Optimizing Performance of Sentiment Analysis through Design of Experiments

Gary S. W. Goh*, Andy J. L. Ang, Allan N. S. Zhang
Singapore Institute of Manufacturing Technology (SIMTech)
*Agency for Science, Technology and Research (A*STAR)*
Singapore, Singapore
*E-mail: *gary_goh@mymail.sutd.edu.sg; garygsw@gmail.com*

***Abstract*—Traditional manual design of analytical processes is challenging as it requires a general analyst to have good grasping of numerous algorithms and the interaction effects between each technique and the data across multiple domains. Especially in an increasingly high data variety/multi-domain environment today, this design process can be very laborious/challenging. In this paper, we describe a design optimization approach using design of experiments to determine a suitable design in a standardized text classification process with high classification performance. We focus on sentiment analysis as a use case for this approach, as standard analytical methods in each phase of the sentiment analysis process have been established; from data pre-processing, feature selection and classification. In our proposed approach, we present an automatic and domain-free technique of using design of experiments to this design process, with the sentiment classification evaluation metrics as the performance criteria for optimization. In addition, we show that several interpretable analyses can be made to better understand the complex interaction effects of various analytical techniques with the data, which then can guide a general analyst to select more appropriate process design parameters for better text classification performance.**

***Keywords*—Text classification; sentiment analysis; high data variety; design of experiments; optimization**

I. INTRODUCTION

The explosion of social media usage today generated an abundance of customer feedback data that are easily available. The increasing accessibility to the internet and vocality of individuals encourage more people to share their opinions and experiences online. This emerging trend provides a new channel for companies to take a peek into their customers' perceptions of their brands and products and to gather valuable insights. Data gathered from social media are viewed as goldmines to market researchers as they represent genuine opinions from the public which can contain important and relevant insights. These online feedbacks also engender strong influence on demand as many more people are depending on them to make better informed purchasing decisions [1]. Such data are acquired from a high variety of social media platforms such as Twitter, Facebook, forums and reviews websites. However, most of the data is typically *unstructured* and requires text mining techniques to reveal meaningful interpretations [2].

Consequently, a text mining application known as sentiment analysis has grown its popularity. Sentiment analysis, sometimes also known as opinion mining, refers to the field of study which aims to automatically read and recognize human's "opinions, sentiments, evaluation, appraisals, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes" using computational means [2]. It is usually applied to large sets of unstructured data to gain an overview of the "public" opinions with respect to an entity of interest. The applications of sentiment analysis are wide-ranging. Sentiment analysis technologies enable companies to effectively listen attentively to their customers' voices online so that they can respond quickly by adapting their marketing strategies, brand position, product development and operations planning [3, 4].

Sentiment analysis can be regarded as a text classification problem as most of the previous works such as [5-7] adopt the task as a text categorization into sentiment classes, e.g. positive, negative, or neutral. Due to its gaining popularity, there have been numerous experimental studies on various pre-processing techniques, feature selection and representations and machine learning models [8-10]. As a result, there exist several useful methods available for the sentiment analysis process.

However, this analytical process is often *linear*, and thus any output in the earlier steps will directly affect the results from subsequent steps. There might also be significant interaction effects between the various techniques as well as with the data, which if without good understanding of the techniques and data, will be very difficult for a general analyst to detect. Furthermore, with the broadening issue of high data variety (one of the Vs in the 5Vs of Big Data) in the social media context, there is no "one-size-fits-all" solution to the process design. In order to address these issues, we present an exploration of the use of *design of experiments* to determine a suitable design in the sentiment analysis process to achieve high performance.

Design of experiments is a systematic statistical method that is intended to be used for "black-box" processes and hence does not require specific domain knowledge to optimize the design process. This allows the proposed approach to be "domain-free". The use of *factorial design*

experiment strategy also offers an efficient way to plan for the required experiments necessary to identify the relationship between factors that affects a process and the output *response* of that process [11]. This strategy, compared to the one-factor-at-a-time, employs Fisher's principles of blocking and orthogonality, which increases the differentiating power of each experiment to quantify the main and interaction effects of the factors. The aim of the experiments is to find the most suitable design configuration that optimizes the response – in this case the response refers to the sentiment classification performance.

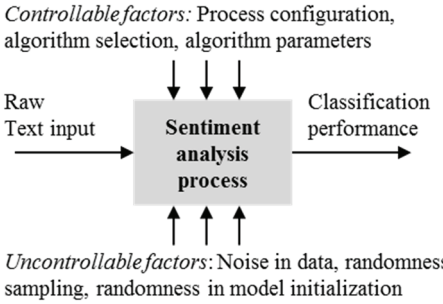

Figure 1. The sentiment analysis process' performance can be generated based on its output given raw text input and it is affected by controllable and uncontrollable factors.

The response of an experiment is dependent on two types of factors as shown in Fig. 1 – controllable and uncontrollable factors. The main emphasis is to investigate the effects of the controllable factors when it is impractical to eliminate the uncontrollable factors. With randomization and replication, the variability to the response caused by uncontrollable factors can be reduced and quantified. To this end, to optimize the response, we simply need to control the statistically significant controllable factors that are identified via hypothesis testing using the results data obtained from the experiment results.

The rest of this paper is organized as follows: Section II covers a brief review of the recent related work with respect to design optimization for machine learning processes; Section III describes the standardized sentiment analysis process that is used within the context of this study; Section IV introduces the datasets and setup of the experiments conducted; Section V discusses the results of the experiments and presents the analysis of the insights; and finally Section VI summarize our findings and suggest some potential future works.

II. Related Work

In the literature, there have been a few extensive work done on the *design and analysis of machine learning experiments* [12-16]. The essence of these studies is largely similar to the gist of this paper – applying statistical analysis on the results from empirical experiments to fine-tune the algorithm configuration in a machine learning process. These studies have also been applied on numerous domains.

Most of them, such as in [13, 14, 16], focus on the selection of a superior classification learning algorithm based on a single dataset, while [15] explore situations beyond just one dataset i.e. global optimization for multiple domains. Our goal differs slightly from theirs, but also complementary as we aim to search for the optimal set of configurations across the entire linear chain of analysis, i.e. from data pre-processing to classification, based on a single dataset or domain. As such, some related theoretical concepts are borrowed from these studies into this paper such as the use of *bootstrapping* as a sampling method. In our research, we also attempt to extend the interpretations of the experiment results to include the investigation of significant interaction effects between the processes in the linear chain of analysis flow.

Our hypothesis is that text pre-processing techniques along with the feature selection step, not just the classification learning algorithm, have considerable effects to the classification performance, especially in text mining applications, as described in [8]. To the best of our knowledge, a study which explores the relevance of experimental design and analysis on text mining applications has yet to be found in the literature. We intend to cover this research gap by using sentiment analysis as a case example.

III. Sentiment Analysis Process

In this section, we describe the details of the sentiment analysis process and the experimental conditions used in the subsequent parts of this paper. In addition, the considerations in the selection of the response variable from a number of evaluation metrics are also briefly discussed.

A. Process sequence

In our exploration, we adopt a sentiment analysis process which mainly includes three major steps: (i) data pre-processing; (ii) feature selection; and (iii) classification. The implementation is carried out in two phases – training phase and operational phase. The entire sentiment analysis process is summarized in Fig. 2.

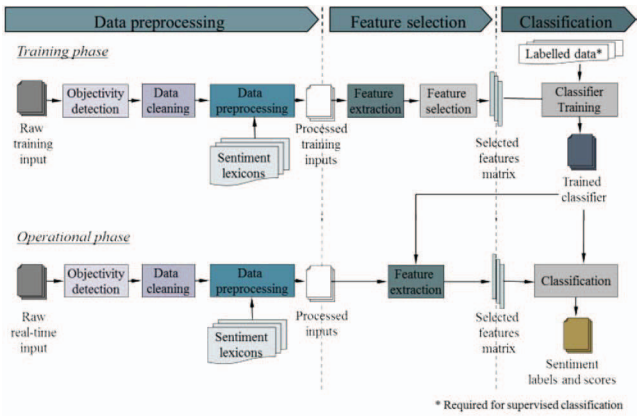

Figure 2. The adopted standardized sentiment analysis workflow can be broken down into three major steps and two phases.

The objective is to find an optimal configuration of the steps in the training phase so that it can generally perform well during the operational phase. In each of the steps, there already exists a myriad of well-known generic analytical algorithms. In accordance to the terminologies used in design of experiments, the configuration of each step corresponds to each controllable factor in the process. For the sake of simplicity, we select 11 controllable factors with a naïve two-level design as a basis for optimization. The selected factors and their level coding are presented in Table I. In this instance, the controllable factors are designed to be categorical, e.g. on or off, as the algorithm choice and their hyper parameters (if any) are fixed. Even though all factors selected are discrete, continuous factors can also be chosen and analyzed using the same methodology.

TABLE I.
11 SELECTED CONTROLLABLE FACTORS AND THEIR TWO-LEVEL CODING CONFIGURATIONS. REFERENCES TO THE ALGORITHM DETAILS (IF ANY) ARE INCLUDED. ANY FIXED HYPER PARAMETER VALUES ARE ALSO STATED.

i	Controllable factors (x_i)	Configuration coding	
		Low (-1)	High (+1)
1	Case preservation	Off	On
2	Repetition handling	Off	On [17]: Fixed length=3
3	Sentence segmentation	Off	On [18]: Punkt
4	Spelling correction	Off	On [19]: Norvig
5	Emoticon handling	Off	On [20]: Vashisht & Thakur
6	Punctuation removal	Off	On
7	Negation handling	Off	On [21]: Das & Chen
8	Stopword removal	Off	On [22]: Porter
9	Stemming	Off	On [23]: Porter
10	Feature selection	Off	On [24]: SAIG[1], top=75%
11	Classification	MNB[2]	SVM[3]: C=10, gamma=0.01

We acknowledge that there are several other workflows pre-defined by other authors in previous work related to sentiment analysis such as in [5, 25], and hence the optimal settings from our results for each step do not represent those from other processes. However, the focus here is to illustrate the concept of using design of experiments to optimize for any sentiment analysis process. As such, focusing on one standardized process is adequate.

B. Response variable

There are a number of evaluation metrics that measures the performance of classification problems in general, such as accuracy, recall, precision, and the F1-score (a harmonic mean between recall and precision). The selection of the response variable depends on the relevance of the metrics to the context. In some cases, recall may seem more critical than precision, or vice versa. If multiple metrics are deemed to be essential, each of them can be assigned a weight so as to obtain a final weighted response. In this paper, we select the F1-score as the main response to optimize; this selection is arbitrary and it is to be subjected to the context of the problem.

[1] **S**parsity **A**djusted **I**nformation **G**ain – a feature selection metric
[2] **M**ultinomial **N**aïve **B**ayes classifier
[3] **S**upport **V**ector **M**achine classifier

IV. EXPERIMENT SETUP

To ensure that the findings of the experiments can be applicable in a high data variety/multi-domain environment, we consider multiple datasets from diverse domains. These datasets are introduced in this section. The sampling procedure and experimental settings are also described in brief.

A. Datasets

Two readily available datasets used in the field of sentiment analysis, one in the domain of beauty products and the other in the domain of movies, are selected in our experiment. In addition, a new dataset that was scraped and prepared on our own from a separate domain of restaurants is also included in this study. The details of the datasets selected are described in Table II.

TABLE II.
DESCRIPTION OF THE THREE DATASETS SELECTED. REFERENCES TO THE SOURCE OF THE DATASETS ARE INCLUDED.

Domain	Source	Total rows	Score ratings
Beauty products	Amazon [26]	252,056	Scale of 1-5
Movies	IMDb [27]	47,582	Positive/negative
Restaurants	Yelp[4]	5,066	Scale of 1-5

All three datasets contain customer reviews and their corresponding score ratings. All reviews with ordinal score ratings of scale 1-5 are translated to positive/negative labels based on a simple rule: reviews with scores greater than 3 is labeled as positive, else negative.

B. Sampling

Multiple samples are necessary to allow for replications in our experiment in order to quantify the effects of the uncontrollable factors and isolate them from the factors' effects. The sampling method is therefore imperative to achieve a good estimation of the experimental error.

The detailed sampling procedure used in this study is as follows. For each dataset, three test sets are randomly chosen from the original dataset without replacement. Following which, the bootstrapping technique is used to generate three train sets by randomly selecting from the remainder of the dataset *with* replacement. In all of the samples, the proportion of positive and negative reviews is equally sampled to ensure an unbiased balanced labeled input to the machine learning model. Due to the varying sizes of the datasets, different sampling sizes are used and they are tabulated in Table III.

TABLE III.
VARYING SAMPLE SIZES SELECTED FOR THE TEST SETS AND THE TRAIN SETS FOR EACH DATASET.

Domain	Test sets size	Train sets size
Beauty products	1000	3000
Movies	1000	3000
Restaurants	300	2000

[4] Please contact the corresponding author for the release of the Yelp restaurants dataset; on demand only.

C. Experimental settings

The procedure of the experiments is described here. Each experiment consists of 2048 runs from a full 2^{11} factorial design with the configuration coding shown in Table I. All runs are executed in parallel to trim off the total execution time of the experiments. After each run, the trained classifier is evaluated based on three test sets of the respective dataset and the mean F1-score is recorded. The same experiment is replicated three times per dataset using the prepared three train sets and evaluated using the same test set for control.

In addition to the steps of the training phase detailed in Fig. 2, a simple pruning of the term-document matrix is done prior to the feature selection step with the following filter: $0.05 \leq document\ frequency \leq 0.95$. The purpose is to remove features that occur too frequently or infrequently so as to reduce the number of features to consider in the feature selection step. The addition of this step is experimented to be negligible to the response value but will help to reduce the total run time required to complete the experiments considerably; from about 10 hours to 20 minutes per dataset.

Lastly, the specifications of hardware used to conduct the experiments are: Processor: Intel(R) Xeon(R) CPU E5-2695 v3 @ 2.30GHz 2.29Ghz; Processor Cores: 20; Storage Size of 100Gb; Memory (RAM): 32GB; and Operating System: Windows Server 2012 R2 Datacenter (64-bit).

V. Results and Discussion

In this section, the interpretations of the results from the experiments are discussed. Firstly, the significance of the effects is investigated by analyzing the relative size of the effects to the response variable. Secondly, we share our insights on the common significant effects across the datasets. Lastly, we examine the optimal solutions within the pre-defined solution space and compare their performances with typical settings.

A. Significance of effects

As it is fairly common to assume that the effects of high order interactions are insignificant, only third order interactions effects and below are taken into account. With 11 factors, there are a total of 231 effects to consider as evaluated by $C_{(11,\ 3)} + C_{(11,\ 2)} + C_{(11,\ 1)} = 231$. Out of the 231 effects, we conduct statistical *t*-tests for each effect to test for significance. The significance level α is set as 0.05. The test hypotheses are written as $H_0: E_j = 0$ and $H_1: E_j \neq 0$ while the *t*-statistic for the j^{th} effect is calculated by:

$$t_j = \frac{E_j - E(E_j)}{\sqrt{Var(E_j)}} = \frac{E_j - 0}{\sqrt{\frac{4s_p^2}{2^k r}}}, \qquad (1)$$

where E_j denotes the size of the j^{th} effect; s_p^2 is the pooled variance from the three replications; k is the number of factors in the experiment which is 11; and r is the number of replications which is 3. The degrees of freedom is determined by $2^k(r-1)$.

From the results of the hypothesis tests, 71, 74 and 77 significant effects from the beauty product dataset, movies dataset and the restaurants dataset have been found respectively. These significant effects can be used to construct a formal mathematical model to describe the black-box's transformation of the inputs to the response variable.

Apart from looking at statistical significance, a simple Pareto analysis of the effects' absolute size can also be performed as shown in Fig. 3. From the Pareto charts, they can reveal very quickly which of the effects are more critical.

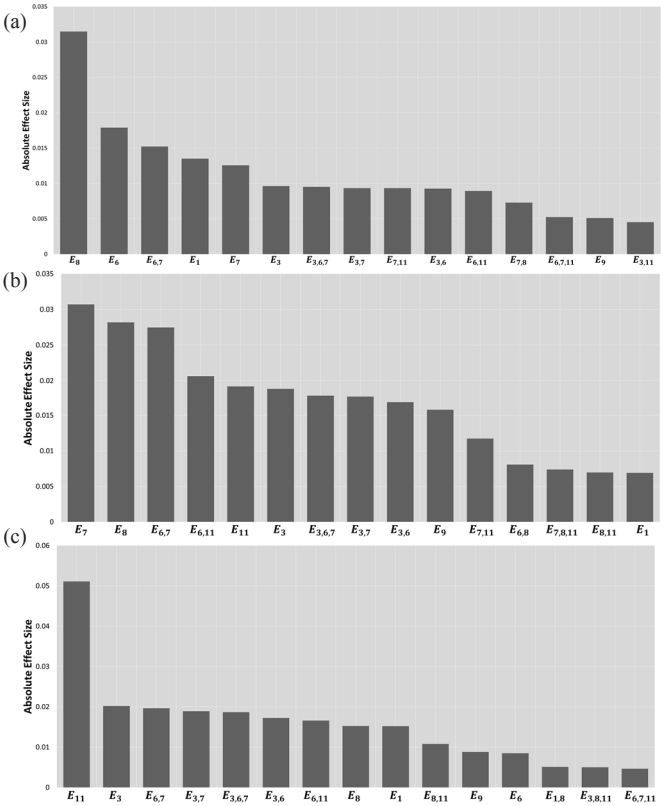

Figure 3. Pareto charts showing the top 15 effects' size for (a) beauty products dataset, (b) movies dataset, and (c) restaurants dataset.

It is clear that out of all the effects, only a small number of them are vital to explain the relationship between the factors and the response output. It is also interesting to note that the most significant effect is not the same for all three datasets; the most significant effect is E_8 (stopword removal), E_7 (negation handling) and E_{11} (classifier) for the beauty products dataset, movies dataset and the restaurant dataset respectively. This supports our claim that every dataset is unique and therefore should be treated differently with proper fine-tuning of the factors in the analysis process.

B. Common significant effects

There are several significant effects observed that are ubiquitous across all three datasets. A set of common significant effects out of the top 15 significant effects identified in Fig. 3 for all three dataset, grouped by their order of the interaction, is reflected in Table IV.

TABLE IV.
SET OF COMMON SIGNIFICANT EFFECTS FROM THE TOP 15 SIGNIFICANT EFFECTS OF EACH DATASET. THEY ARE GROUPED BY THEIR ORDER OF INTERACTIONS.

Main factors	2nd order interactions	3rd order interactions
• E_1 (preserve case) • E_3 (sentence segmentation) • E_6 (punctuation removal) • E_8 (stopword removal) • E_9 (stemming)	• $E_{6,7}$ (punctuation removal + negation handling) • $E_{3,7}$ (sentence segmentation + negation handling) • $E_{3,6}$ (sentence segmentation + punctuation removal) • $E_{6,11}$ (punctuation removal + classifier)	• $E_{3,6,7}$ (sentence segmentation + punctuation removal + negation handling) • $E_{6,7,11}$ (Punctuation removal + negation handling + classifier)

The result suggests that the main factors, x_1 (preserve case), x_3 (sentence segmentation), x_6 (punctuation removal), x_8 (stopword removal), and x_9 (stemming), should always be considered in general.

Similarly, there exist common significant 2nd order factor interactions such as x_6x_7 (punctuation removal + negation handling), x_3x_7 (sentence segmentation + negation handling) and x_3x_6 (sentence segmentation + punctuations removal) across the datasets. These 2nd order factors interactions are expected to be highly related as the algorithms of these steps, Das & Chen and Punkt algorithm, involves identifying the location of the succeeding punctuation as a termination point in the algorithm. This phenomenon can also explain the presence of a common significant 3rd order factor interaction $x_3x_6x_7$ (sentence segmentation + punctuations removal + negation handling).

An attempt to explain the presence of common significant factor interactions such as x_6x_{11} (punctuation removal + classifier) and $x_6x_7x_{11}$ (punctuation removal + negation handling + classifier) is an observable fact that the performance of SVM classifier is usually more sensitive to the presence of noise in the data, as compared to MNB classifier; in view of the fact that when negation handling is performed with punctuation removal, more negated features are introduced and thus this effect results in some unnecessary noise being fed into the data.

C. Optimal solutions

To assess the usefulness of the experiments' optimal solutions within the pre-defined solution space, three separate observations are taken at the end of the experiments for each dataset. Firstly, the optimal setting is observed from the setting of the run with the highest mean F1-score. Secondly, the total average of all of the 2048 runs' mean F1-score is observed; referred to as the average solution. Lastly, the mean F1-score of a particular run with the settings for all controllable factors set as +1 is also observed; referred to as the default solution. The choice of these three observations serves as a comparison between likely options that an average analyst would make without prior information that can be obtained from the experiments. These three separate observations for each dataset are tabulated in Table V.

TABLE V.
MEAN F1-SCORES FOR THREE CASES FOR EACH DATASET.

Dataset	Mean F1-scores		
	Default	Average	Optimal
Beauty products	0.7070	0.7220	0.7637
Movies	0.7772	0.7462	0.8135
Restaurants	0.7753	0.7557	0.8245

Table V demonstrates that the default setting will not always produce better classification performance as compared to the average F1-score obtained from all runs. It is also important to note that the difference in F1-score between optimal and average is substantial (on average about 5.9% points). These observations highlight the value of using design of experiments to search for the optimality even when the pre-defined solution space is small to begin with. They also further reinforce the need for a custom design process for different datasets. To end this section, the optimal solution for each dataset is presented in Table VI.

TABLE VI.
OPTIMAL SOLUTION FOR EACH DATASET.

i	Controllable factors (x_i)	Optimal solution of dataset		
		Beauty	Movies	Restaurants
1	Case preservation	-1	-1	-1
2	Repetition handling	+1	+1	-1
3	Sentence segmentation	+1	+1	-1
4	Spelling correction	+1	+1	-1
5	Emoticon handling	+1	+1	-1
6	Punctuation removal	-1	-1	+1
7	Negation handling	+1	+1	-1
8	Stopword removal	-1	+1	+1
9	Stemming	+1	+1	+1
10	Feature selection	-1	+1	-1
11	Classification	-1	-1	-1

VI. CONCLUSIONS

In summary, this paper proposes the application of design of experiments to determine a suitable design in a standardized text classification process to achieve high performance. This method also proves to be versatile in a high data variety/multi-domain environment as showcased by the application of the approach on three separate datasets from different domains. Although this method is similar to other "grid search"/"brute force" optimization methods, the benefits of adopting the design of experiments methodology include neat interpretations of the possible significant interactions that may be hard to identify by a general analyst. In our case study, we identified some 2nd order and 3rd order factor interactions and offered interesting insights to explain their existence. This is noteworthy as existing process optimization methods usually ignore the effects of interactions. With a stronger understanding of how each step in the entire linear analysis affect each other, it will serve as a good guide to a general analyst to design a suitable process flow for a specific domain.

As for potential future works, further advancements in the practice of design of experiments can be incorporated to improve the current application of optimizing for sentiment

classification. More sophisticated designs such as 3^k or mixed factorial design can be used to include more levels in each factor so as to investigate the presence of any non-linear relationships. Continuous factors, instead of discrete factors, such as the numerical value of hyper parameters can also be added in as one of the controllable factors with the intention of expanding the existing solution space. If the total number of runs has reached to the point where the total execution time of the experiments is of critical concern, alternatives such as fold-over designs and fractional factorial designs can be utilized to reduce the number of runs but at a cost of confounding. If the insignificant factors are known at prior, proper selection of the confounding identity matrix can minimize the confounding effect. Lastly, the Response Surface Methodology (RSM) could be employed to search for the global optimum design using repeated experiments of different factors' configuration coding.

ACKNOWLEDGEMENT

This work is partially supported under the A*STAR TSRP fund 1424200021 and Antuit-SIMTech Supply Chain Analytics Lab.

REFERENCES

[1] J. B. Horrigan, "Online Shopping," 2008.

[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies,* vol. 5, pp. 1-167, 2012.

[3] J. Zabin and A. Jefferies, "Social media monitoring and analysis: Generating consumer insights from online conversations," Aberdeen Group2008.

[4] L. C. Wood*, et al.*, "Using sentiment analysis to improve decisions in demand-driven supply chains," presented at the ANZAM Operations, Supply Chain and Services Management Symposium, Auckland, 2014.

[5] K. Dave*, et al.*, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519-528.

[6] B. Pang*, et al.*, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002, pp. 79-86.

[7] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 417-424.

[8] E. Haddi*, et al.*, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science,* vol. 17, pp. 26-32, 2013.

[9] T. O'Keefe and I. Koprinska, "Feature selection and weighting methods in sentiment analysis," *ADCS 2009,* p. 67, 2009.

[10] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, 1997, pp. 412-420.

[11] R. A. Fisher, *The design of experiments* vol. 12: Oliver and Boyd Edinburgh, 1960.

[12] E. Alpaydin, "Design and Analysis of Machine Learning Experiments," in *Introduction to Machine Learning*, ed: MIT Press, 2009, pp. 475-515.

[13] O. Irsoy*, et al.*, "Design and Analysis of Classifier Learning Experiments in Bioinformatics: Survey and Case Studies," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 9, pp. 1663-1675, 2012.

[14] R. R. Bouckaert, "Choosing between two learning algorithms based on calibrated tests," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 51-58.

[15] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *The Journal of Machine Learning Research,* vol. 7, pp. 1-30, 2006.

[16] O. T. Yildiz and E. Alpaydin, "Ordering and Finding the Best of K > 2 Supervised Learning Algorithms," presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006.

[17] C. Potts. (2011). *Sentiment Symposium Tutorial: Tokenizing*. Available: http://sentiment.christopherpotts.net/tokenizing.html

[18] (2016). *Punkt Sentence Tokenizer*. Available: http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt

[19] P. Norvig. *How to Write a Spelling Corrector*. Available: http://norvig.com/spell-correct.html

[20] G. Vashisht and S. Thakur, "Facebook as a Corpus for Emoticons-Based Sentiment Analysis," *International Journal of Emerging Technology and Advanced Engineering,* vol. 4, pp. 904-908, 2014.

[21] S. Das and M. Chen, "Yahoo! for Amazon: extracting market sentiment from stock message boards," in *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*, 2001.

[22] M. Porter, "List of Porter Stopwords," ed.

[23] M. Porter. (2006). *The Porter Stemming Algorithm*. Available: http://tartarus.org/~martin/PorterStemmer/

[24] B. Y. Ong*, et al.*, "Sparsity adjusted information gain for feature selection in sentiment analysis," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 2122-2128.

[25] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," presented at the Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004.

[26] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 165-172.

[27] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Association for Computational Linguistics*, 2004.