Social Data Science

The Shitstorm Effect:

Investigating changes in brand perception using supervised learning.

A Case study of the Restaurant Chain **Jensens Bøfhus**

Jolien Cremers, Frans Kirpestein, Mads Krarup, David Stoffel

SEPTEMBER 1, 2018

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS

Contents

1 Introduction

In the aftermath of summer in 2014, Jensens Bøfhus, a Danish steakhouse chain, was involved in a lawsuit against a small fish restaurant in Denmark. Jensens Fiskerestaurant, the fish restaurant in question, carried the same name as the steakhouse chain. Jensens Bøfhus claimed to be the sole owner of this trademark and took this issue to court to re-establish their rights. Jensens Bøfhus won the lawsuit, but despite their victory in court, the event caused major uproar on social media, witnessing an increase of negative sentiments in public opinion toward the steakhouse chain (Berlingske, 2015). This negative reaction is referred to as a shitstorm. Presently, Jensens Bøfhus seems to find itself in a financially precarious position. Even though external investors have been called to the rescue to keep the company afloat, it has led substantial loses since 2014. (Rychla, 2015).

Using supervised learning on the reviews from Trustpilot, a social platform which allows to rate businesses by user-generated reviews of subjective experiences, (Hovy et al., 2015) this paper seek to answer the question: *'Are there differences in the public opinion towards the Jensens' brand between the periods before, during and after their victory in court in September 2014?'*

1.1 Social Media and Public Opinion

Social media has significantly changed the world in which we live. We are increasingly involved in an information society through the extensive media and communication technologies that arose since the mid 20th century. According to Cheng Manli, *"the development of science and technology is likely to account for the shift from communication with obstacle, to communication with lesser obstacle and to communication without any obstacle"* (Manli, 2010, p. 7486). The digital age has introduced an environment in which information is no longer bound by limitations of distribution. Interaction and information gathering have assumed a whole new meaning with the modern (electronic) communication technologies at our disposal, hence increasing the flexibility in social interactions (Kaul, 2011).

By means of social media platforms, (public) opinions can be shared in an instance, scrutinized by a widespread array of people and discussed throughout the geosphere. Compared to other forms of media, social media enable people to participate in the feed by sharing comments and opinions about issues (Lee and Chung, 2016). Ideas, behaviors, ethics, concepts and political views, just to name a few, have become increasingly borderless and can be shared and discussed with anyone who wants to participate (Kaul, 2011). The events that transpire in the world are no longer unknown, and local phenomenons can reach national coverage with relative ease (Kaul, 2011). Hence, through the use of social networks we have become increasingly exposed to information outside the scope of our direct environment. *"With these prominent information sharing capabilities, negative news and rumors about an organization can quickly spread online"* (Lee and Chung, 2016, 479). Social Media is not just about content sharing, it

also allows for posting comments and assessing others' opinions and sentiments (Sung and Lee, 2015).

The attitudes shared on social platforms may spill over to many people. There is a potential relation between content on social media and the public opinion towards a topic in society. During the last decade there have been incidents that impacted a companies' market value negatively. Some of the incidents were due to poor managerial leadership, as seen in the overbooked flight incident at American Airlines resulting in decreasing stocks values (Shen, 2017). Another example of company externalities that are critical in influencing the consumers' brand perception, comes from the formerly Russian brand Stolichnaya Vodka. They experienced a massive boycott among the American LGBT community due to Russian Anti gay laws (Shah et al., 2018). Internal and external decision making may influence the degree to which consumers' brand perception is affected. This makes a continuous monitoring of public opinion a business critical task, especially for companies operating in the business to consumers markets.

The degree to which customer behavior is affected can be partially explained by the extent to which the incident is covered and debated in the media. Social media take up an important proportion of the media, and are likely to play an important role in the overall coverage of such debates. As Moon and Jung state, *'a small group of people can easily create an appearance of positive or negative public opinions toward a particular issue/organization (...) not only the news itself is damaging, but the associated opinions/comments shared among users can also make or break the organization'* (Lee and Chung, 2016, p. 480).

1.2 Recommendation systems

Customer satisfaction is crucial for businesses operating on the business to consumers market. A survey by Brightlocal has shown that up to 93% of consumers at least occasionally read reviews of companies to determine whether the business is good or bad. The survey further shows that 67% of the respondents stated that they need to see more than 4 reviews to determine the trustworthiness of a company (BrightLocal, 2017). These customer reviews represent an independent type of sentimental expression on a large scale and is referred to as user-generated content (Liu, 2010). This potentially makes standard recommendation systems and recommendation systems embedded in social networks a strategic focus area for businesses whose revenues are sensitive to the general public opinion.

At Jensens Bøfhus, the costumers' opinion is the most important managerial aspect of their business (Lauridsen, 2016). A large fraction of customer opinion is expressed through the site Trustpilot (Hovy et al., 2015). The monthly review frequency for all of Trustpilot's business categories is over 500.000 reviews (Ingham, 2015), making Trustpilot an interesting data pool for experimenting with Natural Language Processing (NLP) techniques, such as sentiment analysis, to predict overall costumer satisfaction. Both the amount of collectively generated data on Trustpilot and its purpose of business evaluation is in line with our investigation aim.

Previous studies have shown the potential of sentiment analysis on Trustpilot (Hovy et al., 2015) and similar social platforms, e.g. TripAdvisor (Valdivia et al., 2017; Lu et al., 2011). Hovy et al. (2015) examine gender specific words in Trustpilot reviews using $tf\text{-}idf$ tokenization. Valdivia et al. (2017) apply four different sentiment analysis methods (SentiStrength, Bing, Syuzhet, CoreNLPSentiStrength) in order to identify the sentiment polarization of sightseeing reviews and Lu et al. (2011) show the potential of multi-aspect rating predictions using Support Vector Regression and Perceptron Ranking.

1.3 Sentiment Analysis

Sentiment analysis is heavily used within computational linguistics. In contrast to factual based textual data, it retrieves the subjective content of text such as, emotions, opinions, feelings etc. To extract relevant features from text for model predictions, a set of pre-processing steps has to be applied. First a text is transformed into unique tokens (e.g. words). These tokens can then be further transformed by assigning part of speech taggers (POS), such as nouns, verbs and adjectives to each token (Jurafsky and Martin, 2018). We can improve the tokenization further by *stemming* or *lemmatizing* the data set and *removing stop words*. Stemming and lemmatization respectively reduce a word into the root element - or the base form plus infected forms (Asghar et al., 2014). Stop words describe terms that are not related to the sentimental content, and therefore have no explanatory value.

The tokens are subsequently transformed into binary feature vectors that indicate whether a part of text includes or does not include a specific token. The frequency-inverse document frequency (tf-idf) technique can be used to weight tokens relative to their occurrence within the vector. The tf-idf is defined as tf-idf $= tf(t,d) * idf(t,d)$, where $tf(t,d)$ refers to the term frequency and $idf(t,d)$ to the inverse document frequency (Raschka and Mirjalili, 2017). In fact, this method represents a substitute to the stop-word removal approach with a previously constructed word list. Furthermore, it also takes into consideration rare words in the weighting process. In the present paper we applied the tf-idf approach rather than utilizing a pre-defined stop-word list. We do this to avoid the risk of extracting relevant tokens and/or not improving the data set for the analysis as a result of not having a standard stop-word-list (Gentzkow et al., 2017).

The process described above is referred to as a bag-of words model of which the output can be used for fitting machine learning models (Raschka and Mirjalili, 2017). Some of the machine learning models widely used in the NLP community are the Support Vector Classification, Logistic Regression and Naive Bayes (Medhat et al., 2014). In this paper we will therefore use these three methods to perform a supervised classification task on the output of the bag of words model.

2 Data

We noted in the previous sections that we will use customer reviews from Trustpilot to analyze our research question. In this section we more formally introduce the Trustpilot data corpus and its potential ethical concerns associated with it. Finally, we outline the data generating process and perform some preliminary data exploration.

2.1 Data Corpus

Trustpilot offers two different types of data, the review itself representing text data, as well as meta-data of the comment specific users (TrustpilotDeveloper, 2018). Thus, the data generating process will include not only the text body, but also the name of the reviewer, the composition language, the corresponding time stamp, the rating value and the regarded company. The rating value refers to a scale in the range of one to five, which has to be chosen by the user to accomplish the review. On the Trustpilot website reviewers give this rating by choosing the appropriate amount of stars. The minimum amount of stars is one and represents a negative experience. In contrast, five stars refers to the best possible experience (Support, 2014).

2.2 Ethics in Data Generating Process

The innovative techniques of data science open up large possibilities concerning data generating processes and data analysis. Scraping data from the web can be done without the consent of or the notice to the actual data generating object. This creates a potential for information exploitation by the so-called *unseen seer* and results in ethical challenges in various aspects. Therefore, privacy and subsequent principles are taken into account in the present paper: *respect for persons, beneficence, justice* and *respect for law and public interest*.

Respect for persons means, that every individual should be able to make their own decisions with regard to the mediation of their information and that we should keep a focus on their wishes. Beneficence can be understood as a trade-off between the risk of causing harm and the benefit that a certain research can gain. Reaching a fair balance between these two aspects describes the third principle, justice. Finally, respect for law and public interest broadens the responsibility to all possibly affected parties by the research.

Following the recommendation of 'putting ourselves in everyone else's shoes' (Salganik, 2018), we excluded the names of the Trustpilot users and renounced graphics and information that would have made it possible to trace back to the individual level and violate the reviewers' privacy. Furthermore, we insist to stress that our research aim is not to criticize managerial decisions on the firm level. Finally, the data is collected in compliance with Trustpilot's terms of use, allowing the use of data for non-commercial perspectives (Trustpilot, 2014).

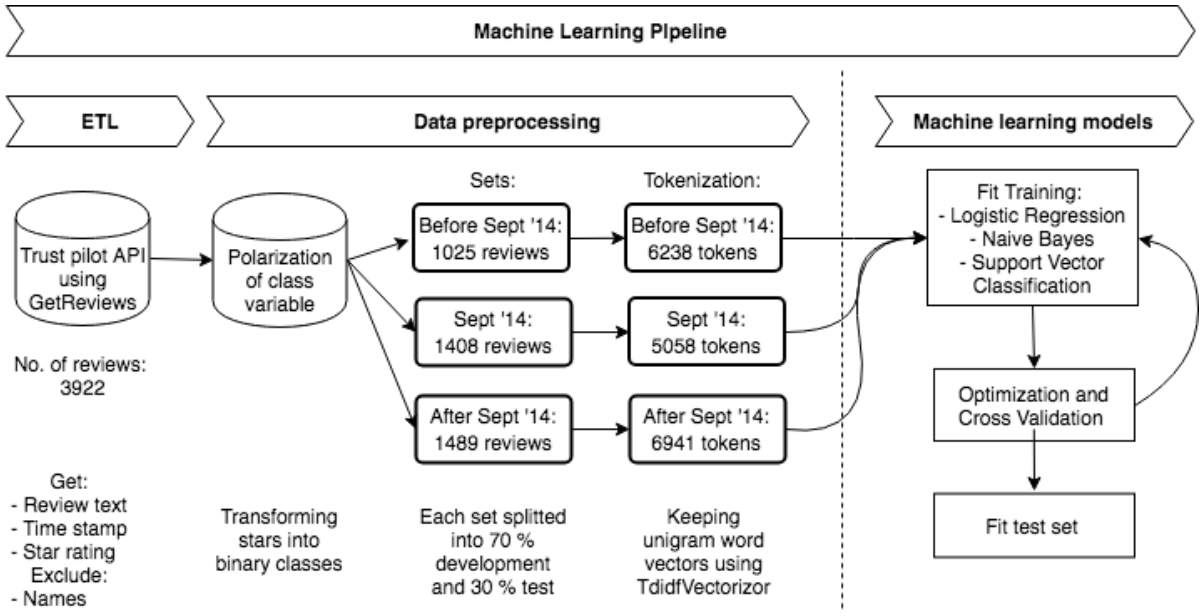


Figure 1: Data generation process pipeline

2.3 Data Generating Process

Figure 1 visualizes the pipeline covering the extract, load and transform task (ETL), data preprocessing and the model selection/validation. In the ETL phase, a review collector named `TrustPilotReader` developed by Prayson (2018), is used to extract and load the reviews from the Trustpilot reviews that were written in Danish, 3922 in total. The script is calling the Trustpilot API which returns 6 dimensions to each review.

As the aim for this paper is to predict either a positive or a negative sentiment toward Jensens Bøfhus, we then polarize the original one-to-five 'star' rating to a binary variable where 1-3 stars are converted to a rating of 0 and 4-5 stars are converted to a rating of 1. The next step in the preprocessing phase is to split the data into 3 sets, namely before, during and after the court case in September 2014. The reviews in each set are then tokenized separately using `nltk`, the output of which is used as input in the `TfidfVectorizer` from the `sklearn` module. This results in 6238 tokens in the period before September 2014, 5059 tokens during and 6941 tokens after September 2014 represented as unigrams. Finally, we split the three data sets (before, during and after September 2014) into a development and test data set. Of the reviews in each set 70% is used as a development set and the remaining 30 % is used as test set.

Subsequently, we optimize 3 machine learning models using 10 fold cross validation. The best model resulting from this process is fitted on the test set. We use the `scikit-learn` module for model deployment. More information on the exact procedure is given in Section 3.1. However, before we optimize the models, we will do some preliminary data exploration, the results of which are presented below.

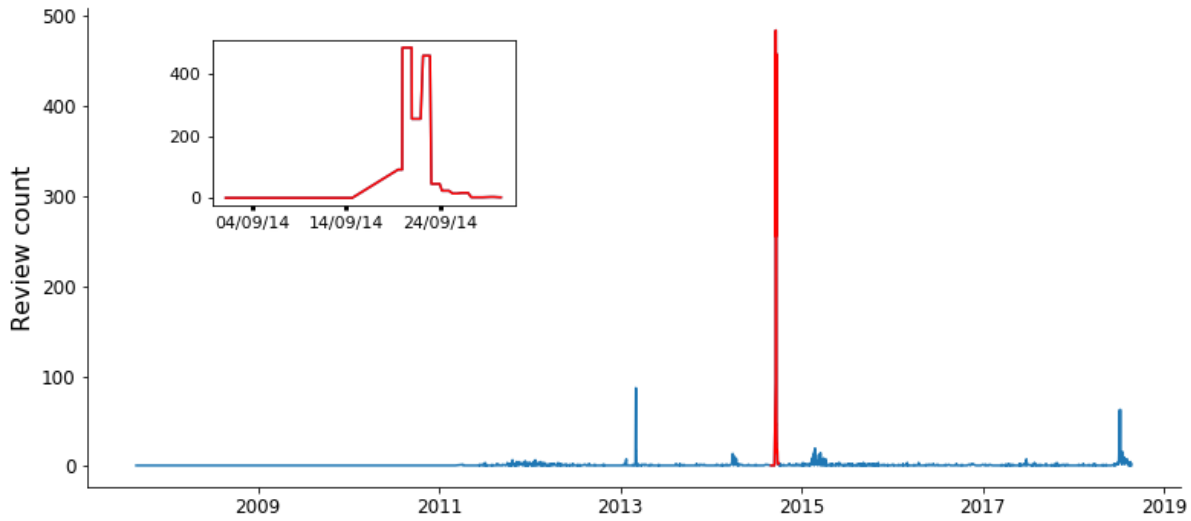


Figure 2: Reviews from Trust pilot plotted over time. Nested line plot indicates reviews from September 2014

2.4 Preliminary Data Exploration

A plot of the amount of reviews per day spanning the time period from 28^{th} of August 2007 to 24^{th} of August 2018 can be seen in Figure 2. The significant spike in number of reviews during the shitstorm in September 2014, is clearly visible. We zoom in on this period (shown in red) in the nested plot. The plot shows that the highest spike of reviews is occurring at the 20^{th} and 22^{nd} of September on which respectively 484 and 458 reviews were written. On the 19^{th} of September the news about the court case outcome became public (Bloch Andersen, 2014), causing an major outrage in the public opinion on Trustpilot. It even resulted in a temporary disabling of creating reviews. Three days later, a large set of newspapers released articles about the indignant reactions from people, causing once again an increase of responses on Trustpilot (Prasz, 2014).

If we look at the cumulative mean ratings over time in Figure 3, we see that there is a dip in the mean rating during the same period as there is an increase in the daily number of reviews. Note that in the period up until 2012 there is quite some volatility in the mean rating, which is probably caused by the fact that up until 2012 less than 500 reviews were written in total. However, over time this volatility disappears.

To investigate the dip in mean rating in September 2014 further, the number of reviews per rating are presented in Table 1. In this table we split the information for the reviews written before, during and after September 2014. The reviews that were written during September 2014, contain a much higher fraction of negative reviews compared with the two other periods. Furthermore, we show the mean and standard deviation of the length of the reviews for the three time periods in Table 2. For all time periods the average amount of characters used in the reviews is lower in the positive class compared to the negative class. However, during September 2014 the length of the negative and positive reviews is almost equal. Additionally,

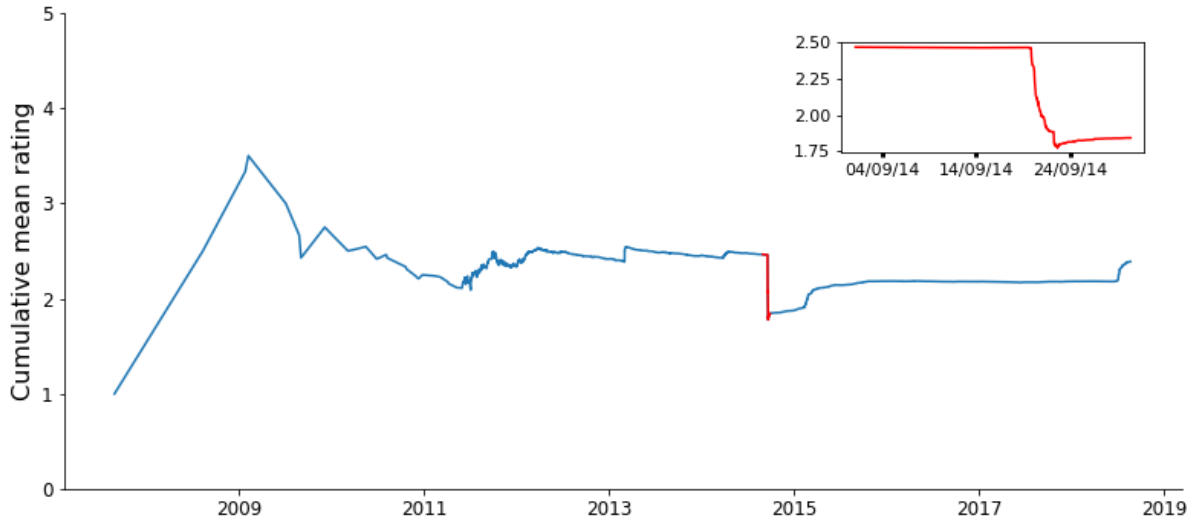


Figure 3: Moving average plot for the ratings (stars) over time. Nested line plot indicates ratings from September 2014

Table 1: Distribution of Trustpilot scores before and after rating transformation

		Count		
Stars	Rating	Before Sept 2014	Sept 2014	After Sept 2014
1 star	Negative	379	1209	438
2 stars		217	63	122
3 stars		137	16	101
Total		**733**	**1288**	**661**
4 stars	Positive	156	26	241
5 stars		136	94	587
Total		**292**	**120**	**828**

the variation in character counts is also smaller. This is in line with the fact that we both have a smaller set of reviews as well as a smaller set of individual tokens to build our models upon during the Sept 2014 set.

3 Sentiment Analysis

In this section we will perform a sentiment analysis on the reviews from Trustpilot in which the aim is to predict the rating a reviewer gave regarding his/her experience with Jensens Bøfhus. First we will outline the methods used and the modelling procedure in Section 3.1. Then we present the results of this procedure in Section 3.2.

Table 2: Character counts of Trustpilot reviews after rating transformation

		Before Sept 2014	Sept 2014	After Sept 2014
Negative	Mean	637.89	283.04	680.95
	SD	584.20	337.12	614.56
Positive	Mean	232.65	203.68	211.14
	SD	313.68	219.60	239.73

3.1 Methods

In order to draw up the relation between public attention and the costumer behavior on Trustpilot, we conduct the subsequent procedure separately on the before, during and after September 2014 data to attain period specific models and predictions. First, we will predict the binary variable 'rating' with the 'tokens' derived from the tf-idf tokenizer by means of three widely used classification methods to predict the sentiment of a review. As mentioned in Section 1.3, these three methods are a logistic regression classifier (LRC), a Bernoulli naive Bayes classifier (NBC) and a support vector classifier (SVC) (Raschka and Mirjalili, 2017; Manning et al., 2008). To fit these classifiers we can use the `sklearn` module in `python` (Buitinck et al., 2013). All three methods will be optimized in a training sample (70% of the reviews) and subsequently fit to a test sample. The accuracy (ACC) of the model will be used as a model-assessment criterion and is defined as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{1}$$

where TP, TN, FP and FN represent the amounts of true positives, true negatives, false positives and false negatives from the predicted ratings of the models respectively.

In the LRC and SVC the L1-norm penalty will be used to reduce the amount of tokens in the final model. Additionally we will optimize the regularization parameter from these two models. In the NBC we can adjust the prior proportions of positive and negative sentiments. By default these proportions are set to the proportions in the data. For this paper we additionally consider a non-informative setting in which we assume that the proportions of positive and negative sentiments are equal. We can also adjust the smoothing parameter α for the prior which accounts for features not present in the training samples and prevents zero probabilities in further computations. Using cross-validation we assess the bias and variance through learning curves as well as optimize the regularization parameter and smoothing parameter in the NBC using validation curves.

After fitting the three optimized classifiers to the test data and evaluating the accuracy, we will further investigate which tokens were the best predictors of the type of sentiment displayed by the reviewer in the time periods *before September 2014, during September 2014* and *after September 2014*. To do so we use the coefficients of the tokens in the LRC, the vector coordinates in the SVC and the empirical log probability of tokens given a predicted senti-

ment, $\Pr(\text{token}_i \mid \text{sentiment})$ in the NBC. Instead of using this probability directly however we compute the following ratios:

$$\frac{\Pr(\text{token}_i \mid \text{negative}) / \Pr(\text{token}_i \mid \text{positive})}{1 / \Pr(\text{token}_i \mid \text{negative})} \tag{2}$$

$$\frac{\Pr(\text{token}_i \mid \text{positive}) / \Pr(\text{token}_i \mid \text{negative})}{1 / \Pr(\text{token}_i \mid \text{positive})} \tag{3}$$

to get the tokens that are best and most discriminating in determining whether a review shows a negative or positive sentiment respectively. In addition to the accuracy we will also compute precision (the ability not to label a negative sample as positive), recall (the ability to find all positive samples) and F1 (weighted average of precision and recall scores) for the optimized models of each time period. Note that we do not consider these three criteria during the optimization procedure.

3.2 Results

The results in the following section represent those of the optimized classification models. The optimized parameters used in the final model and the amount of tokens in the model are presented in Table 3. The regularization parameter represents the inverse regularization strength with higher values representing less regularization. Together with the L1-norm penalty the regularization leads to a decrease in the amount of tokens for which the coefficients and vector coordinates were not set to 0. In the NBC we cannot set a penalty and thus all tokens were retained in the model. The prior of the NBC was set to the proportion of positive and negative ratings in the training data. Note that the complete procedure for the analysis and the data on which it was performed can be found in the following GitHub repository: https://github.com/joliencremers/JensensCaseStudy.

Table 3: Optimized Classification Models

		Before Sept 2014	Sept 2014	After Sept 2014
LRC	regularization parameter	0.6	0.8	0.6
	tokens after L1-norm penalty	18	12	26
SVC	regularization parameter	0.15	0.25	0.10
	tokens after L1-norm penalty	18	22	14
NBC	smoothing parameter	2.5	3.0	3.0
	no penalty	-	-	-

Table 4: Accuracy of the classification models in the three time periods

	Before Sept 2014		During Sept 2014		After Sept 2014	
Classifier	Training	Test	Training	Test	Training	Test
AFINN	0.69	0.72	0.89	0.90	0.60	0.61
Logistic	0.81	0.76	0.93	0.93	0.86	0.81
Naive Bayes	0.81	0.76	0.91	0.92	0.78	0.76
Support vector	0.81	0.77	0.94	0.94	0.84	0.80

3.2.1 Accuracy measures and diagnostic testing

Table 4 shows the obtained accuracies (rounded to two decimals) of the training and test set for the three classification methods that were used. In addition it shows the benchmark accuracy that was obtained by using `AFINN`, a dictionary for sentiment analysis (Nielsen, 2011). Note that we rescaled `AFINN` scores above 0 to 1 (positive sentiment) and scores of 0 and below to 0 (negative sentiment).

We conclude from this table that the accuracy differs between the classifiers and the periods. In September 2014 the accuracy was highest for each classifier. Our benchmark which is obtained by the `AFINN` score exhibits the weakest accuracy in terms of training and test data in all regarded periods compared to the other models used. In the period before September 2014, the LRC, NBC and SVC report a similar accuracy in both, training and test section, which is 81% for the training data, and 76% (respectively 77% for SVC) in the test data. During September 2014 the three classifiers reach an accuracy of over 90% in training and test data, with the SVC showing the best performance at 94% accuracy. In the time period after September 2014 the three classifiers show a different pattern. While the NBC only provides an accuracy in the test set of 76 percent (respectively 78%t for the training data), the SVC shows 80% (respectively 84% for the training data) and the LRC performs best with an accuracy of 81% (respectively 86% for the training data).

Tables 5, 6 and 7 show the precision, recall and F1 scores of all models. We see that for all models both the recall and F1 score perform worst during September 2014 and best after. In general the three classifiers used show a performance that is similar to or better than that of the `AFINN` score. Only during September 2014 the NBC performs significantly worse than the `AFINN` precision scores, the LRC and SVC perform best during September 2014 but also in general perform rather well.

3.2.2 Feature investigation

Figures 4, 5 and 6 show the most important tokens (we display a maximum of 20) for the prediction of the positive and negative sentiments towards Jensens Bøfhus in the time periods *before September 2014, during September 2014* and *after September 2014* according to the three classifiers. For the LRC the most extreme negative and positive regression coefficients

Table 5: Precision of the classification models in the three time periods

	Before Sept 2014		During Sept 2014		After Sept 2014	
Classifier	Training	Test	Training	Test	Training	Test
AFINN	0.41	0.56	0.35	0.35	0.76	0.75
Logistic	0.88	0.79	0.96	0.69	0.87	0.82
Naive Bayes	0.79	0.71	0.25	0.00	0.74	0.71
Support vector	0.90	0.86	0.96	0.90	0.85	0.82

Table 6: Recall scores of the classification models in the three time periods

	Before Sept 2014		During Sept 2014		After Sept 2014	
Classifier	Training	Test	Training	Test	Training	Test
AFINN	0.26	0.35	0.30	0.32	0.41	0.42
Logistic	0.39	0.28	0.26	0.26	0.89	0.82
Naive Bayes	0.42	0.35	0.01	0.00	0.95	0.93
Support vector	0.37	0.26	0.28	0.26	0.88	0.81

Table 7: F1-scores of the classification models in the three time periods

	Before Sept 2014		During Sept 2014		After Sept 2014	
Classifier	Training	Test	Training	Test	Training	Test
AFINN	0.31	0.43	0.32	0.34	0.53	0.54
Logistic	0.54	0.42	0.40	0.38	0.88	0.82
Naive Bayes	0.54	0.47	0.02	0.00	0.83	0.81
Support vector	0.52	0.40	0.43	0.41	0.86	0.81

Table 8: Overlapping tokens between models

		Period		
Model	comparison with	Before Sept 2014	Sept 2014	After Sept 2014
LRC	SVC	all	all	13 of 20
	NBC	7 of 18	7 for 12	0 of 20
SVC	LRC	all	all	all
	NBC	7 of 18	9 of 20	0 of 20
NBC	LRC	7 of 20	7 of 20	0 of 20
	SVC	7 of 20	9 of 20	0 of 20

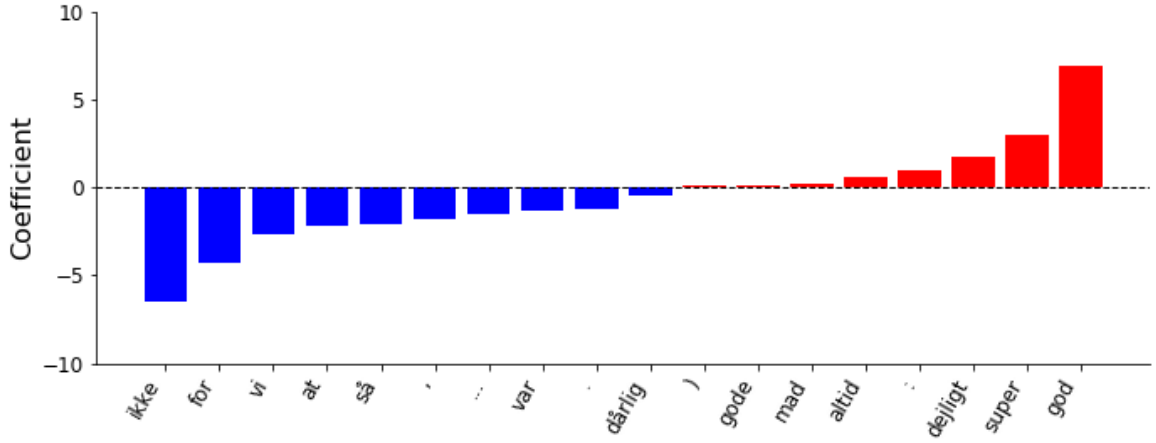
Table 9: Overlapping tokens between periods

	Model		
Comparison Periods	LRC	NBC	SVC
Before Sept 2014 and Sept 2014	6 of 14	1 of 20	9 of 18
Sept 2014 and After Sept 2014	7 of 14	0 of 20	5 of 14
Before Sept 2014 and After Sept 2014	10 of 18	0 of 20	10 of 14

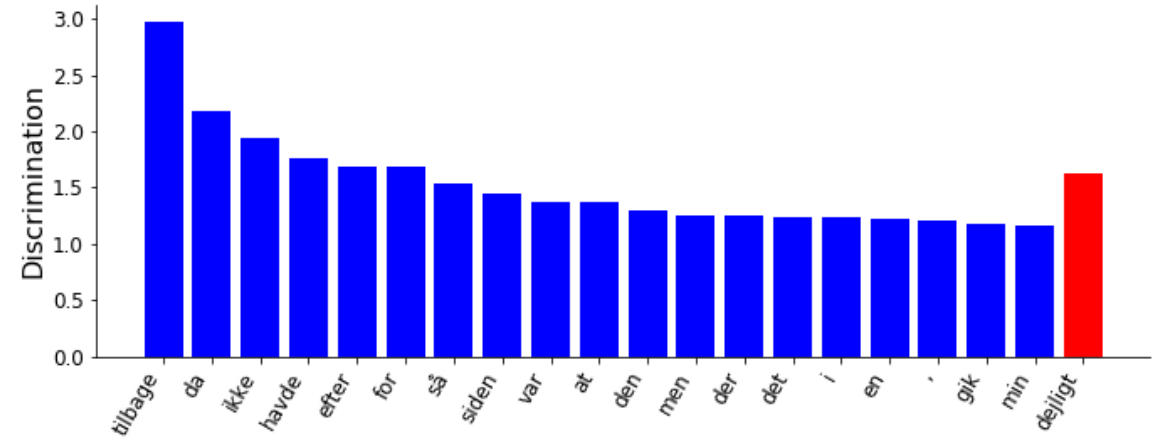
are shown with their corresponding tokens. Note that the exponents of these regression coefficients represent the odds ratio, e.g. for the token 'god' during September 2014 the odds ratio equals $\exp(\beta) = \exp(12.27) = 212936.25$ meaning that an increase of 1 unit in the tf-idf value of 'god' increases the odds of a positive sentiment by about 213,000. For the NBC we see the most extreme positive and negative discrimination values as computed in (2) and (3). Note that these are computed from the raw conditional probabilites $\Pr(\text{token}_i \mid \text{sentiment})$, e.g. $\Pr(\text{dårlig} \mid \text{negative}) = 0.24$ and $\Pr(\text{dårlig} \mid \text{positive}) = 0.05$ during September 2014. For the SVC the most extreme negative and positive vector coordinates are shown with their corresponding tokens.

We see that for the three time periods in Figures 4, 5 and 6 both the number of important tokens as well as whether a token is important for classifying a positive or a negative sentiment differs for the three methods. During September 2014 there are much more tokens that are important for classifying positive sentiments than there are for classifying negative sentiments. Both before and after September 2014 there is a better balance between tokens that are important for classifying positive and negative sentiments except for the NBC. In the NBC only one token is important for classifying positive reviews both before and after September 2014.

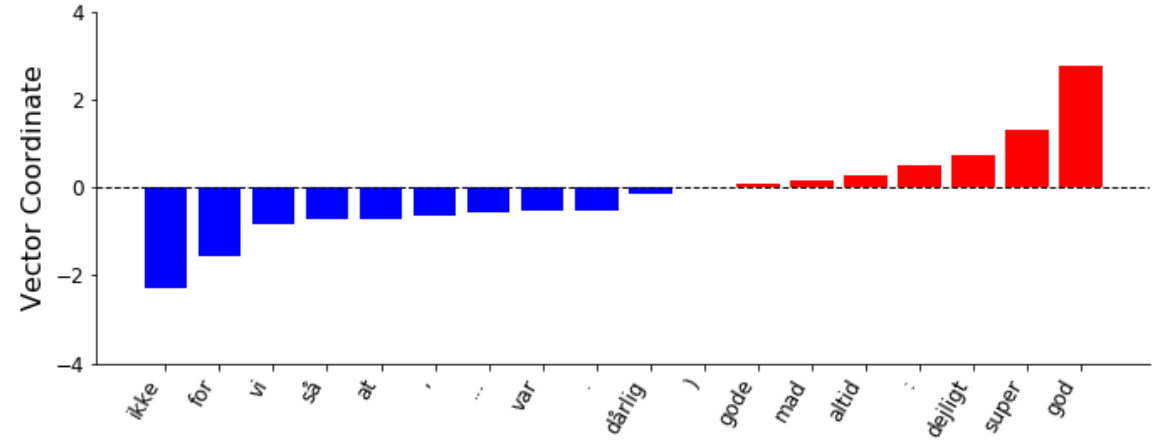
Tables 8 and 9 show the overlap in important tokens between the three classification models and between the three time periods. Table 8 shows that the LRC and SVC have the most overlap whereas especially after September 2014 the NBC does not show a lot of overlap with the other two models. Table 9 shows that the overlap in tokens before and after September 2014 is largest for both the LRC and the SVC. For the NBC there is almost no overlap between any of the periods.



(a) Logistic regression classifier

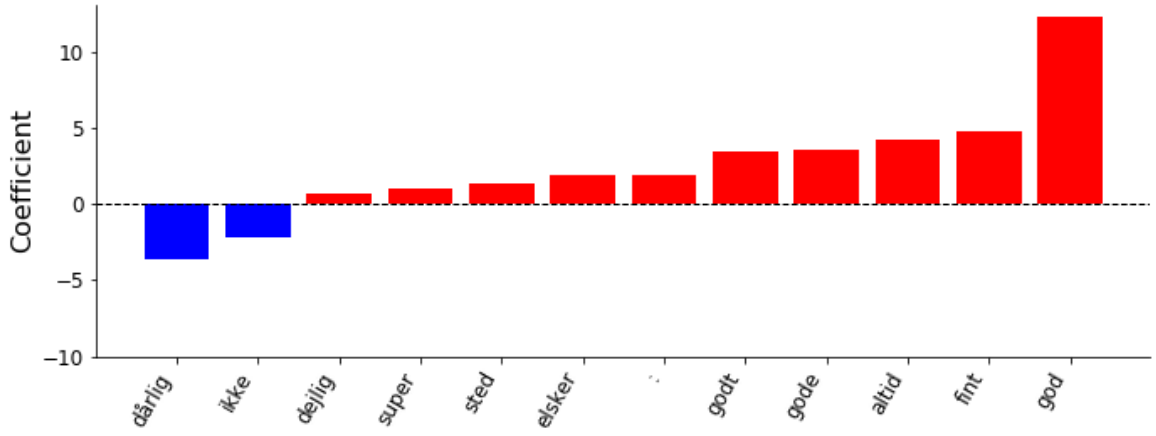


(b) Naive Bayes classifier

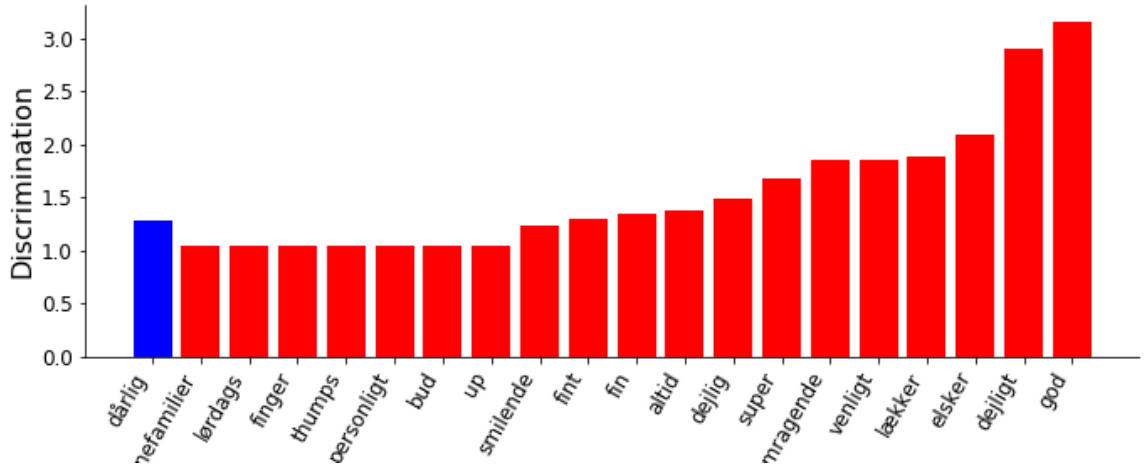


(c) Support vector classifier

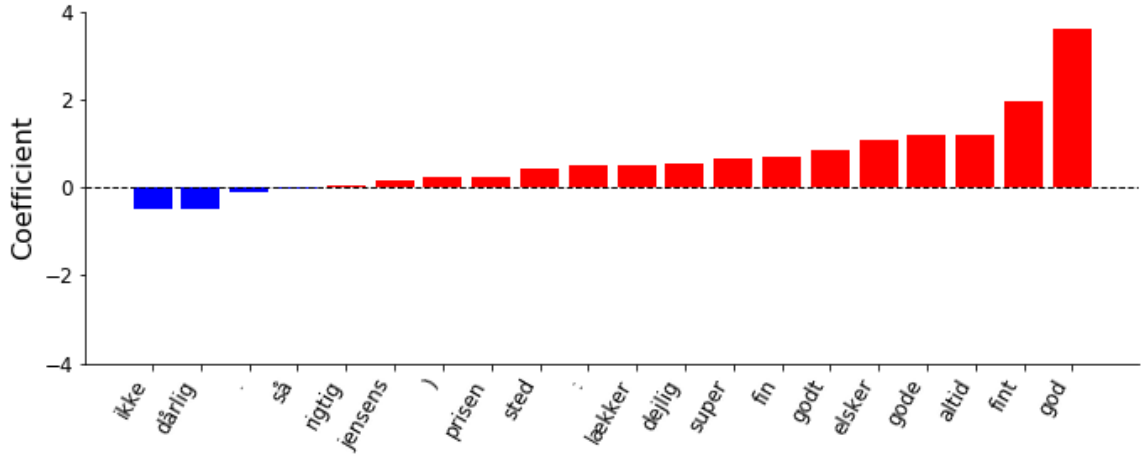
Figure 4: The most important tokens for the prediction of positive and negative sentiments in the logistic regression, naive Bayes and support vector classifier before September 2014. The size of the bars represent the importance and color indicates prediction of a positive (red) or negative (blue) sentiment.



(a) Logistic regression classifier

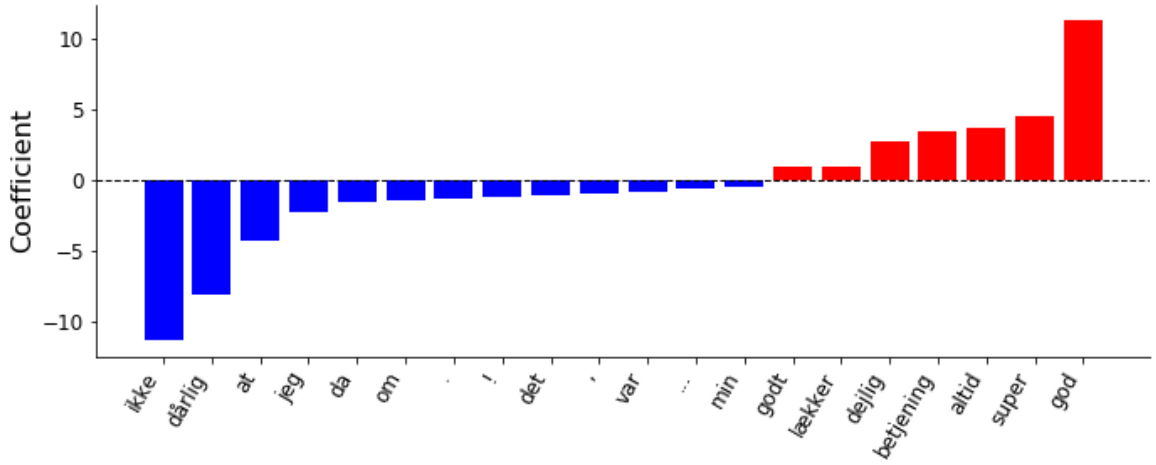


(b) Naive Bayes classifier

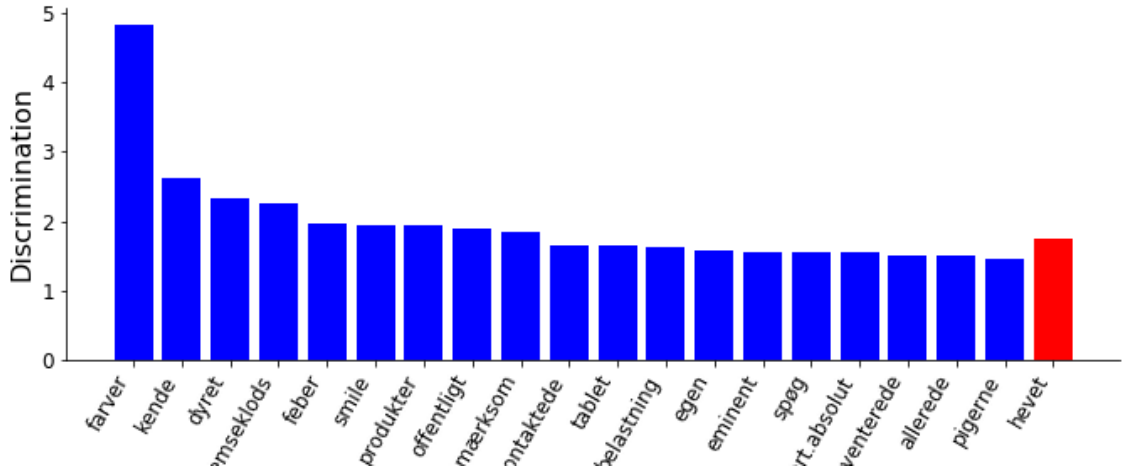


(c) Support vector classifier

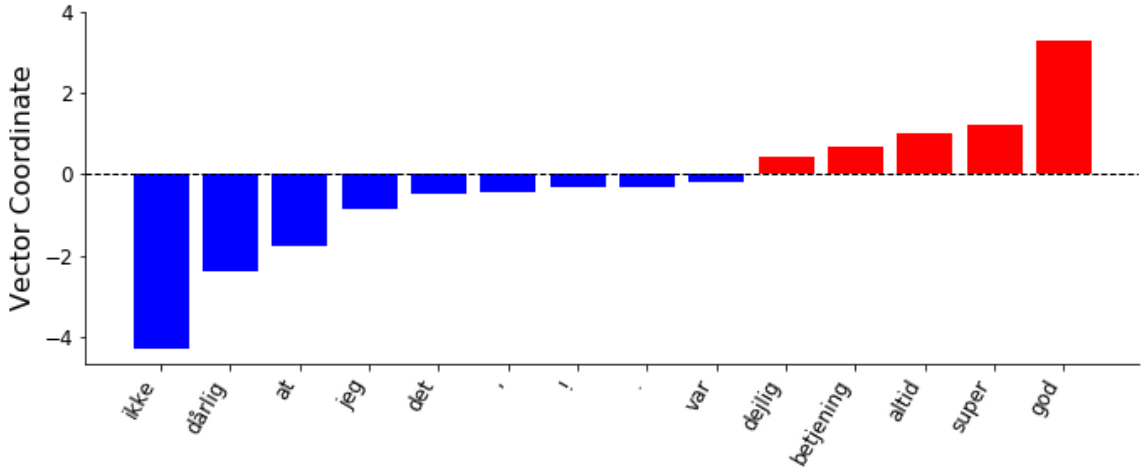
Figure 5: The most important tokens for the prediction of positive and negative sentiments in the logistic regression, naive Bayes and support vector classifier during September 2014. The size of the bars represent the importance and color indicates prediction of a positive (red) or negative (blue) sentiment.



(a) Logistic regression classifier



(b) Naive Bayes classifier



(c) Support vector classifier

Figure 6: The most important tokens for the prediction of positive and negative sentiments in the logistic regression, naive Bayes and support vector classifier after September 2014. The size of the bars represent the importance and color indicates prediction of a positive (red) or negative (blue) sentiment.

4 Discussion

In this paper we examined the effect of a shitstorm following a court case of Jensen Bøfhus in September 2014. We obtained data from a user generated recommendation system; Trust Pilot. Subsequently we applied a set of NLP models to this data and investigated which tokens were deemed to be important in explaining the polarity of customer reviews. In the following section our key results will be discussed, as well as the limitations of the study and ideas for further research.

4.1 Interpreting results

As can be seen in Figure 3, the moving average plot for the star rating takes a significant drop in September 2014. Subsequently, the curve slowly converges to approximately the level it had before September 2014. The court case can thus be seen as causing an impulse response. Due to the high frequency of negative reviews, the convergence to the level before September 2014 happens slowly. It is only in 2018 that the curve has normalized. One should however take care using the court case as a proxy for a permanent decrease in brand reputation among Jensens Bøfhus' customers, as the star rating may reflect other factors that are not present in the Trustpilot data set. In 2018 Jensens Bøfhus has initiated a satisfaction guarantee to stimulate costumer satisfaction and that may have been a driver for the increasing star rating (Jensens, 2018).

When we compare the most influential tokens between the three applied models in our results in Table 8, we see that the models have most overlap in September 2014. During this period most of the tokens are important for predicting positive sentiments. This may due to the fact, that during September 2014 only a very small fraction of reviews were positive. Thus, only a small pool of tokens could have accounted for this sentiment, which is why the models indicate similar patterns. Interesting is that in all time periods both the tokens predicting negative sentiments are the same for the LRC and SVC and the tokens predicting positive sentiments are the same for the LRC and SVC. This increases our belief in the fact that these tokens correctly represent the reviewers' sentiments.

Having a deeper look at the tokens of a model over time, we can recognize different patterns. Firstly, a large amount of tokens, e.g. 'da', 'ikke', 'så' or ')' in the before and after September periods do not immediately seem correspond to a sentiment. Secondly, the NBC has almost no matching tokens between the periods. This might be due to the fact that the NBC re-polarized the tf-idf values to either a 0 or 1 value by default. In further research we may consider using a Gaussian NBC and see whether this increases the overlap in tokens. In contrast to the NBC, the tokens from the LRC and SVC show an overlapping character between the periods. Especially the periods before and after September show a large overlap. This fact might allow the conclusion, that the negative public attention in the lawsuit case of Jensens Bøfhus' in September 2014 did not effect a major change in the reviewers' behavior

and sentiment expression.

4.2 Limitations and future research

As this paper is a case study focusing on one incident one has to be careful generalizing from this study even though the empirical data represent the true population of reviews from one of Jensens Bøfhus communication channels. In order to get a broader view of the shitstorm incident, it is necessary to retrieve information from Jensens Bøfhus' other communication channels such as Facebook. However, due to restrictions in the terms of conditions of the Facebook Graph API (Facebook, 2018), this has not been possible.

Another interesting direction would be to extend the research to an industry study, where several companies operating in the same industry could be analyzed using NLP techniques on data accessible from recommendation systems. Insights about token similarities across a large set of restaurants could potentially explain the customers' perception of tokens, being either truly positive or truly negative within the industry. This could serve as a guideline for strategic focus allocation which is valuable for restaurant industry associations.

Concerning the data quality we should note that we have a relatively low sample size in all time periods but especially in the period before September 2014. Furthermore, the distribution of the polarized ratings is skewed. Where before September 2014 there ratio of negative vs. positive reviews was about 2.5 to 1 this increased to about 10 to 1 during September 2014. Only in the period after September 2014 this ratio is about even at 0.8 to 1. This probably explains why the accuracy and precision during the September 2014 period are highest but recall and F1 scores are lowest. Because there are almost no positive reviews in this period the classifiers have a hard time to actually find them. The NBC actually predicts all of the positive reviews to be negative resulting in recall and F1 scores of 0 in the test set. Note that because the amount of true positives is 0 this also means that the precision score TP/(TP + FP) for the NBC is 0. The fact that there are almost no positive reviews during September 2014 also explains why we only need a couple of tokens to predict a sentiment to be negative (see Figure 5). A larger sample size, more specifically a larger proportion of positive reviews, in further research could help increasing the performance of the classifiers.

Considering the computational part of this paper, the scope of feature extraction is kept at a simplistic level. There are however numerous possibilities extend the scope. We could for example use different sets of n-grams and other types of POS tagging focusing which would enrich the corpus with word classes and simultaneously exclude ambiguous tokens such as 'da', 'ikke', 'så' or ')' seen in Figures 4 - 6. It could further be relevant to include `UNICODE_EMOJI` from the library *emoji* (Kim and Wurster, 2018), as emojies. This may be a relevant proxy for sentiment expression. Furthermore, we have in the present paper polarized the star ratings into a binary rating. In future research multinomial or ordinal classification methods could be used to analyze the star ratings directly instead of polarizing them. However, in this paper

using these models is probably not feasible considering that there are multiple rating - time period combinations of which we have very little observations. A last extension which could be interesting is to include LSA algorithms for topic classification (Landauer et al., 1998) in the analysis, which could allow for a decomposition of reviews split into topic classes.

5 Conclusion

In conclusion, there are differences in the public opinion towards the Jensens' brand between the periods before, during and after their victory in court in September 2014. More specifically we saw an increase in the proportion of negative sentiments being expressed in Trustpilot reviews during September 2014. After this period this proportion seems to have normalized to pre-court victory levels. However, to touch upon the content of the sentiment, it may be beneficial to conduct a more complex sentiment analysis. More specifically, one in which the tokenization procedure is improved and different sets of n-grams are included.

References

Asghar, M. Z., Khan, A., Ahmad, S., and Kundi, F. M. (2014). A review of feature extraction in sentiment analysis.

Berlingske (2015). Sagen er ikke bøf. ProQuest document ID: 1690118098. Copyright - Copyright ©2015 Berlingske Media.

Bloch Andersen, M. (2014). Jensens bøfhus vinder strid om retten til navnet jensen. https://www.business.dk/detailhandel/jensens-boefhus-vinder-strid-om-retten-til-navnet-jensen.

BrightLocal (2017). Local consumer review survey. https://www.brightlocal.com/learn/local-consumer-review-survey/.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Facebook (2018). Facebook platform policy. https://developers.facebook.com/policy/.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.

Hovy, D., Johannsen, A., and Søgaard, A. (2015). User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee.

Ingham, E. (2015). 48 hours on the copenhagen startup scene: Here's what i learned. https://www.forbes.com/sites/edmundingham/2015/11/06/48-hours-on-the-copenhagen-start-up-scene-what-i-learned/#7e41dc169d2c,.

Jensens, B. (2018). Tilfredshedsgaranti. https://www.jensens.com/da/tilfredshedsgaranti/.

Jurafsky, D. and Martin, J. H. (2018). Speech and language processing (3rd ed. draft).

Kaul, V. (2011). Globalisation and media. *Journal of Mass Communication and Journalism*, 1(105):1–6.

Kim, T. and Wurster, K. (2018). emoji 0.5.0. https://pypi.org/project/emoji/.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lauridsen, M. A. (2016). Voldsomt gennemtræk på chefgangen ijensen's bøfhus: Ledelse.der er stor udskiftning på chefgangene iselskaberne i jensen's-konsortiet, der er i økonomisk bekneb. tidligere medarbejdere bebrejder stifter og ejer palle skov jensen for dårlig ledelsesstil. hovedpersonen kalder udskiftningerne »nødvendige«. ProQuest document ID: 1783119254. Copyright - Copyright © 2016 Berlingske Media; Last updated - 2016-04-21.

Lee, M. J. and Chung, J. W. (2016). Reading others' comments and public opinion poll results on social media: Social judgment and spiral of empowerment. *computers in human behavior*, 65:479–487.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops(ICDMW)*, volume 00, pages 81–88.

Manli, C. (2010). The trend of journalism and communication in the age of information globalization. *Procedia Social and Behavioral Sciences*, 2(5):7486–7490.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *An introduction to information retrieval*. Cambridge University Press, Cambridge.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5:1093–1100.

Nielsen, F. (2011). In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*.

Prasz, L. (2014). Vrede sender jensens bøfhus i knæ på trustpilot. https://politiken.dk/oekonomi/virksomheder/art5538159/Vrede-sender-Jensens-B%C3%B8fhus-i-kn%C3%A6-p%C3%A5-Trustpilot.

Prayson, W. D. (2018). Trustpilotreader. https://github.com/Proteusiq/TrustPilotReader.

Raschka, S. and Mirjalili, V. (2017). *Python Machine Learning Second Edition*. Packt Publishing Ltd.

Rychla, L. (2015). Jensen's bøfhus lost millions due to negative facebook campaign. http://cphpost.dk/news/business/jensens-bofhus-lost-millions-due-to-negative-facebook-campaign.html,.

Salganik, M. J. (2018). Bit by bit: Social research in the digital age. *The Times Higher Education Supplement: THE*, (2339).

Shah, G., Rovenpor, J. L., and Jafar, M. (2018). A brand under attack: The boycott of stoli vodka and the power of social media. *Case Research Journal*, 36:1–20.

Shen, L. (2017). United airlines stock drops $1.4 billion after passenger-removal controversy. http://fortune.com/2017/04/11/united-airlines-stock-drop/.

Sung, K. H. and Lee, M. J. (2015). Do online comments influence the public's attitudes toward an organization? effects of online comments based on individuals' prior attitudes. *The Journal of Psychology*, 149(4):325–338.

Support, T. (2014). Trustscore explained. https://support.trustpilot.com/hc/en-us/articles/201748946-TrustScore-Explained-.

Trustpilot (2014). Trustpilot terms of use. https://legal.trustpilot.com/end-user-terms-and-conditions.

TrustpilotDeveloper (2018). Consumer api. https://developers.trustpilot.com/consumer-api.

Valdivia, A., Luzón, M. V., and Herrera, F. (2017). Sentiment analysis in tripdvisor. *IEEE Intelligent Systems*, 32(4):72–77.