

Gain Customer Insights Using NLP Techniques

Kanna Akella, N. Venkatachalam, K. Gokul, Keunho Choi, and Ramachandraprabhu Tyakal
Ford Motor Company

ABSTRACT

Voice of customer is typically captured through multiple connect points like surveys, warranty claims, social media, and so on. Customer verbatim is collected through these connect points to encourage free expression of opinion by customers. Such verbatim data is generally of high value and is typically analyzed using Natural Language Processing (NLP) techniques for translating into influencing actions in manufacturing, customer service, marketing, and product development departments. One of the challenges in analyzing unstructured verbatim data is to map that data onto appropriate concern codes (CCCs), which are typically used in automotive firms for tracking quality and satisfaction metrics. These concern codes map to a hierarchy of function areas in the organization aimed at improving product, service and hence the customer's overall experience. In this paper, we discuss our approach to address the challenge of mapping customer verbatim to concern codes, which is a classical natural language text classification problem.

In this work, verbatim inputs from transactional systems of quality office, warranty claims, issues matrix, surveys, and social media content are used. Diverse sources and format of these textual comments pose challenges with free flow writing, content, language, and abbreviations. We apply traditional approaches of word counts and TF-IDF for representing documents. Considering complexity of word relationships, we adopt word embedding using word2vec to derive document vector representation of text. Machine learning models are then used to classify customer comments into concern codes. As next steps, we plan to use translation techniques to extend this comment classification framework for English to other languages.

Keywords: word embedding, TF-IDF, classification, SVM, SGD, word2vec, text analytics, natural language processing

CITATION: Akella, K., Venkatachalam, N., Gokul, K., Choi, K. et al., "Gain Customer Insights Using NLP Techniques," *SAE Int. J. Mater. Manf.* 10(3):2017, doi:10.4271/2017-01-0245.

INTRODUCTION

Customer service is of prime importance in any product driven organizations like Ford. Customer comments are captured at every connect points with customer. Many internal systems at Ford also capture customer comments. Classification techniques are used on these comment verbatim in order to understand customer buying pattern, satisfaction, and complaint analysis with the goal of improving loyalty and brand value.

In the recent past, companies developed rules based text analytics on various sources of customer verbatim. Several thousand rules had to be written using tools like SAS in order to classify verbatim into predetermined categories called concern codes, CCC. These CCCs provide feedback to specific departments like manufacturing, product development, and marketing, enabling them to take appropriate actions. It in turn helps focus on developing right products to suit customer preferences.

However, concern code categories are frequently updated and new ones added by quality team. With existing rule based engine it is found to be increasingly difficult to maintain since rule writing is slow and

painstaking. That motivated us to try an automate classification process for customer verbatim using Machine learning (ML) methods. When categorization hierarchies change, ML models only require retraining while rules have to be rewritten or extensively edited.

Two methods of representing verbatim have been tried in our effort: (a) Term Frequency-Inverse Document Frequency, TF-IDF representation and (b) vector representation, word2vec [2, 3, 8]. Machine Learning provides advanced models that use word representations and ground truth data to classify verbatim into categories. Machine learning models need large label dataset, (training set) to learn from and to predict with higher accuracy. One drawback has been paucity of labeled data. Techniques to augment training set by automated process is also being addressed here.

The methods of word representation and classification we present here are: (1) use TF-IDF and various models including Support Vector Machine (SVM) with/without Stochastic Gradient Descent (SGD) for classification into categories, and (2) use word2vec representation and SVM-SGD based classification model.

Finally, we explore avenues for further improvement in classification and apply this word embedded classification process to other Content Categorization use cases. Various notations used in the paper are defined in *Definitions/Abbreviations* section at the end.

WORD REPRESENTATION AND CLASSIFICATION MODELS

Use of association rule mining in building text categorization systems was a promising strategy for automating text categorization since 2002 [1]. Googles’ article [2] presented techniques that preserve syntactic and semantic regularities among words that can be learned with high accuracy. Neural network with non-linear hidden layer, is improved by training the model in two ways: (a) continuous word vectors are learned using a simple model called, Continuous bag of words, CBOW, or (b) Neural Network Learning Model (NNLM) is trained from word representation vectors called, Skip-gram.

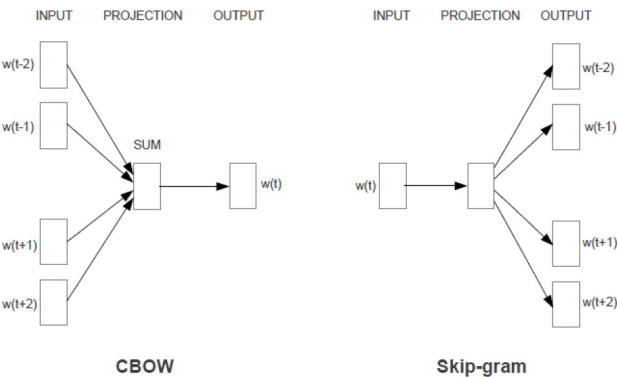


Figure 1. Two step model architecture, continuous word vectors are learned using CBOW and Skip-gram predicts surrounding words given a word [2]

Several variants of Convolutional neural network based models show that pre-trained word vectors perform better than complex deep learning models [3]. Fine tuning the parameters (for example, filter widths) and feature maps seem to improve performance remarkably. A simple method which incorporates character n-grams into skip-gram model that takes into account sub-word information and small training data sets [4] is going to be open sourced soon. Higher representational power is seen to be helping in classification and sentiment analysis on par with deep neural network [5].

HUMAN KNOWLEDGE VS. MACHINE LEARNING

Multiple verbatim data sources ranging from warranty claims, customer survey, call center to social media are used in order to gain customer insights and analyze customer likes or dislikes. The verbatim records were mainly selected from the past five years of data. There are five hundred categories to potentially classify the verbatim records with and that keeps changing due to new/ discontinued technologies/features.

For this experiment, we selected 43 categories based on volume of ground truth data available per category and its business importance. First, we had a focus group of ten subject matter experts, SMEs

assign correct labels using their knowledge of the classification process. The SMEs classified documents by as many labels as deemed necessary. They reviewed 6,840 unique verbatim records, working with twenty knowledge retainers in total. It took about 20 weeks to format knowledge based rules that are executed on SAS based Rules engine.

CLASSIFICATION MODEL USING TF-IDF

TF-IDF is a product of *term frequency*, frequency of the specific word in a document and *inverse document frequency*, which provides the measure of information value a term/word provided in the document.

We split 80% of the data into a training set and 20% into a testing set. First of all, we tested the rules that knowledge retainers extracted from the domain experts. The performance of the rule based model is taken as baseline to compare against other ML classification models. The overall precision and recall ratios were 68% and 67% respectively, which results in 67% of the overall F1 score.

We tested two popular machine learning algorithms for text classification in practice, viz., Naïve Bayesian (NB) and Support Vector Machine (SVM). For the Naïve Bayesian approach, we used two models depending on how to calculate term frequency in a document. The Bernoulli model tests the presence/absence of a term in a document while the Multinomial model takes into account the number of occurrences of a term within the document. Smoothing parameters of both models for the Naïve Bayesian approach was chosen to be 0.01.

We employed two SVM algorithms, depending on the optimization technique used for scalability of the training performance. For the linear SVM, we used L2 regularization and a square of the hinge loss function with dual optimization and 0.00001 of learning rate. We also used a stochastic gradient descent, SGD optimization technique with L2 regularization and a hinge loss function. The learning rate staying same for this model.

In order to represent verbatim text, we adopted two methods: (1) term frequency and inverse document frequency, TF-IDF and (2) word embedding, word2vec. For the word embedding approach, we leveraged bigger verbatim records than the data set in order to learn more precise word relationship. We trained the representation model on more than a million records. Table 1 shows the overall performance of each classification model with different representation techniques.

Table 1. Performance matrix of human knowledge based classification model vs. machine learning based classification models

Classification Method	Precision	Recall	F1-score
Knowledge-Based Rules Engine	68%	67%	67%
Naïve Bayesian – Bernoulli (TF-IDF)	51%	84%	62%
Naïve Bayesian – Multinomial (TF-IDF)	65%	68%	65%
Support Vector Machine (SVM) (TF-IDF)	91%	67%	75%
Stochastic Gradient Descent (SGD) (TF-IDF)	90%	70%	78%