ABSTRACT
Companies invest substantial capital in advertising campaigns for products and services. Advertisements that generate undue controversies can completely destroy an advertising campaign. Given the large investments and the explosively viral nature of the spread of controversies, early detection of potential controversies is of vital importance in deciding the future course of these campaigns. However, it is difficult to estimate the potential of controversies through traditional methods such as customer surveys and market research. In this paper, we develop a controversy detection system based on initial comments on online advertisements posted on YouTube. We extract early YouTube comments on a collection of around 45 Superbowl advertisements. We generate a comprehensive set of over 2500 semantic and linguistic features and evaluate their efficacy in automatically detecting controversial comments. Our results show good accuracy in early detection of controversies. The proposed data-driven approach can complement and greatly aid traditional approaches of market research.

General Terms
Controversy, Classification

Keywords
Controversy, Detection, Classification

1. INTRODUCTION
Advertising campaigns are extremely important for companies to promote their brand, products and services. It has been claimed that advertising perceptions have a significant impact on the customers' decision to accept or reject products and brands [3]. In order to generate brand awareness, increase popularity and tackle intense competition advertising agencies aim to be creative in developing ads that resonate with the majority of the population. Accordingly, a large portion of marketing expenses are spent in developing and distributing advertisements for the television and on the web.

However, in certain cases, the commercials can create controversies - either due to the content or messages that are provided. The presence of controversies can significantly hamper the marketing campaign causing consumers to gain a negative perception or even boycott the product or brand in question[22]. At the same time, controversy can also be used by sellers and manufacturers to “cut through the clutter” and make the presence of the brand felt in an increasingly competitive marketplace[6].

Over the years, television has been a powerful channel for advertising due to the large guaranteed viewership for popular programs. In recent times, this audience-commercial dynamic has changed with the advent of online distribution channels and video-sharing services like YouTube. Unlike television commercials that are ephemeral, once a recording of a televised advertisement is posted on a social media website like YouTube, it not only affords for continued viewership over time but also serve as an invaluable barometer to public reaction towards the advertisements. Furthermore, it also facilitates user sharing of content virally and enables advertisements to gain significant popularity among the masses. On the flip side, the inability to control user generated data in these online social media websites, comes with its own pitfalls like the unauthorized spread of potentially damaging content [24].

One of the most profitable and expensive segment of advertisements is during the Superbowl when commercial spots cost around 4 million for a 30 second spot [12]. This is due to the high volume of viewership 1. Given the high price, it becomes even more important for companies to be able to discern any potential negative publicity from their commercial early in the distribution cycle. In the absence of methods to detect controversial advertisements automatically, companies rely on their test audiences to reveal any potential issues with advertisements. This is obviously not very reliable, since these audiences are a small sample of the general population and there are no well-defined rules in place to determine what makes something controversial. Hence automatic controversy detection would immensely help in marketing, increasing popularity and cultivating a desired brand image. Unfortunately, very little work has been done to capitalize on user generated content to understand the nature,  

1 measured to be around 108 million in 2013
spread and impact of advertisements. Our work aims to fill in this gap. Our objective is to develop an automatic mechanism for early detection of controversy in advertisements using user generated data.

In this paper, we focus on a collection of Superbowl advertisements from 2013 and 2014. We extract early (first 24-hours) comments for each of these videos from YouTube. We first analyze the growth and evolution of the user comments for these advertisements. We study different statistical and temporal evolution features to see if they help us to differentiate non-controversial advertisements from controversial ones. Subsequently, we focus on the problem of automatically detecting controversies based on initial user comments. We extract semantic and linguistic features from the user comments for the advertisements and use them to construct a classifier to automatically detect controversial comments achieving a good accuracy of around 0.83 (area of the ROC curve). Next, we identify terms that are highly associated with controversies from the user comments. Finally, we demonstrate how the classifier can be used on a hold-out testset to identify if a commercial has a high likelihood of generating controversies. Our experimental results on hold-out testsets on the 2013 and 2014 Superbowl demonstrate high accuracy in detecting controversial advertisements.

The paper is organized as follows. In the next section we survey some related work in the area. In Section 3, we describe the dataset which comprises of user comments on YouTube clips of Superbowl commercials. Next we carry out an empirical investigation of comments associated with both these categories. In Section 5 we describe the linguistic and semantic features that we extract and the classification technique. In Sections 6 and 7 we go from detecting controversial comments to classifying advertisements.

2. RELATED WORK
There have been several studies aimed at understanding what makes an advertisement controversial or offensive. Fam and others[6] use results from a survey carried out in four different countries of the Asia-Pacific region. Their study found strong correlations between product groups and reasons of offensiveness and were related to individualism/collectivism dimension (people in collectivist as opposed to individualistic societies are more concerned with doing what is deemed fit by the society); Confucian dynamism (placing more emphasis on morality and family cohesion) and religion. In the same line, [3] found that the responses to offensive advertisements vary across cultures. [7] argue that some controversies are an outcome of complex social negotiations that can be understood only in their cultural, commercial and political contexts and not merely by the scientific facts behind them.

To the best of our knowledge, there is almost no prior work on automatic detection of controversies. However, there has been some research done on the cause and effect of controversial advertisements in mass media and psychology [7, 10, 23, 11, 4]. [10] and [23] explore the relationship between advertisements and offensiveness. [23] study the effect of violent images in advertisements using a survey of university students from six countries. They conclude that gender, country, intensity of religious beliefs, economic inclination, and social/political groups produced the strongest reaction. Interestingly, in our study, we automatically detect violence to be often associated with controversy. In addition, we find that features associated with gender, race, religion, economy, politics tend to have higher likelihood of being associated with controversial comments. Like [11], we find that features related to sexuality are highly predominant in controversial ads.

In [4], the authors study how controversy affects conversation. They claim that the increase in controversy does not necessarily translate into increase in conversations. Using data from an on-line news forum they claim that though controversy increases the likelihood of discussion at moderate levels, additional controversy actually decreases the likelihood of conversation. We compare our findings with this work in Section 7.

While there has been very little work to use social media to predict controversy, there has been some work done in the broader context of categorization and classification of user generated data. [13] use classification to detect spam in social tagging systems. [1] show how the chatter from Twitter.com can be used to forecast box-office revenues for movies, [14] utilize classification of tweets to speculate if they are associated with an interesting topic. [8] employ entropy-based classification to predict the characteristics of retweeting activity: whether it is automatic/robotic activity, news-worthy information dissemination, advertising and promotion, campaigns, and parasitic advertisement. [17, 16, 20, 9] classify twitter users based on their tweet content and their social networks. [9] state that topic modeling approaches can be useful as features for short text like tweets but when content information is already large, topic models are less effective compared to simple tf-idf scores. We on the other hand observe that for the classification of short YouTube comments, tf-idf scores turn out to be the most effective feature and give a better classification when compared to topic modeling approaches.

In the context of research on YouTube videos and their associated comments, [19] try to find the characteristics of a typical YouTube video. For instance they claim that negative comments on YouTube elicited many more replies than positive comments. They found the biggest trigger of discussion to be religion. [15] use comments from YouTube and tweets from Twitter to predict the IMDB ratings of movies. Siersdorfer et al. [18] predict the rating of new comments using Support Vector Machine Classifiers. [5] study qualitatively the effect of new media like YouTube and its effect to controversial advertising, especially if the ad is banned in traditional media. Though theirs is a very small study of 10 people they find that new media does seem to create an additional opportunity for user engagement. Unlike this qualitative study using a very small set of users, our quantitative framework takes a much larger user base into account, by automatically analyzing comments of thousands of users.

3. DATASET
To identify past Superbowl commercials that have been controversial, we performed Google searches to get a list of commercials and then had human judges evaluate the advertisements. From this process, we obtained a list of 18 controversial advertisements and 27 non-controversial (con-
4. GROWTH AND EVOLUTION OF CONTROVERSIAL COMMENTS

In this section, we study the overall statistics and growth of comments in both these categories. Superbowl commercials generate tremendous activity on YouTube with lots of comments. We begin our analysis with an empirical study into these comments, both for controversial and non-controversial advertisements. We investigate simple statistical features of the comments associated with each advertisement video to see if these features help us to distinguish controversial advertisements from non-controversial ones.

4.1 Number of Comments

When we consider the total number of comments across the two categories of advertisements (controversial and non-controversial), we observe a significant difference. We see that more than 67% of controversial advertisements have more than 300 comments. On the other hand only 37% of the non-controversial advertisements have more than 300 comments. The average number of comments for controversial advertisements is 546 while for a non-controversial ad the number is 354. This follows intuition that controversy creates popularity. This seems to suggest that controversies do indeed increase conversation, since the user comments are analogous to conversations.

4.2 Distribution of the Number of Words

We investigate the word distributions of comments belonging to controversial advertisements and compare them to comments belonging to non-controversial advertisements. Figure 1 (a) gives the distributions of the number of words in both cases. We observe that controversial advertisements seem to be more expressive, and have a slightly higher probability of having larger number of words in them.

4.3 Average Word-length Distribution

For each comment we compute the average length of words in it. Then we compare the distribution of word-lengths in controversial advertisements to those in non-controversial or control advertisements. Figure 1 (b) shows this comparison. We observe that the two distributions are almost identical for comments from controversial advertisements and those from non-controversial advertisements.

4.4 Emoticons

User generated content like comments on YouTube videos often contain emoticons. The particular emoticons that people use in comments for videos can be indicative of the nature of the videos. We next examine if emoticons can be useful for detecting controversial comments. For this purpose, we extract all the emoticons from comments pertaining to controversial and non-controversial videos. We consider emoticons pertaining to 5 categories - happiness (smile, grin, cheeky, wink and so on), sadness (frown, cry and so on), annoyance, embarrassment and surprise. We observe that the emoticons pertaining to surprise and embarrassment seem to be statistically significantly more in controversial advertisements as compared to non-controversial advertisements ($\alpha = 0.05$). The relative proportions of emoticons pertaining to the different emotions is shown in Figure 2.

4.5 Temporal Evolution of Advertisements

Next, we want to examine the evolution of comments in both the categories. Accordingly, for each posting of an advertisement on YouTube, we extracted the timestamps when each...
Figure 1: (a) Distribution of the number of words in comments from controversial advertisements (in red) vs. non-controversial or control (in blue). (b) Distribution of the average length of words per comments for controversial (in red) and control (in blue).

Figure 2: The bar charts showing the relative proportion of emoticons (in percentage in logarithmic scale) associated with a particular emotion for comments associated with controversial (red) and non-controversial (blue) videos. We see the emotions associated with surprise and embarrassment are significantly more in advertisements associated with controversial videos than non-controversial ones.

5. DETECTING CONTROVERSIAL COMMENTS
First we describe content-based features that can be extracted from the YouTube user comments. Then we show how these features can be used for classifying comments.

5.1 Feature Categories
We extract over two thousand semantic and linguistic style features that can be associated with a comment. These features can be grouped into 5 categories:

- TF-IDF: We put together all the text from the comments, excluding stop-words, and use term frequency-inverse document frequency (TF-IDF) as the weighting schema. TF-IDF allows us to emphasize the words which are most discriminative for a category. These words include emoticons present in the text.

- Latent Topics: Topic modeling approaches discover topics in large collections of documents. The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) [2]. In this work we fit an LDA model to our training documents where each document consists of a comment. We fix the number of topics $T = 100$ empirically by estimating the log likelihood of a model with $T = 50, 100, 150, 200, 250, 300$ on held out data. We choose the default hyper-parameters ($\alpha = 50/T, \beta = 0.01$). They are optimized during the training by using Wallach’s fixed point iteration method [21]. Using collapsed Gibbs sampling for inference, carry out many iterations of the Markov chain (which consists of topic assignment for a token in the training corpus given the assignment of all the other tokens) until the topic assignments seem to potentially converge (at around 2000 iterations). has potentially converged and we can get estimates of the word distribution of topics ($\phi$) and the topic distribution of documents ($\theta$). The estimated distributions $\phi$ and $\theta$ are predictive distributions and are used to infer the topic distribution for each user in our training and test corpus.

Although these measures show some differences between the two categories, they are not sufficient to classify an advertisement into controversial or non-controversial. Next, we focus on extracting content-based features that can provide other signals to differentiate between the two categories.
Figure 3: The temporal evolution of comments over (a) 5 hours (b) 1 day. Each green (blue) dotted line is shows change in the number of comments with time in a controversial (non-controversial) Superbowl advertisement. The gray(black) line shows the average temporal increase in the number of comments in all controversial (non-controversial) advertisements. As we can clearly observe even after just 5 hours, the average rate of commenting in controversial advertisements is higher than the rate of commenting in non-controversial advertisements. This difference becomes more pronounced as time increases as we see in (b).

- Ngram: Taking just single words as features does not help to map the complete information present in the comments into features. For example the bigram ‘not funny’ has a different semantic interpretation and meaning than the unigram ‘funny’. Hence n-grams help to characterize additional linguistic patterns which are not captured by just taking single words or unigrams into account. Therefore, besides using TF-IDF and latent topics as features, we also extract bi-grams and trigrams delimited by whitespaces as additional features.

- Word Statistic features: We include the statistical features that we described in the previous section, such as the number of words and the average length of words in a comment.

- Frequency of Wikipedia controversial issues: Wikipedia contains an entry on issues interpreted as controversial by the web community. It includes issues like abortion, African American, communism, gun control and so on which has been created using the wisdom of the crowd. We extract this list of controversial issues and for each comment, we calculate the number of controversial issues occurring in it and use it as a feature.

Using this set of features, we perform supervised classification to identify controversial and non-controversial comments.

5.2 Classification

We used logistic regression as our predominant classification technique. For classification, logistic regression gives similar results other methods like Random Forest. However, logistic regression has an added advantage in that it enables us to understand the relative importance of features in the two categories. We evaluated different classification techniques

with the features. We used the F-measure, ROC curve and MCC (Matthews Correlation Coefficient) as evaluation metrics.

Feature Comparison: First we compare the performance of each of the five categories of features (tf-idf, ngrams, word statistics features, frequency of controversial issues, latent topics) features when used separately with the performance obtained when all these features are taken together using 10-fold cross validation. The results using a logistic regression classifier are shown in the Figure 4(a). We see that the tf-idf features seem very efficient in detecting controversial comments irrespective of the evaluation metric considered. When we use only tf-idf features, we obtain a performance that is comparable to performance obtained on combining all features. We find that the Wikipedia controversial issues and the word statistics perform very poorly.

Feature Selection: Since tf-idf features seem to give a substantial performance gain, we focus on these features. We examine how the performance changes on reducing the number of features. We reduce the number of features by including only N top features in terms of information gain. The performance comparison is shown in Figure 4(b) when the evaluation metric is the area under the ROC curve. We observe that the best performance is obtained with around 400 features with the area under the ROC curve being 0.822.

Next we perform feature selection for the combination of all features for classification. Again we use information gain for feature selection, taking the top N features for classification and evaluating the classification performance. We used area under the ROC curve for measuring performance. We can observe from Figure 4(c) that the best performance is obtained when the top 700-800 features are selected based on information gain. The area under the ROC curve in this

3We applied other classifiers such as random forest and ob-
scenario is 0.826. Other evaluation metrics give similar results.

5.3 Controversial Terms
For the classification task described earlier, all comments associated with controversial comments were labeled as true and the rest as false. To have an understanding of controversial terms associated with user comments, we ranked the term features in descending order based on their odds ratios. The algorithm predicts a high likelihood of the top features to be contained in controversial advertisements. Some of the high ranked terms are shown in Table 5.3. We discover the main categories of features (and terms) that have a high likelihood of being associated with controversy are: racist, religious, sexual, choice, negative terms, political, violent, humor, finance and abusive. This correlates with earlier observations on controversial topics by [23] and [11].

6. COMMENTS AND CONTROVERSY
In the earlier classification, we assumed that all comments belonging to a controversial advertisement are controversial and all advertisements belonging to non-controversial advertisement are non-controversial. This is because it is a labor intensive process to go through each comment individually and label it as controversial or non-controversial. However, for a better understanding of the relationship between comments and controversy, we need to weaken this strong assumption as it is not necessary that all comments in a controversial (non-controversial) advertisement are controversial (non-controversial). We try to find the fraction \( f \) of controversial comments in controversial advertisements and in non-controversial advertisements. More specifically, we seek to discover if there can be a threshold \( \tau \) such that if \( f \geq \tau \), then the advertisement has a very high likelihood of being controversial.

To find the fraction of controversial comments in a controversial advertisement versus a non-controversial one, we asked human annotators to label specific user comments as controversial or non-controversial. We used the CrowdFlower system for crowd-sourcing this task. We randomly chose a selection of comments from the Superbowl advertisements and assigned the labeling of these comments as the task. We set the language of the task as English and used users only from United States for labeling the comments, since the Superbowl is primarily popular in the United States. In CrowdFlower Gold Standard data is used to test the accuracy of the task. It comprises of few comments manually annotated by experts as controversial or non-controversial. The gold standard data is regularly inserted in the random selection of comments a user is given to annotate and used to test whether the judgment of the user can be trusted or not trusted. For each comment, we ensured that it was labeled by at least 5 annotators and chose only the comments that achieved consensus among the 5 annotators.

We then determined the fraction of controversial comments in each advertisement. Since we had unequal number of comments after labeling, we picked only the advertisements that had at least 50 comments labeled in consensus by the CrowdFlower reviewers. Figure 5 shows the fraction of controversial comments in each advertisement.

Figure 4: The classification task is to classify comments as controversial and non-controversial (a) Performance comparison of the different categories of features for the classification task using Logistic Regression using F-measure, ROC curve and MCC. (b) Performance comparison as the number of features are increased using are under the ROC curve.

\footnote{www.CrowdFlower.com}
sial comments (C-score) in controversial (in red) and non-controversial (blue) advertisements. The size of the bubble signifies the percentage of controversial (or non-controversial) advertisements having that C-score. As we can see in the plot, if an advertisement has a C-score above 0.3 it is highly likely to be controversial i.e. if there are more than 30% controversial user comments pertaining to an advertisement, then there is a high probability (90%) that it is a controversial advertisement.

7. DETECTING CONTROVERSIAL ADVERTISEMENTS

The next task then is to determine whether our proposed method can detect a controversial advertisement. Detecting controversial advertisements is a two step procedure. The first step comprises of classifying the comments associated with that advertisement as controversial or non-controversial. The next step comprises of determining if the fraction of controversial comments (C-score) is above the determined threshold (0.3). If C-score is less than 0.3, then the advertisement is classified as non-controversial and else it is classified as controversial.

Chen and Berger[4] made the claim that controversy does not increase conversations. If we term advertisements having a C-sore value above 0.6 as highly controversial and those between 0.3 and 0.6 as mildly controversial and study the temporal evolution of the highly and mildly controversial ads, then we find that of all the highly controversial Superbowl ads studied, 75% have less than 400 votes after a day. On the other hand of all the mildly controversial Superbowl ads only 33% have less than 400 votes after the day. Though from our data and the user study, it is difficult to say something conclusive in this direction, it does suggest that mildly controversial articles (measured by 0.3 ≤ C-score ≤ 0.6) might be more discussed than highly controversial articles.

7.1 Predictions for Superbowl 2013

For testing our algorithm, we randomly chose 3 controversial and 5 non-controversial advertisements from the dataset as a hold-out test dataset. The rest of data is used for training. For each advertisement we consider only comments posted within 24 hours after the video was posted. For training the classifier, all comments pertaining to controversial advertisements are taken as controversial and all comments pertaining to non-controversial advertisements are taken as non-controversial. Next the comments in the test set are classified into controversial or non-controversial. Advertisements having more than 30% controversial comments (C-score > 0.3) are labeled as controversial, else, they are labeled as non-controversial. The results are shown in Table 2. As you can see from the table, our system correctly predicts whether an advertisement is controversial or not with an accuracy of 100% for the given test set.

7.2 Predictions for Superbowl 2014

We extracted comments on YouTube (for about a day) for 6 advertisements from Superbowl 2014. We tested the percentage of controversial content in these ads. The ads under consideration were: Budlight, Doritos, Maserati, Coke, Chevrolet, Turbo Tax. When we applied our classification model, as shown in Table 3, we found that most advertisements in Superbowl 2014 have a very low C-score value indicating that they have a low probability of stirring a controversy. This corroborates with the findings of newspaper
8. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown how user comments in YouTube can be mined successfully to automatically detect controversies in advertisements. To the best of our knowledge this is the first work in this direction. We have focused on commercials screened during the Superbowl and have constructed a broad set of linguistic and semantic features using early user comments to build a classifier. The classifier was first applied to classify user comments as controversial achieving a good accuracy of around 0.83. Subsequently, we have used crowd-sourced labeled data to determine how the percentage of controversial comments can be used to indicate the likelihood of the commercial being controversial. Our results on two hold-out testsets for the 2013 and 2014 Superbowls have shown high accuracy suggesting that this can greatly aid traditional approaches of market research in determining the likelihood of a commercial generating controversy.

In this work we have only considered the text in the user comments to derive the semantic and linguistic features. We have seen that TF-IDF features are extremely useful in identifying controversies. We wish to extend this analysis and modeling to other types of features including visual, audio and word-ontologies as possible cues for detecting controversies in advertisements. Also, we plan to build a larger repository of manually annotated comments, so that we can do away with the assumption that all comments associated with controversial advertisements are controversial.

9. REFERENCES


Table 3: Table showing the number and fraction of controversial comments (C-score) in advertisements of the test set. The fifth and the sixth column show that for the test set, the algorithm correctly predicts whether an advertisement is controversial or not.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Controversial Comments</th>
<th>Total Comments</th>
<th>C-score</th>
<th>Is Controversial</th>
<th>Is Controversial (Predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ad1</td>
<td>92</td>
<td>374</td>
<td>0.245989305</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ad2</td>
<td>3</td>
<td>68</td>
<td>0.044117647</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ad3</td>
<td>2</td>
<td>46</td>
<td>0.043478261</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ad4</td>
<td>54</td>
<td>464</td>
<td>0.11637931</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ad5</td>
<td>45</td>
<td>356</td>
<td>0.126404494</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>ad6</td>
<td>163</td>
<td>458</td>
<td>0.355895197</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>ad7</td>
<td>363</td>
<td>686</td>
<td>0.529154519</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>ad8</td>
<td>184</td>
<td>600</td>
<td>0.306666667</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>


