Information Retrieval Journal (2021) 24:205–239
https://doi.org/10.1007/s10791-021-09392-6

Algorithmic copywriting: automated generation of health-related advertisements to improve their performance

Brit Youngmann[1] · Elad Yom-Tov[1] · Ran Gilad-Bachrach[1] · Danny Karmon[2]

Received: 2 June 2020 / Accepted: 23 March 2021 / Published online: 13 April 2021

Abstract
Search advertising, a popular method for online marketing, has been employed to improve health by eliciting positive behavioral change. However, writing effective advertisements requires expertise and experimentation, which may not be available to health authorities wishing to elicit such changes, especially when dealing with public health crises such as epidemic outbreaks. Here, we develop a framework, comprising two neural network models, that automatically generates ads. The framework first employs a generator model, which creates ads from web pages. These ads are then processed by a translation model, which transcribes ads to improve performance. We trained the networks using 114K health-related ads shown on Microsoft Advertising. We measure ad performance using the click-through rates (CTR). Our experiments show that the generated advertisements received approximately the same CTR as human-authored ads. The marginal contribution of the generator model was, on average, 28% lower than that of human-authored ads, while the translator model received, on average, 32% more clicks than human-authored ads. Our analysis shows that, when compared to human-authored ads, both the translator model and the combined generator + translator framework produce ads reflecting higher values of psychological attributes associated with a user action, including higher valence and arousal, and more calls to action. In contrast, levels of these attributes in ads produced by the generator model alone are similar to those of human-authored ads. Our results demonstrate the ability to automatically generate useful advertisements for the health domain. We believe that our work offers health authorities an improved ability to build effective public health advertising campaigns.

Keywords Copywriting · Online advertising · Health · Deep learning · Marketing

This paper builds on our previous work, published in Youngmann et al. (2020), where we considered only one of the pipelines (the translator) presented in this current work. Correspondingly, the current work includes an extension of that work, including the methods and experimental study. In particular, we have included experiments examining the marginal contribution of each of the proposed pipelines, as well as the entire framework, compared with human-authored ads.

✉ Brit Youngmann
t-bryoun@microsoft.com

Extended author information available on the last page of the article

Mathematics Subject Classification 62M45 · 62P10 · 68T50

1 Introduction

Spending on search advertising (also known as sponsored ads) in 2019 was valued at US$36.5 billion in the U.S. (Statista, 2019a) and US$109.9 billion worldwide (Statista, 2019b). 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 user preferences (Garcia-Molina et al., 2011). 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 (See, for example, Fig. 1). Advertisers bid for specific keywords that they perceive as indicating an interest in their product. When these keywords are searched, the purchased 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 is 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 query (namely the keywords), which are provided by the advertiser; advertisers may additionally express preference for demographics (such as age or gender), location, and other user parameters. By submitting bids, which represent monetary values for acquiring the ad slots associated with keywords, advertisers compete with others who chose the same keywords. The advertising system uses the bids to choose which ads to display, allowing advertisers to increase their exposure by bidding higher.

Various ad performance measures can be tracked (Garcia-Molina et al., 2011), 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.

Health authorities, pharmaceutical companies, and other stakeholders in the health domain have recognized that search advertising can assist not only in selling products but also (more importantly) in steering people towards healthier behaviors. Ads have been shown to be effective in nudging people towards less harmful pro-anorexia websites (Yom-Tov et al., 2018), to quit smoking (Yom-Tov et al., 2016), and towards increased physical activity and better food choices (Yom-Tov et al., 2018). The conversion optimization mechanism of advertising engines has been utilized (in conjunction with advertising campaigns) to improve HPV vaccination rates (Mohanty et al., 2018) and to screen for cancer (Yom-Tov, 2018).

However, creating effective advertisements, as measured either by using the above-mentioned performance measures or by eliciting behavioral change (the two are not synonymous (Yom-Tov et al., 2018)) 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 that are available to internet advertising systems. We propose a framework that receives a (health-related) web page promoting healthy behavior or selling some product or service, and generates from it an ad that maximizes a required performance measure. After discussing relevant prior work, we describe the building blocks of this framework.

Ideally, a single model would be enough for the task of automatically generating an ad from a given web page and maximizing its CTR. However, as we demonstrate, rephrasing an existing ad to achieve better performance is an easier task than automatically generating an advertisement from a web page. We therefore divide the ad-generation task into two parts. In the first step, we employ a generator pipeline, which receives as inputs URLs of web pages and generates ads from them. This pipeline includes a context-extraction module, used to extract the relevant parts from the web pages, as well as an out-of-the-box text summarization module, used to generate ads from the extracted text. In the second step, we employ a translator model, which receives the generated ads as input and generates new optimized ads that are expected to achieve high performance. This pipeline includes ad normalization, preprocessing, and a sequence-to-sequence model. For completeness of this work, we also report the results achieved by using solely the generator model as well as the results achieved by using solely the translator model (directly translating the original ads), allowing estimation of each pipeline's marginal contribution.

Importantly, we note that the proposed pipeline is not entirely automated. This is because the generated ads contain tokens that should be substituted by the advertiser, for example, percent reduction in cost from a discount (see further details in Sect. 3.2). However, our goal is to assist health agencies, not replace them. Thus, minor grammatical errors in the generated ads may be manually corrected, or semi-manually corrected using prior work on the automatic transformation of text to proper English (e.g., Simplenlg, 2019). Another important advantage of the semi-automated pipeline is that advertisers can ensure that the semantic meaning of the generated ads is correct and that the generated ad properly reflects the input.

We demonstrate the performance of the models and share our insights into what they have learned in order for them to optimize the ads. Our experiments show that the advertisements generated by our full framework (i.e., the generator followed by the translator model) received approximately the same CTR as the human-authored ads (i.e., the original ads), implying that our framework can assist health authorities to automatically generate effective ads from scratch. The ads produced solely by the generator model received, on average, approximately 28% fewer clicks than the human-authored ads. In comparison, the ads produced exclusively by the translator model received, on average, 32% more clicks than the human-authored ads. This indicates that translating an ad to achieve better performance is an easier task than automatically generating an advertisement from a web page. Our analysis shows that the translator model produces ads reflecting higher values of psychological attributes associated with a user action, including higher valence and arousal, more calls to action, and the amplification of user desires, while the ads produced by the generator model behave similarly to the human-authored ads with regard to these metrics.

2 Related work

Our work draws on past literature in several areas, including headline generation, content extraction, abstractive summarization, 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 areas.

2.1 Headline generation

Ad creation is closely related to the problem of headline generation, which is the task of generating a headline for a given document. Headline generation can be viewed as a text summarization problem with the constraint that only a short sequence of words is allowed to be generated while preserving the essential topics of a document.

State-of-the-art methods for headline generation employ extractive or abstractive text summarization methods (Nallapati et al., 2016, 2017; Xu et al., 2010; Woodsend et al., 2010) (see discussion below). Similar to our proposed pipeline, headline generation techniques that use abstractive methods aim to generate the headline based on the understanding of the documents (Rush et al., 2015), sometimes using the lead sentences as the input text to generate the headlines (Ayana et al., 2016; Takase et al., 2016). Based on the observation that news article headlines in the form of a question may arouse users' curiosity and encourage them to click to find the answer, the authors of (Zhang et al., 2018) proposed a method to automatically generate a question headline for a given news article. Our goal is different from these previous works, as we aim to generate advertisements, which consist of both an headline and a short description (as we explain in Sect. 4.2). Moreover, as we show in Sect. 5.4, our generated ads capture users' attention with the aim of increasing advertising effectiveness.

2.2 Content extraction from HTML documents

Web pages are often cluttered with extraneous information around the body of an article, distracting users from the actual content they are interested in Gottron (2008). This information may include (pop-up or banner) advertisements, unnecessary images, or links scattered around the text. Thus, automatic extraction of *useful and relevant* content from web pages is a challenging task, one which has been extensively addressed in the literature (Sluban & Grčar, 2013; Gupta et al., 2003, 2005; Gottron, 2008; Peters & Lecocq, 2013; Song et al., 2015). Automatic content extraction from web pages has many applications, including enabling users to access web pages more easily over smart mobile devices, rendering speech for visually impaired users, and summarizing text (Gupta et al., 2005).

We note, however, that our task is simpler, as we aim to extract only meaningful and bounded-size text from a web page, such as the main paragraph and title, with the goal of generating an advertisement describing the product or service the page offers. Therefore, as we describe in Sect. 3.1, in this work we adapted the simple, efficient *Arc90 Readability algorithm* (https://github.com/masukomi/arc90-readability), which extracts the most important content from a web page. This commonly used algorithm

was transformed into the Readability.com product, which was incorporated into Safari's Reader view, Flipboard, and Treesaver but is now defunct.

2.3 Abstractive summarization

Text summarization is the process of automatically generating natural language summaries from an input text while retaining its main content. Generating short and informative summaries from large quantities of information is a well-studied task (Dorr et al., 2003; Nallapati et al., 2017, 2016; Tas & Kiyani, 2007; Gambhir & Gupta, 2017), applicable for multiple applications such as news digest creation, search, and report generation.

There are two prominent types of summarization algorithms. First, extractive summarization algorithms form summaries by selecting sentences of the input text as the summary (Dorr et al., 2003; Nallapati et al., 2017). Second, abstractive summarization models build an internal semantic representation of the original text, then use it to create a summary that is close to what a human might express (Liu & Lapata, 2019; Dong et al., 2019; Paulus et al., 2017). Abstraction may transform the extracted content by paraphrasing sentences of the source text. Such transformation, however, is computationally much more challenging than extractive summarization, involving both natural language processing and often a deep understanding of the domain of the original text.

Neural network models for abstractive summarization are typically based on the attentional encoder-decoder model for machine translation (Paulus et al., 2017; Liu & Lapata, 2019). State-of-the-art abstractive summarization techniques (e.g., Liu and Lapata 2019; Dong et al., 2019) employ Transformer-based models that have shown advanced performance in many natural language generation and understanding tasks. In this current work, we employ the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) for text embedding, the latest incarnation of pretrained language models which have recently advanced a wide range of natural language processing tasks (Young et al., 2018). We have also used the model of Liu and Lapata (2019) for summarizing the text extracted from web pages into ads.

2.4 Machine translation

Machine translation (MT) is a subfield of computational linguistics that investigates the use of a machine to translate text from a source language to a target language while retaining the meaning and sense of the original text.

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 (Nyberg & Mitamura, 1992; Nirenburg et al., 1994), statistical MT (Weaver, 1955; Koehn et al., 2007; Brown et al., 1988), and neural MT (Cho et al., 2014b; Papineni et al., 2002).

In rule-based MT systems (e.g., Forcada et al., 2011), 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 (Koehn et al., 2003; Chiang, 2005). Generally, the more human-translated text is available for a given

language pair, the better the translation quality. Two main drawbacks of statistical MT are that it depends upon huge amounts of parallel texts and it is not able to correct singleton errors made in the source language.

State-of-the-art MT systems use neural networks to predict the likelihood of a translation of a sequence of words (Kalchbrenner & Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014a). 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 them. 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 (Luong et al., 2015; Bahdanau et al., 2014). However, as we show, even our simple neural MT model achieves significant improvement in terms of click-through rates (See Sect. 5). 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.

2.5 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 allocated display space 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 (Graepel et al., 2010). As a consequence, the task of CTR prediction has been extensively studied (Tang et al., 2017; Juan et al., 2016; Yan et al., 2014; Zhou et al., 2018), since it impacts user experience, advertising profitability, 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 first two of these are dense, the keywords are sparse, and hence need care in their representation.

Bengio et al. (2003) 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 Tang et al. (2017), Gai et al. (2017) 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., Covington et al., 2016; Cheng et al., 2016; Shan et al., 2016) suggested replacing the transformation functions with a complex multilayer perceptron (MLP) network, which greatly enhances the model's capability. Generally, these methods follow a similar model structure, combining an embedding layer (for learning the dense representation of sparse features) and an 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 Sect. 3.2.4, here we focus on generating ads with higher CTR values than those

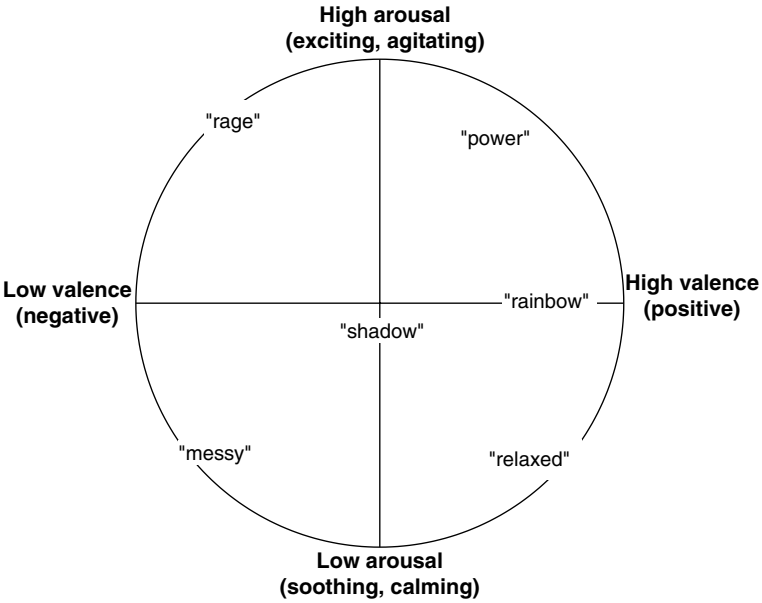


Fig. 1 Different English words mapped on the arousal-valence space. (Words were placed according to Bradley & Lang, 1999)

of their human-authored ads, hence we use a simple ranking model for this task. However, our model is based on insights from the above-mentioned prior work.

2.6 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 (Kensinger, 2004; Lang et al., 1995; Bradley et al., 1992; Conway et al., 1994; Hamann 2001). 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 1 shows an illustration of this, where different affective states are located in the space spanned by these axes (Bradley & Lang, 1999).

Advertising professionals aim to increase advertising effectiveness by creating ads which capture consumers' attention. It has been shown that the inclusion of high arousal and valence sequences in ads increases user attention and interest (Belanche et al., 2014; Lang et al., 1995; Lee et al., 2012).

Arousal is related to body activation level in reaction to external stimuli (Gould & Krane, 1992) and has been associated with simple processes such as awareness and attention and also more complex tasks such as information retention (Holbrook & Hirschman, 1982). Previous work suggests that arousal modulates ad effectiveness and memory decoding (Jeong and Biocca 2012). However, highly arousing contexts can distract individuals from ad processing, making recall more difficult and thus reducing the ability to encode ad content (Shapiro & MacInnis, 2002).

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 to increase memory performance. Namely, non-arousing ads with very positive or very negative valence are better remembered than neutral ones (Kensinger, 2004; Kensinger & Corkin, 2003).

Ads which refer to the desires of users, especially in the form of special textual content, can affect CTRs in sponsored search (Wang et al., 2013). These desires can be mined from ads 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, for example, brand loyalty and luxury seeking.

Numerous past work studied the problem of how to attract users' attention. For example, the authors of Yang et al. (2019) used a neural network–based method to identify the persuasive strategy employed in texts. As another example, the authors of Pryzant et al. (2018) investigated the relationship between the way a search advertisement is written and internet user behavior, as measured by CTR. Their goal was to identify words or phrases to which a search ad's success (or failure) can be attributed. In contrast, our goal is to generate ads that are expected to maximize CTR automatically. Similar to findings of Wang et al. (2013) (which are also supported by our experiments), they found that phrases containing authoritative framing (such as "official site") or conveying a petty advantage (e.g., "free shipping") attract user attention.

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

3 Methods

We address the problem of automatically generating ads for the health domain. As mentioned in the Introduction, we have divided this task into two parts. In the first step, we employ a generator model, which receives as inputs URLs of web pages and generates ads from them. Then, we employ a "translator" model, which receives as input the generated ads from the first step and outputs new optimized ads that are expected to achieve high performance. We note that the second model actually rephrases the input ads and does not translate them into a different language. However, as we are using existing machine translation techniques, we refer to this model as the translator model. In the following, we describe each of these pipelines in detail.

We refer to the advertiser-created ads in our training set as "human-authored" ads and to ads generated by the proposed pipeline models as "generated" ads.

For completeness of this work, we also report the results achieved by the generator model without the last rephrasing step, as well as the results achieved by using solely the translator model (directly translating the human-authored ads), allowing estimation of the marginal contribution of each of the models.

3.1 Generator

Our proposed ads generator pipeline consists of two steps: content extraction and ad generation.

3.1.1 Content extraction

Given an HTML document, our goal is to extract relevant and meaningful content from the document, identifying the most important information about the product/service to be sold. This information may include details about the product/service (e.g., "Get your birth control pills online"), special offers (e.g., discount, free/fast delivery), and a short explanation of the page content (e.g., "Top 10 Foods to Help Manage Diabetes"). Another important restriction is that we want the extracted content to be short and concise. Unlike previous works that aim to separate the main content of a web page from the noise it contains (e.g., advertisements, links to other pages, etc.) (Gupta et al., 2005; Gottron, 2008; Peters & Lecocq, 2013; Song et al., 2015), our goal is to extract only a few main paragraphs from the web page.

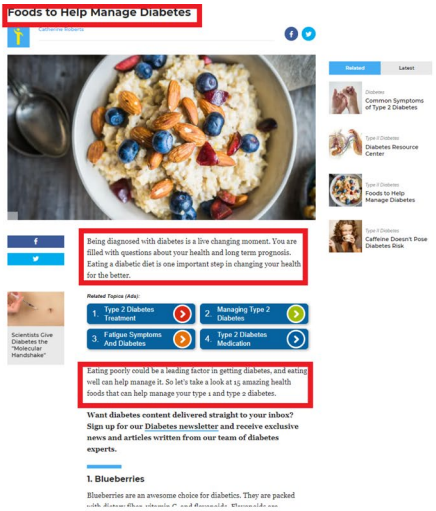
As the web pages have very different structures (e.g., as in Fig. 2), there is no precise "one size fits all" algorithm that could achieve this goal. Nonetheless, in this work, we have adopted a simple yet highly effective algorithm, called the *Arc90 Readability algorithm*, that extracts meaningful content from raw HTML documents. It was developed by Arc90 Labs to make websites more comfortable to read (e.g., on mobile devices). For completeness of this work, we next briefly describe how this algorithm operates.

The Arc90 Readability algorithm is based on two lists of HTML attributes (i.e., ids and classes names). One list contains attributes with a "positive" meaning, while the second list contains attributes with a "negative" meaning. Intuitively, the algorithm operates as follows. For each paragraph (i.e., a p-tag), it adds the parent of the paragraph to a list (if it is not already present), and initializes the score of the parent with 0 points. If the parent has a positive attribute, the algorithm adds points to the parent, otherwise, it subtracts points. Lastly, the algorithm retrieves the top parents with maximal points and extracts their textual content.

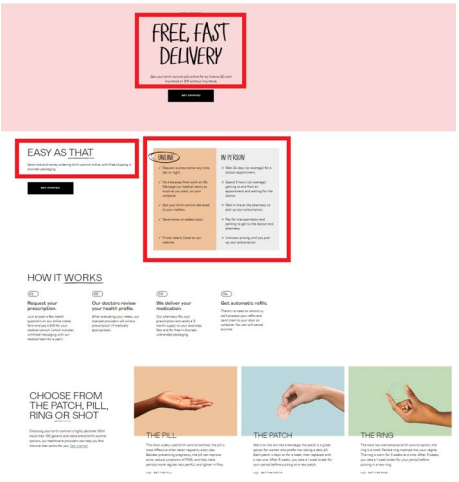
Examples of attributes names we considered in our setting are depicted in Table 1. As for the scoring, for negative attributes, the algorithm subtracts 2 points, and for positive attributes, the algorithm adds 1 point. To illustrate, consider the two web pages presented in Fig. 2. The content extracted from them using the Arc90 Readability algorithm is marked in red. We have extracted, from each page, its title, concatenated with the top 2 parents with the maximal points.

3.1.2 Candidate generation

Given the extracted content from a raw HTML document, our next step is to generate an ad that directs viewers to this page. To this end, we trained a state-of-the-art model for abstractive summarization, named the PreSumm model, presented in Liu and Lapata (2019). The authors of Liu and Lapata (2019) showcased how Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), the latest incarnation of pretrained language models which have recently advanced a wide range of natural language processing tasks, can be usefully applied in text summarization. In that paper,



Diabetes, What To Eat? - 15 Foods For Diabetes. Check Out These 15 Foods That Can Help You Control Blood Sugar.



Order Contraception Online - Safe, effective, convenient. Learn about oral contraceptives from our licensed doctors & free prescription.

Fig. 2 Two examples of web pages along with their displayed ads. Marked in red are the relevant parts that were extracted from the pages

Table 1 Examples of "negative" and "positive" attributes used for the Arc90 Readability algorithm

"Negative" attributes	"Footer", "copyright", "location", "style", "comment", "meta".
	"Content", "text", "title", "body", "article", "page", "description".

the authors proposed a general, off-the-shelf framework for both extractive and abstractive models. Here we focus on abstractive summarization, as it allows paraphrasing a text. As mentioned in the Related Work, texts summarized using this technique tend to look more human-like and produce more condensed summaries. The rationale behind

using the PreSumm model was that it is a top-performing model with good documentation, allowing for easy reproducibility of our results. We note that other summarization models can be employed, and we leave the investigation of the optimal model for future research.

Training data We constructed the training data as follows: For every web page in the training set, we extracted its content and considered its corresponding (human-authored) ad that achieved the highest CTR. The reason for that is to allow the model to learn only how to generate ads which are expected to have high CTR values. We have also experimented with a model trained over all human-authored ads. However, as its results were inferior, we omit them from this presentation. For every (web page, corresponding ad) pair, we generated an example in the training data. For training the model, we used the default parameters of the abstractive summarization model as described in the implementation of Liu and Lapata (2019). We have also experimented with other parameters' values and found the results to be with a very similar level of quality. Thus, for conciseness, we do not report the results obtained while using other parameters.

3.2 Translator

In the previous step, we generated ads from the specified web pages. We note that that task is much harder than the one of rephrasing the ads to maximize CTR. Therefore, to improve the generated ads, we employ a translator model to rephrase them. To this end, we first trained a translator model over the human-authored ads. The proposed ad translation pipeline consists of the following components: (1) normalization, (2) preprocessing, (3) candidate generation, and (4) selection.

3.2.1 Normalization

In order to allow for generalization, we created a model to identify medical conditions and proposed treatments in the text and replace them with generic tokens. This is achieved using a custom named entity recognition (NER) model based on the spaCy library (spacy, 2019). 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 the word's 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 2, the entities we wish to recognize are both "Shingles vaccine" and "shingles" (replacing both of them with the token <CONDITION/TREATMENT>).

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

3.2.2 Preprocessing

We used lemmatization and stemming to reduce the vocabulary size (Korenius et al., 2004). 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 (spacy, 2019) to perform this task. Examples of this process are shown in Table 2.

3.2.3 Candidate generation

We then 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 Sect. 4.1).

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.

Model architecture We employed a simple sequence-to-sequence (seq2seq) model using the PyTorch library (Pytorch, 2019). The 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 the tokens are 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 text as input 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, and 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.

3.2.4 Selection

The translator model can generate multiple candidates, out of which we would like to select the best translation by ranking them. Unfortunately, while the CTR values are known for the input ads, these values are unknown for the output ads. Therefore, given an human-authored or generated 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 Sect. 5, in the vast majority of cases, the generated ads have succeeded to better attract users' interest than their corresponding human-authored ads.

As mentioned in the Introduction, the proposed pipeline is not entirely automated. This is because the generated ads contain tokens that should be substituted by the advertiser (see, for example Table 2). However, our goal is to assist health agencies with ad creation, 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., Simplenlg, 2019). General tokens (such as <CARDINAL> and <CONDITION/TREATMENT>), if they did not appear in the human-authored ad (e.g., Table 2 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 important advantage of the semi-automated pipeline is that advertisers can ensure that the semantic meaning of the generated ads is correct and properly reflects the input.

4 Experimental study

Recall that our goal it to automatically generate ads that attract users' interest. To better assess the marginal contribution of each of our proposed two models, we consider the following baselines:

1 *Human-Authored* the original ads (to be described in Sect. 4.2).
2 *Generator* the generator pipeline (described in Sect. 3.1), without the rephrasing step, which receives *the human-authored ads' URLs* as input and outputs generated ads.
3 *Translator* the translator model (described in Sect. 3.2), which receives *the human-authored ads* as input, and outputs rephrased ads.
4 *Generator + Translator* our full model, which receives as input the URLs of the web pages, generates ads from them, and then rephrases them to improve performance using the trained translator model.

As we shall see, the ads produced by the generator baseline have worst performance than human-authored ads, while the ads produced by the generator + translator model have similar performance to that of human-authored ads. Translation of human-authored had are expected to have higher performance. Not surprisingly, as translating a given ad is a simpler task than generating an ad from scratch, we will show that the ads produced solely by the translator model are expected to get the best performance among all examined baselines. Nonetheless, the translator model receives the human-authored ads as input, and thus

using solely this model requires health authorities to first write ads by themselves, which, as mentioned in the Introduction, is a task requiring expertise and experience.

In the following sections we present the experimental setup.

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, click-through rate (CTR) is the percentage of users viewing a Search Engine Results Page (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 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, and which were shown at least 100 times in that period. The data comprises over 114K advertisements displayed to users who queried in two domains: Medical Symptoms (MS) and Preventive Healthcare (PH) (see Table 3).

To identify ads relevant to these domains, we considered search queries that contain specific (predefined) keywords. Keywords for the MS domain included common medical symptoms (according to Wikipedia (Wikipedia, 2019)). Keywords for the PH domain were extracted from a U.S. government website (Preventive Healthcare, 2019). In the MS domain, for example, we considered all queries containing the word "vertigo" (See Table 3 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.

For the translator, we trained a model separately for each of the two 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 and thus omitted from this presentation (see Sect. 4.1 for an explanation on how quality was estimated).

For the generator model, we trained a single model for both domains. Recall that for the generator model, we consider only the corresponding ad achieving the highest CTR for each web page. Thus, splitting the data here between the two domains would result in insufficient data for efficient training. In the last step of the generator + translator model, i.e., when using the trained translator model, for each generated ad from the MS (resp., PH) domain we have used the corresponding translator model trained over the data of the MS (resp., PH) domain.

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

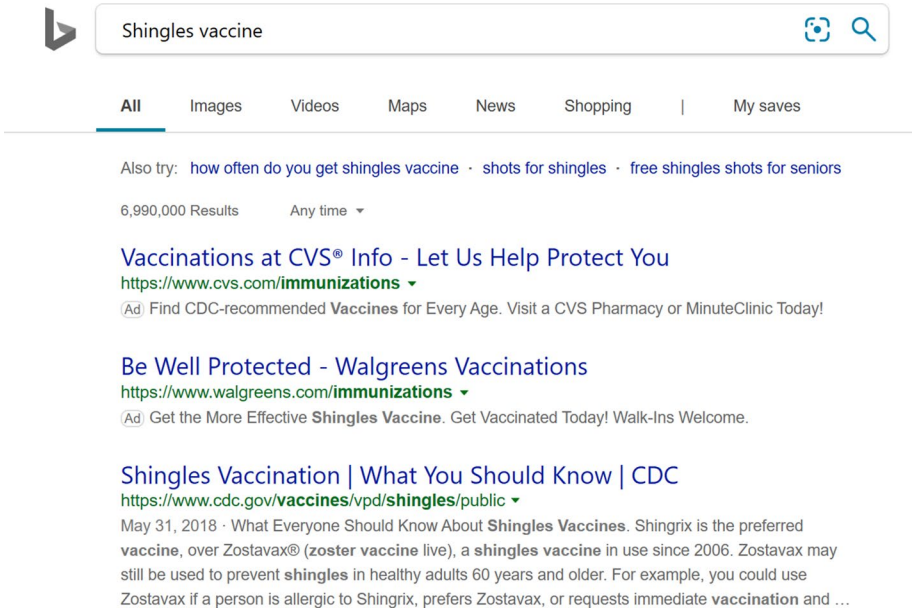


Fig. 3 Example Bing SERP with two search ads. The search query is "Shingles vaccine" and the search ads appear as the two top results

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 Fig. 3. 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 subfields, that is, there are up to three title fields and two 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 Sect. 2). However, in our setting, it is sufficient to be able to predict the relative order of CTRs for ads on a given query, so the model can predict whether a generated ad is more likely to achieve a higher CTR value than its human-authored version. To this end, we employed a learning-to-rank model (LambdaMART (Burges, 2010)) as implemented by the pyltr library (pyltr, 2019).

Training Data The training data consist of the text of the human-authored ads. Previous work has shown the importance of writing a simple, concise, consistent, and easily understood text to increase advertising effectiveness (Grewal & Alagaratnam, 2013; Chebat et al., 2003). To quantify the text of the ads for the LambdaMART model, we represented it using simple features that were chosen so as to quantify the ability of users to understand and decode the content of ads. These features include the sentiment, lexical diversity, and

Table 2 Example (human-authored, preprocessed, and generated) ads

Human-authored 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 3 Extracted Data

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.

the readability index of the text. 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 also considered but were found not to improve ranker performance and were thus excluded in the following analysis. Table 4 provides the full list of features. The training data of the CTR ranker model consists of lists of extracted (human-authored) 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 translator and generator models in the following manner: for every human-authored ad and its corresponding translated and generated ads, we examine how our trained model ranked the three ads. Importantly, when predicting the performance of a translated or generated ad, we ensured that this ad was not in the training data of the corresponding model.

For each ad, we report the rank of (1) the human-authored ad, (2) the translated ad (using the translator model), (3) the generated ad (using only the two first steps of the generator model), and (4) the generated + translated ad (using the full generator model).

To estimate the likelihood of ranker error (where it will rank higher the ad with the lower CTR in a pair), we report the 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 (over the human-authored ads). The average KT across folds was 0.44 ($P < 0.01$). This should be compared to KT equalling 0.36 when randomly ordering the ads.

4.4 Online evaluation

For each of the domains, MS and PH, we selected 10 random ads out of the generated ads for which the ranker ranked the generated ads higher than the human-authored ad. We chose to select ads that were estimated to have higher rank because this more closely mimics practical scenarios, where new ads are likely to be used if they are predicted to have superior performance to that of existing ads.

We tested in a real-world setting the performance of the human-authored advertisement compared to the translated ads, the generated ads, and the generated ads after applying translation to them. Here again, we ensured that each generated/translated ad was not in the training data of the corresponding model. 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 was too long, moving any excess text into the secondary or tertiary title fields. 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 group of ads (i.e., the

Table 4 Features extracted from text ads for the CTR ranker model

Feature	Explanation	Extractor
Flesch-Kincaid readability ease	Indicating how difficult a sentence in English is to understand.	Textstat (2019)
Flesch-Kincaid readability grade	Indicating the number of years of education generally required to understand the text.	Textstat (2019)
# of "difficult" words	According to the Textstat library (2019).	Textstat (2019)
Readability consensus based upon the Dale-Chall Readability Score, the Linsear Write Formula, and the Coleman-Liau Index	Indicating the estimated school grade level required to understand the text.	Textstat (2019)
Vader-Sentiment	Indicating how positive/negative is the sentiment according to the Vader measure.	VaderSentiment (2019)
Lexical diversity	Number of distinct words divided by the number of words in the text.	
# of punctuation marks		spaCy (2019)
# of noun phrases		spaCy (2019)
# of adjectives		spaCy (2019)

human-authored and the three algorithmically generated 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 300 impressions per ad were obtained.

4.5 Components evaluation

Content extraction evaluation We next examined the Arc90 Readability algorithm's performance when used to extract the main content of the web pages. To this end, we considered 50 random web pages. For each of these web pages, we asked five crowdsourced workers from Amazon Mechanical Turk (MTurk, https://www.mturk.com/) to manually point out the two most relevant paragraphs of the page, defined as those which best describe its content. Using the plurality vote aggregation function, two paragraphs for each web page were selected as the most informative paragraphs. Then, we considered the paragraphs selected by the Arc90 Readability algorithm and compared them with the paragraphs selected by the crowd workers. We report here the number of hits (where the paragraphs selected by the Arc90 Readability algorithm were also selected by the crowd workers) and misses (otherwise).

Summarization evaluation Next, we examine the performance of the PreSumm model (Liu & Lapata, 2019), used to generate an ad from the extracted content of a given web page. We report the average F1 ROUGE-1 and ROUGE-L scores[1] (Lin, 2004), where we compare between pairs of original (i.e., author-generated) and generated ads.

Translation evaluation Last, we examine the performance of the proposed translation model. Here, again, we report the average F1 ROUGE-1 and ROUGE-L scores, where we compare between: (1) pairs of original (i.e., author-generated) and translated ads and (2) pairs of original (i.e., author-generated) and generated + translated ads (namely, ads produced by our full pipeline).

4.6 Emotion analysis

We examined three main forms of emotional effect in ads, as described in Sect. 5.4, namely, the call to action, arousal and valence, and thought- and feeling-based effects. For each of these effects, we measured their value in human-authored 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 (Rettie et al., 2005). CTAs can drive a variety of different actions depending on the content's goal and 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 (Rettie et al., 2005). 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 CTAs 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).

[1] Computed using the Python implementation of the ROUGE metric https://pypi.org/project/py-rouge/.

To classify the emotional attributes of the ad we focused on two concepts which have been shown useful to measure emotional experiences (Lang et al., 1995; Kensinger, 2004): **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 marked the CTA verbs therein. Then, we created a model for each of the three quantities using this dataset and applied it to a larger set of ads.

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

The valence and arousal values and the CTA verbs for the dataset were found by asking five 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 (spacy, 2019). The trained model was applied to non-labeled ads to tag all words. Specifically, we have created a new tag called CTA, which was associated to each CTA verb marked by the crowd workers. We then used the trained model on other unlabeled ads, examining the tags of each word. Every word that was tagged as CTA was considered as a new 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 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. (2013). 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 5 lists these effects, their associated user desires, and the keywords that were used to identify these effects in the ads. Wang et al. (2013) 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.

Table 5 Examined user desires under thought-based and feeling-based effects

Effect	User's Desire	Examined Keywords	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 6 Average F1 ROUGE scores

	ROUGE-1	ROUGE-L
Summarization	31.2	28.9
Translation (human-authored vs. translated)	37.8	36.6
Translation (human-authored vs. generated+translated)	28.7	25.8

In Sect. 5.4, 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 pipelines. We also aim to explain what each of the models has learned.

5.1 Components evaluation

Content extraction evaluation As mentioned in Sect. 4.5, we evaluated the performance of the Arc90 Readability algorithm via a user study. We report that in the vast majority of the examined cases, the algorithm and the participants selected the same paragraphs as having the most important content. In particular, in 84% of the cases, the algorithm and the participants chose the same paragraph as the most important one, and in 81% of the cases, the algorithm and the participants chose the same paragraph to be the second most important paragraph. Our results show that identifying the most important paragraph is easier than defining the second most important paragraph. Indeed, in 78% of the cases, three or more (out of five) crowd workers chose the same paragraph to be the most important paragraph for a given web page. In comparison, three or more crowd workers chose the same paragraph to be the second-best paragraph in only 56% of the cases. This implies that even in the cases where the second paragraph selected by the Arc90 Readability algorithm was counted as "miss," it may nevertheless be important.

Summarization & translation evaluation As discussed in Sect. 4.5, we compared the original text with the text produced by our models using the ROUGE-1 and ROUGE-L metrics (Lin, 2004). The results are depicted in Table 6. Observe that the overlap between the generated ads and the human-authored ones is relatively high, suggesting that the model has succeeded in producing ads containing similar content to the original ones. Not surprisingly, the translation model results are higher than the ones obtained by our full pipeline (generated + translated), as the translation model "only" rephrased the original ads and did not generate new ones. As mentioned in the Introduction, translating an ad to achieve better performance is an easier task than automatically generating an advertisement from a web page. Nevertheless, we note that the ROUGE metrics are based only on content (i.e., n-grams) overlap. Thus they may determine if the same general concepts are discussed between an automatically generated summary/translation and a reference one. However, the metrics cannot determine if the result is coherent or the sentences flow together sensibly.

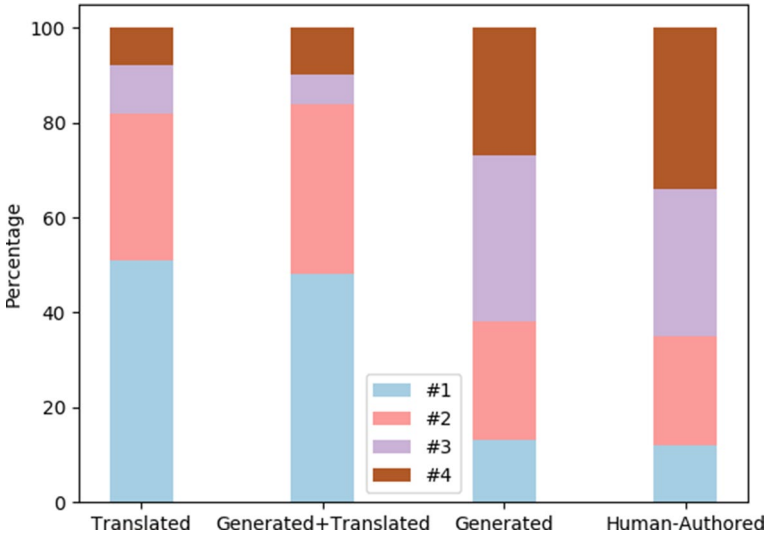


Fig. 4 Percentage of ads produced by each baseline placed at each rank (1–4), according to the CTR ranker

5.2 Offline estimation of result quality

Using the trained CTR ranker, for each human-authored ad, we examine the ranks of its corresponding generated ads (produced by our proposed models). For each (human-authored or auto-generated) ad, the CTR ranker outputs probabilities reflecting its estimated rank. Ads having a difference in probabilities of less than 0.1 were treated as having the same rank. According to the Kemeny–Young method (Wikipedia, 2020), we found that the optimal ranking is: 1. translator 2. generator + translator 3. generator 4. human-authored. The full results are depicted in Fig. 4.

Overall, the CTR ranker predicts that the ads generated by the translator model will have the highest CTR in 51% of the cases. This result is not surprising, since, as mentioned, translating a given ad to improve its performance is a much simpler task than generating an ad from a web page. The second-best competitor is the full generator model (i.e., generated + translated), where the CTR ranker predicts that the ads generated by this model will have the highest CTR in 48% of the cases (in 38% of the cases, the differences in performance between ad pairs were less than 0.1).

Observe that adding the rephrasing step to the generator model significantly improves its performance. Namely, the ads generated only by the generator model (without the last rephrasing step) are predicted to have the highest rank only in 13% of the cases. As can be seen, our results indicate that ads produced by the generator model behave similarly to the human-authored ads. This result is surprising considering the difficulty of the task of generating an ad directly from a given web page.

We note, however, that the CTR ranker cannot be considered as a perfect proxy for experimentation, since it does not precisely predict the ordering of ad pairs by their performance (as mentioned in Sect. 4.3). Thus, because the CTR ranker is imperfect, some of the generated ads which are predicted by the ranker to be less effective than their corresponding human-authored ads may achieve higher CTR values in practice, and vice versa.

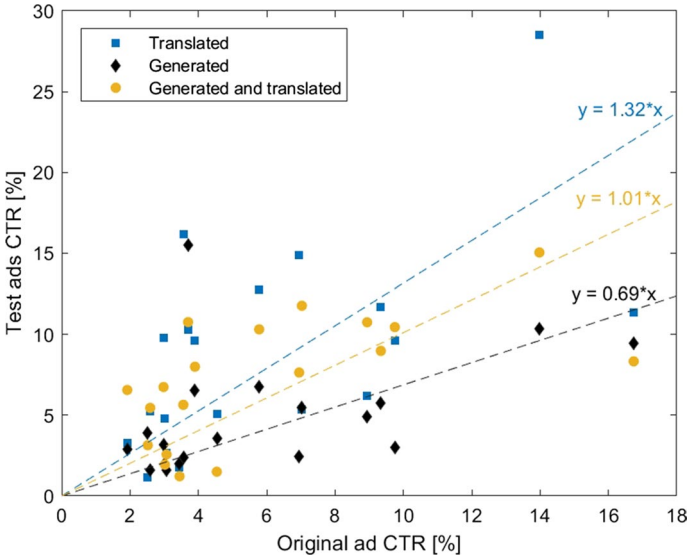


Fig. 5 CTR of the generated ads versus the human-authored ads. Each dot represents one advertisement pair (one human-authored versus one generated by each of the three models). Dotted lines denote linear regression fits

5.3 Online estimation of result quality

Twenty ads were run on the Microsoft Advertising platform between March 3rd and April 28th, 2020, until they were shown at least 300 times each. The average number of times that each ad was shown was 1,740 times, and the maximum was 6,191 times. A post-test found that the average rank at which different versions of the same ad were displayed was not statistically significantly different (Friedman test, $P > 0.05$).

Figure 5 shows the CTR of the generated ads versus those of the human-authored ads, for each type of generated ad. As the figure shows, the generated ads received, on average, approximately 31% fewer clicks than the human-authored ads ($P = 0.046$, Wilcoxon's signed rank test). The translated ads received 32% more clicks than the human-authored ads ($P = 0.040$, Wilcoxon's signed rank test), and the translation of the generated ads had approximately the same CTR as the human-authored ads (not statistically significant).

These results indicate that our proposed full framework, i.e., the generator + translator model, can generate entirely new ads with performance highly similar to the ads that were written by experts. Moreover, the results show that the translator model can significantly improve the performance of the current health-related ads, by simply rephrasing them.

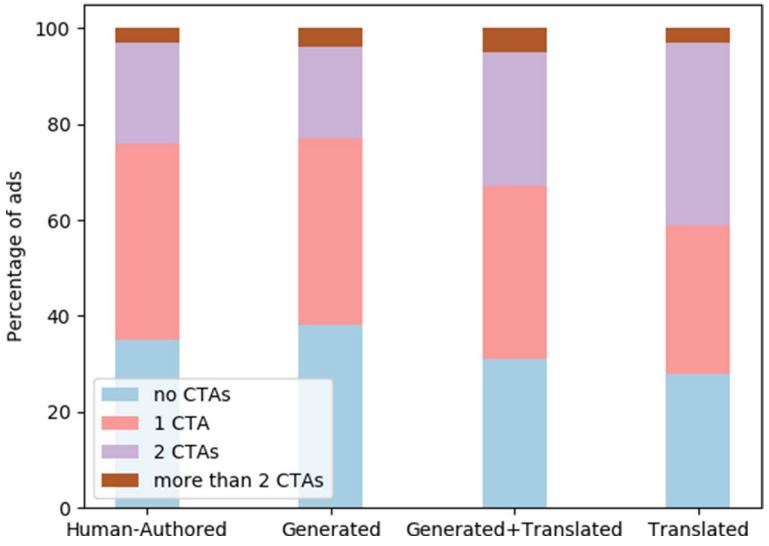


Fig. 6 Predicted number of CTAs

5.4 Emotion analysis results

To obtain an understanding for what the proposed models have learned, we examined the frequency of using calls to action (CTAs) and the emotional attributes of the generated ads.

5.4.1 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 (Wang et al., 2013). Furthermore, Rettie et al., (2005) 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 Sect. 4.6) to identify CTA verbs in both human-authored and generated ads. The performance of the tagging 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 (Fig. 6).

Analysis reveals that 72% of the ads generated by the translator model include at least one CTA verb, compared the human-authored ads, of which only 65% include at least one CTA verb (statistically significant, Wilcoxon's signed rank test, $P < 0.05$). Regarding the generator model, without the last rephrasing step, only 62% of the ads generated by this model include at least one CTA verb, compared with 69% of the ads generated by the full generator + translator model (statistically significant, Wilcoxon's

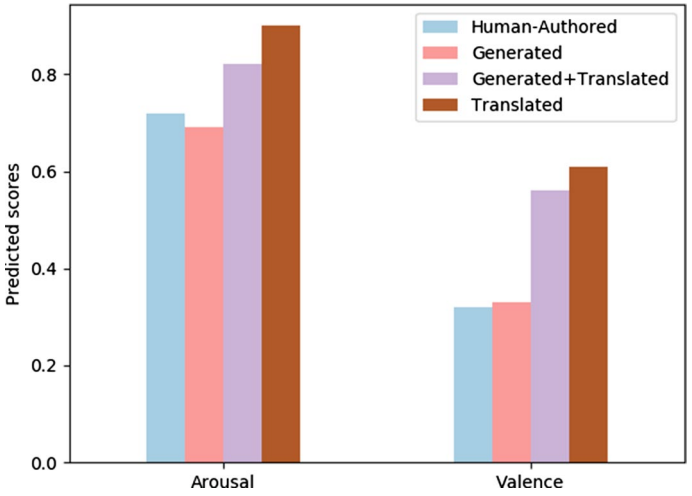


Fig. 7 Predicted arousal and valence scores

signed rank test, $P < 0.05$). According to the Wilcoxon's signed rank test, the difference in the number of CTAs between the human-authored ads and the ones produced by the generator model were found to be not statistically significant. Similarly, the difference in the number of CTAs between the ads produced by the full generator model (i.e., generator + translator) and the translator were found to be not statistically significant.

5.4.2 Arousal and valence analysis

To assess how well the crowd workers agree among themselves regarding the arousal and valence of ads, we report the obtained Krippendorff's alpha coefficients (as annotators have given a score on an ordinal scale of five independent labels per ad). The average Krippendorff's alpha for the arousal scores was 0.82, and the average Krippendorff's alpha for the valence scores was 0.78. 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 then applied to predict the arousal and valence scores of all human-authored and generated ads.

The average predicted arousal and valence scores of the human-authored and generated ads are depicted in Fig. 7. Note that even though these values are measured on a scale of $[-2, 2]$, most of the predicted scores were positive, as are the average scores.

Interestingly, observe that the predicted arousal and valence scores of the ads generated by the translator model are the highest. This implies that the proposed translator model is predicted to increase the arousal and valence of the input ads (i.e., the human-authored ads or the ads generated by the generator model).

Figure 8 shows the predicted valence and arousal scores of 1000 random human-authored ads versus those of the ads generated by the proposed models. The diagonal line indicates equal values for the human-authored and auto-generated ads. Points on or below

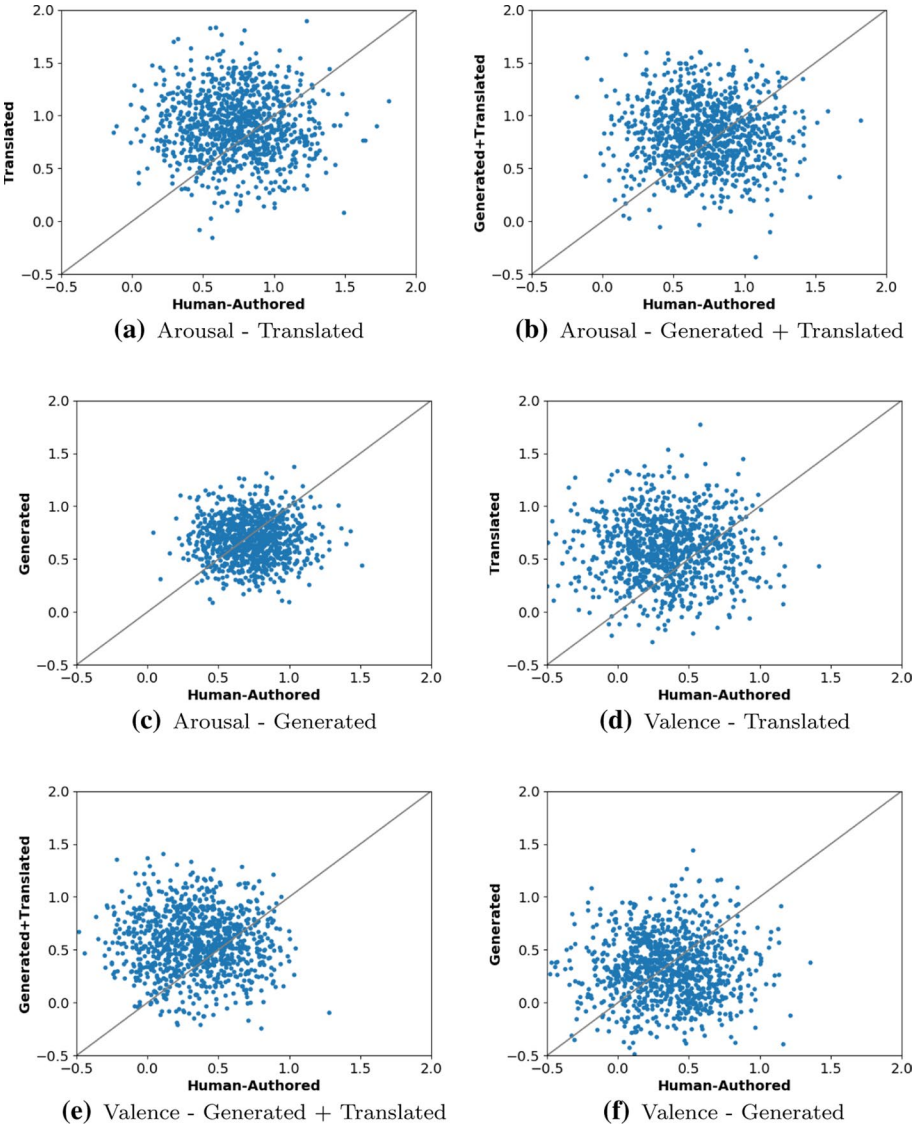


Fig. 8 Predicted arousal/valence scores of 1K random human-authored ads versus those of the ads generated by the baselines. The diagonal line indicates equal values

the grey line denote ads for which the auto-generated ads had equal or lower valence or arousal scores, compared to those of the human-authored ads. As the figures show, for the translator and the full generator (i.e., generator + translator) models, the vast majority of points (i.e., ads) are above the grey lines for both arousal and valence. Thus, in most cases, the ads generated either by the translator or by the full generator models are predicted to have higher arousal and valence scores than their corresponding human-authored ads. In contrast, as can be seen, only slightly more than half of the ads generated only by the

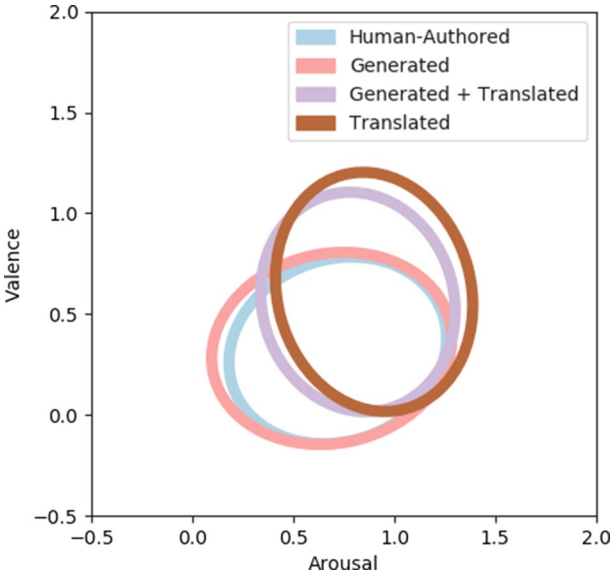


Fig. 9 Predicted valence versus arousal scores of 1K random human-authored ads and their corresponding auto-generated ads. Ellipses denote one standard deviation of the mean

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

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

generator model are predicted to have higher arousal or valence scores than their corresponding human-authored ads.

Figure 9 depicts the predicted valence scores versus the predicted arousal scores of (the same) 1000 random human-authored ads and their corresponding auto-generated ads of the three models. Ellipses denote one standard deviation around the mean. As can be seen, on average, the ads generated by either the translator or by the full generator model are predicted to have higher arousal and higher valence, compared to the human-authored ads. That is, these models increase both the arousal and the valence scores of the ads, compared with their human-authored versions. In contrast, the ads produced by the generator model are rather similar to the human-authored ads and there is almost no change in the predicted arousal and valence scores.

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 Sect. 2, as opposed to arousal, where it has been shown that ads with high arousal result in increased 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,

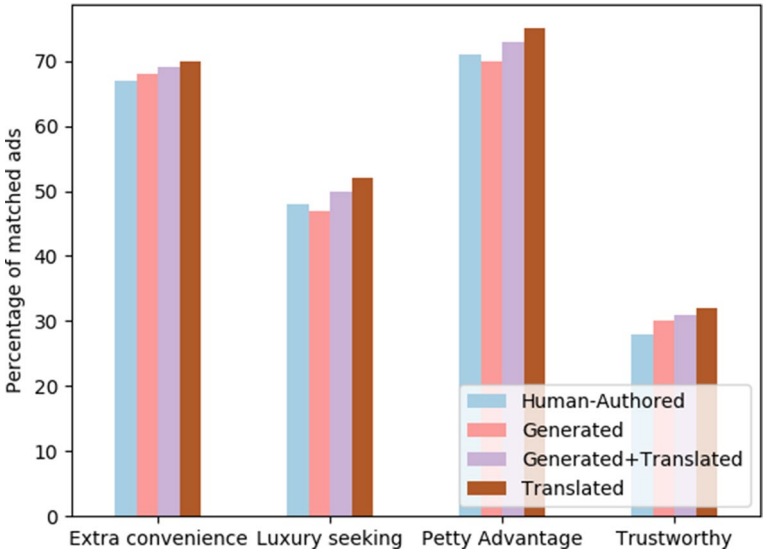


Fig. 10 Percentage of ads containing at least one of the keywords associated with each of the examined user's desires

here we consider the absolute valence scores. The Pearson correlation between the CTR of the human-authored ads and the predicted arousal score was 0.72 ($P < 0.01$), and the Pearson correlation between the CTR of the human-authored ads and the absolute value of the valence score was 0.62 ($P < 0.01$). Namely, we observe a moderate positive relationship between CTR and the arousal and (absolute) valence scores. Thus, as the arousal score increases, it is more likely that users would click on the ads. Similarly, as the absolute valence score increases, so does the chance a user would click on the ad.

5.4.3 Analysis of thought- and feeling-based effects

As mentioned in Sect. 2, the inclusion of specific textual content referring to user desires increases user interest, and consequently CTR (Wang et al., 2013). To investigate if these user desires are associated with increased CTRs, we computed the CTR of (human-authored) ads containing the keywords associated with each desire (see Table 5) 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 5.

Additionally, we examined the percentage of human-authored and auto-generated ads containing at least one of the keywords associated with each of the examined user desires, as listed in Table 5.

The results are shown in Fig. 10. 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 ads created either by the translator or the generator+translator models include more such keywords compared with the human-authored ads, indicating that both models have learned that incorporating specific textual content increases user interest.

Thus, our results support the empirical results of Wang et al. (2013), showing that users tend to click more on ads containing specific patterns and keywords.

6 Discussion

In this study, we presented a system which, when given an health-related web page, generates an optimized ad that is expected 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 models, our work required 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 our proposed pipelines, separately and together, in a real-world scenario by running 20 advertisements in a head-to-head competition and found that our auto-generated ads are expected to behave similarly as the original ad. This surprising result indicated that our proposed framework can assist public health authorities to automatically generate ads. In case the end-user (i.e., health authority) provides an input ad, our results indicate that we can improve its performance by 32% (on average) using solely the translator model.

To investigate what our models have learned, we examined three factors of the original ads compared to those of the automatically generated ads: (1) the use of calls to action (CTAs), which have been shown to be essential in increasing an ad's effectiveness (Rettie et al., 2005; 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 (Belanche et al., 2014); and (3) the inclusion of special textual content that refers to the desires of users (Wang et al., 2013).

Our empirical results indicate that the translation model improved all three factors. Specifically, the ads generated by this model include more CTAs than the original or auto-generated ads, they are predicted to have higher arousal and valence emotions, and they combine more keywords that have been shown to be 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. In contrast, our experimental study shows that the ads produced by the generator model behave similarly to the human-authored ones with regard to these three factors. Namely, this model did not improve these factors. This could be explained by the following two reasons. First, as mentioned, translating an existing ad is an easier task than generating an ad from a given website. As both tasks aim to optimize CTR, it is not surprising that the translator model has shown more success than the generator model. Second, recall that the generator model is trained using a single pair of a web page and an ad (as explained in Sect. 3.1.2). On the other hand, the translator model is trained over all pairs of ads related to the same search query (as explained in Sect. 3.2.3). Therefore, the training data of the translator model is significantly bigger than that of the generator model, which may affect the performance. Nevertheless, as we demonstrated, the advertisements generated by our full framework (i.e., the generator followed by the translator model) behave approximately the same as the human-authored ads (i.e., the original ads) in terms of CTR, implying that our framework can assist health authorities to automatically generate effective ads from scratch.

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 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 apply our algorithm to more domains, such as wellness and education, and 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 were used (see discussion in the Related Work), and if different outputs of the translator and generator models were examined for the same input ad, tuning the models to achieve better results. Lastly, recall that the translator model receives as input ads in their basic form (i.e., after preprocessing, see Table 2), 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., realizing words from lemma form using Simplenlg, 2019).

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 our models.

Funding All work was funded through the authors' salaried employment.

Data availability statement Experimental results will be made available upon reasonable request.

Code availability Our code will be publicly available upon acceptance.

Declarations

Conflict of interest All authors are employees of Microsoft, owner of Bing.

References

Ayana, S. S., Liu, Z., & Sun, M. (2016). Neural headline generation with minimum risk training. Preprint retrieved from arXiv:1604.01904.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Preprint retrieved from arXiv:1409.0473

Belanche, D., Flavián, C., & Pérez-Rueda, A. (2014). Measuring Behavior: The influence of arousal on advertising effectiveness.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3,* 1137–1155.

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 379.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for english words (anew): Instruction manual and affective ratings*. Citeseer: Tech. rep.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., & Roossin, P. (1988). A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics*, Association for Computational Linguistics.

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning, 11,* 81.

Chebat, J.-C., Gelinas-Chebat, C., Hombourger, S., & Woodside, A. G. (2003). Testing consumers' motivation and linguistic ability as moderators of advertising readability. *Psychology & Marketing, 20,* 599–624.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, ACM.

Chiang, D. (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.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. Preprint retrieved frm arXiv:1409.1259

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. Preprint retrieved from arXiv:1406.1078

Conway, M. A., Anderson, S. J., Larsen, S. F., Donnelly, C. M., McDaniel, M. A., McClelland, A. G., et al. (1994). The formation of flashbulb memories. *Memory & Cognition, 22,* 326–343.

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, ACM.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint retrieved from arXiv:1810.04805

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., et al. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems,* 13042–13054.

Dorr, B., Zajic, D., & Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, Association for Computational Linguistics (pp. 1–8).

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., et al. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation, 25,* 127–144.

Gai, K., Zhu, X., Li, H., Liu, K., & Wang, Z. (2017). Learning piece-wise linear models from large scale data for ad click prediction. Preprint retrieved from arXiv:1704.05194

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review, 47*(1), 1–66.

Garcia-Molina, H., Koutrika, G., & Parameswaran, A. (2011). Information seeking: Convergence of search, recommendations and advertising. *Communications of the ACM, 54,* 121–130.

Gottron, T. (2008). Content code blurring: A new approach to content extraction. In *2008 19th international workshop on database and expert systems applications*, IEEE, (pp. 29–33).

Gould, D., & Krane, V. (1992). The arousal–athletic performance relationship: Current status and future directions.

Graepel, T., Candela, J. Q., Borchert, T., & Herbrich, R. (2010). Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. Omnipress.

Grewal, P., & Alagaratnam, S. (2013). The quality and readability of colorectal cancer information on the internet. *International Journal of Surgery, 11,* 410–413.

Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). Deepfm: a factorization-machine based neural network for ctr prediction. Preprint retrieved from arXiv:1703.04247

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003). Dom-based content extraction of html documents. In *Proceedings of the 12th international conference on World Wide Web* (pp. 207–214).

Gupta, S., Kaiser, G. E., Grimm, P., Chiang, M. F., & Starren, J. (2005). Automating content extraction of html documents. *World Wide Web, 8*(2), 179–224.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences, 5,* 394–400.

Holbrook, M. B., & Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of Consumer Research, 9,* 132–140.

Jeong, E. J., & Biocca, F. A. (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, 28,* 285–291.

Juan, Y., Zhuang, Y., Chin, W.-S., & Lin, C.-J. (2016). Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM conference on recommender systems*, ACM.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*.

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences, 15,* 241–251.

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition, 31,* 1169–1180.
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 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*.
Koehn, P., Och, F. J., & Marcu, D. (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.
Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. In *CIKM*, ACM.
Lang, A., Dhillon, K., & Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcasting & Electronic Media, 39,* 313–327.
Lee, W., Xiong, L., & Hu, C. (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, 31,* 819–827.
Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out,* 74–81.
Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. Preprint retrieved from arXiv:1908.08345
Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. Preprint retrieved from arXiv:1508.04025
Mohanty, S., Leader, A. E., Gibeau, E., & Johnson, C. (2018). Using facebook to reach adolescents for human papillomavirus (hpv) vaccination. *Vaccine, 36,* 5955–5961.
Nallapati, R., Zhai, F., & Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI conference on artificial intelligence*.
Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. Preprint retrieved from arXiv:1602.06023
Nirenburg, S., Carbonell, J., Tomita, M., & Goodman, K. (1994). *Machine translation: A knowledge-based approach*. Morgan Kaufmann Publishers Inc.
Nyberg, E. H., & Mitamura, T. (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.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (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.
Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. Preprint retrieved from arXiv:1705.04304
Peters, M. E., & Lecocq, D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22Nd international conference on world wide web* (pp. 89–90).
Preventive healthcare. (2019). https://www.healthcare.gov/preventive-care-adults/
Pryzant, R., Basu, S., & Sone, K. (2018). Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 125–135).
pyltr: Python learning to rank (ltr) toolkit. (2019). https://pypi.org/project/pyltr/
Pytorch. (2019). https://pytorch.org/
Rettie, R., Grandcolas, U., & Deakins, B. (2005). Text message advertising: Response rates and branding effects. *Journal of Targeting, Measurement and Analysis for Marketing, 13,* 304–312.
Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. Preprint retrieved from arXiv:1509.00685
Shan, Y., Hoens, T. R., Jiao, J., Wang, H., Yu, D., & Mao, J. (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.
Shapiro, S., & MacInnis, D. J. (2002). Understanding program-induced mood effects: Decoupling arousal from valence. *Journal of Advertising, 31,* 15–26.
Simplenlg. (2019). https://github.com/simplenlg/simplenlg
Sluban, B., & Grčar, M. (2013). Url tree: efficient unsupervised content extraction from streams of web documents. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 2267–2272).
Song, D., Sun, F., & Liao, L. (2015). A hybrid approach for content extraction with text density and visual importance of dom nodes. *Knowledge and Information Systems, 42*(1), 75–96.

spacy: Industrial-strength natural language processing. (2019). https://spacy.io/
Statista: Global no.1 business data platform. (2019a). https://www.statista.com/outlook/219/109/search-advertising/united-states#market-revenue
Statista: Global no.1 business data platform. (2019b). https://www.statista.com/outlook/219/100/search-advertising/worldwide#market-revenue
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*.
Takase, S., Suzuki, J., Okazaki, N., Hirao, T., & Nagata, M. (2016). Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1054–1059).
Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. *PressAcademia Procedia, 5*(1), 205–213.
Textstat: package to calculate statistics from text. (2019). https://pypi.org/project/textstat/
Vader-sentiment-analysis. (2019). https://pypi.org/project/vaderSentiment/
Wang, T., Bian, J., Liu, S., Zhang, Y., & Liu, T.-Y. (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.
Weaver, W. (1955). Translation. *Machine Translation of Languages*.
Wikipedia. Common medical symptoms (2019). https://en.wikipedia.org/wiki/List_of_medical_symptoms
Wikipedia: The kemeny–young method. (2020). https://en.wikipedia.org/wiki/Kemeny-Young_method#Calculating_the_overall_ranking
Woodsend, K., Feng, Y., & Lapata, M. (2010). Generation with quasi-synchronous grammar. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 513–523).
Xu, S., Yang, S., & Lau, F. (2010). Keyword extraction and headline generation using novel word features. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 24).
Yan, L., Li, W.-J., Xue, G.-R., & Han, D. (2014). Coupled group lasso for web-scale ctr prediction in display advertising.
Yang, D., Chen, J., Yang, Z., Jurafsky, D., & Hovy, E. (2019). Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3620–3630).
Yom-Tov, E. (2018). Clinically verified pre-screening for cancer using web search queries: Initial results. *CoRR*.
Yom-Tov, E., Brunstein-Klomek, A., Mandel, O., Hadas, A., & Fennig, S. (2018). Inducing behavioral change in seekers of pro-anorexia content using internet advertisements: Randomized controlled trial. *JMIR Mental Health, 5,* e6.
Yom-Tov, E., Muennig, P., & El-Sayed, A. M. (2016). Web-based antismoking advertising to promote smoking cessation: A randomized controlled trial. *JMIR, 18,* e306.
Yom-Tov, E., Shembekar, J., Barclay, S., & Muennig, P. (2018). The effectiveness of public health advertisements to promote health: A randomized-controlled trial on 794,000 participants. *npj Digital Medicine, 1,* 1–6.
Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine, 13*(3), 55–75.
Youngmann, B., Yom-Tov, E., Gilad-Bachrach, R., & Karmon, D. (2020). The automated copywriter: Algorithmic rephrasing of health-related advertisements to improve their performance. *Proceedings of The Web Conference, 2020,* 1366–1377.
Zhang, R., Guo, J., Fan, Y., Lan, Y., Xu, J., Cao, H., & Cheng, X. (2018). Question headline generation for news articles. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 617–626).
Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., & Gai, K. (2018). Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, ACM.

Authors and Affiliations

Brit Youngmann[1] · Elad Yom-Tov[1] · Ran Gilad-Bachrach[1] · Danny Karmon[2]

Elad Yom-Tov
eladyt@microsoft.com

Ran Gilad-Bachrach
rani.gb@gmail.com

Danny Karmon
Danny.Karmon@microsoft.com

[1] Microsoft Research, Herzliya, Israel

[2] Microsoft Healthcare NExT, Herzliya, Israel