Using user reviews to evaluate product features

Andreas Lloyd
Barcelona Graduate School Economics
(Dated: June 26, 2016)

I. INTRODUCTION

Increasing presence of social media type platforms on the Internet have made user reviews of various products more common place. There are numerous websites dedicated to reviewing specific products, including online forums. Some companies use these sources to evaluate and improve their products, and gauge the public's reception.

User reviews are generally useful because it is unlikely that someone would intentionally submit a purposefully positive review or ignore some major negative feature. This may not be the case for professional reviewers, who may have some financial incentive. Similarly, user reviews are likely to address issues that arise over longer use of the product, use in larger groups, or use in unexpected contexts. Furthermore, having a large base of reviews to work with, from customers with diverse backgrounds, diversifies and improves the evaluation, with respect to the points above. The issue with this is that it is a hard task to extract proper information from thousands of reviews, and so a text mining approach is interesting.

This report will explore the application of Latent Dirichlet Allocation (LDA) methods to reviews, to try and extract the key features of reviews. Some sentiment analysis will then be performed to try and gauge the public view of the different aspects of the product that arise. To look slightly at the nature of reviews, some relationship between sentiment and frequency score of words, will be established for the different topics.

The data used are reviews of Volkswagen cars from the UK website `www.weloveanycar.com`. There are 4194 reviews available over the three most popular models, Golf, Polo, and Passat. Car reviews were chosen as topics are likely interpretable, and because user feedback and customer loyalty are highly important in this industry.

II. DESCRIPTION OF THE DATA

The corpus of reviews consists of 4194 unique reviews of varying length. The website included a recommendation and rating system for each review, but upon further inspection, the information is not useful, as many reviews have seemingly erroneous values. It was chosen to divide each review into a list of words to simplify the analysis.

Stop word removal was done in two stages. The first was to use a supplied list of common words and contractions that add no value, but also certain words specific to this domain. This includes words like car, polo, golf, etc. Evaluation type words like good, great, bad were also removed for the topic allocation, because these are not informative for topic formation. They were kept for the sentiment analysis, though.

The second stage was to score the words using a straightforward count, and then remove those words that did not occur frequently. This was done so that the LDA could focus only on the key themes of the reviews, and not be clouded by minority opinions. This also helped to ignore spelling mistakes that were somewhat common. There is a downside in that niche opinions are not found, but this is probably not possible to do in an effective manner with LDA, and is left for further work.

Straightforward stemming was used instead of lemmatisation for the topic assignment, because there were no particular words that suffered, and reviews were generally quite similar. Computation time was preferably allocated to running the actual LDA as a result.

III. LATENT DIRICHLET ALLOCATION AND SENTIMENT ANALYSIS

To evaluate a product, a sensible technique is to divide the product into its key features or components. For cars, this is sensible, as a manufacturer has several key target areas, such as loyalty, long term performance, mechanical issues, and so on. Then, it is also important to evaluate these features to get a sense of the public's opinion. To do this, LDA and sentiment analysis were used.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation is the process of assigning each word from the corpus to a given, undefined topic. The only specification about the different topics that is made, is the total number that represents the corpus. This is advantageous because it does not require the topics to be explicitly defined, meaning a reduction in bias of what the researcher "thinks" the topics should be.

For reviews this is useful, because a firm may think that users are highly focussed on a set of 5 main features of a product, but in reality this number could be lower or much higher. In this way LDA lets the reviews speak for themselves, and no important feature will be lost due to initial bias of the researcher. This is in contrast to other methods which require a specific assignment of pre-defined topics.

An example of this could be that a car firm thinks that their engine is strongly associated with the fuel economy of the product, but in reality users associated it more with the mechanical problems. LDA would allow both of these topic assignments, with their own weighting.

Another advantage is that it allows a review to have a mixture of topics or key aspects. This is well represented as an output of the algorithm is an assignment to each review, the probability that it originated from a certain topic. This is crucial to the analysis of reviews, as they often contain varied material.

There is still an important specification that has to be made, however, as the number of topics must be properly specified. If this value is wrong, the reviews may be misrepresented. Certain features could be missed entirely, if the value too low, or could be invented on false grounds, if too high.

There is also a major disadvantage compared to other methods that specify topics, as there is no real way of asking a directed question, such as "what do my customers think of the longevity of my product". This problem is two fold because not only could this topic be totally unrepresented, it could be represented along with something else, and so the evaluation derived would be inaccurate.

LDA is still suitable as a general tool, though, as these sorts of questions can be asked specifically by the company in independent surveys. With proper optimisation, which is required by all algorithms, it is possible that all desired topics would be represented as well.

B. Sentiment analysis

Sentiment analysis is the study of the positivity or negativity of a document. This was implemented in a simple way, by taking the AFINN list of words, which assigns a specific sentiment score to each word contained. Sentiment analysis can be done in a more complex way, whereby negations, bi-grams and more contextual information is used, but that was felt to be beyond the scope of the report. Furthermore, it may not be necessary if a company just wants a general overview of how customers view different features of their product.

The way the sentiment for a topic was analysed, was by first taking the sentiment score for each document. Then, taking the probability of a document given a topic, as a weighting for that document, a normalised sentiment score was calculated for each topic, by summing over all documents.

Then, looking within each topic, a weighted count for each word was created, where the weighting is given again by the probability of that topic to generate the word. This count was then compared to the sentiment of the word to try and see how much more common negative or positive words are in interviews.

These two analyses will be used to try and evaluate the customer perception of the features (topics) of the product.

IV. RESULTS

After several trials, three topics were decided on for the LDA. With only three, there seemed to be more meaning to each of them. When more topics were added, several words often repeated, indicating that there were only a few key themes that people focused on.

A. Topics and their sentiments

Three word clouds are shown in Fig.1 to summarise the three topics, with their sentiments highlighted by the colour. It can be seen that the three topics are quite distinct. The first seems to represent some positive evaluation of the car, along the lines of how it runs and how it is to drive. The second topic represents the negative evaluation and the problems associated with

FIG. 1: The word clouds for the three different topics. Words in green are generally positive, while words in red are generally negative.

breakdowns and servicing the car. The final topic represents some sort of loyalty and how the car compares to purchase history of the customer.

The sentiment score for the three topics are 1.90, 0.96, and 1.98. This is reflected well in the word cloud, as there are clearly more negative words in the second topic, while the others seem more positive. It shows as well that customers generally have a good experience with the car, and their loyalty outlook is positive, but there are persistent issues and the experience with fixing the car is not good. It is also likely that those issues are long term ones, as "year" is a prominent characteristic. Considering that "year" is a key word in two of the topics, it is likely that reviews have a long term perspective. Considering that Volkswagen market long term family cars, this is an interesting aspect.

B. The relationship between presence of words and their sentiment

To try and look at review behaviour with respect to word usage and sentiment, the weighted word count for each topic was regressed against the sentiment for that word. Words with neutral sentiment were ignored. The results for the three topics are seen in Figs. 2, 3, and 4. The coefficients for the regressions are -8.97×10^{-7}, -2.25×10^{-6}, and 4.69×10^{-4} respectively. It is interesting that the first two show a negative correlation, as to say that reviewers are more likely, in general, to use more negative words. Then the last one regression shows that for more long term commentary, a reviewer is more likely to use positive words.

Considering the overall sentiments for the topics, the regression coefficients make sense only if, for the first topic, the positive words are, on average, more strongly positive. From the graph, this seems to be the case, as there is a larger concentration around $+2$ and $+3$, than their negative counterparts. It should be noted that the first regression is the poorest fit to the data, so its coefficient could be erroneous.

It should also be noted that the final coefficient is much larger than the other two, implying that the more long term the review is, the stronger the relationship.

V. CONCLUSION AND FURTHER WORK

It can be said with confidence that text mining techniques have a purpose to serve for review analysis. The three topics had clear meanings, and the associated sentiments also made sense, from both analyses. At a basic level, the results could be used to reveal what customers care about, and how they feel.

There are several improvements that could be made to the method, including better assignment of sentiment scores, more advanced sentiment analysis, and using more reviews. These would all help better define the topics and better assign an associated

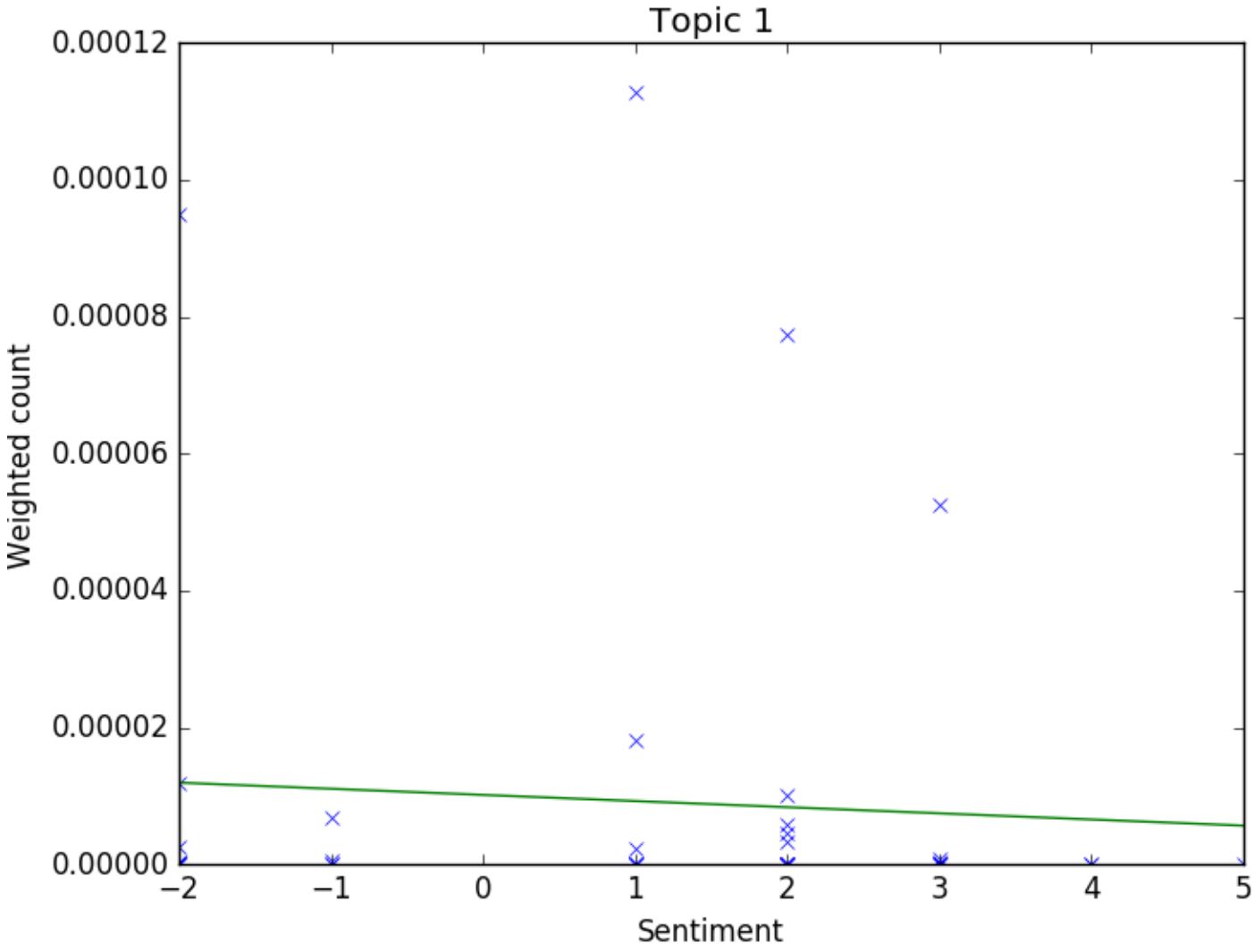

FIG. 2: The regression for topic 1. The slope is -8.97×10^{-7}.

sentiment. This would be needed if a company desired to use the reviews as more than just an indicator of what their customers care about.

As for the use of LDA, while it has performed well, there are more developed methods for opinion mining already well established. Product feature extraction is often done by looking at frequently occurring nouns that are paired with verbs, among other techniques. This is an interesting approach as the features are then given more meaning, and can be reviewed post-analysis, unlike with LDA.

In terms of data cleaning, there could have been some usage of spell-checking methods as well as “bogus review” detection. This would have improved results to some extent, although an initial overview did not reveal any major problems.

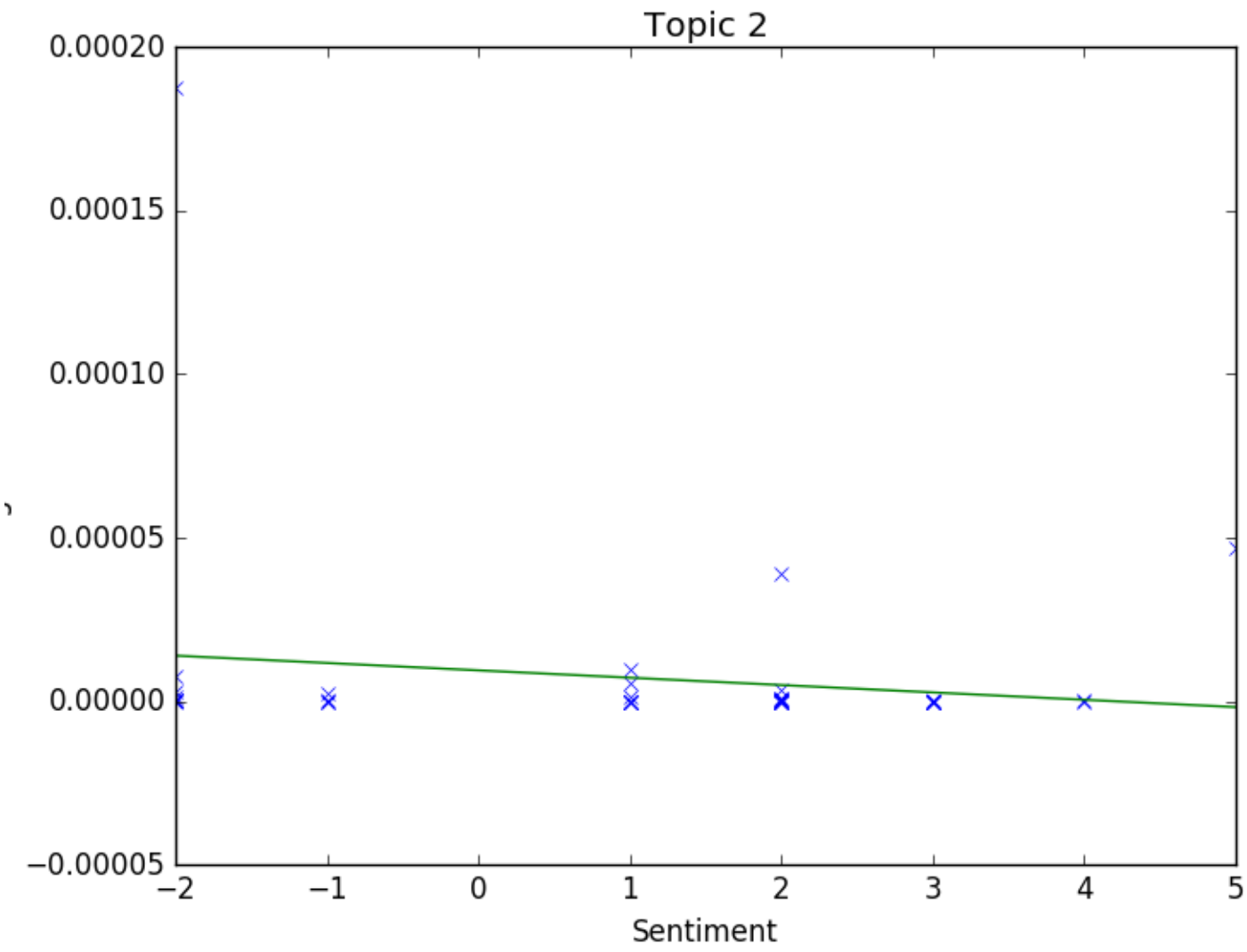

FIG. 3: The regression topic 2. The slope is -2.25×10^{-6}.

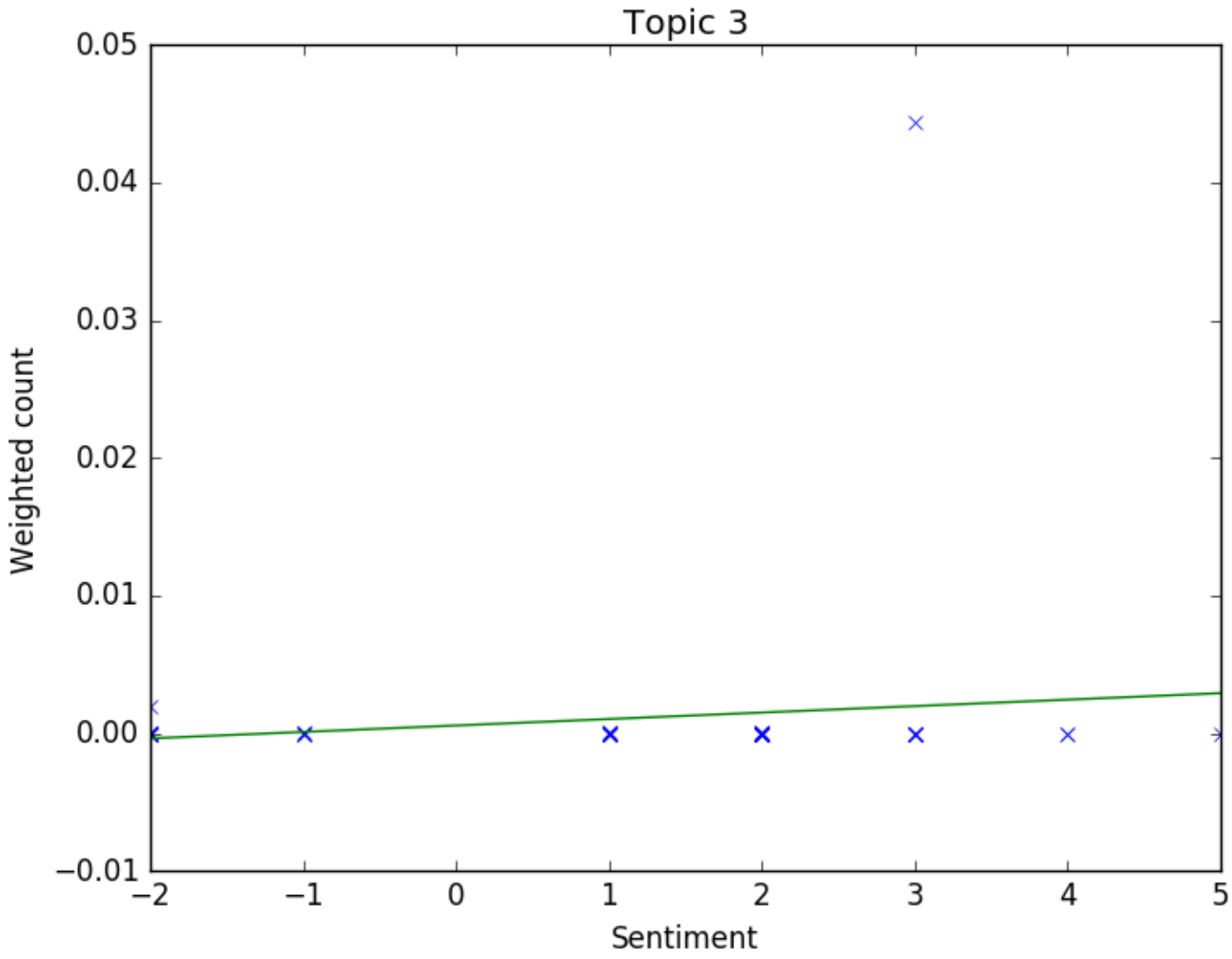

FIG. 4: The regression for topic 3. The slope is 4.69×10^{-4}.