

# Data properties and the performance of sentiment classification for electronic commerce applications

Choi, Y & Lee, H

Published PDF deposited in Coventry University's Repository

**Original citation:**

Choi, Y & Lee, H 2017, 'Data properties and the performance of sentiment classification for electronic commerce applications' *Information Systems Frontiers*, vol 19, no. 5, pp. 993-1012  
<http://dx.doi.org/10.1007/s10796-017-9741-7>

DOI 10.1007/s10796-017-9741-7

ISSN 1387-3326


ESSN 1572-9419

Publisher: Springer

**This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.**

**Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.**

# Data properties and the performance of sentiment classification for electronic commerce applications

Youngseok Choi<sup>1</sup> · Habin Lee<sup>2</sup> 

© The Author(s) 2017. This article is published with open access at Springerlink.com

**Abstract** Sentiment classification has played an important role in various research area including e-commerce applications and a number of advanced Computational Intelligence techniques including machine learning and computational linguistics have been proposed in the literature for improved sentiment classification results. While such studies focus on improving performance with new techniques or extending existing algorithms based on previously used dataset, few studies provide practitioners with insight on what techniques are better for their datasets that have different properties. This paper applies four different sentiment classification techniques from machine learning (Naïve Bayes, SVM and Decision Tree) and sentiment orientation approaches to datasets obtained from various sources (IMDB, Twitter, Hotel review, and Amazon review datasets) to learn how different data properties including dataset size, length of target documents, and subjectivity of data affect the performance of those techniques. The results of computational experiments confirm the sensitivity of the techniques on data properties including training data size, the document length and subjectivity of training /test data in the improvement of performances of techniques. The theoretical and practical implications of the findings are discussed.

**Keywords** Sentiment classification · Opinion mining · Data properties · Comparative analysis · Sentiment orientation approach · Machine learning approach

## 1 Introduction

Due to the sheer volume of digital contents such as customer reviews, blogs and news corpora, sentiment classification has received enormous attention from large number of scholars as well as practitioners. Sentiment classification, also known as sentiment analysis, in electronic commerce is a task of judging the opinions (positive or negative) of customers about products and services (document, sentence, paragraph, etc.) based on computational intelligence such as machine learning. Sentiment classification provides organizations with a tool to transform data into ‘actionable knowledge’ that decision maker can use in pursuit of improved organizational performance. Customer review data can be used for development of market strategy and decision making for product/ service requirements for customer satisfaction, strategic analysis, and commercial planning (Gamon et al. 2005; Ye et al. 2009; Li and Wu 2010; Yu et al. 2013; Kang and Park 2014; Meisel and Mattfeld 2010; Yan et al. 2015; García-Moya et al. 2013). Government and public sector can also take advantages of analysing public sentiment from their blog and social media to obtain citizen feedback on new policy implementation (Ceron et al. 2014; Cheong and Lee 2011).

Due to the strategic importance of sentiment classification, the literature is abundant of many studies that propose various algorithms for sentiment classification to improve its accuracy particularly in business and management research domains (Bai 2011; Duric and Song 2012; Fersini et al. 2014; Kontopoulos et al. 2013; Sobkowicz et al. 2012), computer science (Denecke 2008; Melville et al. 2009; Prabowo and

---

✉ Habin Lee  
Habin.Lee@brunel.ac.uk

Youngseok Choi  
Youngseok.Choi@coventry.ac.uk

<sup>1</sup> Coventry Business School, Coventry University, Priory Street, Coventry CV1 5FB, UK

<sup>2</sup> Brunel Business School, Brunel University London, Uxbridge UB8 3PH, UK

Thelwall 2009; Hung and Lin 2013), and computational linguistics (Mullen and Collier 2004; Pang and Lee 2004; Aue and Gamon 2005; Okanohara and Tsujii 2005; Davidov et al. 2010; Liu and Yu 2014) among others. With more sophisticated machine learning algorithms or auxiliary resources for word polarity, researchers tried to make an improvement in accuracy.

However, in spite of such strategic values of sentiment classification techniques, the literature still lacks studies that provide practitioners and scholars with clear guidance on how and when to apply different sentiment classification algorithms to data obtained from different problem domains. While previous studies are focusing on increasing the accuracy of the algorithms, less effort was made to understand the impact of the linguistic properties of the dataset they use on the performance of the algorithms. The lack of clear guideline on the use of the algorithms against different datasets makes decision makers underutilize their data that may lead to under-optimal and sometimes wrong decisions by neglecting fits between data and algorithms.

Some studies (Pang et al. 2002; Moraes et al. 2013) showed a performance comparison among existing sentiment classification algorithms but they fail to suggest factors that can affect the performance of each algorithms as they just compared the performance with regards to the different test data or the experiment results from previous works. The lack of literatures regarding systematic comparison among various sentiment classification can be a critical barrier to apply the sentiment classification for researchers and practitioners.

Especially in case of machine learning based, the classification performance can directly depends on the quality of features obtained from a training dataset (Cortes et al. 1995; Mitchell 1999; Kira and Rendell 1992) as it usually adopts the bag-of-words approach, also known as bag-of-features approach (Pang et al. 2002). In machine learning, a feature is an individual measurable property of a phenomenon being observed (Bishop 2006), so features in sentiment classification indicate words or phrases in documents that reflect the sentiment of writers (Pang et al. 2002). The use of same classification for different training datasets fails to replicate the same level of high accuracy (Prabowo and Thelwall 2009) as features may vary according to the training dataset used for the classification model. To enhance the performance of sentiment classification, informative features need to be used for training classification and the quality of features extracted from documents in dataset can be closely related to documents' data quality such as subjectivity and length of document. This indicates that, understanding the impact of linguistic properties of data on classification performance is important.

This paper tackles the research gap described above by providing a systematic investigation on the impact of the linguistic properties of training and test data on the performance of diverse algorithms. The comparison will provide scholars

and practitioners with a practical guidance for the choice of algorithms for a given dataset.

Next section provides basic concepts of two most widely used sentiment classification approaches: machine learning and semantic oriented approach. Section 3 then describes the method used in the comparison of selected algorithms from two approaches against selected data properties. In Section 4, the experiment results are presented to show the effect of controlling the linguistic properties of training and test data. A discussion about the theoretical and practical contributions of the study is presented in Section 5. Section 6 summarises the findings of the study and provides future research directions.

## 2 Sentiment classification approaches

The approaches used for the sentiment classification can be categorized<sup>1</sup> into either machine learning or semantic orientation approach. Machine learning approach classifies the polarity of a document based on the features extracted from training data while semantic orientation approach refers a predefined resource like a dictionary that contains word polarities for classification.

### 2.1 Machine learning approach (MLA)

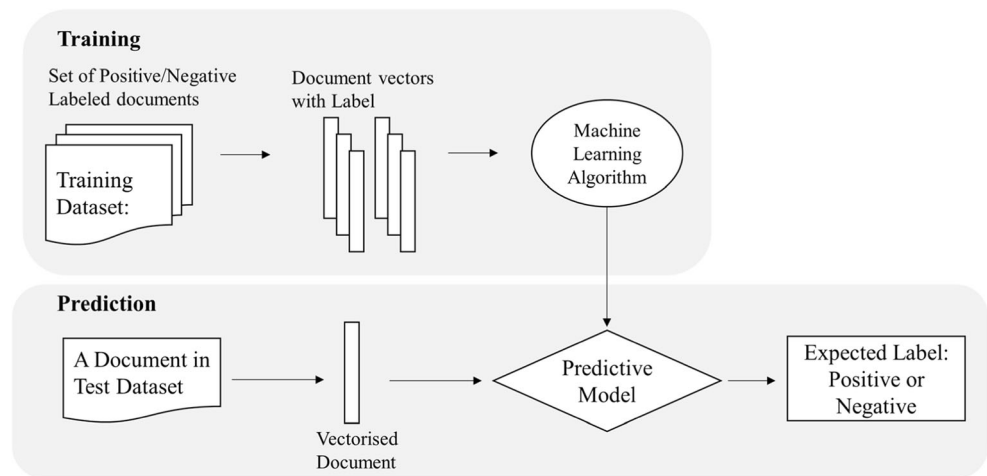
To implement machine learning based sentiment classification, the following standard of bag-of-words framework is usually used. Let  $\{f_1, f_2, \dots, f_m\}$  be a predefined set of  $m$  words (features) that can appear in a document and  $n_i(d)$  be the number of occurrences of  $f_i$  in document  $d$ . Then, each document  $d$  is represented by  $m$ -dimensional document vector  $\vec{d} = (n_1(d), n_2(d), \dots, n_m(d))$ . For example, if the predefined set is defined as  $\{\text{this, that, is, a, car, bicycle}\}$ , then the document "this is a car" can be presented as  $(1, 0, 1, 1, 1, 0)$ .

This bag-of-words framework is used for two step of MLA; training and prediction. As depicted in Fig. 1, labelled documents in training dataset are converted to document vectors during training. Basically, all documents in training dataset have one of two label; positive or negative. Document vectors and their label should be used to classify un-labelled documents in test dataset. During prediction, a document in test dataset is converted to a document vector. This document vector is then fed into the model, which generates predicted labels whether it is positive or negative (Bird et al. 2009).

In the previous research, Naïve Bayesian (Kang et al. 2012; Yoshida et al. 2014), support vector machine (Mullen and

<sup>1</sup> In terms of target contents for analysis, sentiment analysis can be categorized into three level: document, sentence, and aspect/attribute level sentiment analysis. In this paper, we will focus on the document level analysis as this is the most widely used approach in the literature.

**Fig. 1** The flow chart of machine learning based sentiment classification



Collier 2004) and decision tree (Sui et al. 2003) have been used for implementation of sentiment classification and these algorithms are reported to be effective for sentiment classification in the literature (Forman 2003; Dhillon et al. 2003; Sebastiani 2005; Wan et al. 2012; Ur-Rahman and Harding 2012). In this study, we also use these three algorithm for the experiment of impact of data properties on sentiment classification performance.

Naïve Bayes is one of the most popular MLA as it is easy to implement without any complicated iterative parameter estimation schemes (Wu et al. 2008). Based on the bag-of-words model, Naïve Bayesian based sentiment classification defines the likelihood of a document ( $d$ ) to be positive or negative as a sum of total probability over all mixture components, i.e.,  $P(d) = \sum_j P(d|positive)P(positive)$  for positive; where  $P(positive)$  is the probability of the positive and  $P(d|positive)$  is the probability of the document belonging to positive. For balanced training dataset,  $P(positive)$  and  $P(negative)$  are equal to 0.5 as the equal number of documents are used for positive and negative. To compute the likelihood of being positive or negative for given document, Naïve Bayesian approach applies the so-called “naïve assumption” that all words are independently used in all document, (Melville et al. 2009), which implies that  $P(w_i) = P(w_i|w_j)$  where  $w_j$  can be any other words. Based on this assumption, the probability of a document  $d$  being generated in positive is  $P(d|positive) = \prod_i P(w_i|positive)$ , where  $i$  is the number of words in a document.

The Naïve Bayes classification rule uses Bayes’ theorem to compute the probabilities of a document belonging to class  $c_j$  as follow,

$$P(positive|d) = \frac{P(positive)\prod_i P(w_i|positive)}{P(d)}$$

and the label with the highest likelihood is predicted, i.e.,

$$\operatorname{argmax}_{label} P(label)\prod_i P(w_i|label)$$

where label = {positive, negative}.

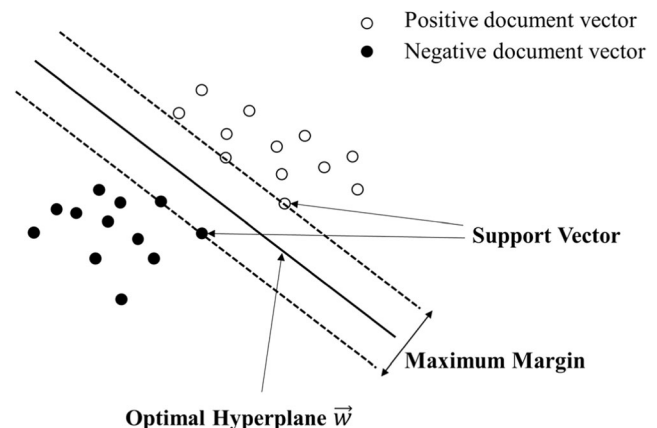
In this research, we adopt Multinomial Naïve Bayes as a classifier as it works well for text data that can easily be turned into numerical data, such as word-counts in a text.

Support vector machines (SVM) is considered as one of the best text classification methods (Xia et al. 2011). SVM is a statistical classification method based on the structural risk minimization principle of the computational learning theory. In sentiment classification case, SVM seeks a classification surface that can separate positive and negative document vectors from training dataset. All documents in training dataset can be vectorised using bag-of-word framework and then these document vectors can mapped into vector space as depicted in the Fig. 2.

The process for finding a hyperplane, presented by vector  $\vec{w}$  in Fig. 2, corresponds to a constrained optimisation problem to maximise the margin and the solution can be written as:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \alpha_j \geq 0,$$

where  $\alpha_j$ ’s are obtained by solving a dual optimisation problem and  $c_j \in (1, -1)$ , which means positive and negative. Those  $\vec{d}_j$  such that  $\alpha_j$  is greater than zero are called *support*



**Fig. 2** An illustration of SVM

vectors, since they are the only document vectors considered to find the hyperplane  $\vec{w}$ . Classification of test document vector consists simply of determining which side of  $\vec{w}$ 's hyperplane they fall on (Pang et al. 2002). In this study, we use radial basis function (rbf) as a kernel function for classifier.

Decision Trees Classifier (DTC) is a non-parametric supervised learning method used for classification and non-parametric regression analysis. The goal of DTC is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data features. DTC has been used for sentiment analysis as it is easy to understand the algorithms and its result (Pang et al. 2002). However, it is not commonly applied to long document classification because a long document tends to be converted into a high dimensional document vector.

DTC for sentiment classification usually work top-down, by choosing a feature (i.e., word) at each step that best splits the documents in training dataset (Rokach and Maimon 2005). For metrics to choose features for each step, we use Gini impurity (Kuh and De Mori 1995), which is a measure of how often a randomly chosen document from training dataset would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the subset (see Fig. 1 for labelling). Gini impurity can be computed by summing the probability  $f_i$  of each item being chosen times the probability  $1 - f_i$  of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category. To compute Gini impurity for a training dataset, let  $f_i$  be the fraction of items labelled with value  $i$  (in this case, positive and negative) in the set.

$$I_G(f) = \sum_i f_i (1 - f_i) = \sum_i (f_i - f_i^2) = 1 - \sum_i f_i^2 = \sum_{i \neq k} f_i f_k$$

## 2.2 Semantic orientation approach (SOA)

The SOA is based on identifying and selecting sentiment words in test documents (Wang et al. 2014). The main idea of this approach is to classify the sentiment of words in a document to infer its semantic orientation, i.e. whether the document has positive or negative opinion. To classify the sentiment of words, this approach usually uses external data sources such as corpus that has massive text data containing sentiment expression and dictionary that shows the polarity of large number of words. The corpus and dictionary can provide either the polarity score between  $-1$  and  $1$  ( $-1$  means extreme negative and  $+1$  means extreme positive) or sentiment polarity of words (i.e., positive or negative). Studies based on SOA may use learning algorithms to construct dictionary and semantic network between words from corpus therefore can be considered as a learning approach as well. More common approach is to use predefined sentiment lexicons such as

WordNet and SentiWordNet that provides lists of sentiment words. Synonyms, antonyms and hierarchies in WordNet with sentiment can be used to determine the polarity of documents (Andreevskaia and Bergler 2006; Das and Bandyopadhyay 2011). SentiWordNet, which was built upon WordNet, also have been used for external resource to score the polarity and classify its polarity (Devitt and Ahmad 2007; Denecke 2008).

Compared with the Naïve Bayes based sentiment classification, a semantic orientation specification based on lexicon of sentiment words is quite simple and intuitive. In this study, we use SentiWordnet (Baccianella et al. 2010; Esuli and Sebastiani 2006) as a lexicon resource for sentiment classification. It is the most popular open sentiment lexicon resource that is used for automatic sentiment classification. Many sentiment classification tasks extract sentimental words directly from SentiWordNet to avoid a manual sentiment lexicon or building new lexicon from the massive data using additional learning approach (Hung and Lin 2013). To classify the sentiment of a given document using the lexicon resource such as SentiWordNet, the elements of a vectorised document should be tagged grammatically, i.e., part-of-speech tag. In corpus linguistic and computational linguistics, *part-of-speech tagging* (POS tagging) is a process that makes up a word in a text (corpus) as corresponding to a particular part of speech, such as noun, verb, adjective, etc. The automatic POS tagging has a long history in computational linguistic studies and now its tagging accuracy reached to over 97%<sup>2</sup> (Manning 2011; Toutanova et al. 2003). POS tagged document vectors can be easily scored by comparing adjectives/adverbs/verbs in documents and those in SentiWordNet that contains the polarity scores of words (for example, adjective 'bad' has  $-0.625$  and 'worst' has  $-0.75$  in SentiWordNet). The pseudo-code representing overall procedure for SOA is presented in Fig. 3.

Though machine learning based approach usually outperforms other approaches, SOA using predefined lexicon is used for several reasons. Firstly, it is one of the most realistic ways to realize sentiment classifier without training dataset. Usually it is hard to have reliable training dataset for classification in practical situation for both researchers and practitioners. Human-rated and tagged data can be seen as the most reliable data but it requires too much effort and human resources. In addition, to replicate the complex classification algorithms is also very challenging while calculating the sentiment score of document based on lexicon is relatively easy to realize. Secondly, as a semantic lexicon such as SentiWordNet contains the general sense of word, this approach is free from domain dependency so it can show the moderate performance regardless of the application domain of test data (Denecke 2009).

<sup>2</sup> Stanford's POS tagger is the state-of-art tagger for English POS tagging.



**Fig. 3** Pseudo-code presenting the procedure of SOA

```

FOR every document in the TestDataSet:
    FOR each sentence in the document:
        TaggedSentence = POS(sentence)
        FOR SentiCandidate (adverb, adjective, and verb) in TaggedSentence:
            PolarityScore += LookupSentiWordNet(SentiCandidate)
            TotalCandidateCount++
        AveragePolarity = PolarityScore/TotalCandidateCount
        IF(AveragePolarity>0): RETURN POSITIVE
        ELSE: RETURN NEGATIVE
    
```

**2.3 Pros and cons of existing sentiment analysis method and performance of sentiment analysis**

As we reviewed in previous two subsections, MLA and SOA has different computational procedure to decide the sentiment of document and each method has their pros and cons accordingly. SOA doesn't need a training data for its classification but it needs external source instead. In the case of lacking enough training data for given domain, SOA can be applicable with external sources and implemented very easily by matching the words and external sources. Also, SOA is free from domain dependency as it considers the general words for expressing sentiment (Denecke 2009). But SOA can be faced with the problem resolving the semantic of word with multiple meaning (i.e., "mean" as a verb is neutral while its adjective is negative). Another concern is uncertainly regarding the classification of long document with both positive and negative sentiment.

Generally, if enough training data is given, MLA is preferable and known to show superior accuracy than SOA (Aue and Gamon 2005). But it is also known that MLA has domain dependency problem for training classifier. If the training data has different domain from test data, the accuracy can be dropped comparing to the classification with training using training data from same domain (Denecke 2009). Another variable of MLA can be a choice of classification algorithms as each algorithm has its strengths and weaknesses. MLA based on decision tree is very fast and robust to noise but has weaknesses to process long document that can cause complex tree. SVM based MLA can be also robust to noise and good for long document classification due to its strength in

high dimensional processing while it takes longer time than other algorithms such as decision tree and Naïve Bayes classifier (Pang et al. 2002). Table 1 shows the summary of pros and cons of each approaches.

Since sentiment analysis has been paid attention from academics and practitioners, many open source library has been developed for various programming language such as Java, python, Ruby, etc. Each library has different algorithm coverages so user can select the library according to the programming language and algorithms. Tables 2 and 3 shows the comparison and algorithm coverages of existing open source library. Note that we only consider the libraries that are still managed and updated by creators as some of libraries stop their update.

Table 4 summarises how this study extends existing studies on the sensitivity of sentiment analysis performance. Compared to existing studies, this research provides comprehensive investigation between data properties and the performance of sentiment classification algorithms covering four algorithms and four different datasets. This study conducts not only the performance comparison among basic sentiment classification algorithms, but also multi-dimensional comparisons based on data properties. Followings are unique findings from this study compared to existing studies. For MLA, we unveiled the existence of optimal word-count of documents for training the classification model and the effectiveness of documents with higher subjectivity as training datasets. The minimum training size for good performance was also validated using four datasets. We also found that the performance of SOA depends on the subjectivity and length of documents in test dataset. The result tells that SOA works well for shorter

**Table 1** Pros and Cons of SOA and MLA

Approach	Pros	Cons	Note
SOA	<ul style="list-style-type: none"> <li>- No need for training sets</li> <li>- Applicable without domain constraints</li> <li>- Easy to implement</li> </ul>	<ul style="list-style-type: none"> <li>- Uncertainty in long document classification</li> <li>- Difficulties in processing words with multiple meaning</li> </ul>	
MLA	<ul style="list-style-type: none"> <li>- Generally superior performance than SOA with training data</li> </ul>	<ul style="list-style-type: none"> <li>- Need for training data from same domain of test data</li> </ul>	<ul style="list-style-type: none"> <li>- Decision tree based classification is fast but may have problem for long document</li> <li>- SVM based classification has its strength in long document but may take longer time</li> </ul>

**Table 2** Existing Open-source Libraries for Sentiment Classification

Library Name	Supporting Language	Supporting Algorithms	Note
NLTK (Natural Language Toolkit) <sup>a</sup> (Bird 2006)	Python	Naïve Bayes, Maximum Entropy	scikit-learn library <sup>b</sup> can be used to apply more MLA.
CLiPS <sup>c</sup> (Smedt and Daelemans 2012)	Python	SOA	Supporting a <i>part-of-speech</i> tagging and including SentiWordNet
Stanford NLP library <sup>d</sup> (Manning et al. 2014)	Java	Most of MLA and SOA with SentiWordNet	Supporting a <i>part-of-speech</i> tagging
Weka <sup>e</sup> (Hall et al. 2009)	Java	MLA	Weka only supports the ML classifier. For the sentiment classification, another text processing library is needed.
tm library <sup>f</sup> (Feinerer 2015)	R	SOA	Other ML libraries are needed for MLA
sentimentalizer <sup>g</sup>	Ruby	SOA	

<sup>a</sup> [www.nltk.org](http://www.nltk.org)

<sup>b</sup> [scikit-learn.org/](http://scikit-learn.org/)

<sup>c</sup> [www.clips.ua.ac.be](http://www.clips.ua.ac.be)

<sup>d</sup> [Nlp.stanford.edu](http://Nlp.stanford.edu)

<sup>e</sup> [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

<sup>f</sup> [cran.r-project.org/package=tm](http://cran.r-project.org/package=tm)

<sup>g</sup> <https://rubygems.org/gems/sentimentalizer>

documents and higher subjectivity. All these findings regarding data properties were absent from previous studies as they focus on simple performance comparison among existing algorithms using limited datasets without data-driven perspective.

Barbosa and Feng (2010) and Pak and Paroubek (2010) shows the impact of subjectivity words on the performance of sentiment classification but only short documents such as Tweets were used in their experiments. Due to the experiment using limited dataset, they cannot investigate potential influence of document length on classification accuracy. Aue and Gamon (2005) have tackled the domain dependency problem of sentiment classification performance and the impact of training size on accuracy but they do not explain a reason why different test datasets produces different classification performances. They do not explain what different data properties their datasets have and how the properties make differences on performance. Ranade et al. (2013) report

decreased classification performance when a sentiment statistical scoring model based on sentence length and sentiment words is used for longer documents. However, the results are based on only online debate articles and only SOA is tested.

Similar to this study, there are other studies that also compare existing approaches. Pang et al. (2002) are one of the first scholars who provide the comparison of the performances of different MLA algorithms (Naive Bayesian, Maximum Entropy and Support Vector Machine) by collecting the results from existing studies that use the same dataset (movie reviews). They, however, do not provide the details of their experiment settings such as training size and the properties of the data they use therefore difficult to generalise the findings. They also do not investigate the impact of narrative and objective words on the performance which are one of important contributions of this study.

Tang et al. (2009) provide a referential review on sentiment classification studies as they discuss and introduce related

**Table 3** Summary of existing studies on the performance of sentiment analysis

Research	Classification Algorithms	Data Properties	Datasets used
Barbosa and Feng (2010)	SVM	subjectivity	Twitter Dataset
Pak and Paroubek (2010)	Naïve Bayes and SVM	subjectivity	Twitter Dataset
Aue and Gamon (2005)	Naïve Bayes	n/a	Car Review Dataset
Ranade et al. (2013)	SOA	Document length	Online Debate Dataset
Pang et al. (2002)	Naïve Bayes, SVM, and Maximum Entropy	n/a	Movie Review
Moraes et al. (2013)	SVM and Neural Network	Document length	Movie Review Data
This study	SAO and MLA (Multinomial Naïve Bayes, SVM, and Decision Tree)	Training size, document length, and subjectivity	Movie Review, Twitter, Hotel Review, and Amazon product Review Dataset

**Table 4** Data properties selected for the research

Data properties	Description	Algorithm	References
Document length/ Words count	The quantity of information depends on the length, or words count and information about author's sentiment can affect the quality of training as well as classification accuracy of test datasets.	MLA, SOA	(Abbasi et al. 2008), (Davidov et al. 2010), (Moraes et al. 2013), (Thelwall et al. 2010), (Stieglitz and Dang-Xuan 2012), (Ranade et al. 2013)
Document subjectivity	Subjective words can be the critical cues for sentiment polarity determination.	MLA, SOA	(Pang and Lee 2004), (Pang et al. 2002), (Lin and He 2009), (Liu 2010)
Training size	In the case of ML-based sentiment classification, the training size has a significant influence on the classification performance.	MLA	(Aue and Gamon 2005), (Ye et al. 2009), (Barbosa and Feng 2010), (Pak and Paroubek 2010)

issues and main approaches. They provide a performance comparison based on the reviews of existing studies, however, without discussing factors affecting the performance differences. Though their performance comparison table shows that the same MLA can show different performances with the same IMDB dataset (for example, they showed Naive Bayesian's performance varies between 65.9 ~ 81.5% according to other studies), they do not pinpoint the factors that cause such performance differences. Vinodhini and Chandrasekaran (2012) report the similar findings with Tang et al. (2009).

Moraes et al. (2013) provides results of empirical comparison between SVM and ANN using same datasets. Their findings show that the performance of SVM and ANN can be affected by word-count of documents in test datasets, however, they do not consider the properties of training datasets such as training size, subjectivity and word-count of documents that are considered in this study. As the performance of MLA strongly depends on the way of training a classification model, data properties of documents in a training dataset also need to be considered.

As reviewed above, most of existing sentiment classification studies provide simple performance comparison in terms of the "accuracy" of algorithms without considering the role of "data properties and setting" that may cause differences in the performance. Without the details of the data setting and the experiment control, the accuracy of algorithms cannot be replicated as the performances may vary according to the nature of the data used and the way experiments are conducted.

### 3 Method

#### 3.1 Selection of data properties for comparison

This study investigates the impact of linguistic properties of data such as word-count, size of training dataset and subjectivity of document on the classification performance of two approaches: SOA and MLA. The importance of these properties in sentiment classification is stated in related studies in

Table 4, however, none of them provides a comprehensive comparison of performances of different approaches on datasets with different properties.

Based on literature review, the most common data used for sentiment classification in electronic commerce and social media includes news documents, SNS messages, Blog documents, and customer review on products and services. It can be classified according to two dimensions: length and subjectivity. Studies on strategic decision for marketing usually apply sentiment classification to short SNS messages (Hennig-Thurau et al. 2015), product reviews in e-commerce platform (Yang and Chao 2015; Hu et al. 2014), or the Internet forums. News documents and financial columns, which are longer than social media messages and product reviews, can be used for financial decision support and risk assessment (Wu et al. 2014).

The quality and properties of training dataset are of considerable importance for the performance of MLA. Training datasets need to be defined by analysts in such a way that they are typical and representative of each individual class and both quality and size of training dataset are of key importance (Kavzoglu 2009). The size of training dataset is one of the critical properties determining the accuracy of supervised learning. Usually, more training dataset can improve the classification accuracy as too small training dataset can cause over-fitting. However, in practice it is not desirable to organise a very large dataset. To secure a reliable training dataset in a certain domain is a challenging task. For this reason, deciding the optimum size that can produce reasonably high level of accuracy is a common problem in machine learning studies.

A training dataset also needs to have informative and relevant features of their class where they belong to for accurate classification (Kotsiantis 2007). Too much noise and missing features in a training dataset can cause significant diminution in the performance of a MLA based classification algorithm. For this reason, training a classifier with an appropriate dataset is considered as important as implementing sophisticated algorithms in machine learning field and many studies try to improve the performance of classifiers by improving just the



quality of used training dataset (Weiss and Provost 2003; Batista et al. 2004; Sheng et al. 2008). Wrongly labelled training dataset (e.g., some positive documents in negative training dataset, and vice versa) causes a poor classification performance. The use of a training dataset with human-generated labels can solve this problem but not realistic when a massive training dataset is needed (Gamon 2004). One of the alternative approaches is to use polarized reviews from e-commerce platforms; for example, 1 star reviews for negative and 5 star reviews for positive training datasets using a 5 scale rating system (Pang and Lee 2005).

The length of documents in training dataset may be significant in determining the number of informative features. Long documents are likely to have many words that have sentiment even though we cannot confirm that most of sentiment words are informative for classifying the sentiments of the whole documents. The uncertainty of consistency between the sentiment of single words and that of whole document can be found in some theories related to the human utterance and behaviour. By the *Politeness Principle* (Leech 1983), people tend to mix opposite opinion to show their politeness and to emphasize their opinion. The example:

“This movie has a fantastic scale and a perfect location for a fantasy movie. But there’s no theme so I can’t understand what the director want to say. The plot is also awful. I do not want to recommend this movie to my friends.”

This review is obviously negative review but the paragraph has both positive and negative sentiment. By POS tagging, the words that have semantic orientation will be added as features for classifier but all the words are not informative. This example shows that why we need to analyse the effectiveness of long document as a training dataset.

In SOA, using a predefined semantic lexicon, the property of the test data can affect the accuracy of classification. The document with simple expression of sentiment is more likely to be classified correctly as people tend to use short words and use symbols in their comments to express their opinion (Khan 2011). If a document has both positive and negative words, the polarity score of the document may not be good enough for classification as the score is usually derived from the summation of scores of each individual word (Kim and Song 2013).

### 3.2 Data

One of the difficulties associated with sentiment classification of web contents is that datasets tend to be highly imbalanced as there is general tendency that users are willing to submit positive reviews while hesitant to submit negative ones (Liu and Yu 2014). For this reason, we select balanced datasets such as IMDB movie reviews (Maas et al. 2011) that are used

for many sentiment classification studies due to its balanced amount of data between positive and negative reviews. Twitter datasets (Mukherjee and Bhattacharyya 2012) and hotel review datasets (Lu et al. 2011; Pontiki et al. 2014) has been used in the previous studies and verified to be balanced datasets. Amazon review datasets contain reviews for small electronics and collected using script crawling for this research. All the repeated data has been filtered to avoid redundancy and non-English data also has been removed using NLTK (Bird 2006) language detection function. For the data from IMDB, Hotel Review, and Amazon, we assigned the positive and negative to the document following to the star rating form review authors as Pang et al. (2002) and Pang and Lee (2005) did. Twitter dataset was adopted from (Mukherjee and Bhattacharyya 2012)) as it contains manually tagged sentiment but we eliminated hashtag to make same experiment environment as the other review dataset had no hashtag.

The datasets cover various domains and contexts of e-commerce and social media as shown in Table 5. For the classification models, we adopt a SentiWordNet based word scoring for SOA and Naïve Bayesian, Decision Tree, and Support Vector Machine classification model for MLA. For the implementation of all classification methods and corresponding experiments, Python 3.0 (v.3.4.3) was used with text processing and machine learning libraries such as scikit-learn (v.0.71.1)<sup>3</sup> (Pedregosa et al. 2011), Anaconda (v.2.3.0 for 64 bit),<sup>4</sup> and CliPS (v.2.6) (Smedt and Daelemans 2012), etc.

## 4 Experiment results

### 4.1 The sensitivity of SOA on data properties

To investigate the sensitivity of the performance on word-count in test documents, we split the documents in the datasets into several groups having different word-counts except twitter dataset as it contains only short documents due to the word-count limitation of the twitter service. Table 6 shows how the accuracy of SOA changes as the word-count increases. The overall accuracy of the SOA was around 0.65 ~ 0.75, which shows the similar level of performance with other studies that use the same approach (Ohana and Tierney 2009).

From the experiment results, we can see that SOA based classification shows better accuracies when the test documents have less words. The accuracy of the results of documents with less than 200 words is significantly higher than those in other groups except hotel dataset. Hotel datasets results show the higher accuracy for longer document but it has small portion for longer document. For IMDB and amazon datasets, the

<sup>3</sup> We used `svm`, `naive_bayes`, and `tree` modules in scikit-learn library for the implementation of MLA.

<sup>4</sup> <https://www.continuum.io>

**Table 5** Datasets Specifications

Dataset	Size (positive/negative)	Domain/Context	Language
IMDB Dataset	10,000/10,000	Movie Review	English
Twitter Dataset	4000/4000	Social Data	English
Hotel Review Datasets	8000/8000	E-commerce/ Service	English
Amazon Review Datasets	6000/6000	Product Review (small electronics)	English

accuracy of short document (word count 0 ~ 100, 101 ~ 200) show higher accuracy than other longer documents. So we can conclude that SOA show better performance for those with less than 200 words than other documents.

Though the results do not imply the inverse linear relationships between word-count and accuracy, we can conclude that SOA is more appropriate for shorter documents (less than 200 words count).

Next experiment was designed to test the impact of subjectivity of test documents on the performance of SOA. In this experiment, we tagged POS and found the words that had semantic orientation based on SentiWordNet and finally found the document’s subjectivity by averaging the subjectivity

scores of the tagged words. Documents that have many strong polarity words such as “best” or “extremely” were expected to have higher subjectivity score.

As shown in the experiment result in Table 7, SOA shows the better performances for documents with higher subjectivities. That is, if the authors of documents use a strong negative words to express their sentiment, only few positive words might be found from the documents. This makes the document with strong subjective words easier to be correctly classified.

## 4.2 The sensitivity of MLA on data properties

### 4.2.1 Training size and document length

**Table 6** The accuracy of the SOA with regard to the word-count of test documents

Word-count in Document	Number of Document	Accuracy	Accuracy Difference
IMDB Dataset			
0 ~ 100	2557	0.7560	
101 ~ 200	9474	0.7052	-0.0508
201 ~ 300	3669	0.6721	-0.0331
301 ~ 400	1812	0.6440	-0.0281
401 ~ 500	996	0.6285	-0.0155
500~	1492	0.6635	0.0350
Amazon Review Dataset			
0 ~ 100	7637	0.6816	
101 ~ 200	4577	0.6552	-0.0263
201 ~ 300	2164	0.5924	-0.0628
301 ~ 400	825	0.5200	-0.0724
401 ~ 500	394	0.4797	-0.0403
500~	163	0.4724	-0.0073
Hotel Review Dataset			
0~100	5112	0.62.03	
101~200	3774	0.58.29	-0.374
201~300	1475	0.61.83	0.0354
301~400	733	0.61.94	0.0011
401~500	353	0.72.24	0.1030
501~	134	0.73.13	0.0090
Twitter Dataset			
0~100	8000	0.7943	

Firstly, we controlled two properties of training datasets, i.e., the size and the length of documents. Using the all datasets in Table 5, we increase the size of training dataset from 1% to 5% of the whole dataset. For each size, the training dataset was randomly selected from the whole dataset from each sentiment. The test was repeated 20 times for each size and for the MLA techniques. Fig 4 shows the average accuracy of classification result for different training sizes.

As Fig. 4 shows, the accuracy tends to increase as the size of training dataset increases. The training dataset below 1% of test dataset size do not guarantee the MLA’s maximum capability. Meanwhile, increment of training size over around 1% cannot improve the performance of classification. With more than 2% size of test dataset, most of the test results show over 0.75 accuracy. This result implies that the features for classification can be obtained from small amount of training dataset. Considering that the supervised learning has its basis on bag-of-words approach, this result shows that people usually use limited set of words and phrases to express their sentiment. Multinomial Naïve Bayes classifier shows the best performance while Decision Tree classifier generally lower than others. However, the latter shows the best performance for Twitter datasets.

To clarify the impact of word-count of document in training dataset on the accuracy of supervised classification, we controlled the word-count of document in training dataset and observed the variation in accuracy. For each interval of words count, we repeated the test 10 times. The results are shown in Table 8. Note that the experiments for word-count 0–50 of IMDB dataset and word-count 250–300 of Amazon Review Dataset were not conducted as both cases have less than 500

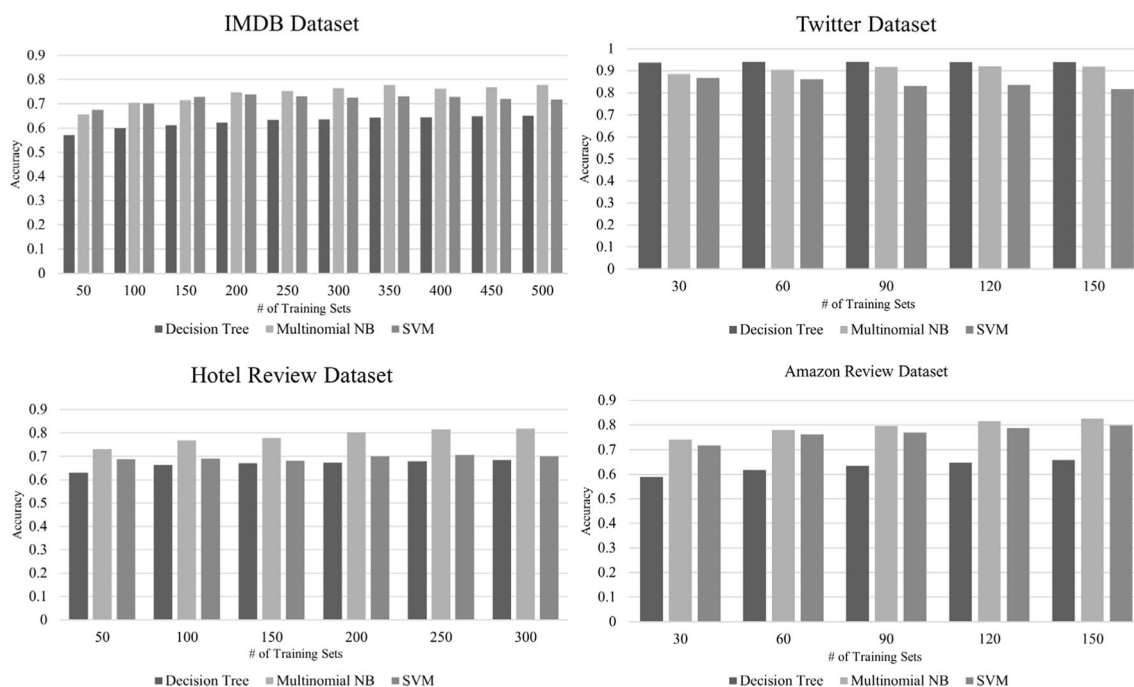
**Table 7** Accuracy of SOA with regard to subjectivity of test dataset

Subjectivity	Number of Document	Accuracy	Accuracy Difference
IMDB Dataset (Average subjectivity of dataset =0.531)			
0 ~ 0.5	7243	0.6520	
0.5 ~ 0.7	12,026	0.7112	0.0592
0.7 ~ 1.0	731	0.8114	0.1002
Overall	20,000	0.6931	
Hotel Review Dataset (Average subjectivity of dataset =0.544)			
0 ~ 0.5	4591	0.4907	
0.5 ~ 0.7	10,087	0.6851	0.1943
0.7 ~ 1.0	1322	0.8434	0.1582
Overall	16,000	0.6424	
Twitter Dataset (Average subjectivity of dataset =0.598)			
0~0.5	2693	0.6082	
0.5~0.7	1628	0.8428	0.2345
0.7~1.0	3679	0.9092	0.0664
Overall	8000	0.7943	
Amazon Review Dataset (Average subjectivity of dataset =0.519)			
0~0.5	4980	0.5149	
0.5~0.7	6116	0.6848	0.1699
0.7~1.0	904	0.6925	0.0077
Overall	12,000	0.6148	

reviews for those document length range so it was unable to training the classifier.

For all datasets and classification methods, the maximum accuracy can be attained with the documents around 50 ~ 150 words in training datasets. High Accuracy can be obtained

with documents around 50 ~ 200 words, and training dataset with documents that are shorter and longer than that showed lower accuracies. This means that there is an optimal length for document in training dataset for sentiment classification. Too short documents usually do not have enough features for



**Fig. 4** The sensitivity of MLA accuracy on the size of training dataset

**Table 8** Accuracy variation with the word-count of documents in the training dataset

Word-count	IMDB Dataset (Training size 500, Test size 10,000) <sup>a</sup>			Hotel Review Dataset (Training size 500, Test size 10,000)			Amazon Review Dataset (Training size 500, Test size 10,000)		
	DT	M- NB	SVM	DT	M- NB	SVM	DT	M- NB	SVM
0–50				0.6484	0.7919	0.6363	0.6288	0.8113	0.7343
50–100	0.6611	0.8189	0.7151	0.6874	0.8459	0.767	0.6408	0.8206	0.8006
100–150	0.6576	0.8168	0.7332	0.6797	0.8471	0.7999	0.6381	0.8143	0.7964
150–200	0.6429	0.7959	0.7331	0.6696	0.8389	0.7941	0.6406	0.8172	0.777
200–250	0.6354	0.7557	0.7334	0.6668	0.8308	0.796	0.6202	0.8141	0.7845
250–300	0.6392	0.7347	0.7141	0.6617	0.839	0.7914			

We excluded the twitter dataset for this experiment as all twitter documents have less than 50 words

DT Decision Tree, M-NB Multinomial Naïve Bayes, and SVM Support Vector Machine

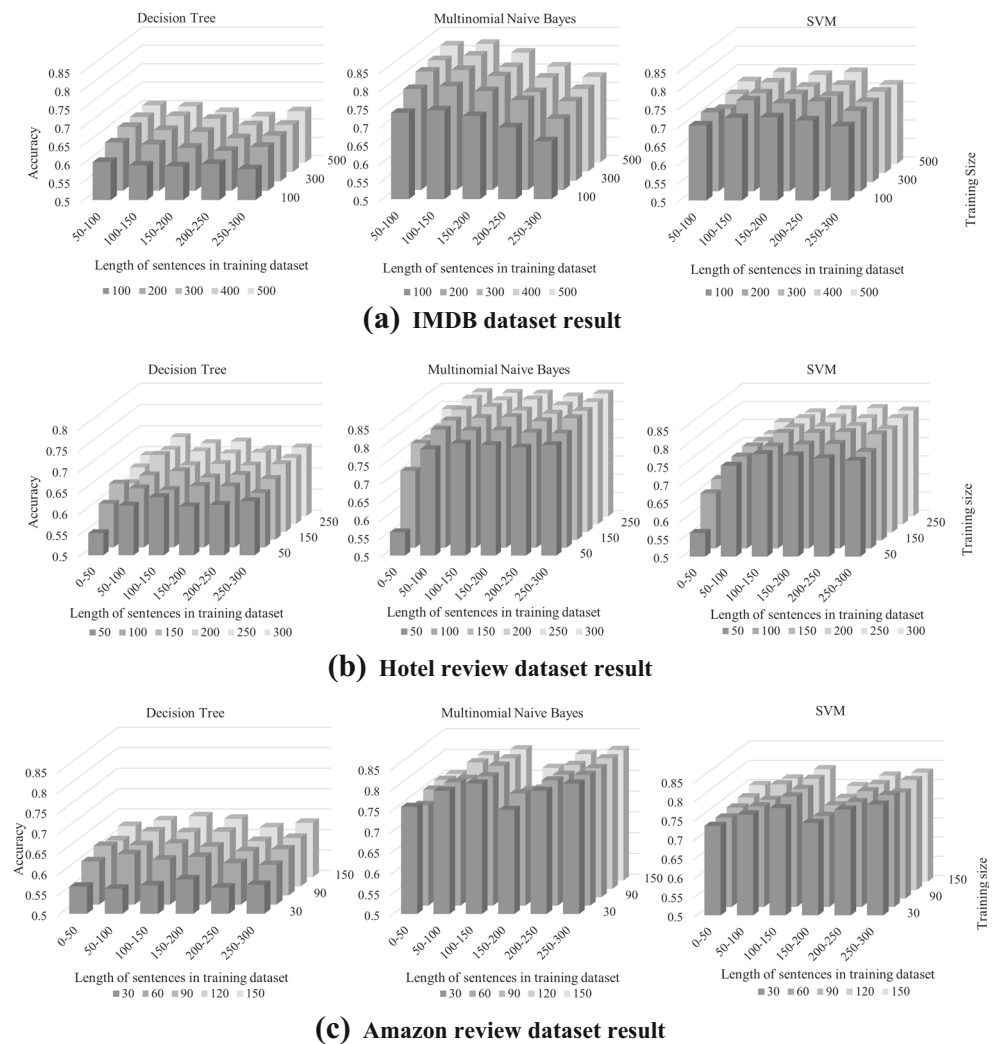
<sup>a</sup> IMDB Dataset has only small portion of short documents so we exclude documents with less than 50 words

classification and too long documents have noise features that hinders the accuracies. These results are also consistent with the Politeness Principle (Leech 1983) in normal utterance. Long documents usually have more features but not all

features are consistent and informative for sentiment classification.

For more sophisticated analysis of the impact of training size and word-counts on classification performance, we

**Fig. 5** The sensitivity of MLA performance on training size and word-counts



controlled two variables simultaneously. For every size of training dataset, we repeated the classification test with different range of word-counts. Fig 5 shows the experiment results.

The results show that training dataset with more than 3% training size that have 100 ~ 200 words performs best consistently. This result is consistent with previous experiment results that showed the optimal word-count of documents for training. Multinomial Naïve Bayes classifier showed best performance among the MLA with higher than 80% accuracy. The general comparison among the classifiers will be presented in Section 4.3.

#### 4.2.2 Training size and subjectivity

The aim of this experiment is to ascertain whether the training dataset with a higher average subjectivity can make better classification performance. We increased the training size to 5% of test dataset size and controlled the subjectivity of document in training dataset as well. The details of this experiment are shown in Fig. 6.

The results indicate that training with highly subjective document is more effective for increasing the performance for all training sizes. For most training sizes, classification

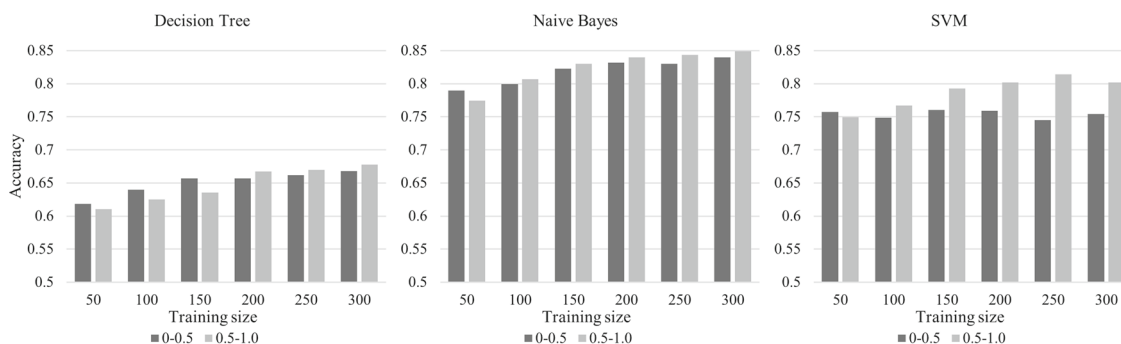
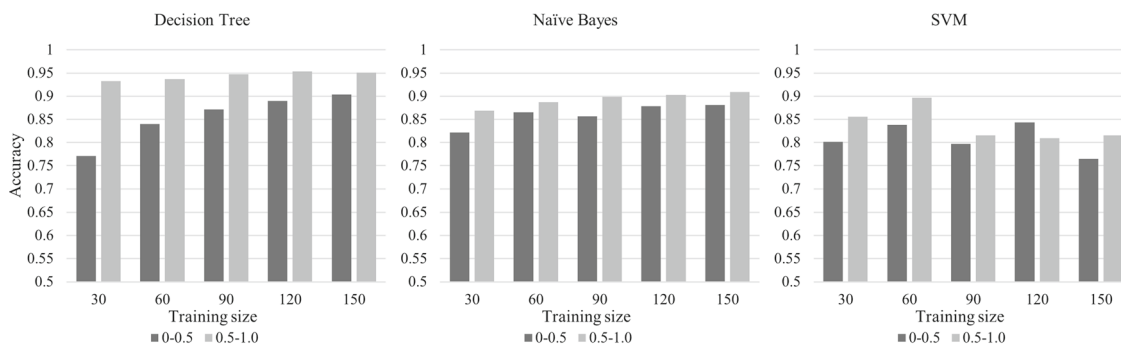
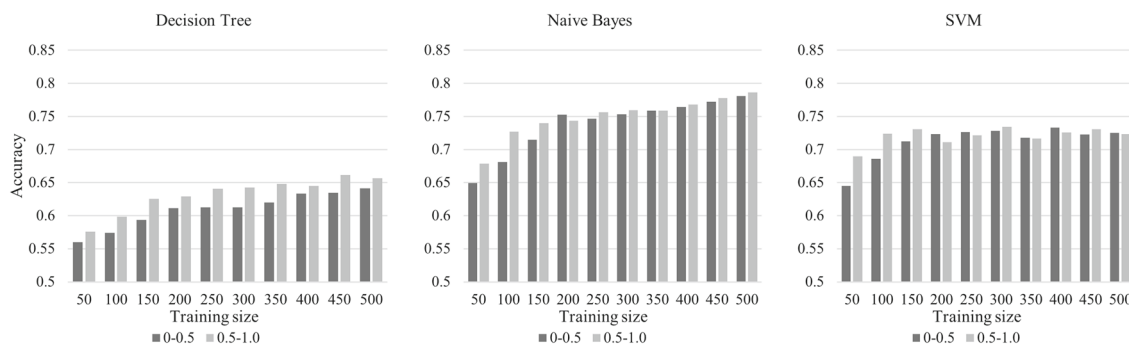
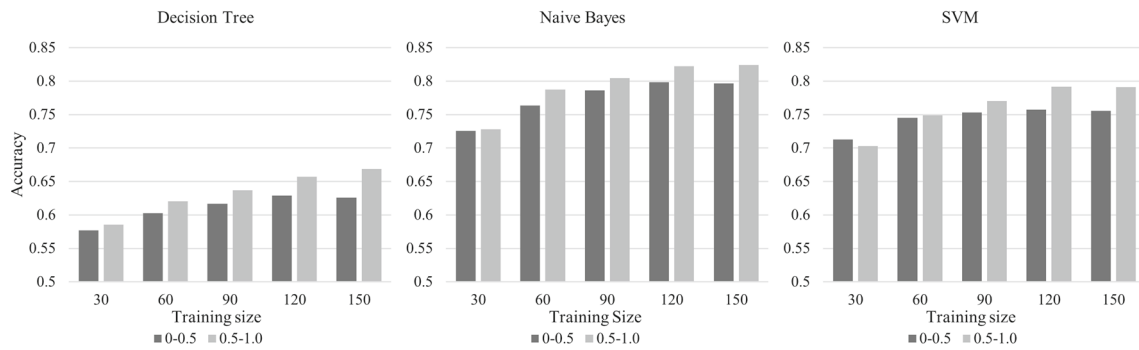


Fig. 6 The impact of subjectivity of training datasets with the change of training size





(d) Amazon review dataset result

Fig. 6 (continued)

with higher subjectivity produces higher accuracy with statistical significance for all datasets except twitter dataset and shows a good performance for most training size.

Every experiment using training datasets with higher subjectivity shows accuracy over 75% and a better performance than experiments with training datasets with lower subjectivity. In contrast, the accuracy of classification based on the training datasets with lower subjectivity documents tends to depend on the training size. The accuracy is higher in classification with larger training size. Overall accuracy of each subjectivity level also shows the difference by more than 8% points. For all dataset except twitter, Multinomial Naïve Bayes classifier outperforms other classifiers and shows the better performance as training size increases while other classifiers' performance starts to flatten. In addition to classifier's performance, the properties of datasets also have an impact on the accuracy. All the results using twitter dataset showed very high performance (over 90%) compared to others. This is closely related to the average subjectivity of test dataset. In the case of twitter dataset, its average subjectivity is 0.598, which is higher than other datasets (Hotel Review dataset – 0.544, Amazon dataset – 0.519, IMDB dataset – 0.508).

4.2.3 Document length and subjectivity

Next experiment was to find out the impact of subjectivity and document length on accuracy. This experiment has been performed using only IMDB dataset as the other datasets do not have enough number of documents which belong to all document length and subjectivity. IMDB dataset has 10,000 training data for both positive and negative so it can cover full spectrum with regard to the word-count and subjectivity constraints. We tested three times for different training size (200, 300, 400 documents). The results confirm the accuracy variation when we control all the properties of training data – training size, word-count of training document, and its subjectivity. A conclusion from the experimental results is that the longer

documents do not guarantee the higher accuracy even though the documents have higher subjectivity. For all three cases of training dataset size and for all three MLA approaches, having more than 400 words in the training documents fails to show better accuracy than training with middle length documents

Table 9 The distribution of documents in test data with regards to word-counts and subjectivity

word-count of documents in test dataset	0–0.5	0.5–1.0	
<b>IMDB Dataset</b>			
0–100	4.21%	8.38%	12.59%
100–200	16.68%	30.58%	47.26%
200–300	6.90%	11.63%	18.53%
300–400	3.56%	5.57%	9.13%
400–500	1.92%	3.09%	5.01%
500-	3.29%	4.22%	7.51%
	36.55%	63.45%	100.00%
<b>Amazon Review Dataset</b>			
0–100	16.66%	25.60%	42.26%
100–200	13.14%	18.51%	31.65%
200–300	5.26%	7.07%	12.33%
300–400	2.57%	3.58%	6.15%
400–500	1.14%	1.83%	2.98%
500-	1.93%	2.71%	4.64%
	40.70%	59.30%	100.00%
<b>Hotel Review Dataset</b>			
0–100	7.89%	39.54%	47.43%
100–200	9.54%	19.19%	28.73%
200–300	5.57%	8.02%	13.59%
300–400	2.64%	2.58%	5.23%
400–500	1.26%	1.24%	2.49%
500-	1.36%	1.18%	2.53%
	28.26%	71.74%	100.00%
<b>Twitter Dataset</b>			
0–100	31.65%	68.35%	100.00%

**Table 10** The comparison of the accuracies of MLA and SOA

word-count of documents in test dataset	SOA			Multinomial NB		
	Subjectivity		Overall	Subjectivity		Overall
	0–0.5	0.5–1.0		0–0.5	0.5–1.0	
<b>a. IMDB Dataset</b>						
0–100	0.7102	0.7875	0.7616	0.7755	0.8442	0.8212
100–200	0.6676	0.7295	0.7076	0.7851	0.8250	0.8109
200–300	0.6345	0.6896	0.6691	0.7846	0.8220	0.8081
300–400	0.5997	0.6801	0.6488	0.7851	0.8005	0.7945
400–500	0.5651	0.6629	0.6254	0.7786	0.8201	0.8042
500-	0.6484	0.6730	0.6622	0.7793	0.8140	0.7988
Overall	0.6525	0.7185	0.6944	0.7830	0.8239	0.8090
	SVM			Decision Tree		
	0–0.5	0.5–1.0	Overall	0–0.5	0.5–1.0	Overall
0–100	0.7411	0.7916	0.7747	0.6283	0.6985	0.6750
100–200	0.7452	0.7781	0.7665	0.6250	0.6567	0.6455
200–300	0.7114	0.7623	0.7433	0.6273	0.6612	0.6486
300–400	0.6742	0.7161	0.6997	0.6559	0.6595	0.6581
400–500	0.6328	0.7099	0.6803	0.6354	0.6434	0.6404
500-	0.6941	0.7275	0.7129	0.6575	0.6825	0.6716
Overall	0.7209	0.7649	0.7488	0.6323	0.6644	0.6527
<b>b. Hotel Review Dataset</b>						
0–100	0.7102	0.7875	0.7616	0.7755	0.8442	0.8212
100–200	0.6676	0.7295	0.7076	0.7851	0.8250	0.8109
200–300	0.6345	0.6896	0.6691	0.7846	0.8220	0.8081
300–400	0.5997	0.6801	0.6488	0.7851	0.8005	0.7945
400–500	0.5651	0.6629	0.6254	0.7786	0.8201	0.8042
500-	0.6484	0.6730	0.6622	0.7793	0.8140	0.7988
Overall	0.6525	0.7185	0.6944	0.7830	0.8239	0.8090
	SVM			Decision Tree		
	0–0.5	0.5–1.0	Overall	0–0.5	0.5–1.0	Overall
0–100	0.7411	0.7916	0.7747	0.6283	0.6985	0.6750
100–200	0.7452	0.7781	0.7665	0.6250	0.6567	0.6455
200–300	0.7114	0.7623	0.7433	0.6273	0.6612	0.6486
300–400	0.6742	0.7161	0.6997	0.6559	0.6595	0.6581
400–500	0.6328	0.7099	0.6803	0.6354	0.6434	0.6404
500-	0.6941	0.7275	0.7129	0.6575	0.6825	0.6716
Overall	0.7209	0.7649	0.7488	0.6323	0.6644	0.6527
<b>c. Amazon Review Dataset</b>						
0–100	0.5633	0.7171	0.6565	0.8049	0.8298	0.8200
100–200	0.4959	0.6479	0.5848	0.7907	0.7870	0.7886
200–300	0.5547	0.6639	0.6173	0.7781	0.7842	0.7816
300–400	0.5422	0.6721	0.6179	0.7532	0.7814	0.7696
400–500	0.6788	0.7455	0.7199	0.7664	0.8318	0.8067
500-	0.6509	0.7292	0.6966	0.8017	0.7723	0.7846
Overall	0.5465	0.6879	0.6303	0.7924	0.8055	0.8002
	SVM			Decision Tree		
	0–0.5	0.5–1.0	Overall	0–0.5	0.5–1.0	Overall
0–100	0.7844	0.7969	0.7920	0.6763	0.6621	0.6677

**Table 10** (continued)

word-count of documents in test dataset	SOA		Overall	Multinomial NB		Overall
	Subjectivity			Subjectivity		
	0–0.5	0.5–1.0		0–0.5	0.5–1.0	
100–200	0.7806	0.7861	0.7838	0.6088	0.6538	0.6351
200–300	0.7353	0.7559	0.7471	0.6482	0.6380	0.6423
300–400	0.6883	0.7488	0.7236	0.5649	0.6605	0.6206
400–500	0.7372	0.7864	0.7675	0.7007	0.6455	0.6667
500-	0.7414	0.7415	0.7415	0.7026	0.7662	0.7397
Overall	0.7674	0.7829	0.7766	0.6458	0.6608	0.6547
d. Twitter Dataset	SOA		Overall	Multinomial NB		Overall
	Subjectivity			Subjectivity		
	0–0.5	0.5–1.0		0–0.5	0.5–1.0	
	0.8693	0.8950	0.8869	0.8938	0.9305	0.9189
	SVM		Overall	Decision Tree		Overall
	0–0.5	0.5–1.0		0–0.5	0.5–1.0	
0.8013	0.8270	0.8156	0.9273	0.9656	0.9535	

(100 ~ 250 words) and any improved effects of training with higher subjectivity documents disappears. Usually, training using documents with higher subjectivity (0.5 ~ 1.0) and middle length word-count shows the best performance (around 0.8 of accuracy) compared to other cases. This implies that too short documents do not have enough features while too long documents have features which are noisy and can cause misclassifications. This result shows the existence of an optimal document length for training datasets.

### 4.3 General performance comparison - MLA and SOA

The last experiment is to investigate the performance difference between MLA and SOA with regards to the properties of test data. Without controlling the data properties of training dataset, 5% of test data size are used for extracting features for each positive and negative (IMDB – 500 training and 10,000 test documents, Hotel review – 400 training and 8000 test documents, twitter – 200 training and 4000 test documents, and Amazon review – 300 training and 6000 test documents for each positive and negative). All experiments are performed 20 times with random sub-sampling, which is equivalent to 20-fold validation, to derive reliable results. The distribution of test documents with regard to the subjectivity and word-count is summarized in Tables 9 and 10 shows the accuracy comparison based on the test data properties.

Multinomial Naïve Bayes and Support Vector Machine show better performances than SOA for all dataset cases. Both approaches show better accuracy for high subjectivity test data than for low subjectivity test data. MLA shows moderate performance for the test documents with more than 100 words. It fails to show

good performance for short documents with less than 100 words while the performance of SOA is best among other for that range. Short documents containing small but consistent sentiment words are more likely to be correctly classified using SOA as it can cover wide range of words based on dictionary while MLA misses some words due to the limitation of training dataset size. For a training dataset with more than 200 words, MLA shows the better performance than SOA. In this case, MLA is more effective because the longer documents are likely to have both positive and negative sentiment words that causes misclassification. This indicates that only few features play an important role for correct classification for long documents.

The running time for each experiments is presented in Table 11. Multinomial Naïve Bayes classifiers take less running time for training and test in all cases while SVM take longer time for both training and test due to the high dimension problem.

### 4.4 Summary of experiment results

Key findings from experiments are summarised in the Table 12.

## 5 Discussion

This study proposed a data-centric view on the performance of sentiment classification by clarifying the effect of linguistic properties of data on the accuracy of representative sentiment analysis algorithms, i.e., SOA and MLA. Experiments in this research covered the data from e-commerce domain and part of social media and we can derive the meaningful implication for

**Table 11** The comparison of the running times of used algorithms

	Decision Tree		Multinomial Naïve Bayes		SVM		SOA
	Training	Test	Training	Test	Training	Test	
IMDB Dataset	1.7175	11.5816	0.5444	10.4597	2.8626	53.0191	96.1212
Hotel Review Dataset	1.1233	8.7047	0.3222	7.8711	1.5191	26.5369	54.3441
Twitter Dataset	0.1986	0.2937	0.0235	0.2992	0.046	0.81	5.3138
Amazon Review Dataset	1.1278	6.7353	0.6374	4.9226	1.3784	17.4996	52.4596

\*The specification of machine used for testing: Intel core i7–5500 processor with 8GB system memory, the running program has been coded in Python 3.4 with Anaconda and scikit-learn Open source library

the sentiment classification of document from these domains. The experiment results provide decision makers with a strategy that can elevate the accuracy of sentiment classification by choosing appropriate algorithms for their datasets as follow.

**The performance of SOA** SOA can be easily implemented with predefined semantic lexicon such as WordNet and SentiWordNet. For this reason, many studies and practical applications adopt this approach. However, the experiment results indicate that its performance is affected by the properties of test data. The accuracy for short document with less than 100 words is higher than longer documents for IMDB and twitter datasets. The average accuracy for short documents was 75% or higher than that. The level of subjectivity of test documents also needs to be considered in the application of SOA. For the test data with the subjectivity higher than 0.7, the accuracy was higher than 80% for IMDB, Hotel review, and twitter datasets while the accuracy for test data with lower average subjectivity (lower than 0.7) was around 70%. The results imply that SOA can be

effective for classification of short documents such as twitter and short reply, but not for long documents like news, blog body texts. Especially, for some practical cases in which training dataset does not exist for supervised learning-based classifier, SOA can be a good alternative with moderate accuracy for short test documents.

**Training scheme of MLA for best performance** The results from experiments in 4.2 show how the properties of documents in training dataset can affect the performance of MLA. The size of training data should be at least 3% of test data training dataset for both sentiments and no significant improvement in accuracy was found even if the training size was increased to 4 ~ 5%. An important finding from the experiment is the existence of optimal length of documents in training dataset. The classification results using short training documents and too long training documents tend to show the lower accuracies than the results from optimal length (100 ~ 200 words) of training documents as short documents

**Table 12** The summary of experiment results

Approach	Data properties	Findings
SOA	Words count	• works better for shorter documents (less than 200 words in document) than longer documents (more than 200 words in document)
	Subjectivity	• works better for documents with high average subjectivity (higher than 0.7) than documents with low average subjectivity (lower than 0.7)
MLA	Training size, document length	• More than 2% of test dataset size is required. • Ideal document length for training dataset is 50 ~ 150 words for all datasets and all MLA algorithms
	Training size, subjectivity	• Larger size of training dataset is required if the average subjectivity of document in training datasets is lower than 0.5 • Documents with higher subjectivity (0.5 ~ 1.0) are more suitable for training data than document with lower subjectivity (0 ~ 0.5)
	Document length, subjectivity	• Documents with higher average subjectivity (0.5 ~ 1.0) and 100 ~ 250 words counts can be best for training dataset.
General Performance		• In general, Multinomial Naïve Bayes and SVM outperform SOA • SOA works as good as MLA for very short document (0 ~ 100) with higher average subjectivity (0.5 ~ 1.0). • For document with higher average subjectivity (0.5 ~ 1.0) SOA outperforms Decision Tree. • Decision Tree fails to show better performance than other MLA such as Multinomial Naïve Bayes and SVM.

lack enough features for classification while long documents struggle with noises. By simply controlling the word-counts of training dataset, the performance of ML classifier can be improved. Another linguistic properties of data affecting the improvement of performance is the average subjectivity of training documents. Training using documents with higher subjectivity shows the better performance than documents with lower average subjectivity. The optimal word-count plays a role in the training with documents with higher subjectivity. Even for the document with higher average subjectivities, the algorithm is not effective for relatively longer documents. Documents with 100 ~ 200 words and higher average subjectivities (larger than 0.7) are the most effective training dataset.

**Which approach is better for given dataset** The experiment results from this study confirm the findings from other studies (Pang et al. 2002; Tang et al. 2009; Vinodhini and Chandrasekaran 2012) that report the superiority of MLA over SOA. The overall accuracy of Naïve Bayesian classifier is higher than SOA even without any control of training dataset. However, as we can see from Table 9, for the short test document with less than 100 words, SOA shows better performance than Decision Tree Classifier for test data cases with lower and higher average subjectivities. The results imply that we can select appropriate algorithms depending on the dataset. If the length of test data is relatively short, user do not need to implement complex machine learning algorithm as the semantic orientation can show good performance for short document.

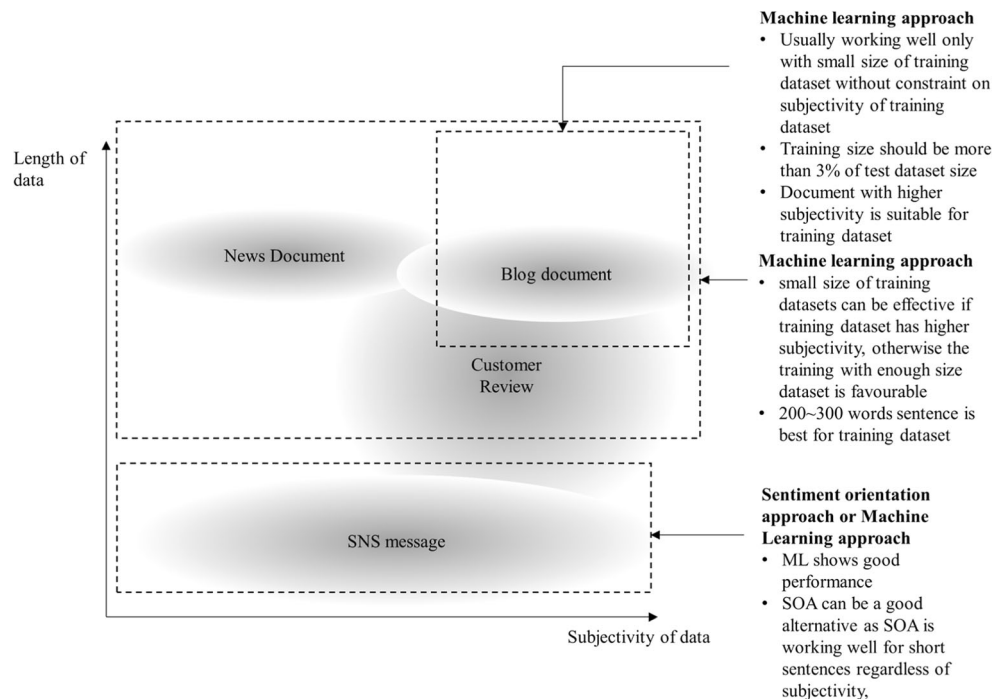
**MLA – pros and cons** If enough number of training data is available, Multinomial Naïve Bayesian or SVM can be good

choice in general. Considering the running time, Multinomial Naïve Bayesian also can be better than SVM as the latter has limitation to handle high dimensional feature space. SOA has also limitation in terms of running time but it can be a good option if test documents are short and has a high subjectivity with enough emotional expressions like twitter.

**The effectiveness of increasing training size** Considering all experiment results from MLA, the effectiveness of increasing training size tends to be valid only for Multinomial Naïve Bayesian, which is statistical approach. The other ML showed only marginal effect from expanding training data.

Based on the discussion above, we can derive a guideline for sentiment classification application according to properties of data to be used for a research and practical application. For short document from SNS such as Twitter and Facebook, SOA can show better performance comparing to Naïve Bayesian. Without expending effort to collect a massive training dataset, we can adopt the sentiment orientation approach for short documents. News documents and blog documents usually have longer documents than SNS message and customer reviews. The length of customer reviews varies according to the domain or platform but likely to have high subjective words in the document to express their opinion on product or service. For documents with high subjective words such as customer review, blog document and online forum text, an MLA is expected to show a good performance even with small size of training dataset. For longer documents, the “proper training” of a classifier is critical for the better performance. The guidance for selecting the sentiment classification approach and application scheme is depicted in Fig. 7.

**Fig. 7** Guidance for the application of sentiment classification for various data types





## 6 Conclusion and future work

Sentiment classification of text data is the starting point of transforming unstructured qualitative data into quantitative data that can be used for decision making in e-commerce. Therefore sentiment classification has attracted large amount of attention from various research areas including computational intelligence, machine learning, and computational linguistics. Existing studies, however, do not paid much attention to the role of linguistic properties of datasets used in sentiment classification but instead concentrate on proposing more sophisticated and complex algorithms to improve the performance.

The findings of this study suggest that researchers and practitioners need to consider the properties of datasets they have when they choose a sentiment classification algorithm. The findings also support the contention that appropriate control of training datasets and algorithms that match to the datasets is as important as finding a sophisticated algorithm. In this regard, the study proposes practitioners and scholars with guidance on applying different sentiment classification algorithms. The study shows that the performance of classification can be improved by controlling data properties of documents in training datasets.

Future studies need to deal with other dimensions of the linguistic property of data that can affect the performance of different algorithms. Also, we only use the data from e-commerce and twitter so future research can cover the data from news and the other social media context. The comparison of accuracies between classifications of negative and positive documents can be an important topic considering that the results of the experiments of this study show different performances according to different level of the sentiment of test datasets. Domain dependency between training and test dataset also need to be investigated from the view point of data properties. By conducting further experiments with data from various domain, we can try to clarify why the domain dependency exists and how we can resolve the problem. The data-centric view on the performance of sentiment classification can provide not only guidance for application of sentiment classification algorithms, but also an easy way to obtain better performance with simple algorithms via controlling training datasets.

**Acknowledgement** This study was supported by Global Research Network Program through the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (Project no. NRF-2016S1A2A2912265) and partially by EU funded project Policy Compass (Project no. 612133).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3), 12.
- Andreevskaia, A., & Bergler, S. (2006). Mining WordNet for a fuzzy sentiment: sentiment tag extraction from WordNet glosses. In *EACL* (Vol. 6, pp. 209–216).
- Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. In *Proceedings of recent advances in natural language processing (RANLP)* (Vol. 1, pp. 2.1, Vol. 3.1).
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC* (Vol. 10, pp. 2200–2204).
- Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4), 732–742. doi:10.1016/j.dss.2010.08.024.
- Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36–44). Association for Computational Linguistics.
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20–29.
- Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69–72). Association for Computational Linguistics.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Sebastopol: O'Reilly Media, Inc..
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, 16(2), 340–358.
- Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, 13(1), 45–59 1387–3326.
- Cortes, C., Jackel, L. D., & Chiang, W.-P. (1995). Limits on learning machine accuracy imposed by data quality. In *KDD* (Vol. 95, pp. 57–62).
- Das, D., & Bandyopadhyay, S. (2011). Document level emotion tagging: machine learning and resource based approach. *Computación y Sistemas*, 15(2), 221–234.
- Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241–249). Association for Computational Linguistics.
- Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *IEEE 24th International Conference on Data Engineering Workshop (ICDEW) 2008* (pp. 507–512). IEEE.
- Denecke, K. (2009). Are SentiWordNet scores suited for multi-domain sentiment classification? In *IEEE Fourth International Conference on Digital Information Management (ICDIM) 2009* (pp. 1–6).
- Devitt, A., & Ahmad, K. (2007). Sentiment polarity identification in financial news: a cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, June 2007, 984–991.
- Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of machine learning research*, 3, 1265–1287.
- Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems*, 53(4), 704–711. doi:10.1016/j.dss.2012.05.023.

- Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: a publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417–422). Citeseer.
- Feinerer, I. (2015). Introduction to the tm Package Text Mining in R. <http://cran.r-project.org/web/packages/tm/index.html>.
- Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, 68, 26–38. doi:10.1016/j.dss.2014.10.004.
- Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3, 1289–1305.
- 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* (pp. 841). Association for Computational Linguistics.
- Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI* (pp. 121–132). Springer Berlin Heidelberg.
- García-Moya, L., Kudama, S., Aramburu, M. J., & Berlanga, R. (2013). Storing and analysing voice of the market data in the corporate data warehouse. *Information Systems Frontiers*, 15(3), 331–349 1387–3326.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10–18.
- Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, 43(3), 375–394. 0092–0703.
- Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision support systems*, 57, 42–53 0167–9236.
- Hung, C., & Lin, H.-K. (2013). Using objective words in SentiWordNet to improve word-of-mouth sentiment classification. *IEEE Intelligent Systems*, 28(2), 0047–0054.
- Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach. *Expert Systems with Applications*, 41(4), 1041–1050.
- Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5), 6000–6010.
- Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software*, 24(7), 850–858. doi:10.1016/j.envsoft.2008.11.012.
- Khan, A. (2011). Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and Blogs. *International Journal of Computer Science & Emerging Technologies*, 2(4), 539–552.
- Kim, H. J., & Song, M. (2013, November). An ontology-based approach to sentiment classification of mixed opinions in online restaurant reviews. In *International Conference on Social Informatics* (pp. 95–108). Springer International Publishing.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning* (pp. 249–256).
- Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10), 4065–4074. doi:10.1016/j.eswa.2013.01.001.
- Kotsiantis, S. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31, 249–268.
- Kuh, R., & De Mori, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5), 449–460.
- Leech, G. N. (1983). *Principles of pragmatics*, Vol. 30. New York: Taylor and Francis Group.
- Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2), 354–368. doi:10.1016/j.dss.2009.09.003.
- Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375–384). ACM.
- Liu, B. (2010). *Sentiment analysis and subjectivity, handbook of natural language processing, chemical rubber company (CRC) press*. New York: Taylor and Francis Group.
- Liu, C., & Yu, N. (2014). Feature selection for highly skewed sentiment analysis tasks. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)* (pp. 2–11). Dublin, Ireland.
- Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models (pp. 81–88). IEEE.
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 142–150). Association for Computational Linguistics.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational linguistics and intelligent text processing* (pp. 171–189). Springer Berlin Heidelberg.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)* (pp. 55–60).
- Meisel, S., & Mattfeld, D. (2010). Synergies of operations research and data mining. *European Journal of Operational Research*, 206(1), 1–10.
- Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1275–1284). ACM.
- Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, 42(11), 30–36.
- Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633.
- Mukherjee, S., & Bhattacharyya, P. (2012). Sentiment analysis in twitter with lightweight discourse analysis. In *Proceedings of the 24th International Conference on Computational Linguistics* (pp. 1847–1864).
- Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *EMNLP* (Vol. 4, pp. 412–418).
- Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. In *9th IT & T Conference* (p. 13).
- Okanohara, D., & Tsujii, J. I. (2005). Assigning polarity scores to reviews using machine learning techniques. In *Natural Language Processing–International Joint Conference on Natural Language Processing 2005* (pp. 314–325). Springer.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of International Conference on Language Resources and Evaluation* (Vol. 10, pp. 1320–1326).
- Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.
- Pang, B., & Lee, L. (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 115–124). Association for Computational Linguistics.

- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (Vol. 10, pp. 79–86). Association for Computational Linguistics.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of International Workshop on Semantic Evaluation* (pp. 27–35).
- Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: a combined approach. *Journal of Informetrics*, 3(2), 143–157.
- Ranade, S., Gupta, J., Varma, V., & Mamidi, R. (2013). Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 7). ACM.
- Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers—a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(4), 476–487 1094–6977.
- Sebastiani, F. (2005). Text categorization. In *Encyclopedia of Database Technologies and Applications* (pp. 683–687). IGI Global.
- Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). *Get another label? Improving data quality and data mining using multiple, noisy labelers*. Paper presented at the proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas.
- Smedt, T. D., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13, 2063–2067.
- Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4), 470–479.
- Stieglitz, S., & Dang-Xuan, L. (2012). Impact and diffusion of sentiment in political communication: an empirical analysis of public political Facebook pages. In *Proceedings of the 20th European Conference on Information Systems (ECIS)*.
- Sui, H., Khoo, C., & Chan, S. (2003). Sentiment classification of product reviews using SVM and decision tree induction. *Advances in Classification Research Online*, 14(1), 42–52.
- Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7), 10760–10773. doi:10.1016/j.eswa.2009.02.063.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (Vol. 1, pp. 173–180). Association for Computational Linguistics.
- Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: a business oriented approach. *Expert Systems with Applications*, 39(5), 4729–4739.
- Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6), 282–292.
- Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, 39(15), 11880–11888.
- Wang, H., Yin, P., Zheng, L., & Liu, J. N. (2014). Sentiment classification of online reviews: using sentence-based language model. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(1), 13–31.
- Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 315–354.
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37.
- Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(8), 1077–1087.
- Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6), 1138–1152 0020–0255.
- Yan, X., Wang, J., & Chau, M. (2015). Customer revisit intention to restaurants: evidence from online reviews. *Information Systems Frontiers*, 17(3), 645–657 1387–3326.
- Yang, H.-L., & Chao, A. F. Y. (2015). Sentiment analysis for Chinese reviews of movies in multi-genre based on morpheme-based features and collocations. *Information Systems Frontiers*, 17(6), 1335–1352 1387–3326.
- Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3, Part 2), 6527–6535. doi:10.1016/j.eswa.2008.07.035.
- Yoshida, S., Kitazono, J., Ozawa, S., Sugawara, T., Haga, T., & Nakamura, S. (2014). Sentiment analysis for various SNS media using Naïve Bayes classifier and its application to flaming detection. In *IEEE Symposium on Computational Intelligence in Big Data (CIBD), 2014* (pp. 1–6). IEEE.
- Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decision Support Systems*, 55(4), 919–926.

**Youngseok Choi** is a Lecturer in Business Analytics in Coventry Business School, Coventry University. Prior to join Coventry University, he had worked on an EU-funded research project for few years in Brunel University London. He got BEng in electrical engineering and computer science and Ph.D. in Management Information System from Seoul National University. His research interest covers big data analytics and data-driven decision modelling based on machine learning approach. His refereed articles have appeared in academic journals including International Journal of Electronic Commerce, Annals of Operation Research, Journal of Organizational Computing and Electronic Commerce, Information Systems Frontier, and so on.

**Habin Lee** is a Professor in Data Analytics and Operations Management at Brunel Business School, Brunel University London. He obtained MSc and Ph D in Management Science from Korea Advanced Institute of Science and Technology. His research interests include applications of data analytics including data mining, fuzzy cognitive maps, multi-agent simulation and so on. He published articles on international journals including *Management Science*, *European Journal of Operational Research*, *Transportation Research Part E*, *Annals of Operational Research*, *Information Systems Frontier* among others.