AdBERT: An Effective Few Shot Learning Framework for Aligning Tweets to Superbowl Advertisements

Debarati Das[1], **Roopana Chenchu**[1], **Maral Abdollahi**[2], **Jisu Huh**[2] and **Jaideep Srivastava**[1]
Department of Computer Science, University of Minnesota, Twin Cities[1]
Hubbard School of Journalism and Mass Communication, University of Minnesota, Twin Cities[2]
{das00015,vuppa007,abdol022,jhuh,srivasta}@umn.edu

Abstract

The tremendous increase in social media usage for sharing Television (TV) experiences has provided a unique opportunity in the Public Health and Marketing sectors to understand viewer engagement and attitudes through viewer-generated content on social media. However, this opportunity also comes with associated technical challenges. Specifically, given a televised event and related tweets about this event, we need methods to effectively align these tweets and the corresponding event. In this paper, we consider the specific ecosystem of the Superbowl 2020 and map viewer tweets to advertisements they are referring to. Our proposed model, AdBERT, is an effective few-shot learning framework that is able to handle the technical challenges of establishing ad-relatedness, class imbalance as well as the scarcity of labeled data. As part of this study, we have curated and developed two datasets that can prove to be useful for Social TV research: 1) dataset of ad-related tweets and 2) dataset of ad descriptions of Superbowl advertisements. Explaining connections to Sentence-BERT, we describe the advantages of AdBERT that allow us to make the most out of a challenging and interesting dataset which we will open-source along with the models developed in this paper.

1 Introduction

The joint consumption of television programming and social media participation has become increasingly popular, leading to the rise of Social TV ecosystems (Proulx and Shepatin, 2012; Benton and Hill, 2012; Cesar and Geerts, 2011). Twitter has become an integral outlet for TV viewers, with a whopping 85% of users tweeting while watching television programming (Midha, 2014). Marketers, television networks, and social media platforms have explored this rising potential of Social TV ecosystems. For example, Twitter and content providers on television networks collaborated recently (Crook, 2016) to create social TV experiences, and companies such as Nielson (Talkwalker, 2020) are investing in technologies to quantify and analyze social TV audiences.

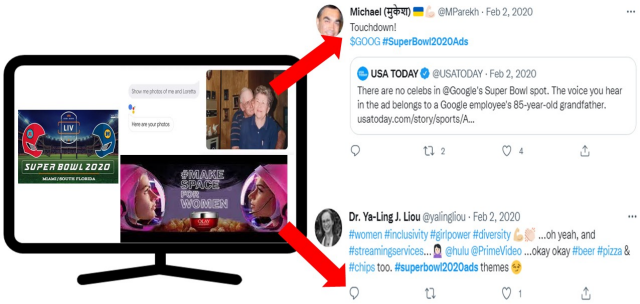

Figure 1: Mapping the the tweets referring to advertisements telecast during the Superbowl event. For e.g., the tweet mentioning *#diversity* and *#inclusivity* is mapped to the advertisement by `Olay` featuring the hashtag *#makespaceforwomen*.

Social TV research is still in its infancy stage (Liaukonyte et al., 2015). However, a few studies (Diakopoulos and Shamma, 2010; Fossen and Schweidel, 2017) have already explored the impact of television on social media word of mouth (WOM) and "impactful" factors (celebrity presence) that influence the volume of social media WOM. Similarly, identifying "attention-grabbing" moments in media (e.g., the performance of the speaker during the presidential debate or a funny ad during the Superbowl), can help gauge the reaction of the viewers'. Hence, it becomes necessary to build tools that capture these "attention-grabbing" moments and analyze the subsequent responses. These tools are not only necessary for program recall and re-contextualization (Wang, 2006), but also for the design of more personalized recommendations in the future. (Pyo et al., 2014).

With a large amount of social buzz generated online, analyzing the responses of the viewers towards televised events has now become much easier as opposed to earlier slow and costly methods that involved surveying the viewers. However, this also comes with its technical challenges. For instance, given a televised event and an associated set of social media posts, an approach that effectively maps the posts to the parts of the event they are referring to (Figure 1) is necessary. This raises two follow-up questions: (a) discretely atomizing the event into segments and (b) identify if the tweet focuses on a specific event segment or the event as a whole. For ex-

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 171–179

ample, tweets can be related to the commercials during the break or the game as a whole during the Superbowl broadcast. Therefore, a method that can align the tweets and their related televised events is an essential building block toward answering fundamental questions regarding the event's influence on the viewer's social TV activity. Machine learning based methods towards this end have attempted event segmentation (Galley et al., 2003); however, they analyze events and tweets independently. This is a big drawback as the event influences the viewer's response; hence there is a need to *jointly* model tweets ad televised content information.

Our research study considers the specific social TV advertising ecosystem during the high-stakes Super Bowl sporting event. In this event, since the ads telecast and audience responses are on different media channels (i.e. TV and Twitter, respectively) over a fixed time, we view the problem as a closed system consisting of two interacting sets - the set of stimuli (advertisements) and the set of responses (tweets). Our research focuses on modelling the function that maps these two sets to each other.

$$\forall a \in A \text{ (Set of all ads)}$$
$$\forall t \in T \text{ (Set of all tweets)}$$

Estimate a function $f : T \rightarrow A$ *where the mapping is 1-1.*

Tweet to Ad mapping is non-trivial problem as the viewer's tweet could be about multiple aspects of the advertisement, such as its creative elements or the brand making the advertisement. For example, if we consider the tweet,*"The pepsi ad was so amazing"*, this is a simplistic case as it is easy to map that the viewer is talking about the advertisement by `Pepsi`. However, in the case of another tweet, *"Mc hammer is still making money with songs low key"*, it is not easy to understand that this tweet is even ad-related (MC Hammer is a celebrity who featured in the `Cheetos` Advertisement). Moreover, we are trying to capture ad-related tweets against the general noise of the Superbowl-related tweets. In this setting, WOM is much less for ad-related tweets (limited representation) and the viewers are also likely to talk about some ads more than others (class imbalance).

Our mapping methodology **AdBERT** is an effective few-shot learning framework that establishes semantic relatedness between an advertisement and a tweet under the constraints of class imbalance and limited class representation. Once this mapping is established, it can be used as an essential building block in an audience engineering pipeline that can help incorporate a feedback loop to an advertisement and aid in downstream tasks such as ad-engagement measurement and sentiment analysis. As a by-product of our experiments, we also developed a manually annotated rich dataset of ad-related tweets and a manually annotated dataset of Superbowl ad descriptions, which can be used for further research in social TV literature.

2 Related Work

With the rise of social TV technologies, research has been done to examine how media multitasking affects viewers' response to advertisements and how advertisers can leverage this behaviour (Hu et al., 2017). (Lewis and Reiley, 2014) find a sudden increase in online searches for brands shown during Superbowl commercials immediately after the ad is telecast. While the aligning of real-time social media responses to TV advertising has been explored in recent years (Hill et al., 2012), their methods to map the tweets to the advertisement is based on the underlying assumption that a person tweets about an advertisement as soon as they see it, which need not be true (Murphy et al., 2006). Our proposed method relies on content mapping, which would *capture tweets about an advertisement irrespective of its time of airing*.

Though (Hu, 2021) consider the primacy effect, their topic-model based approach method cannot be applied to televised segments for which no transcripts are readily available. Advertisements broadcast on TV are usually tiny time segments for which auto-generated transcripts are not meaningful, as they could be theme songs or even a catchphrase. However, even this kind of short TV content is impactful enough to generate significant WOM. Our approach to solving this correspondence problem with its unique challenges draws inspiration from some previous research works (Devlin et al., 2018; Chang et al., 2019; Thakur et al., 2020) which use different encoders for pairwise sentence scoring tasks and (Reimers and Gurevych, 2019) which inspires the idea for joint learning. Our approach aligns tweets with their corresponding TV advertisement through *jointly learning from both the advertisement information and the tweets*.

3 Superbowl 2020 Dataset

Ad Name	Ad Description
`Audi`	Maisie Williams, Frozen, etron, sportback, traffic, letitgo
`Doritos`	Sam Elliott, Lil Nas X, Old Town Road, cowboy, cool ranch dancer, billy ray cyrus, wild west, wild-wildwest, makeyourmove
`Weather Tech`	pets, golden retriever, dog

Table 1: Subset of the created ad information dataset, that contains descriptive phrases or words describing each advertisement in the Superbowl 2020.

3.1 Data Collection

To validate our idea computationally, our objective was to collect a data set that would provide us with a high density of advertisement-related tweets. This meant that a timed sporting event where a lot of advertisements are shown to consumers (who happen to respond to these advertisements) would be perfect for our study. Hence, the Super Bowl 2020 event was chosen as a use case because it is a high-stakes national sports event in the

US watched by a massive audience. This event attracts multiple advertisers who spend millions of dollars to place their ads during this game to attract consumers' eyeballs and spark social media conversations about their ads and brands.

We collected tweets using the Twitter streaming API via the AIDR (Imran et al., 2014) tool from the start of the broadcast (Feb 2nd, 5:30 PM CST) to the end of the day. For this purpose, we used a set of event-related keywords (#superbowlads, Superbowl 2020 etc.) and brand-related keywords (Nike, Pepsi, Olay, etc.). While the data was being collected, the search terms on AIDR were modified in real-time to include words and catchphrases ad-specific to the Super Bowl. The underlying idea is that the audience could be reacting to the brand's message (e.g. *#makespaceforwomen* is a catchphrase of the commercial broadcast by `Olay`) or specific elements of the commercial (e.g. celebrity *Katie Couric* was present in the `Olay` commercial). This was done to ensure that most tweets mentioning ad-specific features were collected. This collection is preferable to scraping user responses to online advertisements as such a method would be bottle-necked by fewer responses to each advertisement.

3.2 Data Preparation

Firstly, we create an **ad information dataset**, a subset of which is shown in Table 1. To create this dataset, three authors watched all of the Superbowl advertisements and made lists of phrases describing unique elements (celebrity, hashtag, tagline, etc.) they noticed in each ad. These lists were then combined to create a comprehensive set of phrases that describe each advertisement from the "annotator's point of view". These ad-related phrases are intended to be unique with respect to each advertisement, to differentiate ads as best possible and are agreed upon by all three authors.

Secondly, we prepare the **tweet-ads dataset**. As most of the originally collected data (around 1.1 million tweets) were event-related tweets, we had to first filter the general Superbowl-related noise to capture the candidate ad-related tweets to be used as training data. After removing the Twitter-specific symbols and artifacts during the initial tweet pre-processing stages, we remove retweets and duplicate tweets to retain only original tweets made by users. To narrow down on candidate tweets that are possibly ad-related, we developed some heuristics (e.g. checking for the presence of brand names). Another heuristic relied on the ad information dataset collected (mentioned above) and checked for the presence of a high degree of overlap between ad-related information and the tweet by using the Jaccard Index measure (Niwattanakul et al., 2013). For example, some of the phrases that describe the brand `Olay` in the ad information dataset are *{Olay, #makespaceforwomen, Katie Couric, Lilly Singh}*. This kind of Jaccard-based heuristic could capture candidate tweets mentioning any of these ad-related features.

A random sample of these candidate tweets was chosen for manual labeling. The tweets were labeled such that each tweet was assigned to the advertisement it referred to or labeled as "none" if the tweet was Superbowl related. Tweets mentioning multiple ads were disregarded in the sample.

For this annotation task, three authors went through a common training session, where it was agreed that the annotation would be based on the common ad information dataset (Table 1) as well as their own personal notes on viewing the commercial. This annotation task involves matching the tweets to the nearest advertisement given the mention of specific elements in the tweet. Since the advertisements are quite different in terms of these elements, the degree of subjectivity in this task is low and we did not require multiple annotations per tweet. The only advertisements which were similar were the ones from a common brand, and for these cases, we combined the advertisements to represent one ad class. Statistics of the resulting tweet-ads dataset are given in Table 2.

Collected no of tweet samples	1114931
No of candidate ad-related tweets (post filtering)	111652
No of tweet samples - training	4656
No of tweet samples - test	1165
No of ad categories	61

Table 2: Statistics about the Superbowl 2020 ad-tweets dataset

4 Main Technical Challenges

A tweet could be a response to either the advertisement's creative elements (for example, a cute retriever in the `WeatherTech` ad) or the advertisement as a whole. Therefore, **detecting the ad-relatedness** requires a holistic understanding of the advertisement's content.

Identifying ad-relatedness can be viewed through a *semantic relatedness* lens such that we try to establish a relationship between the tweet and the advertisement description. However, the short length of tweets and their characteristic lingo adds to the complexities of identifying semantic-relatedness. While the tweet is a short sentence, the ad description is a comma-separated list of key phrases or words. Hence a semantic gap exists between the twitter lingo and the advertisement descriptions ("audience-annotator" gap.)

Identifying ad-relatedness can also be seen through the lens of *multi-class classification*, which involves scoring a set of candidate labels given an input context. The Superbowl Dataset shows the unique characteristic of **class imbalance** with 15 popular or controversial commercials having high representation in the dataset and we call these -"majority classes". For example, the `Hulu` advertisement featured Superbowl superstar

Tom Brady and was a viral ad and hence, a "majority" class. Our threshold for a majority class is that the number of samples for that class should be at least 40. 46 other commercials exist in our data with lesser than 40 samples for the model to train on and understand patterns in these cases. We call these classes - "minority classes". Each class in our training data also suffered from **limited representation**, with the average number of samples in a majority class being 122 and in a minority class being 17.

5 Experiments

From a *semantic relatedness perspective*, we can try to map the text and ad information into a common feature space wherein a dot product, cosine or (parameterized) non-linear function is typically used to measure their similarity.

SentenceBERT (Reimers and Gurevych, 2019) is a bi-encoder model, which applies BERT independently on the two inputs, followed by mean pooling on the output to create separate fixed-sized sentence embeddings. As the representations are separate, the bi-encoders is able to cache the encoded candidates and reuse these representations for each input resulting in faster prediction times than cross-encoders. However, The tweets and ad description information in our dataset are not in the same vector space because the tweet has a sentence structure, while the advertisement information is a set of key phrases describing the ad from the annotator's point of view.

Therefore we consider the *multi-classification perspective* where we can try to score a set of candidate ad descriptions given an input tweet. This kind of multi-class classification can be done via Classical Machine Learning approaches (Debole and Sebastiani, 2004) such as **Logistic Regression** and **Multilayer Perceptron (MLP)** with TF-IDF vectorization of features. In these approaches, words characteristic to an ad are given greater weight than words that frequently appear across all the ads. Our implementation of the MLP has 12 hidden layers each with dimension of 6000. The model is trained for categorical entropy loss with a batch size of 20 and number of epochs as 50.

Deep learning based methods like **BERT** (Devlin et al., 2018) uses a cross-encoder (Wolf et al., 2019; Vig and Ramea, 2019) where a special SEP token separates the input and label candidate and multi-head attention is applied over all input tokens. In our implementation of BERT for multi-class classification, we fine-tune (Sun et al., 2019) the pre-trained 'Bert-base-uncased' model with 12 layers from Transformers library (Wolf et al., 2019) to identify if a tweet can be identified as related to a Super Bowl commercial or not. If the tweet is "Superbowl event related" and does not relate to any ad, it is categorized as a 'none' class. Else, the tweet is classified as 'ad-related'. For all the tweets classified as ad-related, we compute the embeddings from BERT and run a softmax on similarity scores to identify the ad class. The model is trained on 4656 tweets and 61 classes. We use a batch size of 32, a learning rate of 2e-5 and the number of epochs as 4. We also use the epsilon parameter eps with a value 1e-8 to prevent any division by zero in the implementation.

In our implementation of SentenceBERT, we fine-tune the pre-trained "nq-distilbert-base-v1" model using the joint learning setup described in Section 6 and using cosine similarity loss. We use number of epochs as 30, warmup_steps as 100 and evaluation_steps as 500. During test time, we compute the maximum cosine similarity of the input tweet against all of the ad descriptions to get the ad class assigned to the tweet, but with the embeddings obtained from the fine-tuned SentenceBERT model.

6 AdBERT : Proposed Joint Learning Approach

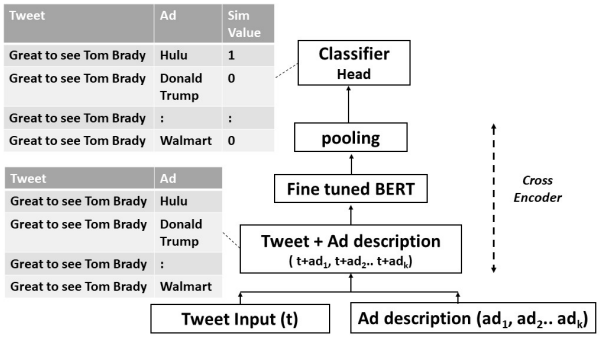

Figure 2: AdBERT Architecture

Our AdBERT approach frames the multi-class classification problem of mapping a tweet to its respective advertisement as a binary classification and semantic relatedness task. As we faced a problem of a limited labeled dataset, we required a better training signal from our dataset. In order to solve this problem, we use an approach utilizing class verbalizers as seen in similar research works for few shot learning (Aly et al., 2021; Pappas and Henderson, 2019; Obamuyide and Vlachos, 2018). In our case study, we propose learning from both the tweet as well as the textual descriptions of each ad class, which is a part of our ad information dataset (Table 1). This means that instead of using label IDs as we did in earlier experiments with BERT, we concatenate tweet text with contextual descriptions about the ad labels. The key phrases of the ad description are concatenated together into a single sequence, which is the contextual description of the ad.

Specifically, the input to the model is a $<$*tweet, ad-description*$>$ pair, and the output is either 1 (if the tweet is related to the ad in the included ad-description) or 0 (otherwise). Therefore, given N tweets and K ad cat-

		Without Ad Information								
		LogReg			MLP			BERT		
	#classes	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Majority	15	0.88	0.84	0.85	0.94	0.76	0.84	0.59	0.72	0.64
Minority	46	0.66	0.52	0.58	0.55	0.36	0.43	0.60	0.49	0.53

Table 3: Given a multi-classification setting where the input is tweet information and the output is ad class, this table reports the weighted average Precision, Recall and F1 score metrics of each model, grouped by majority and minority classes.

egories, the AdBERT model would be trained on NK instances. For each tweet in the original dataset, $K-1$ tweet-ad pairs correspond to negative combinations, and one pair corresponds to the positive class label. The hyperparameters for training remain the same as that used in our experiments with BERT. This architecture is illustrated in Figure 2.

This kind of joint learning training strategy is able to handle the **class imbalance** problem, as the model also learns from the "negative" combinations. The training strategy we describe makes no assumptions about the number of ad categories and is easily extensible. Including new ad categories or adapting to newer ad themes would only require a modification in the ad descriptions with little to no fine-tuning of the classifier architecture. We also do not need to handle explicitly the "not ad-related" case here, as tweets not referring to any ad are automatically classified as 0 in all cases.

The cross encoder in AdBERT takes as input to the network both the tweet and the ad description separated by a SEP token and multi-head attention, is applied across all tokens of the inputs. Compared to a bi-encoder, the cross-encoder offloads the similarity computation to the self-attention matrices and hence is able to better learn to identify ad-relatedness. This implies that both inputs are compared simultaneously and helps solve the **ad-relatedness** problem.

The problem task reformulation we suggest, where we append the label information to the tweet and assist the cross encoder, also solves the **limited representation** problem, thus allowing our model to behave as an effective few-shot learning framework.

7 Results and Discussion

7.1 Quantitative Analysis

We implement the Logistic Regression, MLP and BERT models described in Section 5, where the only input to the model is the tweet information, and the output is the ad class. Table 3 reports the weighted average precision, recall and F1 score metrics of each model, grouped by the majority and minority classes for this multi-classification setting. In the second round of experiments, we implement our benchmarks but supplemented with ad information as per the joint learning strategy described in Section 6. Table 4 reports the metrics of each model, grouped by the majority and minority classes for this setting. Our model, **AdBERT** is a joint learning strategy using a modification of BERT, where the model learns from both the tweet and the ad descriptions.

In our models, we argue that recall is the more important performance metric than precision, given our focus on identifying all true ads. This is because, in the context of Twitter, ad mentions are rare with less than 1% of all tweets even mentioning ad names, with our dataset further highlighting that. For these reasons, we argue that while precision is relevant, it is not critical since false positive ads can be filtered out in downstream tasks, so there is limited harm in falsely identifying ads while there is significance in correctly identifying ads which may not be readily identified using current methods.

In the setting where there is no ad information (Table 3), we observe that Logistic Regression (0.84 Recall) and MLP (0.76 Recall) do well when it comes to prediction of the majority classes. This must imply that there are inherent data patterns in the tweets that can be captured just using TF-IDF features. However, with minority classes both models do quite poorly (0.52 Recall for LogReg and 0.36 Recall for MLP) and cannot handle the class imbalance or limited representation problem. BERT in the multi-class classification setting is comparable (0.72 Recall) to the classical machine learning models with the majority classes.

In the setting where we include ad information (Table 4), we see that the performance of the classical machine learning models goes down as expected. Classical models are known to be sensitive to class imbalance (Atla et al., 2011; Mirylenka et al., 2017; Santiago et al., 2012; Cervantes et al., 2017) and with the joint learning strategy, there is an increase in the size of training data and class imbalance and noise become more pronounced.

In the earlier experiment with BERT, we used just the tweet input, so the cross-encoder in BERT could not be completely harnessed to map the relationship between the tweet and the ad labels. Therefore, the joint learning strategy of **AdBERT** shows very high performance across all metrics across both majority and minority classes. **AdBERT** also does much better than SentenceBERT in the joint learning setting with our data (Recall of 0.75 vs 0.41 for minority classes). This is because the cross-encoder offloads the similarity

		With Ad Information											
		LogReg_JL			MLP_JL			SentenceBERT			AdBERT		
	#classes	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Majority	15	0.66	0.32	0.43	0.67	0.34	0.45	0.65	0.73	0.68	**0.91**	**0.91**	**0.91**
Minority	46	0.42	0.23	0.29	0.36	0.15	0.21	0.39	0.41	0.39	**0.74**	**0.75**	**0.74**

Table 4: Given a binary classification setting where the input is tweet information concatenated with ad description and the output is 1 (ad-related) or 0 (not-ad-related), this table reports the weighted average Precision, Recall and F1 score metrics of each model, grouped by majority and minority classes. AdBERT outperforms all other baselines.

computation to the self-attention matrices in all the layers and can better identify ad-relatedness as compared to the bi-encoder of SentenceBERT. Bi-encoder based methods usually achieve lower performance than the cross-encoders method and require a large amount of training data. The reason is cross-encoders can compare both inputs simultaneously, while the bi-encoders have to independently map inputs to a meaningful vector space which requires a sufficient amount of training examples for fine-tuning. The cross-encoder approach is typically not computationally feasible, but it is in this scenario, as the number of ad labels is much less than the number of tweets.

7.2 Qualititative Analysis

This section discusses the different types of tweet-to-ad correspondence we observed in our Superbowl 2020 Dataset and how AdBERT handles them.

When tweet mentions the advertiser's product : In many cases, the tweet responses directly mention the advertising brand or the product of the advertiser. Consider an example tweet, *"pepsi is totally copying #nooriginality"*. AdBERT is easily able to establish this kind of mapping, with the tweet directly mentioning `Pepsi` in its response. TF-IDF based models would also be effective for these cases.

When a tweet is about advertisement's creative elements : Sometimes the tweets are motivated by the creative elements in the commercial, such as a celebrity's presence. In these cases, the tweet content is not enough to map the tweet to the correct advertisement, and additional commercial-related information is necessary to establish context for the mapping. Consider the tweet, *"gotta let it go doritos right away"* to which our model gives `Doritos` a score of 0.98 F1 and `Audi` a score of 0.95 F1. This happens because the `Audi` commercial featured actor Maisie Williams singing "Let it Go", and this aspect of the commercial is learned from the ad information (Table 1). As a result of the combined contextual BERT embeddings of the tweet response and ad information, `Doritos` has a higher probability, and the tweet is eventually mapped to the `Doritos` commercial.

Consider another tweet, *"@google almost got canine cancer! who is one actual sucker for golden retrievers?"*. Our model maps this tweet to the commercial for `Weather Tech`, which featured a golden retriever. Although the word 'google' exists in the sentence, context is given preference over mere word matching, and the AdBERT classifier correctly identifies the appropriate ad mapping.

These examples justify the poor results of TF-IDF based models and establish the need for context-rich models like AdBERT for effective mapping.

When a tweet is about multiple commercials: In our test dataset, we observed several tweets mentioning multiple commercials. For example, the tweet, *"who is the cool ranch doritos with lil nasx or ellen"* is referring to two advertisements : `Doritos` featuring celebrity Lil Nas X, and `Amazon Alexa` featuring celebrity Ellen Degeneres. AdBERT is able to understand that most of this tweet is about the `Doritos` ad and gives it a score of 0.98 F1 vs `Amazon alexa` with 0.67 F1. This is because of the combined learning from ad information input and tweet content input.

When a tweet is about similar commercials : AdBERT demonstrates a certain degree of confusion when the tweet is about similar commercials (when you cannot distinguish based on brand or commercial content). This is evident in the case of the tweet, *"good on you michelob"*. Our model assigns similar scores to commercials `Michelob 6 for 6-pack` (0.87) and `Michelob lite` (0.98) for this tweet. This is probably because the tweet only mentions the brand name, and there is no further information to narrow it down. Similarly, tweets corresponding to `Bud light seltzer` and `Tide bud knight` show a degree of overlap in classification. This is perhaps because both ads are associated with the word 'bud'.

Table 5 describes the true annotated label vs the model predicted ad label for some examples from our tweet-ads dataset and further illustrates the impact of including ad information for joint learning. The ad information that is appended jointly with the tweet text, describes creative elements in the advertisements (such as a celebrities, taglines, etc.) even while the tweet might not have any direct reference to the ad class. For example, *"post malone absolutely best ad so far"* cannot be mapped to an ad category without additional context that the celebrity Post Malone was present in the `Budlight-seltzer` ad. Table 5 also illustrates how multiple minority category advertisements were mapped accurately by AdBERT.

Tweet text	True ad class	Predicted ad class	Predicted Ad category	Ad information
post malone absolutely best ad so far	BudLight Seltzer	BudLight Seltzer	Majority	bud light bud light seltzer post malone anheuser-busch inbev hard seltzer postmalone budlightbud-weiser alcohol
john cena with a super bowl wrap i'm ready to let it go man	Michelob	Michelob	Minority	anheuserbusch inbev mich-elob ultrabeer jimmy fal-lon working gym john cena usain bolt brooks koepka kerri walsh jen-nings worth enjoy low carbs jimmyfallon usain-bolt workingout gymbody alcohol
so far companies have spent millions dollars in ads starring people like molly ringwald	Avocados-from-Mexico	Avocados-from-Mexico	Minority	molly ringwald avocados mexico avonetwork avo-carriermollyringwald food
mc hammer still making money with songs low key	Cheetos	Cheetos	Minority	mc hammer cant touch popcorn cheetos cheetos thing cheetle cant-touchthismchammer food
arya stark nostalgic that frozen winter is never coming	Audi	None	Majority	maisie williams frozen etron sportback traffic letitgo
someone please explain josh jacobs win?	None	Kia	Minority	josh jacobs running back kia seltosraiders give ev-erything joshjacobs kiasel-tos car

Table 5: This table describes the true ad class vs the predicted ad class for some tweets from our tweet-ads dataset. We can observe that jointly learning ad information and the tweet text, led to more successful mapping of the tweets to their ads in both majority and minority represented ad categories.

As a counter example, consider the tweet, *"Arya stark nostalgic that frozen winter is never coming"*. This tweet refers to the character played by actor Maisie Williams in the popular series, Game Of Thrones. "Arya stark" was not included as a relevant ad-related phrase in our ad information data. Since neither the ad description nor the tweet data captured 'Arya stark' as a feature of the `Audi` ad, this tweet did not get classified correctly.

Similarly, *"Someone please explain josh jacobs win"* is annotated as `None` but the model predicts the ad class `Kia`, because Josh Jacobs is a football player mentioned in the ad information for this ad class. This is an ambiguous tweet as it could be related to Josh Jacob's performance in the Superbowl or his racing against the `Kia` car in the advertisement. Thus, false positives and false negatives in the prediction indicate towards issues with using manually annotated class verbalizers.

8 AdBERT used in Downstream Tasks

Our model serves as an essential part of multiple audience engineering pipelines in a social TV setting. In the research by (Lu et al., 2022), the aim is to examine the influence of the viewer's temporary affective states during Superbowl ad exposure. In order to compute the viewer's affective state, a key step is to be able to understand which advertisement impacted the user's affective state, thereby making them tweet in a specific way. This is done using AdBERT, which proves to be superior than time-based tweet-ad alignment. This is because the advertisements are typically very short (10 to 20s) and the user is more likely to tweet much later (Murphy et al., 2006) than during this brief time period. Similarly, in the work by (Kim et al., 2021), ad-related tweets derived through AdBERT are analyzed for evidence of gender-targeted advertising during the Super Bowl.

9 Limitations and Future work

Since our current AdBERT approach uses a fine-tuned Bert-base-uncased model, using fine-tuned BERTweet (Nguyen et al., 2020), which is a pre-trained language model for English Tweets, seems like a suitable next step. AdBERT uses additional information about the ad classes for joint learning. Three authors manually annotate this information in this research, but manually generated class verbalizers heavily depend on domain specific prior knowledge and finding appropriate label descriptions automatically is a challenging research problem that can be further explored. Similar and multiple advertisement mentioning commercials pose a problem in ad-tweet mapping and can be further disambiguated by considering the tweet's timestamp in addition to the tweet content.

The joint learning strategy described in AdBERT can also be extended to other social TV datasets. For example, in the Social TV ecosystem of Presidential Debates telecast on television, tweets could be mapped to segments of the debates. This could have multiple downstream implications such as viewer stance detection, viewer engagement analysis etc.

10 Conclusion

In this paper, we develop a model, AdBERT, that aligns tweets to the advertisements they refer to in the context of the Social TV ecosystem of Superbowl 2020. This problem is technically challenging because of the difficulties in establishing ad-relatedness of a tweet, class imbalance in the dataset and limited representation for each ad class. We find that framing this multi-class classification problem as a binary classification and semantic relatedness task results in superior F1 performance compared to our baseline models. In the joint learning setting, the model learns from both the input and label information together, leading to better classification even in lesser represented classes. Thus our model generalizes well despite the class imbalance and limited labelling problems in the dataset. AdBERT makes no assumptions about the number of ad categories and is easily extensible. Our model can be highly useful as a step toward incorporating feedback into advertisements and analyzing viewer engagement and attitudes. As a by-product of this research, we also developed a dataset of ad-related tweets and a dataset of ad descriptions of Superbowl ads, which can be used to further Social TV research.

Acknowledgements

We thank Dr. Xinyu Lu (Shanghai International Studies University) and Chandramouli Shama Shastry (Dalhousie University) for multiple enlightening discussions and brainstorming sessions.

References

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging type descriptions for zero-shot named entity recognition and classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528.

Abhinav Atla, Rahul Tada, Victor Sheng, and Naveen Singireddy. 2011. Sensitivity of different machine learning algorithms to noise. *Journal of Computing Sciences in Colleges*, 26(5):96–103.

Adrian Benton and Shawndra Hill. 2012. The spoiler effect?: Designing social tv content that promotes ongoing wom. In *Conference on Information Systems and Technology, Arizona*, pages 1–26.

Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodriguez, Asdrúbal López, José Ruiz Castilla, and Adrian Trueba. 2017. Pso-based method for svm classification on skewed data sets. *Neurocomputing*, 228:187–197.

Pablo Cesar and David Geerts. 2011. Past, present, and future of social tv: A categorization. In *2011 IEEE consumer communications and networking conference (CCNC)*, pages 347–351. IEEE.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2019. X-bert: Extreme multi-label text classification with bert. *arXiv preprint arXiv:1905.02331*.

Jordan Crook. 2016. *Twitter signs deal with NFL to live stream Thursday Night Football*. https://tinyurl.com/76zjdd42.

Franca Debole and Fabrizio Sebastiani. 2004. Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1195–1198.

Beth L Fossen and David A Schweidel. 2017. Television advertising and online word-of-mouth: An empirical investigation of social tv activity. *Marketing Science*, 36(1):105–123.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Shawndra Hill, Aman Nalavade, and Adrian Benton. 2012. Social tv: Real-time social media response to tv advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, pages 1–9.

Yuheng Hu. 2021. Characterizing social tv activity around televised events: A joint topic model approach. *INFORMS Journal on Computing*, 33(4):1320–1338.

Yuheng Hu, Tingting Nian, and Cheng Chen. 2017. Mood congruence or mood consistency? examining aggregated twitter sentiment towards ads in 2016 super bowl. In *Eleventh International AAAI Conference on Web and Social Media*.

Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 159–162.

Eunah Kim, Debarati Das, Roopana Chenchu, Mayura Nene, Jisu Huh, and Jaideep Srivastava. 2021. *Consumer Responses to Gender-Targeted Advertising: Computational Research Analyzing the 2020 Super Bowl Commercials*. Presented at the 2021 American Academy of Advertising.

Randall A Lewis and David H Reiley. 2014. Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on yahoo! *Quantitative Marketing and Economics*, 12(3):235–266.

Jura Liaukonyte, Thales Teixeira, and Kenneth C Wilbur. 2015. Television advertising and online shopping. *Marketing Science*, 34(3):311–330.

Xinyu Lu, Debarati Das, Jisu Huh, and Jaideep Srivastava. 2022. Influence of consumers' temporary affect on ad engagement: A computational research approach. *Journal of Advertising*, 51(3):352–368.

Anjali Midha. 2014. *Study: Exposure to TV Tweets drives consumers to take action - both on and off of Twitter*. https://tinyurl.com/32jcf5u9.

Katsiaryna Mirylenka, George Giannakopoulos, Themis Palpanas, et al. 2017. On classifier behavior in the presence of mislabeling noise. *Data mining and knowledge discovery*, 31(3):661–701.

Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication*, 11(2):522–535.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78.

Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Mike Proulx and Stacey Shepatin. 2012. *Social TV: how marketers can reach and engage audiences by connecting television to the web, social media, and mobile*. John Wiley & Sons.

Shinjee Pyo, Eunhui Kim, et al. 2014. Lda-based unified topic modeling for similar tv user grouping and tv program recommendation. *IEEE transactions on cybernetics*, 45(8):1476–1490.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

José Hernández Santiago, Jair Cervantes, Asdrúbal López-Chau, and Farid García Lamont. 2012. Enhancing the performance of svm on skewed data sets by exciting support vectors. In *Ibero-American Conference on Artificial Intelligence*, pages 101–110. Springer.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Talkwalker. 2020. *Talkwalker acquires Nielsen Social*. https://tinyurl.com/356pbt4t.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Alex Wang. 2006. Advertising engagement: A driver of message involvement on message effects. *Journal of advertising research*, 46(4):355–368.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.