SYMBIOSIS
INSTITUTE OF COMPUTER STUDIES
AND RESEARCH

TEXT ANALYTICS CASE STUDY

MSC(CA)DATA SCIENCE

BY

Jigisha Barbhaya(18030142009)
Shashank Shivam(18030142030)

Twitter Sentiment Analysis
A Case Study in the Automotive Industry

Abstract

Sentiment analysis is one of the fastest growing areas which uses the natural language processing, text mining and computational linguistic to extract useful information to help in the decision making process. In the recent years, social media websites have been spreading widely, and their users are increasing rapidly. Automotive industry is one of the largest economic sectors in the world with more than 90 million cars and vehicles. Automotive industry is highly competitive and requires that sellers, automotive companies, carefully analyze and attend to consumers' opinions in order to achieve a competitive advantage in the market. Analysing consumers' opinions using social media data can be very great way for the automotive companies to enhance their marketing targets and objectives. In this paper, a sentiment analyses on a case study in the automotive industry is presented. Text mining and sentiment analysis are used to analyze unstructured tweets on Twitter to extract the polarity, and emotions classification towards the automotive classes such as Mercedes, Audi and BMW. We can note from the emotions classification results that, "joy" category is better for BMW comparing to Mercedes and Audi. The "sadness" percentage is larger for Audi and Mercedes comparing to BMW. Furthermore, we can note from the polarity classification that BMW has 72% positive tweets compared 79% for Mercedes and 83% for Audi. In addition, the results show that BMW has 8% negative polarity compared 18% for Mercedes and 16% for Audi.

Introduction

Others' opinions have always been an important piece of information for consumers when it's time to make buying decision. Long before awareness of the World Wide Web became widespread, people often rely on their friends' recommendations and specialized magazines or websites as the main sources of information. But with the growth of the web over the last decade, the social media nowadays provides new tools to efficiently create and share useful information . This made it possible to find out about experiences and the opinions almost everywhere (blogs, forums, social networks, news portals, and content-sharing sites, etc.). Researches indicate that using the social media sites is considered as the best way to grow a business in terms of money, time, effort and other resources. Although these opinions are meant to be helpful, the massive availability of such opinions and their unstructured nature make it difficult for companies to benefit from them. To solve this issue, a number of techniques for analysing data generated by users on social media sites have been developed. Sentiment analysis which is known as opinion mining is one such recent techniques. Sentiment analysis uses natural language processing, text mining and computational linguistic to extract useful information and knowledge from source data. The purpose of sentiment analysis is to classify polarity from a source text into positive, neutral and negative. Text mining is a crucial step in sentiment analysis where unstructured data are analysed and scored based on how much it relates to a specific concept, in order to be classified later based on its given score. Automotive industry is one of the largest and highly competitive economic sectors in the world. Due to the high competition, automotive companies are moving toward using social media sites to reach further customers and advertise their products in considerably short time.

Twitter is one of the highest growing social media websites in the world. Twitter is a micro blogging services which

enables users to tweet within any topic with a maximum length of 140 characters. As of June 2015[1], Twitter has more than 500 million users, out of which more than 302 million are active users. With an average of 500 million tweets created daily; twitter became one of the greatest sources of information that is available on the Internet [4]. Thus, twitter data can be very useful for automotive marketers because it can be used for mining consumers' opinions and reviews in the automotive industry using sentiment analysis. This can provide useful insights to help companies in creating a competitive advantage over their competitors.

This research applies sentiment analysis to analyse peoples' opinions and reviews about three automotive companies: Mercedes, Audi, and BMW. To do so, tweets are extracted from twitter and processed using text mining techniques. These tweets are then used in the sentiment analysis to classify tweets based on the sentiment that is expressed in a text . At the end, tweets are classified into three categories: positive sentiment, negative sentiment, or neutral sentiment. As the attempts to apply applying sentiment analysis in the automotive industry, to the best of our knowledge, are very few, the results of this research can provide further insights about the importance of analysing the consumers' reviews and opinions in this industry. The remainder of this paper is organized as follows: Section II presents the research work related to this research. Section III presents the methodology. Section IV presents a demonstration of the method on the case study and discusses the results. Section V concludes the paper with a summary and an outlook on future research direction.

Methodology

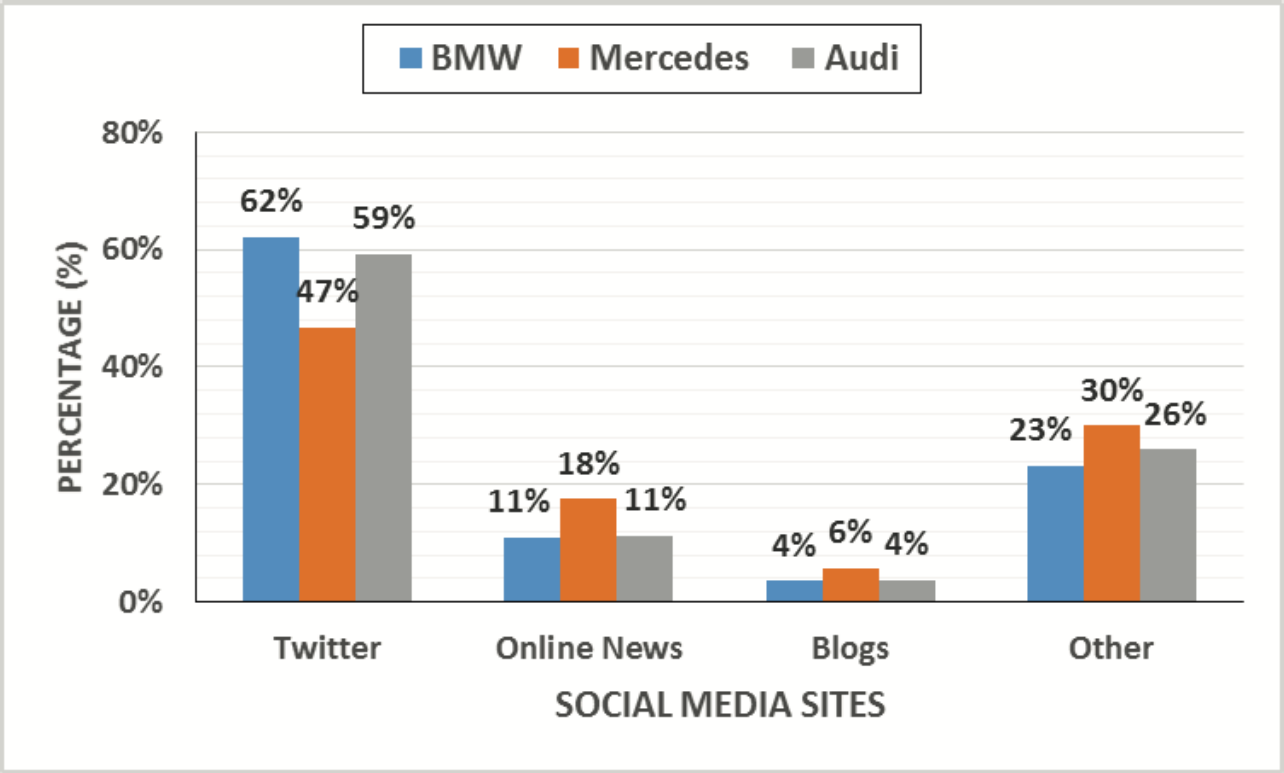
As the usage of social media sites grows and extends, the companies can use social media sites to assess their state in the market as well as their competitors. This can be done by studying the data generated by users on these sites. Such data tells about users' opinions and comments about these companies' products or services. Thus, in this paper we will study the automotive industry in social media, and try to answer the following questions:

- What is the rate of using these companies' data by users?
- What is the percentage of negative reviews and comments compared to the positive ones?
- Who is the leader in automotive sector based on polarity classifications of reviews and comments?

While the social media provides a great engagement of users, and leads to incredibly high level of communication between the user and the seller, still there are some industries that do not engage in social media. The automotive industry represents a great example of engagement in social media, as published in 2014 CMO council report: 1 out of 4 - which equals 23%- of car buyers has discussed other users' experiences and reviews before purchasing their car. 38% of cars' costumers said that they will use social media in the next purchase. 84% of the car's customers use Facebook with a 24% of them using social media sites to purchase their last car and in the range of October 2012- April 2013 an amazing increase in the number of clicks of automotive Ad's on Facebook occurred to jump up from 16% to 39%. In this paper, we will first discuss the level of engagements

in social media of these three automotive manufacturers. We extracted the engagements percentage from the Talkwalker API. BMW, Mercedes and Audi are defined to be of the largest automotive brands in Europe, it's very critical to discuss the level of their engagement in social media. Figure 1 shows the engagement percentage in different social media sites.

Figure 1. Social Media Sites engagement percentage



As we can note in Figure 1, BMW has the largest engagement percentage in twitter with a percentage of 62%. Mercedes also has the largest engagement percentage throw online news, Blogs, and Other with 18%, 6%, and 30%, respectively. Audi also has engagement percentage through twitter comparing to Mercedes with a percentage of 59% (Audi), and 47% (Mercedes).

A. Data collection

In this paper, we collected data from twitter using the twitter API. The corpus had 3000 tweets, tweets are extracted using R[4].

B. Data pre-processing

Tweets are filtered to be in English language. The corpus contains three types of cars: Mercedes, Audi, and BMW. Each type is represented by 1000 tweets. The tweets are extracted based on the search query using "@" annotation followed by the car's type. To build a good experiment, Dataset of each car's type was extracted from twitter pages and users. After that, we have started to prepare the extracted datasets by cleaning them from any unnecessary characters such as retweets and usernames' symbols, hashtags, numbers, punctuations, stop words, whitespaces and html links. In this paper, we applied the following text mining pre-processing techniques:

- Tokenization: that reads the text that will be mined and removes all tabs and punctuations between words and replaces them with a white space,

- Filtering: that will remove words such as: stop words, extremely repeated words and rarely repeated words,

- Lemmatization: which will be used to transform all the verbs to the infinite tense and all the nouns to the singular form.

- Stemming: will be used to return all the words to their basic forms where it will remove the plural 's' from the nouns and the 'ing' from the verbs.

C. Sentiment Analysis Models

We used the classification algorithm **Naive Bayes** (NB) to classify the polarity and emotions in the sentiment analysis. The NB algorithm is simple, easy to implement and efficient with acceptable accuracy. Furthermore, two sentiment models are investigated based on polarity lexicon , and emotions lexicon . The NB algorithm is a simple probabilistic model that assumes all the data attributes are independent. The probabilistic model uses the Bayes theorem to solve the classification problems such as the maximum posterior probability of the class label given the attributes set is calculated. Bayes theorem is given by the following equation:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Where C is a Class label, X is the attributes set, while P(C) and P(X|C) are the prior probability of the class and the conditional probability of the attributes given the class.
The first sentiment model uses NB classifier, which is trained by the training data set, and makes use of Wiebe's polarity lexicon . The training data set is annotated to three classes: positive, neutral and negative tweets.
The NB polarity classifier uses polarity lexicon based on the matching criteria between the tweet words and lexicon words. When the training process is finished and the model is well trained, the second step begins to test the model using testing data set, which is not labeled. The testing process is used to assess the accuracy of the built model. The last step is to validate the model and extract the polarity percentages for the three categories; positive, negative, and neutral.
The second NB classifier is trained on training data set and makes use of emotions lexicon using the Strapparava emotions

lexicon. The training data set is annotated to seven classes: anger, disgust, fear, joy, sadness, surprise, and unknown tweets. Like the polarity classification, the matching criteria between the tweet words and emotions lexicon words.

Results

The tweets collected about BMW, Mercedes, and Audi contains the @BMW tag, @Mercedesbenz, and @Audi, respectively. Each tweet is analysed and classified to be positive or negative or neutral tweet based on a query term and polarity classification. Table I, Table II, and Table III contain some tweet samples about BMW, Mercedes, and Audi, respectively and the polarity classifications.

TABLE I: TWEETS' SAMPLES (BMW)

Tweet	Polarity Classification
#BMW Nice car, you can try it?"	Positive
Elegance and sportiness united in one vehicle: the new #BMW #series Coupé	Positive
such a bad car #BMW	Negative

TABLE II: TWEETS' SAMPLES (MERCEDES)

Tweet	Polarity Classification
@MercedesBenz Intelligent innovation and safety as never before. Preview of the future of the #EClass	Positive
Amazing @MercedesBenz 300 SLR	Positive
@MercedesBenz That's not what we'd expect. Please contact your local Workshop so that our Technicians inspect the issue.	Negative

TABLE III: TWEETS' SAMPLES (AUDI)

Tweet	Polarity Classification
@audi Probably one of my worst decisions was buying an	Negative
Proud to own an Audi @audi	Positive
@audi Sorry RPM but this is rubbish. There is so much great motor sport happening and you dish up crap	Negative
@Audi Excellent SUV from Audi! Beautiful Car!	Positive

Polarity classification for BMW, Mercedes, and Audi are shown in Figure 2. The figure shows that BMW has 72% positive tweets compared 79% for Mercedes and 83% for Audi. Furthermore, the figure shows that BMW has 8% negative polarity compared 18% for Mercedes and 16% for Audi. This gives a good indication for customers seeking to buy cars from the manufacturers that have a good reviews and comments from users owning this car and it gives indications to competitors that Audi is a huge competitor.

Fig 2. Polarity Classification for BMW, Mercedes, Audi

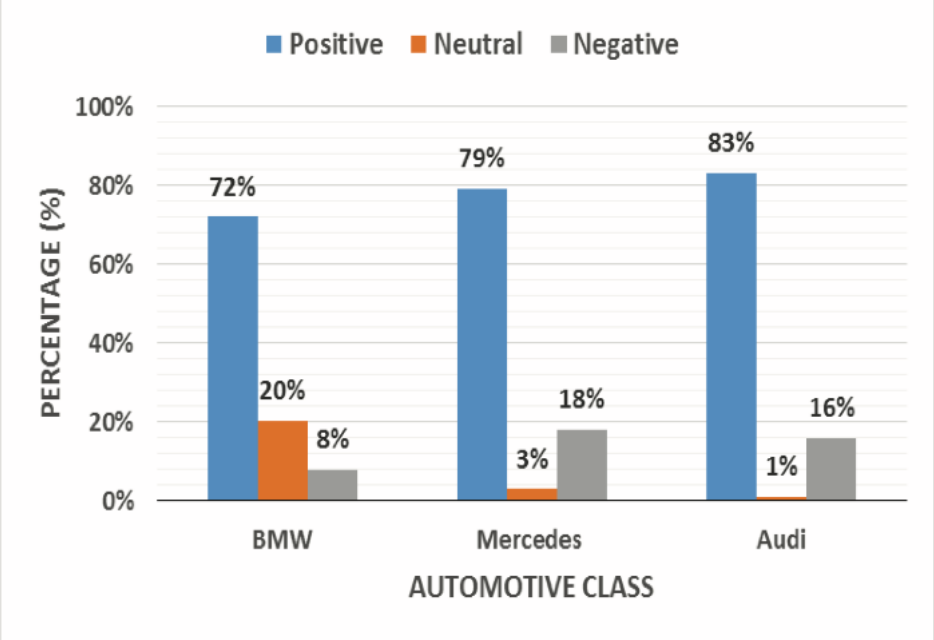
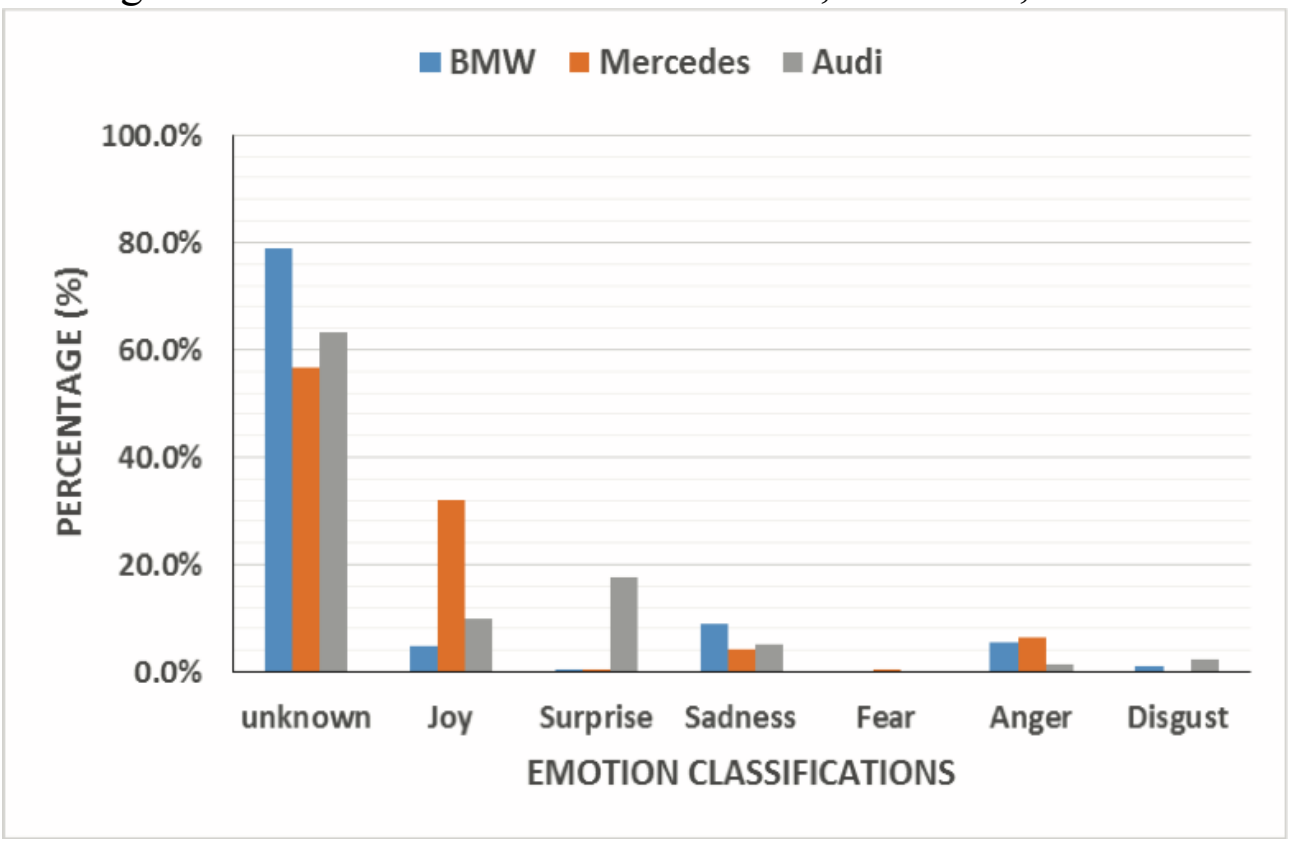


Figure 3 shows emotions classification results for three automotive companies. BMW emotion classifications are 79% labeled as “unknown”, 5% “Joy”, 0.5% “Surprise”, 9% “Sadness”, 0% “Fear”, 5.5% “Anger” and 1% for “Disgust”. Mercedes emotions categories are 56.6% labeled as “Unknown”, 31.9% “Joy”, 0.5% “Surprise”, 4.1% “Sadness”, 0.4% “Fear”, 6.4% “Anger” and 0.1% for “Disgust”. Audi emotions categories are 63.2% labeled as “Unknown”, 10% “Joy”, 17.7% “Surprise”, 5.1% “Sadness”, 0.2% “Fear”, 1.3% “Anger” and 2.4% for “Disgust”. These results give a good indicator for customers seeking to buy cars and help them to take a right decision. We can note that, “joy” category was better for BMW comparing to Mercedes and Audi. This is can be due to the fact that positive reviews are not necessary to be “Joy” always, other categories can be also determined as a positive, since it has no negative implication.

Fig 3. Emotion Classifications for BMW, Mercedes, and Audi



Conclusion

Sentiment Analysis is considered one of the most attractive fields that encourage to study and apply in various sectors. In this paper, sentiment analysis models are applied on three of most leading automotive industry companies to extract the polarity and emotions (opinions) of customers around each company, which are very useful information that helps in marketing. The results showed that Audi's positive polarity was higher (83%) than other companies. On the other hand, the negative polarity of Audi is less than all other companies. This means that for example offers in Audi's page would circulate to higher number of satisfied people than in BMW and Mercedes.

Furthermore, the analysis results show that that the percentage of positive reviews in Audi are the most among the three companies with a percentage of 83%. In addition, Audi negative polarity is less than others with a percentage of 16%. We can conclude that, the Audi users have more satisfaction comparing to the other users. This will help the users that welling to buy a car to compare between the three of the companies based on the previous users' opinions. In addition, the emotions classification results were consistent with the polarity classifications, and give more information about each polarity class.

References

- Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis."IEEE Intelligent Systems 2 (2013): 15-21.

- Edosomwan, Simeon, et al. "The history of social media and its impact on business."Journal of Applied Management and entrepreneurship 16.3 (2011): 79-91.

- Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." Decision Support Systems 48.2 (2010): 354-368.

- Lima, Ana CES, and Leandro N. de Castro. "Automatic sentiment analysis of Twitter messages." Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on.IEEE, 2012.

- Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis."Foundations and trends in information retrieval 2.1-2 (2008): 1-135.