A New Framework of Feature Selection Approach for Sentiment Analysis

Mochmad Wahyudi\textsuperscript{1,}\textsuperscript{*}, Muhammad Zarlis\textsuperscript{2}, Herman Mawengkang\textsuperscript{3}, and Syahril Efendi\textsuperscript{4}

\textsuperscript{1,2}Graduate School of Computer Science, Universitas Sumatera Utara
\textsuperscript{2}Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
\textsuperscript{3}Department of Mathematics, Universitas Sumatera Utara
E-mail: mochamad.wahyudi@students.usu.ac.id

Abstract. Undoubtedly that the huge business data could make data analysis becomes more complicated such that the decision-making process would be out of reach. This condition happens. In the fields of consumer buying behavior, A well-known method called sentiment analysis can help in extracting information about the up-to-date trends and is able to increase market value of product through improving its quality. One of the approaches in solving the sentiment analysis is feature selection technique. However, this technique contains a combinatorial behavior and the analysis of the huge data can experience uncertainty parameter. This paper describes a framework for solving the sentiment analysis based on feature selection approach using a stochastic combinatorial programming.

1. Introduction

The opinion of the people may be considered to be significant for most of the management in the decision-making phase. The media and the internet are the main outlet of people’s thoughts. Due to the increasing number of digital records, it is both time-killing and difficult to identify and review the data you want. One of the key factors why sentiment analysis of different phrases and domains is becoming an increasingly researched sector for many scientists is the growing number of Internet users.

Sentiment analysis can be regarded as a technique to process information in such a way could conclude the customer’s opinion and give responses in the form of text obtained from social media sources. It is a process of recognizing and classifying the opinions of customers in the text format. It is also known as opinion mining. It decides whether the substance of documents are either in positive or in negative sense. It also labels that how customer feels towards any product, brand or any other thing. In relation with this, customers uses positive words like- ‘great’, ‘very easy’, ‘remarkable’, etc., negative words would be - ‘bad’, ‘fraud’ etc. By evaluating the data, organizations found out the feedbacks of customers which could be positive or negative about the product. Analysis of collected data over time give perceptions into trends, while analysis of individual cares in real companies address and answer customer issues quickly.

A range of conversations on accessible assets, comparison databases and assessment campaigns were also selected during the evaluation phase. Various views, such as the internet review pages and the privately operated journals, the accessibility and reputation of the many
opportunities generated and the barriers posed by the substantial transition to ICT patterns, are used to identify points of view. The increase in perception extraction and sentiment analysis regulates the direction in which viewpoints, document ambiguity and feelings are viewed. A concept for refining opinion-mining for online consumer feedback has been introduced in [1]. The latest generational subject design of the Combined Aspect / Sentimental models has been developed in order to obtain the aspects and aspects of the underlying sentimental literature extracted from online consumer feedback. The aspect-based sentiment lexicon contains identical views of the aspect-conscious opinion polarities linked to particular aspects. Lexicons obtained depending on the aspect may contain a variety of considerations, such as recognition of aspects, aspect-dependent extractive opinions overview and aspect-specific sensory classification considerations, which are introduced in the form of opinion mining tasks at the stage of the aspect. The experimental analysis of the JAS model shows that the model is competent and efficient.

In [2] the authors proposed a model based on Sentiment Analysis and Opinion Mining from Social media. In the paper, they proposed a technique based on 3 sections in sentiment analysis such as pre-processing, feature extraction and classifications. Pre-processing method is applied to escalate the accuracy of the system. The important factor in sentiment analysis is pre-processing technique which is used to get the results precisely. In this pre-processing technique, unigrams and bigrams are utilized to skill the personalities of the system.

In [3] introduced a technique based on Improved Feature Extraction and Classification-Sentiment Analysis. The authors addressed the suggested technique by comparing with the preceding existing system. Here, to conquer all the problems in the existing system the authors proposed a machine learning classification to reach the feature extraction model will be more effective and competent one. Since, the following categories are based on feature-based sentiment analysis such as, feature extraction, sentiment classification and sentiment evaluation. The sentiment analysis model has been attained for movie reconsider datasets. The performance of the introduced technique is match up with the existing system and the result will be much better.

2. Feature Selection Technique
Removing the related function reduces the efficiency of classification. The appearance of irrelevant features often affects the efficiency of classification [4]. Redundancy between characteristics is defined by measuring the link between features [5] and eliminating redundant characteristics improves the efficiency of classification. Feature generation features that are numbered hundreds and usually do not transmit significant numbers of such features due to class irrelevance or duplication and should be excluded in order to boost classification performance [4], [6]. A feature is exceptional because the predictive skill of the classifier is exceptionally discerning and modified because it assigns a fresh sample to a class [7] for sentiment analysis.

In addition to the skilful approaches to extractive features and weight assignment, the efficiency of machine learning models depends on efficient selection techniques. The objective of the Feature Selection is to select the most appropriate and discriminatory features by excluding the noisy and unnecessary classification features. The efficiency of the machine learning method is reduced when higher features are vectors, chaotic, insignificant features [8]. Techniques for the selection of practical vector characteristics, such as information gain (IG) [9], mutual information (MI) [9], etc., or transformation features techniques, such as singular value decomposition (SVD), LDA, etc., are used. Selection processes have chosen the essential characteristics on the basis of the term excellence attribute and select the top characteristics if they lie above the cap and minimize the non-relevant characteristics. Feature transformation processes focus on transforming high-dimensional vectors into a low-dimensional feature domain in which the reduced feature vector, such as the earlier characterization vector, is made up of all features.
In [10], dimensional reduction was performed using the Latent Semantic Analysis Technology (LSA) and the reduced feature vector was used to enhance sentimental analysis efficiency with the Classifier Support Vector Machine (SVM). Feature selection that is suited to transformation features due to flexibility and procedural efficacy is a common way of minimizing vector length.

Numerous feature selection technologies, including MI, IG, DF (document frequency), CHI (chi-square), etc., have been studied for sentiment analysis in [6], [7], [11], [12]. DF (document frequency) is the most basic strategy for the selection of functions. Document Frequency is a commonly used form of sentiment analysis [13] that focuses on the most frequent expressions in texts used to create a practical vector. In the study of Chinese language documents in [7], the output of four techniques for the collection of features, namely MI, IG, CHI and DF, was checked. They used 5 algorithms for machine learning, such as cluster classification, KNN (k-nearest neighbor), NB (naive Bayes), winnow classification, and SVM (vector supported). Compared to other methods, the experimental findings indicate IG efficiency in the functional range and SVM algorithm reliability is significantly higher. IG and GA (genetic algorithm) were evaluated on the basis of a dataset film analysis and proposed a hybrid approach called EWGA (entropy weighted genetic algorithm) to improve the analytical precision of emotions. A new approach to the selection of features called document frequency difference was developed in [15] and a proposed system comparison with other methods of selection of features for nostalgic review was carried out. A log-like approach to selecting key features is used for sentiment analysis in [16].

The CPPD (Categorical Probability Proportion Difference) method in the [11] feature selection framework selects features based on the importance of the function and class specific features. The CPPD includes two separate approaches, the CPD (Categorical Proportional Different) and the Probability Proportional Difference (PPD) techniques. The Categorical Proportional Difference (CPD) approach is based on the estimation of the degree of the class-indiscriminating contribution word and only the maximum contribution word is included in the classification [17]. The PPD methodology focuses on measuring the degree to which a word belongs to a specified class and the variance is a test of the ability to differentiate between them. Applicants are chosen for a greater degree of belonging in terms of ranking. The advantage of CPD technology is that it tests its class defining properties for a concept that is a key feature of a significant trait. Delete terms that are negligible for classification terms that appear equally in all groups and with high document frequency, including certain stop words (i.e. a, a, etc.). The advantage of the PPD methodology is that it may exclude words such as unusual, low-volume words that are not necessary for sentiment analysis. For sentiment analysis of text details, Wang et al. [18] has developed a discriminatory Fisher ratio-based method for selecting functionality.

3. Framework of Stochastic Programming Model

An optimization problem in which some of the data are subject to uncertainty, due to random impacts, or statistical variations can be called a stochastic programming (SP) problem. Eventually, SP conveys a particular framework to include a path dependence of the stochastic process within a realm of an optimization model. Additionally, it allows so many states and actions, along with constraints, time-lags etc.
The SP model considered in this paper can be written mathematically as follows.

\[
\begin{align*}
\min_y f^1(x) + Q(x) \\
g^1(x) &= 0, \\
h^1(x) &\leq 0, \\
g^1 : R^n_{\Omega} \rightarrow R^m_e, \\
h^1 : R^n \rightarrow R^m_1, \\
x &\in R^n_+
\end{align*}
\]

Where

\[
Q(x) = E_\xi Q(x, \xi(w)) \xi
\]
\[
Q(x, \xi(w)) = \min_y f^2(y(w), w)
\]
\[
g^2(x, y(w), w) = 0 \\
h^2(x, y(w), w) \leq 0 \\
g^2 : R^{n_1+n_2}x_\Omega \rightarrow R^{y_e} \\
h^2 : R^{n_1+n_2}x_\Omega \rightarrow R^{y_i} \\
y &\in Y
\]

\(\Omega\) is a probability space provided with \(\sigma\)-algebra \(F\) is probability measure, \(\xi\) is random variable which has probability measure and \(f^1, f^2, g^1, g^2, h^1, h^2\) nonlinear function, differentiable but non-convex. \(x\) signifies first stage variable, \(y(w)\) is the second stage variables. Set \(Y\) is expressed as the union of subsets \(Y_R\) dan \(Y_Z\), where \(Y_R \in R^{n_2}_e\) and \(Y_z \in R^{n_2}_i\).

The main feature of SP two-stage is a recourse action. The decision set is split into two parts. Decision before problem parameters are known, stated as first stage; and decision after the uncertainty are revealed.

The framework for using stochastic modeler in feature selection technique for sentiment analysis is described in Figure 1.

4. Conclusion
This paper describes a framework to be used in feature selection technique for sentiment analysis based on stochastic programming. From Fig. 1 it can be seen that the approach includes a stochastic modeler in such a way to obtain the optimum result for the sentiment analysis.
References

[1] Xu, X., Cheng, X., Tan, S., Liu, Y., and Shen, H. 2013. Aspect-Level Opinion Mining of Online Customer Reviews. Communications, China, Vol.10, No.3, pp.25-41.

[2] Mathapati, J. S. H Manjula & Venugopal K R. 2016. Sentiment Analysis and Opinion Mining from Social Media: A Review. published in Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 16, Issue 5, Version 1.0

[3] Trupthi, J. M., Suresh Pabboju and G. Narasimha. 2016. Improved Feature Extraction and Classification-Sentiment Analysis. published in International Conference on Advances in Human Machine Interaction (HMI-2016), March 03-05, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India.

[4] Yu L, Liu H. 2004. Efficient Feature Selection Via Analysis of Relevance And Redundancy. J Mach Learn Res 5(1): 1205–1224.

[5] Peng H, Long F, Ding C. 2005. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238.

[6] Agarwal B, Mittal N. 2013. Optimal Feature Selection for Sentiment Analysis. In: Proceedings of the 14th international conference on intelligent text processing and computational linguistics (CICLing-2013), vol 7817, no 1, Samos, pp 13–24.

[7] Aphinyanaphongs Y, Fu LD, Li Z, Peskin ER, Efstatiadis E, Aliferis CF, Statnikov A. 2014. A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization. J Assoc Inf Sci Technol 65(10):1964–1987.

[8] Manning CD, Raghvan P, Schutze H. 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge.

[9] Wang L, Wan Y. 2011. Sentiment Classification of Documents Based on Latent Semantic Analysis. In: Proceedings of the international conference on advanced research on computer education, simulation and modeling (CESM), Wuhan, pp 356–361.

[10] Agarwal B, Mittal N. 2012. Text Classification Using Machine Learning Methods-A Survey. In: Proceedings of the 2nd international conference on soft computing for problem solving (SocPros-2012), vol 236, no 1, Jaipur, pp 701–710.

[11] Pang B, Lee L. 2008. Opinion Mining and Sentiment Analysis. Foundations and trends in information retrieval, vol 2, no 1–2. Now Publishers, Hanover, pp 1–135.

[12] Tan S, Zhang J. 2008. An Empirical Study of Sentiment Analysis for Chinese Documents. Expert Syst Appl 34(4): 2622–2629.

[13] Pang B, Lee L, Vaithyanathan S. 2002. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Prague, pp 79–86

[14] Abbasi A, Chen H, Salem A. 2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Trans Inf Syst 26(3): 1–34.

[15] Nicholls C, Song F. 2010. Comparison of Feature Selection Methods for Sentiment Analysis. In: Proceedings of the 23rd Canadian conference on advances in artificial intelligence. LNCS, vol 6085, no 1, Ottawa, pp 286–289.

[16] Gamon M. 2004. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and The Role of Linguistic Analysis. In: Proceedings of the 20th international conference on computational linguistics, Geneva, pp 841–848.

[17] Simeon M, Hilderman R. 2008. Categorical Proportional Difference: A Feature Selection Method for Text Categorization. In: Proceedings of the 7th Australasian data mining conference, Glenelg, pp 201–208.