to body activation level in reaction to external stimuli [27]. Arousal has been related to simple processes such as awareness and attention, but also to more complex tasks such as information retention [32]. Previous work suggests that arousal modulates ad effectiveness and memory decoding [33]. However, highly arousing contexts can distract individuals from ad processing, making recall more difficult, thus reducing the ability to encode ad content [50].

In comparison to the considerable number of studies investigating the effect of arousal on memory and attention, relatively few studies have examined the effect of valence. The few studies which have examined its effect suggest that valence is sufficient

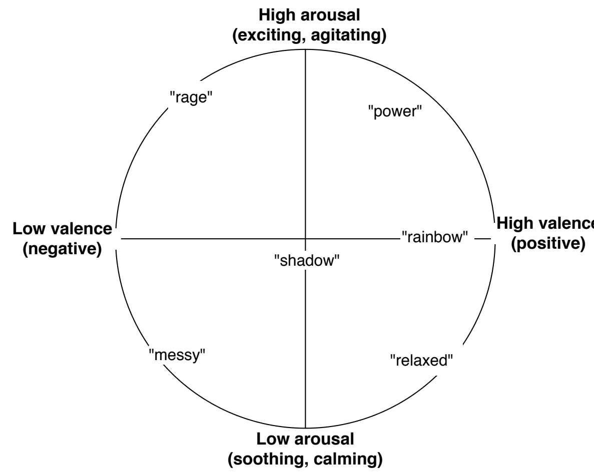


Figure 2: Different English words mapped on the arousal-valence space. (Words were placed according to [14]).

to increase memory performance. Namely, non-arousing ads with very positive or very negative valence are better remembered than neutral ones [36, 37].

Ads which refer to the desires of users, especially those which take the form of special textual content, can affect CTRs in sponsored search [52]. These desires can be mined from ad themselves using *thought-based* effects and *feeling-based* effects. Thought-based effects are primarily related to win/loss analysis (e.g., trade-off between price and quality), while feeling-based effects are more subjective, and include, e.g., brand loyalty and luxury seeking.

In our analysis, we examine both the arousal/valence emotions in the original ads compared to those in the generated ads, as well as the thought-based and feeling-based effects of the original and generated ads.

3 METHODS

We address the problem of rewriting better advertisements as one of a translation from one form of English to another form, where the latter form is more likely to elicit higher response than that of the former. The proposed ad translation pipeline consists of: (i) Normalization (ii) Preprocessing (iii) Candidate generation (iv) Selection. The pipeline receives as input ads in their original form and generates new ads which are expected to be optimized to achieve higher performance. In the following we describe each of these blocks in detail.

We refer to the ads in our training set which were created by advertisers as "original" ads and to ads modified (rephrased) by the translator model as "generated" ads.

3.1 Normalization

In order to allow for generalization we created a model to identify medical conditions and proposed treatments and replace them with generic tokens. This is achieved using a custom Named Entity Recognition (NER) model based on the SpaCy library [5]. Every mention of an entity that corresponds to a medical condition or to a treatment was replaced by the generic token <CONDITION/TREATMENT>.

Specifically, the SpaCy library provides a default NER model which can recognize a wide range of named or numerical entities, including persons, organizations, languages, events, etc. Apart from these default entities, SpaCy also allows the addition of arbitrary classes to the NER model, by training it to recognize new classes.

The training data for the custom NER consists of sentences, target word(s) in each sentence, and its label. SpaCy also supports the case where a sentence contains more than one entity. For example, for the ad displayed in the first row of Table 1, the entities we wish to recognize are both "Shingles vaccine" and "shingles" (replacing them with the token <CONDITION/TREATMENT>).

The NER was trained by manually labeling 300 sentences from each domain (see Section 4.2 for a definition of the domains), splitting each set for training (225 sentences) and test (75 sentences). The trained NER model successfully recognized the entities in 92% and 94% of the test cases, in each of the two domains.

3.2 Preprocessing

We used lemmatization and stemming to reduce the vocabulary size [40]. These two standard techniques transformed inflectional forms and derivationally-related forms of a word to a common base form. We also replaced entity names with their types as follows:

- A mention of a person name was replaced with the token <PERSON>.
- A mention of a geographic location, such as "U.S." was replaced with <GPE>.
- Organization names (e.g., "Herbalife") were replaced with <ORG>.
- Absolute or relative time units, such as "24/7", were replaced with <DATE>.
- Monetary values, such as "5$" were replaced with <MONEY>.
- Numbers were replaced with the token <CARDINAL>.

We used the SpaCy library [5] to perform this task. Examples of this process are shown in Table 1.

3.3 Candidate generation

We trained a sequence-to-sequence translator model that learns how to transcribe an input text to an output text, such that the latter corresponds to an ad with a higher CTR value than the input (see Section 4.1).

3.3.1 Training data. The training data was constructed as follows: For every search query q we extracted all pairs of ads $a_{\text{low}}, a_{\text{high}}$ that were presented on the SERP such that a_{low} generated a lower CTR than a_{high} for the same q. For every such pair, we generated an example in the training data where a_{low} is the source text and a_{high} is the target text. Note that this process assumes that ads displayed in response to the same query are likely to promote similar products or services.

Since we tested our methods in two domains (see Section 4.2), we trained a model separately for each of these domains. Separate models were used because of the observation that the vocabulary of ads within a domain was more similar than between domains. For example, over 35% of ads displayed for the first domain contain one of the words "remedy", "symptom", "treatment", or "pain", while over 27% of the ads of the other domain contain one of the words "help", "advice", "control", or "tip". We also examined the results while training a single model for both domains, however, the results were inferior (see Section 4.1 for an explanation on how quality was estimated).

3.3.2 Model architecture. We employed a simple sequence-to-sequence (seq2seq) model using the PyTorch library [3]. Data was first tokenized. Then, we built the vocabularies for the source and target "languages". Note that even though both the source and the target sentences are in English, they include different sets of sentences, and hence the frequencies of tokens is different.

The model contains three parts: The encoder, the decoder and a seq2seq model that encapsulates the encoder and decoder.

For both the encoder and the decoder we used a 2-layer Long Short-Term Memory (LSTM) model. The encoder takes as input text and produces a context vector, while the decoder takes the context vector and produces one word at a time. The complete seq2seq model receives the input text, uses the encoder to produce the context vector, then uses the decoder to produce an output text.

The optimizer, which is used to update the parameters in the training loop, was set to be the Adam optimizer, and the loss function was set to be the cross-entropy-loss function, which calculates both the log softmax as well as the negative log-likelihood of the predictions.

3.4 Selection

The translator model can generate multiple candidates, out of which we would like to select the best translation by ranking them. Unfortunately, while for the input ads the CTR values are known, for the output ads these values are unknown. Therefore, here, given an original ad, we consider the first ad generated by the translator as the candidate ad. An interesting direction for future research would be to compare between multiple generated ads and to learn how to select the best candidate. Nonetheless, as we show in Section 5, in the vast majority of cases, the generated ads have succeeded to better attract users interest than their corresponding original ads.

We note that the proposed pipeline is not entirely automated, and the generated ads contain tokens that should be substituted by the advertiser (see, for example Table 1). However, as mentioned in the Introduction, our goal is to assist health agencies, not completely replace them. Thus, minor grammatical errors in the generated ads may be manually corrected (e.g., replacing words from their lemma form to the required tense), or revised using prior work on the automatic transformation of text to proper English (e.g., [4]). General tokens (such as CARDINAL and CONDITION/TREATMENT), if they did not appear in the original ad (e.g., Table 1 example 1), may be replaced with the most common original corresponding tokens in the dataset. For example, the most common number for the token CARDINAL is 10. Another advantage of the semi-automated pipeline is that advertisers can ensure that the semantic meaning of the ads are maintained in the generated ads.

4 EXPERIMENTS

Recall that the goal of the proposed system is to optimize a search ad by rephrasing it so as to increase users' interest, while keeping

Origin ad	After preprocessing	Generated Ad
Singling Out Shingles Vaccine - 13 Health Facts. Check out 13 health facts about shingles on ActiveBeat right now.	single out <CONDITION/TREATMENT> - health fact. check out <CARDINAL> health fact about <CONDITION/TREATMENT> on <ORG> right now.	<CONDITION/TREATMENT> - everything you need to know. discover <CARDINAL> fact on <CONDITION/TREATMENT>. get expert advice now!
Best Remedy For Cough - Updated 24/7. Search for best remedy for cough. Browse it Now!	good remedy for <CONDITION/TREATMENT> - update <DATE>. search for good remedy for <CONDITION/TREATMENT>. browse it now!	home remedy for <CONDITION/TREATMENT> - see top <CONDITION/TREATMENT> home remedy. try this <CARDINAL> effective <CONDITION/TREATMENT> remedy that can help you. read more here!
What Does Dark Urine Mean? - Causes Of Dark Urine - Visit Facty, Stay Healthy. See Causes of Dark Urine Color. Learn About What Causes Different Colors Of Urine.	what do <CONDITION/TREATMENT> mean? - cause of <CONDITION/TREATMENT> - visit <ORG>, stay healthy. see cause of <CONDITION/TREATMENT>. learn about what cause different color of <CONDITION/TREATMENT>.	<CONDITION/TREATMENT> - sign to never ignore. <CARDINAL> common <CONDITION/TREATMENT> cause. understand how to avoid <CONDITION/TREATMENT> and stay healthy.

Table 1: Example (original, preprocessed and generated) ads.

the meaning and sense of the original ad the same. Toward that end, we next present the experimental setup used in our work.

4.1 Offline and online metrics

User interest can be measured in a variety of ways, including clicks, conversions, and future behaviors. In online advertising, the click-through rate (CTR) is the percentage of users viewing a SERP who click on a specific ad that appears on that page. CTR measures how successful an ad has been in capturing users' interest. The higher the click-through rate, the more successful the ad has been in generating interest. Moreover, clicks are the most commonly-used measure of ad performance.

While the goal of many health campaigns is to induce long-term behavioral change, optimizing for this goal introduces challenges such as a delayed and sparse rewards. Therefore, in this work we used CTR as the main success metric.

4.2 Data

We extracted English-language search advertisements displayed by customers of Microsoft Advertising between January 1st, 2019 and March 31st, 2019, which were shown at least 100 times. Data is comprised of over 114K advertisements, displayed to users who queried in two domains: Medical Symptoms (MS) and Preventive Healthcare (PH) (see Table 2).

Domain	# of advertisements	# of search queries	Example keywords
MS	46061	2788	nasal, vertigo, fatigue, earache, cough.
PH	68021	6091	weight loss, stop smoking, vaccine, safe sex.

Table 2: Extracted Data.

To identify ads relevant to these domains, we considered search queries that contain specific (pre-defined) keywords. Keywords for the MS domain comprised of common medical symptoms (according to Wikipedia [54]). Keywords for the PH domain were extracted from a US government website [1]. In the MS domain, for example, we considered all queries containing the word "vertigo" (See Table 2 for more examples). We excluded search queries with fewer than 5 search ads during the data period. The extracted search ads were displayed for over 8K unique search queries. On average, 14 different search ads were displayed for each search query.

Search ads in Microsoft Advertising include the following elements: (1) A title, which typically incorporates relevant, attention-grabbing keywords; (2) A URL, to provide users with an idea of where they will be taken once they click on the ad; (3) A descriptive text that highlights the most important details about the product or service, and is used to persuade users to click on the ad. See, for example, the two search ads displayed for the search query "Shingles vaccine", shown in Figure 1. In our work we only considered ads' textual components, i.e., the title and the description, concatenated to form a single paragraph. Both the title and description fields are limited to a maximum character length. Additionally, both fields have several sub-fields, that is, there are up to 3 title fields and 2 description fields, where text needs to be entered into at least one of the title and one of the description fields.

4.3 Offline evaluation

Predicting the exact CTR value of a given ad is challenging (as mentioned in Section 2). However, in our setting, it is sufficient to be able to predict the relative order of CTRs for ads on a given query. This is so that the model can predict whether a generated ad is more likely to achieve a higher CTR value than its original version. To this end, we employed a learning-to-rank model (LambdaMART [16] as implemented by the pyltr library [2]).

The training data consist of the text of the ads. To quantify the text we represented it in features such as the sentiment, lexical diversity and the readability index of the text. These features were chosen so as to quantify the ability of users to understand and decode the content of ads. Previous work has shown the importance of writing a simple, concise, consistent and easily understood text to increase advertising effectiveness [17, 29]. For example, we consider the readability ease of the ads, as measured by the Flesch-Kincaid readability score. Other features such as token counts and word embedding were considered, but were found not to improve ranker performance and were thus excluded in the following analysis. Table 3 provides the full list of features.

The training data consists of lists of extracted ads, along with a partial order between ads in each list. The latter is induced by their CTR values. Each list corresponds to ads shown in response to a different search query.

The CTR ranker model is used to quantify the effectiveness of the translation model in the following manner: For every original ad and its corresponding generated ad we examine how our trained model ranked the two ads. If, for a given original ad, the generated ad was ranked higher by the ranker, we count this as a positive example (the generated ad is better than the original). Otherwise, it is a negative one.

4.4 Online evaluation

Ten random ads from the MS domain and 10 from the PH domain were selected, for which the ranker estimated that a generated ad had a higher rank than the original ad. We chose to select ads which were estimated to have higher rank because this mimics more closely practical scenarios, where new ads are likely to be used only if they are predicted to have superior performance to that of existing ads.

The text of none of the generated ads appeared in the training data. We tested the performance of both the original advertisement and generated ad in a real-world setting. Note that ads were formatted to fit text length requirements of the ads system (maximum length of the Title and Description fields) by fitting the first sentence of the generated ad into the title field and if it were too long, moving any excess text into the secondary or tertiary title field. Similarly, the Description field was set. Any minor grammar errors in the generated ads were manually corrected.

The ads were run as a new advertising campaign (without any history that could bias the experiment) on the Microsoft Advertising network. Each pair of ads were placed into the same ad group. The campaign was set for automatic bidding with the goal of maximizing CTR, with a maximum bid of US$1 per click. The ads were shown until at least 20 impressions per ad group were obtained.

4.5 Emotion analysis

We examined three main forms of emotional affect in ads, as described in Section 5.3, namely, the Call-to-Action, arousal and valence, and thought- and feeling-based effects. For each of these affects we measured their value in original and generated ads, and showed the change in their values in each of these ads.

The Call-to-Action (CTA) of an advertisement is that part of the ad which encourages users to do something [48]. CTAs can drive a variety of different actions depending on the content's goal. CTAs are essential in directing a user to the next step of the sales funnel or process. Previous work has shown that an ad used just to support a product/service, without a CTA, might be less effective [48]. An ad may contain more than one CTA. For example, in the ad: "Dry Cough relief - More information provided. Browse available information about dry cough relief. Check here." The CTA are "browse available information" and "check here". Here we focus on the verbs of the CTAs (in this example, "browse" and "check"), as they are easier to automatically identity (using, e.g., a part-of-speech tagging model).

To classify the emotional attributes of the ad we focused on two concepts which have been shown useful to measure emotional experiences [36, 41]: **Arousal** and **Valence**. Arousal refers to the intensity of an emotion (how calming or exciting it is) and valence deals with the positive or negative character of the emotion. An ad with positive connotations (such as joy, love, or pride) is said to have high valence. Negative connotations (including death, anger, and violence) have low valence. Similarly, the more exciting, inspiring, or infuriating an ad is, the higher the arousal. Information that is soothing or calming produces low arousal.

To quantify CTA, arousal and valence we first created a dataset of ads labeled for their arousal and valence levels, and well as marking the CTA verbs therein. Then, a model for each was created using this dataset, and applied to a larger set of ads.

The dataset was created from a random sample of 265 advertisements, comprising of generated and original ads, as follows: 83 original MS-related ads, 82 generated MS-related ads, 50 original PH-related ads, and 50 generated PH-related ads.

The values of valence, arousal and the CTA verbs for the dataset were found by asking 5 crowdsourced workers from Amazon Mechanical Turk (MTurk, https://www.mturk.com/) to label the 265 ads. Specifically, workers were asked to mark CTA verbs and to estimate (separately) the arousal and valence scores of each ad on a 5-point scale in the range of $[-2, 2]$, where -2 is the lowest arousal/valence score and 2 is the highest score. A score of 0 suggests that the ad is neutral with respect to arousal/valence experience.

Using the training set we created three models: one to identify CTA verbs, and another two to estimate valence and arousal in non-labeled ads.

CTA verbs in other (non-labeled) ads were identified by implementing a custom Part-Of-Speech (PoS) tagger using the SpaCy library [5]. The trained model was applied to non-labeled ads, to tag all words. Every word that was tagged as CTA was counted as a CTA verb.

As for the valence and arousal scores, two models were trained: One to predict the average arousal score reported by MTurk workers and the other to predict the average valence score. The features of these models were set to be the tokenized text of the ads, using the ML.NET tool (https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet).

All models were constructed using the Microsoft ML.NET machine learning tool. We examined multiple models, including linear regression (with stochastic gradient decent training), boosted trees and random forest. The best results were achieved with the Boosted Trees and Random Forest regression models, for the arousal and

Feature	Explanation	Extractor
Flesch-Kincaid readability ease	Indicating how difficult a sentence in English is to understand.	Textstat [8]
Flesch-Kincaid readability grade	Indicating the number of years of education generally required to understand the text.	Textstat [8]
# of “difficult" words	According to the Textstat library [8].	Textstat [8]
Readability consensus based upon the Dale-Chall Readability Score, the Linsear Write Formula, and the The Coleman-Liau Index	Indicating the estimated school grade level required to understand the text.	Textstat [8]
Vader-Sentiment	Indicating how positive/negative is the sentiment according to the Vader measure.	VaderSentiment [9]
Lexical diversity	Number of distinct words divided by the number of words in the text.	
# of punctuation marks		SpaCy [5]
# of noun phrases		SpaCy [5]
# of adjectives		SpaCy [5]

Table 3: Features extracted from text ads for the CTR ranker model.

valence scores, respectively. In the Boosted Trees model the number of trees was set to 100 and the learning rate was set to 0.2. In the Random Forest regression model the number of trees was set to 500. The performance of the models on training data was evaluated using 5-fold cross-validation.

To examine the thought-based and feeling-based effects of the ads, we considered user desires under both effects, as proposed in Wang et al. [52]. We used the phrases proposed in that paper, which were extracted from general ads by human experts, adapting them to those which we found appeared frequently in ads related to the MS and PH domains. Table 4 lists these effects, their associated user desires, and the keywords that were used to identify these effects in the ads. Wang et al. [52] used a novel algorithm to mine these user desire patterns. Here we used predefined keywords to conclude if a given ad encompasses one of the examined user's desires.

In Section 5.3 we discuss how the proposed translator model has learned to take advantage of these user desires, to increase the likelihood a user will click on an ad (i.e., to increase users' interest).

5 RESULTS

In this section we provide results of our efforts to validate the effectiveness of the proposed model. We also aim to explain what the proposed translator model has learned.

5.1 Offline estimation of result quality

Overall, the CTR ranker predicts that the generated ads will have a higher CTR in 81% of MS-related ads and 76% PH-related ads.

We note, however, that the ranker cannot be considered a perfect proxy for experimentation, since it does not precisely predict the ordering of ad pairs by their performance. To estimate the likelihood of ranker error (where it will rank higher the ad with the lower CTR in a pair) we report Kendall's Tau (KT) rank correlation of the CTR ranker. This measure was chosen as it considers the number of concordant and discordant pairs of the prediction, compared to true performance. The CTR ranker model was evaluated using 5-fold cross validation. The average KT across folds was 0.47 ($P < 0.01$ in the MS domain and 0.45 ($P < 0.01$) in the PH domain. This should be compared to KT equalling 0.38 and 0.35 for the two domains, respectively, when randomly ordering the ads.

Because the ranker is imperfect, some of the generated ads which are predicted by the ranker to be less effective than their corresponding original ads may achieve higher CTR values in practice, and vice versa.

5.2 Online estimation of result quality

Ten ads from each ad group were run on the Microsoft Advertising platform until they were shown at least 20 times each. One of the PH ads failed to receive any impressions and was removed from the analysis. The average number of times that each ad was shown was 984 times, and the maximum 8824.

Figure 3 shows the CTR of the generated ads versus those of the original ads, for both domains. In 15 of 19 cases generated ads had a higher CTR than the original ads. On average, the generated ads had a CTR which was 68.2% higher than that of the original ads (statistically significant, signrank, $P = 0.011$).

The average per-advertisement improvement in CTR (micro-averaged improvement) was 25.4%. Note, however, that this average refers to only 15 ads, since, as Figure 3 shows, 4 of the ads had a non-zero CTR for the generated version, but zero for the original version.

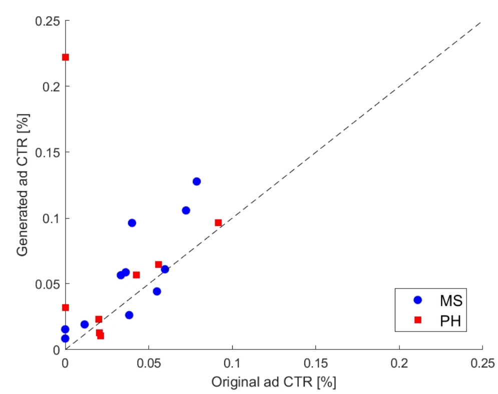


Figure 3: CTR of the generated ads versus the original ads. Each dot represents one advertisement pair (one original versus one generated). The diagonal line shows equal CTR.

Effect	User's Desire	Examined key words	Example ad
Thought-based	Petty advantage	discount, deal, coupon, $x\%$	"Science diet coupons - Up to 60% Off Now. Christmas Sales! Compare..."
Thought-based	Extra convenience	delivery, payment, shipping	"Unbearable Smokeless Coals - Great Range, Fast Delivery..."
Feeling-based	Trustworthy	official, guarantee, return	"Jenny Craig Official Site - A Proven Plan For Weigh Loss..."
Feeling-based	Luxury seeking	top, most, best, good	" Best Remedy For Cough - Updated 24/7..."

Table 4: Examined user desires under thought-based and feeling-based effects.

5.3 Emotion analysis results

To obtain an understanding for what the proposed translator model learned, we examined the most frequently chosen Call-To-Actions (CTAs) and the emotional attributes of the generated ads.

5.3.1 CTAs analysis. Previous work has shown that users tend to click on ads that can convince them to take further actions, and the critical factor is if those ads can trigger users' desires [52]. Furthermore, Rettie et al. [48] demonstrated a relationship between the level of relevance and action taken, such that when users found ads relevant they were significantly more likely to take action (i.e., to click on the ads). Thus, an ad used solely to inform of a product or service, without containing a CTA, might be less effective.

We used the custom PoS tagger (see Section 4.5) to identify CTA verbs in both original and generated ads. The performance of the model was evaluated using 5-fold cross validation on the labeled ads (i.e., the ads that were labeled by the crowd workers), and we report the average score across all runs. On average, the PoS tagger correctly tagged 93% of the CTA verbs.

Analysis reveals that 72% of the generated ads include at least one CTA verb, compared with only 65% of the original ads including at least one CTA verb (statistically significant, chi-square test, $P < 0.05$). The most frequently used CTA verbs in the original and in the generated ads are shown in Table 5. We note that in the MS domain, where users are seeking information on a medical symptom, a desired result contains information on the symptom and ways to treat it. Examining the table, the two most common CTA verbs in the generated ads in the MS domain are "learn" and "treat". Namely, the translator has learned that these CTAs are very effective in triggering users' desires, and therefore the use of them leads to an increase of the CTR. Similarly, in the PH domain, where, for example, users seeking information on ways to quit smoking or to lose weight, the translator often used CTAs which are relevant to this goal (e.g., "quit", "stop", "lose").

Domain	Original ads	Generated ads
MS	read, learn, treat, browse, check.	learn, treat, find, discover, get.
PH	get, learn, visit, read, quit.	quit, get, stop, learn, lose.

Table 5: Top 5 CTA verbs in the original and generated ads.

5.3.2 Arousal and valence analysis. To asses how well the crowd workers agree among themselves in the arousal and valence of ads, we computed the average (valence/arousal) score for each advertisement and compare the standard deviation of the scores around this average value. These are compared with the standard deviation of random scores. A lower standard deviation for crowd worker scores (compared to random) indicates greater agreement among workers.

The standard deviation of the arousal scores in the MS domain was 0.3 (compared to 0.8 for random). In the PH domain the standard deviation was 0.2 (compared to 0.7 for random). The standard deviation of the valence scores in the MS domain was 0.1 (0.7 for random scores) and in the PH domain it was 0.3 (compared to 0.8 for random). Thus, workers were able to agree on the valence and arousal scores of ads to a very high degree.

Using 5-fold cross-validation on the crowdsourced-labeled data, the models predicted arousal and valence from ad text with an accuracy of $R^2 = 0.62$ and 0.42, respectively.

The trained models were than applied to predict the arousal and valence scores of all original and generated ads.

The averaged predicted arousal and valence scores of the original and of the generated ads from both domains (MS and PH) are depicted in Figure 4. In the Figure, the vertical axis represents the averaged predicted arousal and valence scores. Note that even though these values are measured on a scale of $[-2, 2]$, most of the predicted scores were positive, and so are the average scores.

Interestingly, observe that the predicted arousal and valence scores of the generated ads in both domains are higher than the predicted scores of the original ads. This implies that the proposed translator model is predicted to increase the arousal and valence of the input ads (i.e., the original ads).

Figure 5 shows the predicted valence and arousal scores of 1000 random original ads from each domain (MS in red, PH in orange) versus those of the generated ads. The diagonal line indicates equal values for the original and generated ads. Points on or below the grey line denote ads for which the generated ads had equal or lower valence or arousal scores, compared to those of the original ads. As the figures show, for both arousal and valence the vast majority of points (i.e., ads) are above the grey lines. Thus, in most cases, the generated ads are predicted to have higher arousal and valence scores than their corresponding original ads, according to our models. For example, in more than 700 of 1000 of the MS-related original ads, the predicted valance scores of the corresponding generated ads were higher (Figure 5(b), red points).

Figure 6 depicts the predicted valence scores versus the predicted arousal scores of (the same) 1000 random original ads (in red) and their corresponding generated ads (in orange), from each domain

Attribute	MS	PH
Arousal	0.73	0.71
Valence (absolute)	0.61	0.64

Table 6: Pearson correlation between the CTR of ads and the predicted arousal score and the absolute value of the valence score, for ads in both domains. Results are statistically significant, $P < 0.01$ for all combinations.

(MS and PH). Ellipses denote one standard deviation around the mean. As can be seen, on average the generated ads (orange points) are predicted to have higher arousal and higher valence, compared to the original ads (red points). That is, the translator model increases both the arousal and the valence scores of the generated ads, compared with their original versions.

Last, we examine the correlation between users' interest (measured, in our setting, by the CTR values) and the predicted values of the arousal and valence. Specifically, we report the Pearson correlation between CTR and these values. We note that, as mentioned in Section 2, as opposed to arousal, where it has been shown that ads with high arousal increase user attention, ads with very positive (i.e., a score close to 2) or very negative (i.e., a score close to -2) valence are similarly likely to increase users' interest, compared to ads with neutral valence scores (i.e., a score close to 0). Therefore here we consider the absolute valence scores. The results are given in Table 6. As can be seen, we observe a moderate positive relationship between CTR and the arousal and (absolute) valence scores. Namely, as the arousal increases it is more likely that users would click on the ads. Similarly, as the absolute valence score increases so is the chance a user would click on the ad.

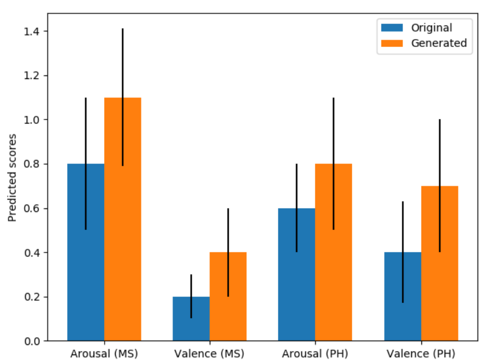
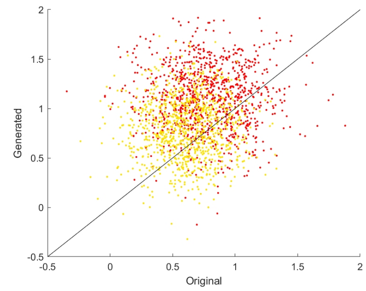
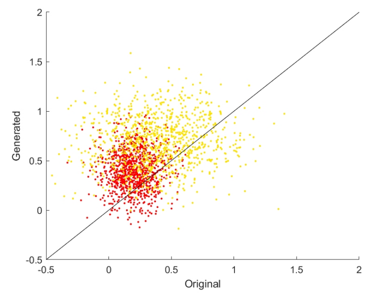


Figure 4: Averaged predicted arousal and valence scores. Vertical lines show one standard deviation of the scores. Differences for each of the 4 values is statistically significant (paired signtest, $P < 10^{-5}$)

5.3.3 Analysis of thought- and feeling-based effects. As mentioned in Section 2, the inclusion of specific textual content referring to user desires increases user interest, and consequently CTR [52]. To investigate if these user desires are associated with an increased CTRs in the MS and PH domains, we computed the CTR of ads containing the keywords associated with each desire and compared them with the average CTR of all ads that were displayed to the same search query. The results are shown in Table 7. Indeed, one can see that the likelihood a user will click on an ad increases if it contains one of the keywords mentioned in Table 4.



(a) Arousal



(b) Valence

Figure 5: Predicted valence/arousal scores of 1k random original ads from each domain (MS in red, PH in orange) vs. those of the generated ads. The diagonal line indicates equal values.

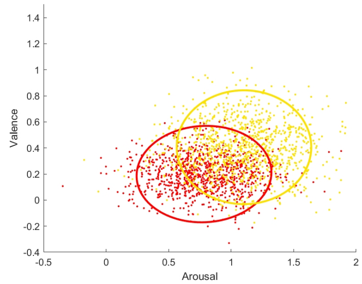
Effect	Percentage of matched ads	Change in CTR
Petty advantage	7.2	48.5%
Extra convenience	6.8	35.7%
Trustworthy	2.8	28.1%
Luxury seeking	4.9	33.2%

Table 7: Change in CTR between ads matched with the patterns of each effect and other ads that were displayed in response to the same query.

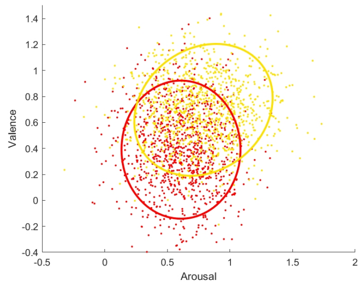
Additionally, we examined the percentage of original and generated ads containing at least one of the keywords associated with each of the examined user desires, as listed in Table 4.

The results are shown in Figure 7. Here, the vertical axis represents the percentage of ads containing at least one of the keywords associated with each of the examined user desires. Observe that in all cases, the generated ads include more such keywords, indicating that the translator has learned that incorporating specific textual content increases user interest.

Thus, our results support the empirical results of Wang et al. [52], showing that users tend to click more on ads containing specific patterns and keywords.



(a) MS



(b) PH

Figure 6: Predicted valence versus arousal scores of 1k random original ads (red) and generated ads (orange) from each domain (MS and PH). Ellipses denote one standard deviation of the mean.

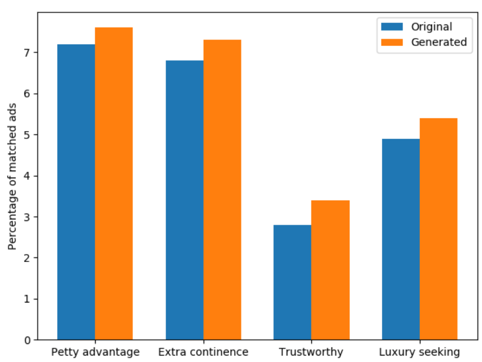


Figure 7: Percentage of original and generated ads containing at least one of the keywords associated with each of the examined user's desires. Differences for each of the 4 values is statistically significant (paired signtest, $P < 10^{-2}$).

6 DISCUSSION

In this study we presented a system which translates a given health-related advertisement it into an optimized ad that is more likely to be clicked by users. An immediate application of this work is to offer such a service to public health authorities, which may improve population health by more effectively eliciting positive behavioral change.

Beyond training of the translation model, our work requires several supporting models for evaluating the resulting ads. These include a ranking model to estimate improvement in CTR, and models for assessing psychological attributes of original and generated ads. Moreover, we tested the translation model in a real-world scenario by running 20 advertisement pairs in a head-to-head competition, and found a 68% performance improvement in the generated ads, compared to the original ones.

To investigate what our model has learned, we examined 3 factors of the original ads compared to those of the generated ads: (1) The use of Calls-to-Action (CTAs), which have been shown to be essential in increasing the ads effectiveness [48]; (2) The estimated arousal and valence scores of the ads, where previous work has shown that the inclusion of high arousal and valence sequences in ads increases user attention and interest [11]; and (3) The inclusion of special textual content that refer to the desires of users [52].

Our empirical results indicate that the translation model improved all 3 factors. Specifically, the generated ads include more CTAs than the original ads, they are predicted to have higher arousal and valence emotions, and they combine more keywords which have been shown related to user desires. Thus, the translation model has, without explicit guidance, learned which psychological factors should be enhanced in ads so as to elicit higher CTR.

Our work enables advertisers, especially in the health domain, to create advertising campaigns without advertising expertise. This new area of work includes several future directions for research.

First, here we focused on improving CTR. As was discussed in the Introduction, other measures of ad performance, including conversion and future behavior, may be more useful for health authorities. Training a translation model to improve these will likely require training on sparser data, but the resulting model may be of better use to health authorities, and comparing the psychological attributes it affects will be of interest to marketing experts.

Another interesting direction for future research will be to generate ads directly from product or service web pages, to assist health authorities in writing effective ads without using (possibly expensive) experts and experiments. Additionally, future work will apply our algorithm to more domains, such as wellness and education and will build a more general model, that is, one which is suitable for any health-related ad, not to one of a specific domain therein.

On the technical side, we assume that performance may be improved if a more complex translator model was used (see discussion in the Related Work), and if different outputs of the translator model were examined for the same input ad, tuning the model to achieve better results. Lastly, recall that the translator model receives as input ads in their basic form (i.e., after pre-processing, see Table 1), and therefore the generated ads are also in this form. Future work will improve the automatic transformation of the ads to proper English (e.g., realizationing words from lemma form using [4]).

Additionally, it may be possible to enable advertisers to manually tune the preferred psychological values of the output, by intervening within the intermediate layers of the translation model.

Finally, here we focused on textual ads. Much of the advertising industry uses imagery and audio. It will be interesting to try and improve those through translation models.

REFERENCES

[1] 2019. Preventive Healthcare. https://www.healthcare.gov/preventive-care-adults/.
[2] 2019. pyltr: Python learning to rank (LTR) toolkit. https://pypi.org/project/pyltr/.
[3] 2019. PyTorch. https://pytorch.org/.
[4] 2019. SimpleNLG. https://github.com/simplenlg/simplenlg.
[5] 2019. spaCy: Industrial-Strength Natural Language Processing. https://spacy.io/.
[6] 2019. Statista: Global No.1 Business Data Platform. https://www.statista.com/outlook/219/109/search-advertising/united-states#market-revenue.
[7] 2019. Statista: Global No.1 Business Data Platform. https://www.statista.com/outlook/219/100/search-advertising/worldwide#market-revenue.
[8] 2019. Textstat: package to calculate statistics from text. https://pypi.org/project/textstat/.
[9] 2019. VADER-Sentiment-Analysis. https://pypi.org/project/vaderSentiment/.
[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[11] D Belanche, C Flavián, and A Pérez-Rueda. 2014. The influence of arousal on advertising effectiveness. *Measuring Behavior 2014* (2014).
[12] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* (2003).
[13] Margaret M Bradley, Mark K Greenwald, Margaret C Petry, and Peter J Lang. 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition* (1992).
[14] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report. Citeseer.
[15] Peter Brown, John Cocke, S Della Pietra, V Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Roossin. 1988. A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics.* Association for Computational Linguistics.
[16] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* (2010), 81.
[17] Jean-Charles Chebat, Claire Gelinas-Chebat, Sabrina Hombourger, and Arch G Woodside. 2003. Testing consumers' motivation and linguistic ability as moderators of advertising readability. *Psychology & Marketing* (2003).
[18] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems.* ACM.
[19] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics.
[20] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
[21] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[22] Martin A Conway, Stephen J Anderson, Steen F Larsen, Carol M Donnelly, Mark A McDaniel, Allistair GR McClelland, Richard E Rawles, and Robert H Logie. 1994. The formation of flashbulb memories. *Memory & Cognition* (1994).
[23] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems.* ACM.
[24] Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation* (2011).
[25] Kun Gai, Xiaoqiang Zhu, Han Li, Kai Liu, and Zhe Wang. 2017. Learning piece-wise linear models from large scale data for ad click prediction. *arXiv preprint arXiv:1704.05194* (2017).
[26] Hector Garcia-Molina, Georgia Koutrika, and Aditya Parameswaran. 2011. Information seeking: convergence of search, recommendations and advertising. *Commun. ACM* (2011).
[27] Daniel Gould and Vikki Krane. 1992. The arousal–athletic performance relationship: Current status and future directions. (1992).
[28] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. Omnipress.
[29] Perbinder Grewal and Swethan Alagaratnam. 2013. The quality and readability of colorectal cancer information on the internet. *International Journal of Surgery* (2013).
[30] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
[31] Stephan Hamann. 2001. Cognitive and neural mechanisms of emotional memory. *Trends in cognitive sciences* (2001).
[32] Morris B Holbrook and Elizabeth C Hirschman. 1982. The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of consumer research* (1982).
[33] Eui Jun Jeong and Frank A Biocca. 2012. Are there optimal levels of arousal to memory? Effects of arousal, centrality, and familiarity on brand memory in video games. *Computers in human behavior* (2012).
[34] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems.* ACM.
[35] Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.*
[36] Elizabeth A Kensinger. 2004. Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences* (2004).
[37] Elizabeth A Kensinger and Suzanne Corkin. 2003. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & cognition* (2003).
[38] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion.*
[39] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Association for Computational Linguistics.
[40] Tuomo Korenius, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. 2004. Stemming and Lemmatization in the Clustering of Finnish Text Documents. In *CIKM.* ACM.
[41] Annie Lang, Kulijinder Dhillon, and Qingwen Dong. 1995. The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcasting & Electronic Media* (1995).
[42] Woojin Lee, Lina Xiong, and Clark Hu. 2012. The effect of Facebook users' arousal and valence on intention to go to the festival: Applying an extension of the technology acceptance model. *International Journal of Hospitality Management* (2012).
[43] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
[44] Salini Mohanty, Amy E Leader, Emily Gibeau, and Caroline Johnson. 2018. Using Facebook to reach adolescents for human papillomavirus (HPV) vaccination. *Vaccine* (2018).
[45] Sergei Nirenburg, Jaime Carbonell, Masaru Tomita, and Kenneth Goodman. 1994. *Machine translation: A knowledge-based approach.* Morgan Kaufmann Publishers Inc.
[46] Eric H Nyberg and Teruko Mitamura. 1992. The KANT system: Fast, accurate, high-quality translation in practical domains. In *Proceedings of the 14th conference on Computational linguistics.* Association for Computational Linguistics.
[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics.* Association for Computational Linguistics.
[48] Ruth Rettie, Ursula Grandcolas, and Bethan Deakins. 2005. Text message advertising: Response rates and branding effects. *Journal of targeting, measurement and analysis for marketing* (2005).
[49] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* ACM.
[50] Stewart Shapiro and Deborah J MacInnis. 2002. Understanding program-induced mood effects: Decoupling arousal from valence. *Journal of Advertising* (2002).
[51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems.*
[52] Taifeng Wang, Jiang Bian, Shusen Liu, Yuyu Zhang, and Tie-Yan Liu. 2013. Psychological advertising: exploring user psychology for click prediction in sponsored search. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.
[53] Warren Weaver. 1955. Translation. *Machine translation of languages* (1955).
[54] Wikipedia. 2019. Common medical symptoms. https://en.wikipedia.org/wiki/List_of_medical_symptoms.
[55] Ling Yan, Wu-Jun Li, Gui-Rong Xue, and Dingyi Han. 2014. Coupled group lasso for web-scale ctr prediction in display advertising. In *International Conference on Machine Learning.*
[56] Elad Yom-Tov. 2018. Clinically verified pre-screening for cancer using web search queries: Initial results. *CoRR* (2018).

[57] Elad Yom-Tov, Anat Brunstein-Klomek, Or Mandel, Arie Hadas, and Silvana Fennig. 2018. Inducing behavioral change in seekers of pro-anorexia content using internet advertisements: Randomized controlled trial. *JMIR mental health* (2018).
[58] Elad Yom-Tov, Peter Muennig, and Abdulrahman M El-Sayed. 2016. Web-based antismoking advertising to promote smoking cessation: a randomized controlled trial. *JMIR* (2016).
[59] Elad Yom-Tov, Jinia Shembekar, Sarah Barclay, and Peter Muennig. 2018. The effectiveness of public health advertisements to promote health: A randomized-controlled trial on 794,000 participants. *npj Digital Medicine* (2018).
[60] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.