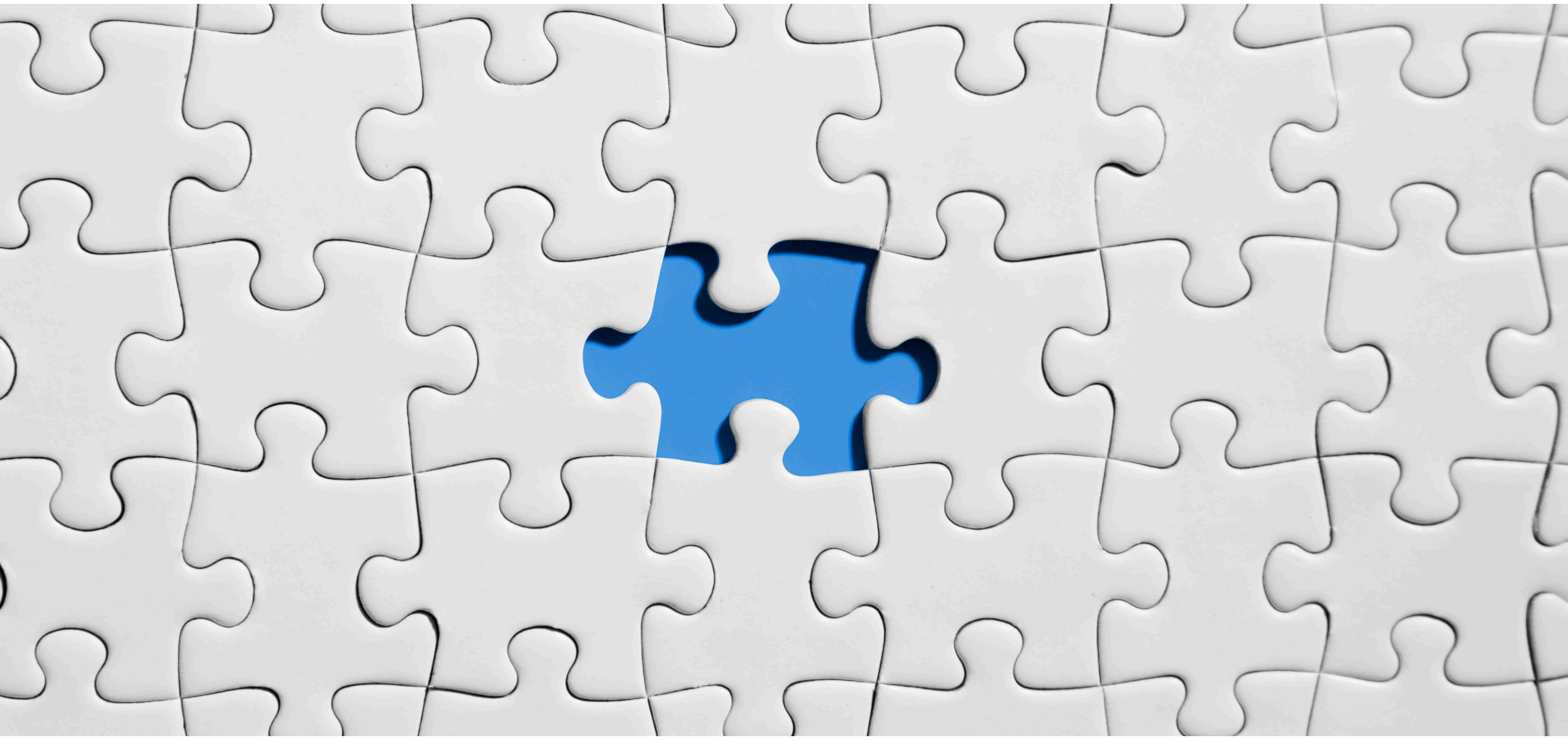
CUSTOMER SENTIMENT ANALYSIS



Sucheta Dhar

Suchetadhar.ss@gmail.com

Prafful Kumar

Prafful.k@dell.com

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Introduction

In recent years, researchers have spent a great deal of time exploring “how a customer thinks” and “what is his/her approach to buying a product” or “Is the customer happy with the product or services provided by the business”. To answer this, researchers can leverage the power of mathematics and artificial intelligence. Sentiment analysis or opinion mining is used to automate detection of subjective information such as opinions, attitudes, emotions, and feelings. Reviews of products/services are an essential source to help them. Thus, online reviews can save the researcher's time and produce promising results in understanding a customer’s future needs and can lead to better customer satisfaction.

This Knowledge Sharing article focuses on analyzing customer reviews on certain products using Natural Language Processing (NLP). Later we will build a machine learning model to classify these customers into returning customers or if the customers are happy, neutral or not satisfied with the product. This model can be deployed or used in analyzing Dell Technologies customers.

The Big Data revolution has altered the way analysts, data scientists, etc. approach any problem in the area of business decision making. With the computing power and abundance of data present in this era, we can learn behaviors of people/customers by merely following their online trail, analyzing it and making predictions that may help to give actionable insights and influence decisions of a probable customer.

Understanding people’s emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analyzing customer feedback – from survey responses to social media conversations – brands are able to listen attentively to their customers, and tailor products and services to meet their needs.

1.1 NLP and Sentiment Analysis

Sentiment analysis domain – also known as Opinion Mining – enables analysts/scientists to mine through the text gathered via various sources and glean how the subject feels. The area depends heavily on Natural Language Processing (NLP) techniques. NLP allows a machine to process a natural human language and translate it to a format that the machine understands. The query processing capabilities of search engines required to add context to the terms entered by users and in turn present a set of results that the user can choose from.

Digital media represents a huge opportunity for businesses of any type to capture the opinions, needs and intent that users share on social media. In fact, the number of Google searches, messages and emails sent in 60 seconds is truly impressive (2,315,000 Google searches, 44,000,000 WhatsApp messages, more than 150,000,000 emails). Truly listening to a customer's voice requires deeply understanding what they have expressed in natural language; NLP is the best way to understand the language used and uncover the sentiment behind it.

In brief, Sentiment analysis is a machine learning technique that automatically analyzes data and detects the sentiment of text. By identifying the sentiment towards products, brands or services, businesses can understand how their customers are talking in online conversations.
Sentiment analysis takes various forms, from models that focus on polarity (positive, negative, neutral) to those that detect feelings and emotions (angry, happy, sad, etc.).

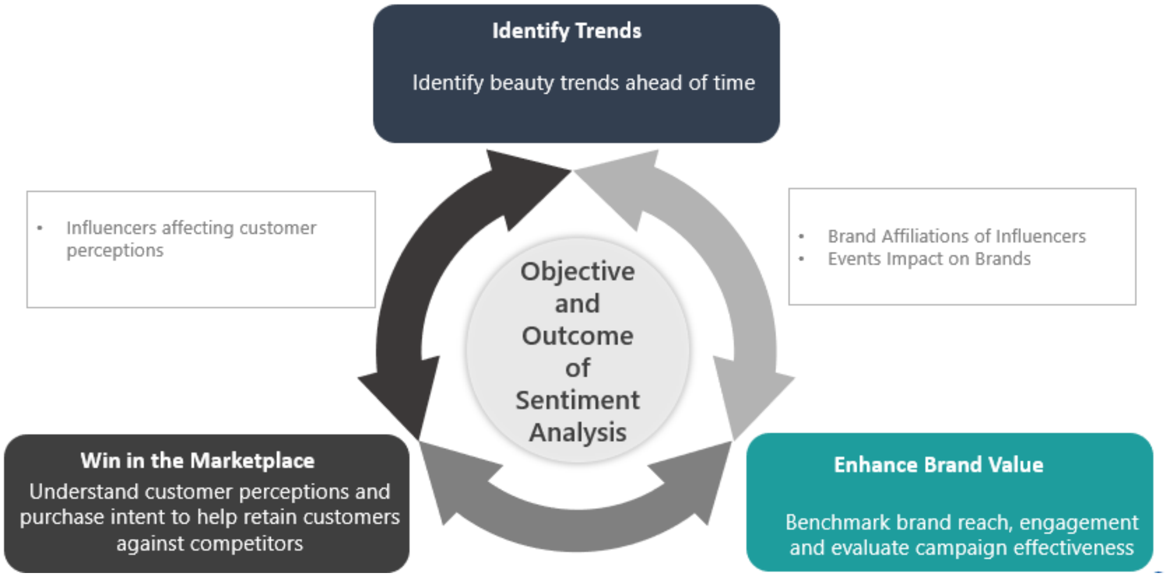

Identify Trends
Identify beauty trends ahead of time
• Influencers affecting customer perceptions
• Brand Affiliations of Influencers
• Events Impact on Brands
Objective and Outcome of Sentiment Analysis
Win in the Marketplace
Understand customer perceptions and purchase intent to help retain customers against competitors
Enhance Brand Value
Benchmark brand reach, engagement and evaluate campaign effectiveness

Fig.1 Objective and outcomes of Sentiment Analysis

1. Process Flow

This section specifies the process or the steps that were involved in performing the customer sentiment analysis.

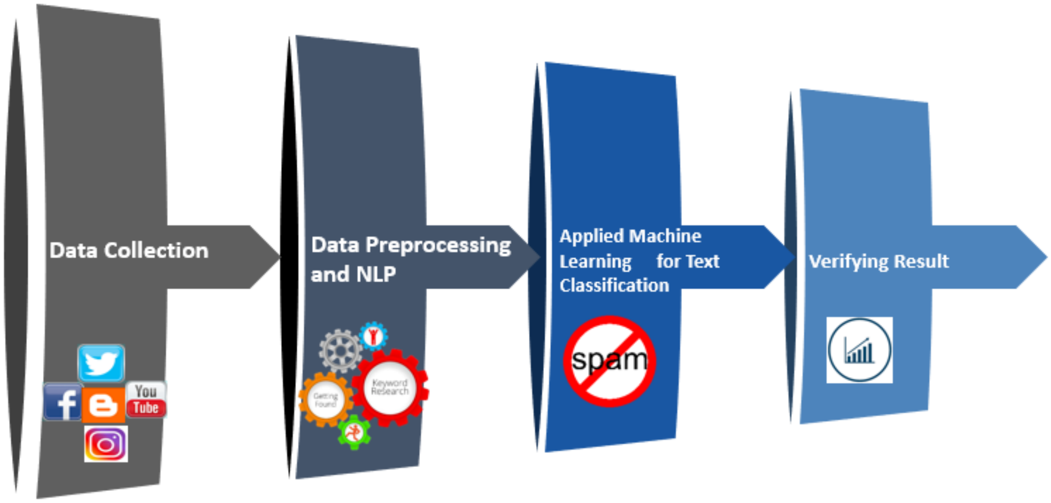

Data Collection
You Tube
Data Preprocessing and NLP
Keyword Research
Getting Found
Applied Machine Learning for Text Classification
spam
Verifying Result

Fig 2: Process Flow

1.1 Data Collection

To perform the following analysis, we have used data made available by Amazon Fine Foods in the form of customer reviews.

The Data includes:
- Reviews from Oct 1999 - Oct 2012
- 568,454 reviews
- 256,059 users
- 74,258 products
- 260 users with > 50 reviews

We removed the neutral comments manually and converted the rating of 5 and 4 to good rating and ratings of 2 and 3 were classified as bad reviews for training the model.

80% of the data was used to train the model and 20% of the data was used to test the model.

1.2 Data Pre-Processing and NLP

Now that we have data large enough to perform the analysis we will have to make the data ready to be used to understand the customer's sentiments. Data preprocessing/cleaning is an important step to scale the data and reduce its dimensionality for efficient model building.

We will perform the following data pre-processing.

2.2.1 Remove HTML tags

As the data is scraped from the web we saw certain instances of HTML tags in the reviews, Since these tags are not useful for our NLP tasks, it is better to remove them.

2.2.2 Remove extra whitespaces

The reviews also have white spaces. Therefore, we will have to remove white spaces as they do not provide any viable information.

2.2.3 Remove special characters

We also removed special characters, i.e. !, @, #, $, %, etc. to reduce the size of the corpus.

2.2.4 Lowercase all texts

We converted the sentence case and upper case to lower case so as to have a standard format case throughout the reviews

.

2.2.5 Convert number words to numeric form

One of the steps involve conversion of number words to numeric form, e.g. seven to 7, to standardize text. To do this, we use the word2number module.

2.2.6 Remove stopwords

Stopwords are very common words that occur in a sentence. Words like “we” and “are” probably do not help in NLP tasks such as sentiment analysis or text classifications. Hence, we can remove stopwords to save computing time and efforts in processing large volumes of text.

2.2.7 Lemmatization

Lemmatization is the process of converting a word to its base form, e.g. “caring” to “care”. We use spaCy’s lemmatizer to obtain the lemma, or base form, of the words.

The data preprocessing task is important so that additional noise in the data can be removed, enabling only the data that is relevant to our NLP task to be obtained.

Natural Language Processing is applied to preprocessed data. NLP can be defined as a set of Feature generation techniques to convert text to a numeric vector so that our machine/model can understand the data we are pushing into it. In this task, we will have to create a data corpus so that the classification technique can be applied to classify these reviews to understand the emotion or notion behind each review.

We have applied a technique called Word2Vec (simply put, word-to-vector) to convert the preprocessed documents into a vector data corpus.

2.2.8 Word2Vec (Word-to-Vector)

Word2vec is a word embedding technique, i.e. a way to convert a piece of text in a numerical format that our machines can read. Technically, Word2Vec is a shallow, two-layer neural network which is trained to reconstruct linguistic contexts of words. Simply put, it is a technique that retains the human-like understanding of each word by creating a relationship among each word. Similar words will have close relationships and dissimilar words will have no/minimal relationship. It takes as its input a large corpus of words and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space. Word2Vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text.

2.2.8.1 Continuous Bag of Words

CBOW predicts target words (e.g. 'mat') from the surrounding context words ('the cat sits on the'). Statistically, CBOW smoothes over a lot of the distributional information (by treating an entire context as one observation). For the most part, this turns out to be a useful thing for smaller datasets.

2.2.8.2 Skip Gram

Skip-gram predicts surrounding context words from the target words (inverse of CBOW). Statistically, skip-gram treats each context-target pair as a new observation, and this tends to do better when we have larger datasets.

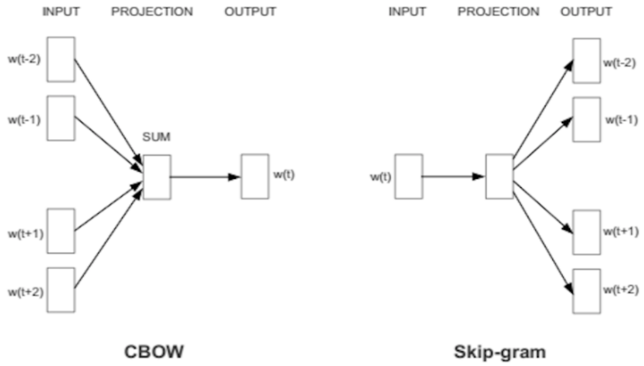
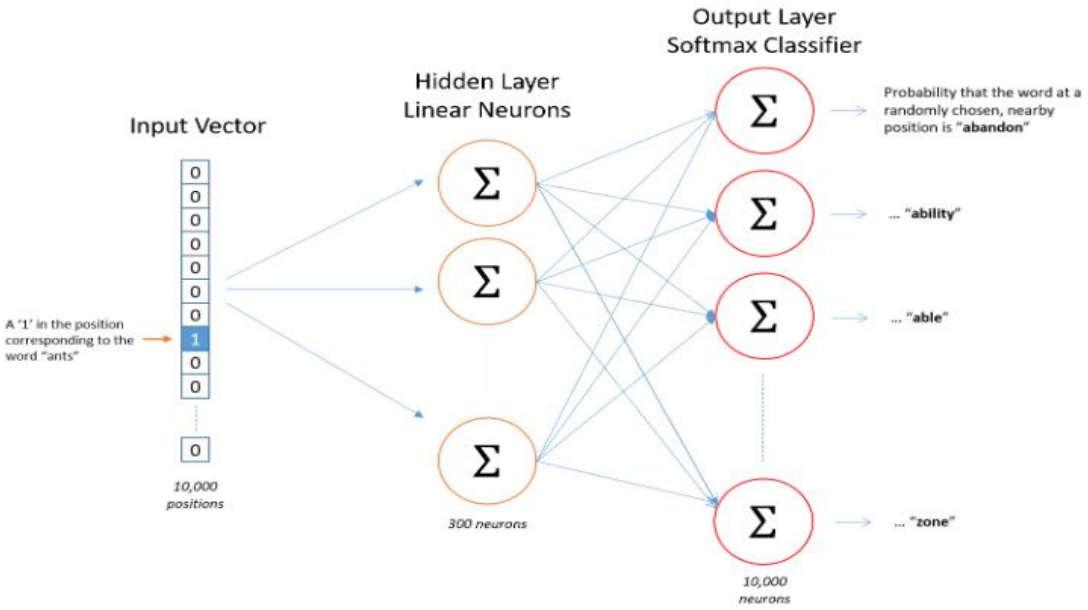


Fig 3. Bag of Words Technique

2.2.8.3 Architecture of Word2Vec

In simple terms, we take a large input vector, compress it down to a smaller dense vector and then instead of decompressing it back to the original input vector as you do with autoencoders, you output probabilities of target words. The hidden layer is a standard fully-connected (Dense) layer whose weights are the word embeddings. The output layer outputs probabilities for the target words from the vocabulary.

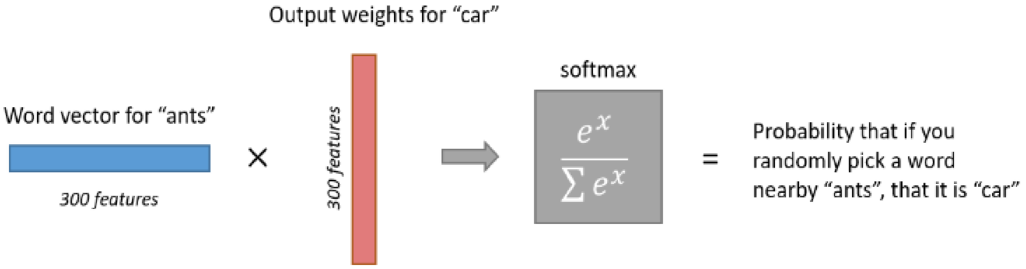


The input to this network is a one-hot vector representing the input word, and the label is also a one-hot vector representing the target word. However, the network's output is a probability distribution of target words, not *necessarily* a one-hot vector like the labels.

The output layer is simply a softmax activation function.

Below is a high-level illustration of the architecture.

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$



After, converting the word in the vector form we pushed the data into our two-class classification model to classify the reviews as per customers sentiments (Good, Bad review).

2.3 Applied Machine Learning for Text Classification

We tried three methods for classifying the reviews into (Good, Bad) reviews. Below are the three algorithms.

2.3.1.1.1 Decision Tree

2.3.1.1.2 Random Forest

2.3.1.1.3 Logistic Regression

We applied these machine learning algorithms to find the best algorithm that classifies the review texts into good or bad reviews and compared the model performance parameters to find the algorithm that best classifies the reviews.

Model Performance parameters:

Model	**Accuracy**	**AUROC**
Decision Trees	71.33	0.64
Random Forest	74.901	0.69
Logistic Regression	81.88	0.75

We found that logistic regression outperformed the other two algorithms.

2.4 Result analysis and Conclusion

After the analysis, we were able to establish that the Logistic regression algorithm was able to classify review text into good or bad reviews with 82% accuracy. Though a human can classify all the reviews with 100% accuracy, it would be a time-consuming task and would require unwanted human effort. Whereas, by applying the power of machine learning and modern computing, similar tasks of analyzing sentiments of humans can be done with an accuracy of 82%.

Using the machine learning algorithm enables us to analyze the sentiments of the Dell Technologies customer and better understand their notion of Dell Technologies services and products, ultimately leading to greater customer satisfaction.

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.