Virtual Market Survey on Automobile Infotainment Gadgets

Gautham R
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
gauthamr.ba05@reva.edu.in

Suresha HP
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
sureshhp.ba05@reva.edu.in

Mutturaj Uppaladinni
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
mutturajuppaladinni.ba05@reva.edu.in

Suresha K
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
sureshak.ba05@reva.edu.in

Abstract
When technology and socio-economic patterns evolve, so will our way of gathering feedback from customers. Today, it is easy to collect actual customer behavior data, to the point where analysis (or data mining) is much more challenging than data collection. Twitter is another great source of ready-to-use information that can be analyzed using text analysis. In the modern era, the virtual marketing survey is all about results. Because of the enormous amounts of data available on Twitter, marketing is becoming increasingly common for many companies as it has a direct connection to increasing revenue.

Significant work was done in the fields of Automotive Gad gets and Twitter Analytics in (In-Vehicle Technology Survey[1],Twitter Sentiment Analysis[2], Twitter Brand Sentiment Analysis[3], Twitter Product Assessment Data [4], Product Review Data[5] and Consumer Insights[6]). However, the work done on Automobile Infotainment Gadgets Virtual Market Survey is limited. In this paper, they focus on the issue of defining Virtual Market Survey in Automobile Infotainment Gadgets using Artificial Intelligence and Machine Learning methods to collect Automobile Infotainment Public Opinion Tweets.
Publicly available tweets provide useful and in-depth insights into the real-time virtual market survey of automotive infotainment gadgets and provide motivation through online influence to others.
We collected data from the desired source (here, Twitter) to perform a virtual market survey. The Tweets are collected for the desired time period of analysis using the streaming API of Twitter.
The obtained text format is translated according to conveni ence. With standard Text Analysis, this information underg oes various steps of pre-processing. The data is tokenized and cleaned before analysis. Before analyzing, the data is tokenized and cleaned.
The cleaned text data is used for data exploration. The class ification models are based on SVC, Decision tree, Random Forest, Gradient boosting and Gaussian Naïve Bayes.
After the creation of the models, we work to train and valid ate the performance of the model using Confusion Matrix, Precision, Recall, F1-Score and Receiver Operating Characteristic Curve for each Supervised Classification Technique as evaluation metrics to understand the Classifier performance.

Once the future trend of the infotainment gadgets has been identified based on the real-time virtual market survey conducted,the results of this virtual market survey could have positive impact on the automotive Industry.

Keywords **–** Virtual Market Survey, Twitter, Infotainment Gadgets, Automobile, Automobile Tweet Classification, Model, Confusion Matrix, Receiver Operating Characteristic curve, Support Vector Machine, Support Vector Classifier, Decision Tree Classifier, Ensemble Learner, Natural language processing (NLP)

1. Introduction

With increased competition and continually growing capital, it is increasingly important for the automotive industry to recognize well-defined market opportunities before committing resources to developing a new product in the area of Infotainment Gadgets.
Analytics can play a very important role in the development of successful new products with increased availability of consumer data and significantly improved hardware and software. Creating market-related quality products can mean the difference between large profits and large losses. The intense competition makes it increasingly difficult to achieve success.
Target marketing analytics provides a means to better identify opportunities that are more likely to succeed because they are based on consumer data, not just observation and interpretation.
Strictly matching demographic and social economic backgrounds in today's market is not enough to effectively market technology items for automotive infotainment.
The automotive marketplace is becoming increasingly well-informed, and there are more competitors to satisfy the needs of that marketplace. Therefore, one-dimensional analysis of non-emotive variables is not enough even for such an approach, the modern automotive infotainment market is far too sophisticated.
Numerous customer surveys and product feedback are available from Twitter nowadays since the consumers are very likely to post their experience about the product over the Internet.

Virtual Market Survey on Automobile Infotainment Gadgets

Large quantities of surveys are available for individual ite ms that make it hard for consumers to peruse each one to make a choice. Subsequently, it is a vital undertaking to collect this data, separate client evaluations and arrange them.

Virtual marketing task takes the advantage of NLP to analyze huge amount of Tweets in order to gather the opinion about the product posted by different users. This process incorporates various strategies, including computational etymology and information retrieval (IR).

The main objective of this paper is to study the existing sentiment analysis methods of Twitter data and provide virtual market survey on infotainment gadgets.

Different sentiment-analysis approaches were defined, including supervised, unsupervised and hybrid approaches used for Twitter data. Lastly, the latter's discussions and co mparisons are emphasized.

In this paper, we examine Twitter users views on automotive infotainment gadgets (Dashcam), covering a range of global consumer brands, and related topics through their Tweets sentiment analysis.

Since Twitter messages are short, defining the target for an opinion is quite difficult.

As a preliminary study, here we take a special approach bas ed on the observation that if a user interacts with an industr y or brand, such as a response or a retweet, the feeling in th e related text probably shows the opinion of this user towar ds that industry or brand.

Thus, by aggregating these interactions from the number of users, the distribution of feelings that reveal the general att itude of twitter users about the industry or brand.

Here we have studied different techniques for sentiment analysis like Decision tree classifier, SVC algorithm, Random Forest Classifier, Gradient Boosting Classifier and Gaussian Naïve Bayes for the sentiment analysis.

In this field, multiple researchers have done different work. events such as detection of earthquakes using social sensors , summarization of events, interpretation of public feelings on Twitter, etc. [8]. As time goes on, these are all the advances in research. Analysis of feelings has therefore become a popular field of research work.

Our dataset is comprised of over 200 different Dashcam related tweets over a 22-month span.

In particular we are seeking to answer the following questions:

- What is the user's opinion about vehicle infotainment gadgets?
- What is the feeling of users against different brands inside infotainment gadgets?
- What are the frequently mentioned issues on social medi a as users communicate with brands?
- What's their thinking about these topics?
- How does the general behavior of users on social media differ when interacting with brands from their behavior?

Our main findings are: Brand here is defined as an independently recognizable product line or business unit and helpful in carrying out a Virtual Market Survey that helps to reduce time and effort for the business.

2. Literature Survey

The target for the Virtual Market Survey while conducting senti ment analysis on Tweets is essentially to correctly identify the Tweets in different classes of sentiment.

Different approaches have developed in this field of researchwh ich recommend methods to train a model and then test it to verify its effectiveness. [2] [13]

Digital consumer survey refers to the study of emotions that has been treated as a natural language processing task at many gra nularity scales [3] [12]

Sentiment Classification can be achieved by classifying a functi on as positive as well as negative as positive as well. [4] [14]

A corpus-based approach is based on assigning each word to the emotional affinity and then identifying each of them from the huge corpus to the probabilistic score. The corpus-based approach is used to test the word happiness variable whic h helps to evaluate the tweets, blogs or posts ' positivity or nega tivity. The method is to pick up from the knowledge base each tweet post or blog and assign a corresponding positivity factor based on the frequency of the positive or negative words and then decide on the overall points for the entire tweet. [5] [11]

Approach based on machine learning uses the technique of classifying text into categories.

Unsupervised learning: It does not consist of a category and they do not provide with the correct targets at all and therefore rely on clustering.

Supervised learning: It is based on labeled dataset and thus the labels are provided to the model during the process. These labeled datasets are trained to get meaningful outputs when encountered during decision making. [7]

A bag-of-words is a representation of text that describes the occurrence of words within a document. The occurrence of words is represented in a numerical feature. It is a way of extracting features from the text for use in modelling, such as with machine learning algorithms. The approach is very simple and flexible and can be used for extracting features from documents. But there is some complexity on two cases i.e., one is on designing the vocabulary of known words and the other is on scoring the presence of known words. [9]

The confusion matrix gives the better understanding of how correctly the Tweets are classified into pre-fined classes.[10]

Accuracy: Accuracy is a rate computed from confusion matrix, which tells us how often our classifier is correct

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Where

True positive (TP) – model predicted positive and is actually positive

True Negative (TN) – model predicted negative and is actually negative

False Positive (FP) –model predicted positive but is actually negative

False Negative (FN) – model predicted negative but is actually positive

Virtual Market Survey on Automobile Infotainment Gadgets

3. Different Classes of Sentiment Analysis used for Virtual Survey

a. Positive Sentiments: This is the category of Tweets shared by the customers who are satisfied and glad about the product, and it is from this category of Tweets where we are likely to get the information about features of the products that attract the customer for the Virtual Survey.

b. Negative Sentiments: This is the category of Tweets shared by the customers who are dissatisfied with the product and it is from this category of Tweets where we are likely to get the information in the Virtual Survey about features of the products that needs to be improved to attract the customers.

c. Neutral Sentiments: This is the category of Tweets shared by the customers used in Virtual Survey from which we are unlikely to get any information about these Tweets.

4. Proposed Solution

4.1 Creation of a Dataset:
The dataset extracted using Twitter API were manually categorized into positive, negative or neutral tweets. Figure 1 shows the categorization of the Tweets.

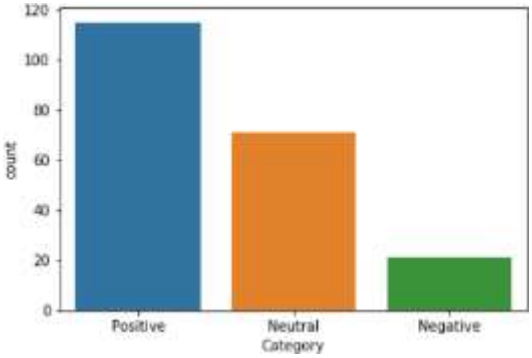

Figure 1: Category of Tweets

4.2 Pre-Processing of Tweets:
In this step the tweets are preprocessed to remove any special symbols [!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~]: , URLs and numbers which do not add any meaning during analysis.

4.3 Word Frequency and Word Cloud:

Figure 2 shows word frequency from the collected Tweets of Dashcam.

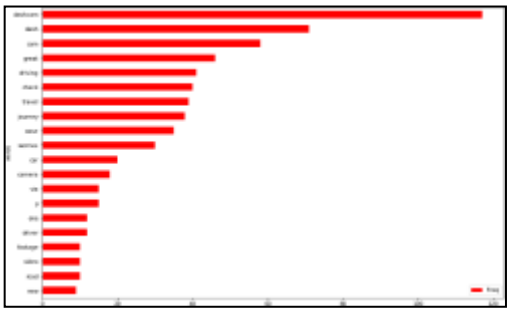

Figure 2: Word Frequency

4.4 Word Cloud:
Figure 3,4,5 shows the Word Cloud to see the most frequent words from the collected Tweets – all, positive , and negative words.

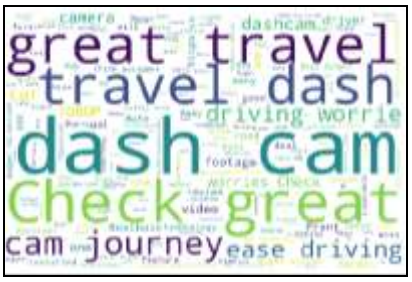

Figure3: word cloud all words

Figure4: word cloud positive words

Figure5: word cloud negative words

4.5 Sentiment Classification
After cleaning Tweets, manual labelling, classification is done using Decision tree classifier, SVC algorithm, Random forest classifier, Gradient Boosting Classifier and Gaussian Naïve Bayes and select the best performing classifier.

5. Classification Techniques

Below listed are the different classifiers that can be used for the Virtual Market Survey

5.1 Decision Tree Classifier:
A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition based on the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking.

5.2 Support Vector Classifier:
SVC is a Supervised Machine Learning algorithm which can be used for classification. Classification is predicting a label/group. SVM performs classification by finding the hyper-plane that differentiates the classes we plotted in n-dimensional space. SVM draws that hyperplane by transforming our data with the help of mathematical functions called **"Kernels".** Types of Kernels are **linear, sigmoid, rbf, non-linear, polynomial etc.**

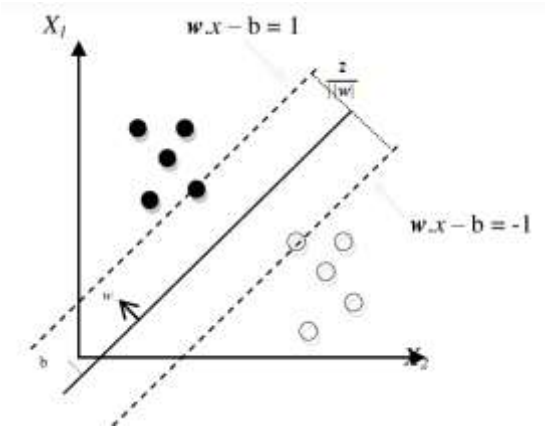

Figure 6: Optimal separating hyperplane between two classes

5.3 Random Forest Classifier:
Random Forests are an example of an ensemble learner built on decision trees.

5.4 Gradient Boosting Classifier:
Gradient boosting classifier is a machine learning technique for classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

5.5 Gaussian Naïve Bayes:
Gaussian Naive Bayes classifier is a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 7: A and B are events

6. System Architecture
The methodology over which our system is based on is as follows:

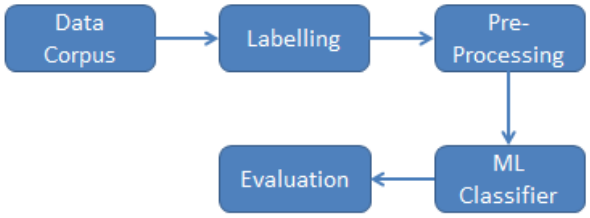

Figure 8: Data Flow

Architecture should look like as follows which will help to analyses market product review.

Figure 9: System Architecture

7. Evaluation
Since we have selected product domain, there is no need of analyzing subjective and objective Tweets separately. This evaluation helps to identify the market survey of product virtually. This shows how context or domain information affects sentiment analysis. These classifiers are tested using python programming.

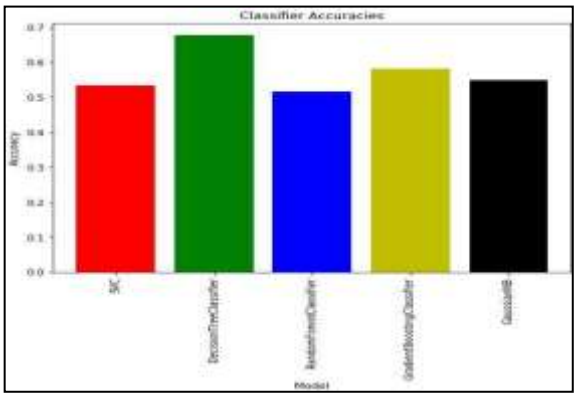

Figure 10: Comparison of model accuracies

Decision Tree Classifier is having higher accuracy of 67.7% which is higher than other classifiers.

Accuracy of DecisionTreeClassifier is 0.6774193548387096

	precision	recall	f1-score	support
Negative	0.33	0.50	0.40	4
Neutral	0.52	0.71	0.60	17
Positive	0.85	0.68	0.76	41
accuracy			0.68	62
macro avg	0.57	0.63	0.59	62
weighted avg	0.73	0.68	0.69	62

8. Conclusion

Microblogging nowadays has become one of the major types of the communication. The large amount of information contained in Twitter web-site makes them an attractive source of data for Virtual Market Survey In our research; we have presented a method for collection of Tweets that can be used to train a sentiment classifier which is able to determine positive, negative and neutral sentiments of the Tweets Survey.

As the future work, we have planned to collect Twitter data for the product review for Virtual Survey which saves substantial amount of time and will help companies to make quick and optimized business decision making. We have also planned to use the obtained data to build a virtual market survey tool which will be helpful to reduce time and effort to analyze the market feedback on Automotive Infotainment Gadgets.

References

[1] Raj K. Kamalanathsharma, Hesham A, Rakha, Ismail H. Zohdy (1 June 2015).”Survey on In-vehicle Technology Use: Results and Findings”.International Journal of Transportation Science and Technology, Volume 4, Issue 2, 1 June 2015, Pages 135-149

[2] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani (May 2017).“Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python.”. International Journal of Computer Applications (0975 – 8887) Volume 165 – No.9

[3] M. Ghiassi, J. Skinner, D. Zimbra(2013). “Twitter Brand Sentiment Analysis: A Hybrid System using N-gram Analysis and Dynamic Artificial Neural Network, Expert Systems with Applications”. vol. 40, 2013

[4] Prof. Sudarshan Sirsat, Dr.SujataRao, Dr.Bharti Wukkadada(2019) ”Sentiment Analysis on Twitter Data for product evaluation” IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719 PP 22-25

[5] Xing Fang , Justin Zhan (2015).“Sentiment analysis using product review data”. Journal of Big Datavolume 2, Article number: 5

[6] Wilas Chamlertwat, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, Choochart Haruechaiyasak(2012).” Discovering Consumer Insight from Twitter via Sentiment Analysis”. Journal of Universal Computer Science, vol. 18, no. 8 (2012), 973-992

[7] Thida win Phyo Thu Zar Tun (2017) “Sentiment Orientation system of a automotive reviews using multinomial Naivebayes classifier at document level”. International Journal of Advances in Electronics and Computer Science ISSN:2393-2835

[8] Alexander Pak, Patrick Paroubek (2015) “A Survey on Sentiment Analysis on Twitter Data Using Different Techniques”. International Journal of Computer Science and Information Technologies, Vol. 6 (6), 2015, 5358-5362

[9] Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof. Dr. D. R. Ingle (2018) “Sentiment Analysis in Twitter”. International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 01 | Jan-2018

[10] Kishori K. Pawar , R. R. Deshmukh, “Twitter Sentiment Classification on Sanders Data using Hybrid Approach”, IOSR Journal of Computer Engineering (IOSR-JCE) (2015)

[11] Sarah E. Shukri, Rawan I. Yaghi, Ibrahim Aljarah, Hamad Alsawalqah, “Twitter Sentiment Analysis: A Case Study in the Automotive Industry” 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), (2015)

[12] Neelima, Dr. Ela Kumar, “IndiSent Analysis in Twitter using Machine Learning Methods”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2015

[13] Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari, “Sentiment Analysis of Twitter Data Using Hadoop”, International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015 ISSN 2091-2730

[14] Prerna Mishra, Dr. Ranjana Rajnish, Dr.Pankaj Kumar, “Sentiment Analysis of Twitter Data:Case Study on Digital India”, Amity University, October 2016

[15] Rajni Singh, Rajdeep Kaur , “Sentiment Analysis on Social Media and Online Review”, International Journal of Computer Applications (0975 – 8887) Volume 121 – No.20, July 2015