Unbalanced sentiment classification: an assessment of ANN in the context of sampling the majority class

Rodrigo Moraes[1], João F. Valiati[2], and Wilson P. Gavião Neto[3]

[1]Unidade de Graduação - UAGRAD , Universidade do Vale do Rio dos Sinos - UNISINOS, Av. Unisinos, 950, São Leopoldo, RS, Brazil.
[2]Artificial Intelligence Engineers - AIE, Vieira de Castro street, 262/206, Porto Alegre, RS, Brazil.,
[3]School of Engineering & IT , Centro Universitário Ritter dos Reis - UNIRITTER - Laureate International Universities , Av. Manoel Elias, 2001, Porto Alegre, RS, Brazil.

Corresponding author:
Wilson P. Gavião Neto[3]

Email address: wilson_gaviao@uniritter.edu.br

ABSTRACT

Many people make their opinions available on the Internet nowadays, and researchers have been proposing methods to automate the task of classifying textual reviews as positive or negative. Usual supervised learning techniques have been adopted to accomplish such a task. In practice, positive reviews are abundant in comparison to negative's. This context poses challenges to learning-based methods and data undersampling/oversampling are popular preprocessing techniques to overcome the problem. A combination of sampling techniques and learning methods, like Artificial Neural Networks (ANN) or Support Vector Machines (SVM), has been successfully adopted as a classification approach in many areas, while the sentiment classification literature has not explored ANN in studies that involve sampling methods to balance data. Even the performance of SVM, which is widely used as a sentiment learner, has been rarely addressed under the context of a preceding sampling method. This paper addresses document-level sentiment analysis with unbalanced data and focus on empirically assessing the performance of ANN in the context of undersampling the (majority) set of positive reviews. We adopted the performance of SVM as a baseline, since some studies have indicated SVM as being less subject to the class imbalance problem. Results are produced in terms of a traditional bag-of-words model with popular feature selection and weighting methods. Our experiments indicated that SVM are more stable than ANN in highly unbalanced (80%) data scenarios. However, under the discarding of information generated by random undersampling, ANN outperform SVM or produce comparable results.

INTRODUCTION

Nowadays, a large number of users' opinions on products and services is available on the Internet and marketing research have studied the power of consumers' reviews on purchasing decisions in e-commerce (Lee et al., 2008; Cheung and Thadani, 2012; Hu et al., 2006; Park et al., 2007; Zhang and Tran, 2011). Some results have indicated that the anonymity provided by the web motivates honest negative reviews (Joinson, 2001; Woong Yun and Park, 2011), which can have a strong influence on reversing a positive purchase decision (Liu, 2006; Markey and Hopton, 2000; Lee et al., 2008; Verhagen et al., 2013).

In order to deal with a large number of textual reviews, Opinion Mining and Sentiment Analysis (OMSA) research area aims to analyze opinions automatically (Liu, 2012). Many studies in the literature have successfully proposed to use Machine Learning (ML) techniques to classify reviews as expressing a positive or negative sentiment (Pang et al., 2002; Turney, 2002; Pang and Lee, 2004; Blitzer et al., 2007; Fersini et al., 2014). However, realistic contexts challenge ML-based approaches since the ratio of positive and negative reviews is unbalanced (Nassirtoussi et al., 2014; Blitzer et al., 2007; Li et al., 2011a,a, 2012; Mountassir et al., 2012; Wang et al., 2013). Especially in the e-commerce domain, negative reviews are substantially less frequent than positive ones (Schlosser, 2011; Li et al., 2011a; Burns et al., 2011), which may result in a poor classification performance. To overcome this problem, popular approaches balance the input data by (i) *undersampling* the majority class or (ii) adding samples to the minority class, which is known as *oversampling* (He and Garcia, 2009). Although both techniques have complementary advantages, only the undersampling approach is computationally feasible in some contexts that involve a large amount of data. In addition, undersampling has shown better results than the random oversampling in sentiment classification (Li et al., 2011a,b; Wang et al., 2013). As a disadvantage, undersampling may cause learning algorithms to miss relevant information on the majority class, and the sensitiveness of an algorithm to this scenario may support the choice for a given approach.

Support Vector Machine (SVM) is a learning algorithm commonly employed in the sentiment classification literature (Ravi and Ravi, 2015; Tsytsarau and Palpanas, 2012; Tang et al., 2009) while Artificial Neural Networks (ANN) has attracted less attention as an approach for sentiment learning (Bespalov et al., 2011; Chen et al., 2011; Claster et al., 2010; Zhu et al., 2010), even though some results have indicated that SVM does not outperform ANN in several contexts (Moraes et al., 2013; Ghiassi et al., 2013; Ravi and Ravi, 2015). Although some studies have compared SVM with ANN under different levels of data imbalance (without balancing data) (Moraes et al., 2013), a comparative study involving ANN and SVM under the same context of loss of information, which is caused by the adoption of an undersampling approach, is still unclear in the sentiment classification literature, as discussed in section .

In this paper, we address unbalanced document-level sentiment classification and focus on empirically assessing the performance of ANN in the context of undersampling the (majority) set of positive reviews. By involving SVM as a baseline, our research question is about investigating in *which circumstances ANN tend to be less/more affected by an undersampling method?* The contributions of our work are:

1. A performance assessment of an ANN-based method under a context of data undersampling, including a comparison with the well-established SVM, which is potentially less prone to the class imbalance problem (Sun et al., 2009b; Japkowicz and Stephen, 2002).

2. A performance assessment of SVM on the benchmark dataset of Movies reviews (Pang and Lee, 2004) in the context of losing potentially critical information, which is caused by an undersampling method. As discussed in section , although SVM have been widely used in sentiment learning studies, there has been little discussion about the impact caused by a preceding sampling method on their performance.

3. An empirical analysis of both ANN and SVM as a function of the number of selected features (i.e., terms), which is supposed to involve an increasing number of noisy terms due to the discard of samples caused by an undersampling approach.

This paper is organized as follows. In order to approach a standard framework of experiments, Section presents an overview of usual techniques in sentiment analysis. Section presents an overview of ANN and SVM and their susceptibilities to unbalanced data. Section discusses the literature and justifies the contributions of our work. Our experimental framework is reported in Section and results are discussed in Section . Section summarizes our conclusions.

USUAL TECHNIQUES

Figure 1 shows an overview of steps and techniques commonly used in sentiment classification approaches. We follow the popular *bag-of-words* model in which documents are represented as vectors, whose entries correspond to individual terms of a vocabulary.

Pre-processing techniques involve removing *stopwords*, which are common terms like articles and prepositions, and reducing term variations to a single representation by applying *stemming* techniques (Weiss et al., 2004). Popular *stemmer* algorithms for the english language are Snowball (Porter, 2001), Porter (Porter, 1980) and the Lovin (Lovins, 1968).

Supervised techniques, which are adopted in the classification step (see Figure 1), are not usually adapted to deal with realistic contexts in which the ratio of positive and negative reviews are unbalanced. Techniques to deal with the problem of unbalanced datasets fall into two major categories: *data sampling* and learning algorithm modification (López et al., 2012), which happen as part of pre-processing and classification steps, respectively. As a pre-processing technique, data sampling aims to balance datasets

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

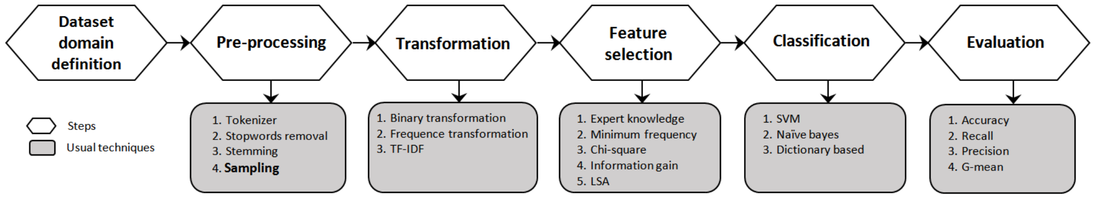


Figure 1. Steps and techniques that are commonly found in sentiment classification approaches.

by reducing the number of samples in the majority class (undersampling) or increasing the minority class (oversampling). Undersampling leads to data loss, while oversampling increases training time and may cause the effect of over-fitting (Tian et al., 2016). Li et al. (2011a) has reported random undersampling as a better choice when compared to (i) random oversampling and (ii) a cost-sensitive learning solution, which involves algorithm modifications (López et al., 2012).

Next, a numerical representation is computed from textual data. *Binary* representation is widely used and only takes into account presence or absence of a term in a document. The number of times a term occurs in a document (i.e., *term frequency*) is also used as a weighting scheme for textual data (Li et al., 2009; Paltoglou and Thelwall, 2010). TF-IDF (*Term Frequency - Inverse Document Frequency*) is one of the most popular representations and considers not only term frequencies in a document, but also the relevance of a term in the entire collection of documents. The classic TF-IDF$_{t,d}$ (Manning et al., 2008) assigns to term t a weight in document d as

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t, \quad \text{where} \quad \text{IDF}_t = \log \frac{N}{\text{DF}_t}, \tag{1}$$

TF$_{t,d}$ is the number of occurrences of term t in document d, N is the number of documents in the collection and DF$_t$ is the number of documents in the collection that contain term t. Essentially, TF-IDF avoids assigning high scores to terms that occur too often in the dataset.

Another stage commonly found in sentiment classification approaches is feature selection. It can make learning algorithms more efficient/effective by reducing the amount of data to be analyzed as well as identifying relevant features to be considered in the learning process. Usual feature selection methods are *Document Frequency* (Pang et al., 2002; Dang et al., 2010; Bai, 2011), *Mutual Information* (Turney, 2002; Li et al., 2009), *Information Gain* (Abbasi et al., 2011; Li et al., 2009; Riloff et al., 2006; Abbasi et al., 2008), *Chi-square* (Abbasi et al., 2011; Li et al., 2009) and *Latent Semantic Analysis* (LSA) (Bespalov et al., 2011). None of them has been widely accepted as the best feature selection method for sentiment classification or text categorization, however, information gain has often been competitive (Abbasi et al., 2011; Xia and Zong, 2010; Li et al., 2009; Forman, 2003; Yang and Pedersen, 1997). It ranks terms by considering their presence and absence in each class (Berry and Kogan, 2010). A high score is assigned to terms that occur frequently in a class (and rarely in the others) as follows (Weiss et al., 2010):

$$IG(t) = \sum_{k=1}^{C} P(c_k) \log \frac{1}{P(c_k)} \quad - \sum_{t \in \{t_p, \overline{t_p}\}} P(t) \sum_{k=1}^{C} P(t|c_k) \log \frac{1}{P(t|c_k)}, \tag{2}$$

where $P(c_k)$ is the prior probability of a document occurring in class c_k, $P(t)$ is the probability of term t occurring or not in a document, i. e. $P(t_p)$ and $P(\overline{t_p})$ respectively. $P(t|c_k)$ is the conditional probability of term t occurring or not in a document of class c_k and C is the number of classes.

In general, sentiment analysis approaches in the literature can be differed in terms of the adopted approach for feature selection. On the other hand, Support Vector Machines has been widely used in the classification stage (Ravi and Ravi, 2015; Tsytsarau and Palpanas, 2012). Learning algorithms like SVM and ANN are also known as classifiers. Since documents are represented as vectors, a classifier aims to learn a decision boundary to assign them to one of C classes.

Classification performance metrics are usually based on a confusion matrix. Table 1 is a confusion matrix whose entries are given as a function of two typical classes in document-level sentiment classification, positive and negative documents. Accuracy is usual as a performance metric. However, when the quantification is applied over an unbalanced binary problem, it may lead to a biased interpretation against

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

the minority class (Barranquero et al., 2015). Therefore, recall and precision, as defined in Equations 3 and 4, are adopted to measure the classification performance on each class (Moraes et al., 2013).

Table 1. Confusion matrix.

	Predicted	
	Positive documents	Negative documents
Actual positive documents	# True Positive samples (TP)	# False Negative samples (FN)
Actual negative documents	# False Positive samples (FP)	# True Negative samples (TN)

$$\text{recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{precision} = \frac{TP}{TP+FP} \tag{4}$$

Recall, as defined in Equation 3, is also known as Positive Recall, True Positive Rate or Sensitivity (Li et al., 2011a). Negative Recall, also called as True Negative Rate or Specificity, combined with Positive Recall constitute Geometric Mean (G-Mean), as defined in Equation 5 (He and Garcia, 2009; Kubat and Matwin, 1997). G-Mean is appropriate to the unbalanced context (Barranquero et al., 2015) and has been widely used in the unbalanced learning literature (Wu and Chang, 2003; Guo and Viktor, 2004; Wu and Chang, 2005; Su and Hsiao, 2007; Li et al., 2011a; Romero et al., 2013).

$$\text{G-Mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{5}$$

ANN, SVM AND UNBALANCED DATASETS

This section provides a brief review of the fundamental aspects of the supervised classifiers ANN and SVM. We adopted the performance of the SVM classifier as a baseline to evaluate results with ANN, since it is a learning algorithm commonly employed in the sentiment classification literature (Ravi and Ravi, 2015; Tsytsarau and Palpanas, 2012; Tang et al., 2009). Instead of providing a detailed description of these approaches, we focus on reviewing concepts of ANN and SVM with the purpose of discussing issues in the context of learning from unbalanced datasets.

Artificial Neural Networks

Neural networks derive features from linear combinations of input data, and then model the output as a nonlinear function of these features (Hastie et al., 2001). As a result, ANN have been one of the most popular forms of learning system (Russell and Norvig, 2009).

Typically, neural networks are represented as a network diagram, which is composed of linked nodes or neurons. Usually, neurons are simple mathematical models that produce an output value in two steps. First, each neuron computes a weighted sum of its inputs, and an output value is computed by applying an *activation function* to the sum (Luger, 2008). An activation function can be a nonlinear function, which ensures that the entire network can estimate a nonlinear function, like a nonlinear decision boundary.

Multilayer Perceptron (MLP) is an usual type of neural networks in which nodes are arranged in layers, namely the input, the hidden and the output layer of nodes (Hastie et al., 2001). Each connection has an associated weight, which is estimated by minimizing a global error function in terms of a gradient descent training process (Haykin, 2008).

In case of training on unbalanced data, the gradient descent direction may be dominated by the majority class and the training error may be minimized only for this class (Sun et al., 2009b; Anand et al., 1993). Therefore, the training process may terminate before the error for the small class decrease (Sun et al., 2009b).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Support Vector Machines

SVM is a supervised learning method with many qualities, and performs classification more accurately than most other algorithms in many areas. Researchers have reported that SVM is perhaps the most accurate method for text classification (Liu, 2011), and therefore it is widely used in sentiment learning tasks (Tsytsarau and Palpanas, 2012).

SVM is a linear method of finding an optimal hyperplane to separate two classes. When classes cannot be linearly separated, the input data space is transformed into a higher-dimensional space so that data can be linearly separable and suitable for the linear approach. *Kernel functions* are typically used to make this transformation (Huang et al., 2006). This makes it possible to determine a nonlinear decision boundary, which is linear in the higher-dimensional feature space, without computing the parameters of the optimal hyperplane in a feature space of possibly high dimensionality (Haykin, 2008). Hence, the solution can be written as a weighted sum of the values of a kernel function, which is usually evaluated only at some data points (Horváth, 2003).

As a supervised classification approach, SVM seeks to maximize the distance to the closest training points from either class to achieve better generalization on test data (Hastie et al., 2001). The solution rely solely on those training data points that are at the margin of the decision boundary. These points are the *support vectors*. Instead of minimizing a global error function in a gradient descent process, which suffers from the existence of multiple local minima solutions, the parameters of the optimal separating hyperplane can be obtained by solving a convex optimization problem.

SVM is potentially less susceptible to the class imbalance problem than other learning algorithms, since the hyperplane between classes is supposed to be calculated with respect to only a few support vectors and the class sizes may not affect the class boundary too much (Sun et al., 2009b; Japkowicz and Stephen, 2002). Although some studies have shown good results of standard SVM algorithm on unbalanced datasets (Sun et al., 2009a), many others have reported that SVM is sensitive to a class imbalance scenario (Wu and Chang, 2003; Akbani et al., 2004), even in sentiment classification tasks (Moraes et al., 2013). Wu and Chang (2003) and Akbani et al. (2004) have discussed some possible reasons to explain what makes SVM sensitive to class imbalance. As an approach to overcome the problem, Akbani et al. (2004) have also shown that the undersampling strategy may discard samples at the class boundary, which may negatively affect the orientation of the separating hyperplane estimated by the SVM algorithm.

Despite the difficulties of both ANN and SVM to deal with unbalanced data, and the fact that an undersampling strategy may lose valuable information, satisfactory results have been reported for different natures of data in the literature (Wang and Japkowicz, 2010; Sun et al., 2009a). Additionally, Moraes et al. (2013) have indicated that SVM requires a high number of support vectors to classify sentiment (positive versus negative reviews), which means that the results may be more dependent on the class sizes, and consequently resulting in a worse performance of SVM when compared with ANN in some contexts.

RELATED WORK

Some studies have emphasized the spread of positive reviews in e-commerce. Schlosser (2011) have found that 80% of e-commerce reviews are positive, which agrees with the findings of Kim et al. (2012) in the sense that 99.1% of customers feedback on eBay are positives. On the other hand, although the number of negative reviews is lower than the positive ones, their strong influence on purchasing decisions has been confirmed (Chevalier and Mayzlin, 2006; Sen and Lerman, 2007). Verhagen et al. (2013) discussed the importance of negative posts and their usefulness for both consumers and companies that monitor their products and image. Cheung and Lee (2012) investigated the restaurant domain and built a psychology model, which found that the act of sharing negative experiences can save others consumers from uncomfortable situations and affect their intentions to post reviews.

In recent years, an increasing number of studies have proposed methods to automate the task of classifying product or services reviews as being positive or negative (Ravi and Ravi, 2015). Regardless of the development in this research field, a practical issue has attracted little attention: the *imbalance between positive and negative reviews* mainly found in the e-commerce environment. The imbalance imposes challenges to learning-based methods, like SVM, that have been performed successfully in balanced data contexts (He and Garcia, 2009; Lane et al., 2012; Wang et al., 2013).

Burns et al. (2011) addressed sentiment classification on unbalanced datasets, however the experiments have involved neither SVM nor ANN. Li et al. (2011b) and Li et al. (2011a) conducted experiments

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

on various domains of unbalanced reviews, like books, DVDs, electronics and kitchen. Although the analysis has involved random undersampling and SVM, they have not combined in a single approach and, therefore, there were no results produced by applying the undersampling technique followed by SVM. In contrast, Wang et al. (2013) and Vinodhini and Chandrasekaran (2017) reported results that combine SVM and a preceding data undersampling method. To the best of our knowledge, these works are the only studies under such a scenario in the sentiment classification literature. However, Wang et al. (2013) focused on comparisons of feature selection approaches with only SVM as the learning algorithm, and the experiments have not involved the popular benchmark dataset of Movies reviews (Pang and Lee, 2004). Vinodhini and Chandrasekaran (2017) have also involved only SVM in their experiments.

Perhaps the most conclusive experiments that compare the sensitiveness of an ANN-based method with SVM-based approaches for unbalanced sentiment learning are reported in (Lane et al., 2012) and (Moraes et al., 2013). Lane et al. (2012) discussed results on a broad setup of experiments, which involves unbalanced datasets, different types of features and comparisons between learning algorithms like Naïve Bayes, SVM and even ANN (Radial Basis Functions - RBF). Despite the variety of techniques under comparison, the datasets used in the experiments are slightly unusual in the context of sentiment classification literature, since the input data consists of documents collected from newspapers and magazines, which were probably well-written by journalists, in contrast to regular consumers reviews commonly found in e-commerce. In addition, class labels were manually assigned by trained analysts in terms of *favourability* scores, which may be different from a rating assigned by the own author of a review. As an interesting result, the Naïve Bayes learning algorithm has outperformed SVM and ANN in the task of distinguishing between documents with generally positive and negative favourability, which is contrary to many studies in the sentiment classification literature (Ravi and Ravi, 2015). Moraes et al. (2013) has also compared ANN with SVM in the task of learning sentiment from unbalanced datasets, however the algorithms were tested directly on unbalanced data and the experiments have not involved any technique to mitigate the effects of class imbalance.

Based on the literature review above, our work contrasts with previous works as follows:

1. The effects of a preceding data sampling method on the performance of ANN have not been discussed in the context of sentiment learning from unbalanced data;

2. The combination of SVM with a preceding undersampling technique is a popular approach to deal with unbalanced datasets (Sun et al., 2009a). However, the impact caused by an undersampling method on the classification performance of SVM and the conclusions of Akbani et al. (2004), which has shown that an undersampling strategy may negatively affect SVM, have not been clearly and completely addressed in the sentiment classification literature.

3. Consequently, a comparison between ANN and SVM has not also been discussed so that we can answer the following question in the context of sentiment classification literature: *Which one tend to be less/more affected by an undersampling method?*

EXPERIMENTAL FRAMEWORK

Our evaluation methodology involves two scenarios in which ANN is compared with SVM. First, we evaluate the classifiers' performance on highly unbalanced datasets, with a data imbalance ratio around 80% and less negative than positive reviews, i.e. #Neg/#Pos ≈ 0.2. The goal of the second scenario is to assess how the classifiers' performance is affected by randomly undersampling the (majority) set of positive reviews.

Both ANN and SVM classifiers were parameterized empirically in a grid search fashion guided by better values of accuracy. We report the best result obtained among the different combination of parameters. We used a classical Feed-Forward Neural Network (Multi-Layer Perceptron) with the Back-Propagation algorithm. A single hidden layer was used and the number of neurons M was selected from the set $M \in \{15, ..., 55\}$. In addition, we used the scaled conjugated gradient to speed up the convergence to a solution (Müller, 1993), as implemented in the Matlab software, and a non-linear function was adopted as the activation function. The SVM classifier was trained by using the LIBSVM software package (Chang and Lin, 2011) with a nonlinear kernel (radial basis function) and default parameter values, except for the cost constant c whose values were selected from the interval $c \in [10^{-1}, 10^3]$.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

We adopted a 10-fold cross-validation and each test fold consisted of 100 positive and 100 negative reviews. To generate the imbalance, a fraction of each training fold was randomly removed. Based on Burns et al. (2011) and Li et al. (2011a), which have collected datasets with the original unbalanced rate around 80%, we considered just 180 reviews for the (minority) negative class, and the positive class consisted of 900 training reviews.

For each training set, we ranked/selected terms by using the Information Gain (IG) technique (Yang and Pedersen, 1997), and evaluate the performance of the learning methods as a function of an increasing number of selected terms.

We adopted the Geometric Mean (G-Mean) to measure the classifiers' performance, as defined in Equation 5. G-Mean is high when the values of both True Positive Rate and True Negative Rate is high as well as the difference is small (Kubat et al., 1997). In addition, we adopted the recall and precision metrics to measure the performance of the classifiers on each class. In order to evaluate how different the performance is between SVM and ANN, we applied the Student's t-test with 5% of significance (Alpaydin, 2010).

Datasets and preprocessing

Our experiments involve four datasets of different domains, which include the classical movie reviews dataset broadly used in the literature, as proposed by Pang and Lee (2004). The other three datasets are reviews about GPS devices, books, and cameras collected from amazon.com, and each of them consists of 1,000 positive and 1,000 negative reviews randomly selected from the data source. The ground truth was obtained according to the customer 5-stars rating. Reviews with more than 3 stars were defined as being positive and reviews with less than 3 stars were labeled as being negative. Reviews with 3 stars are not included in our datasets.

The preprocessing of the datasets consisted of removing stopwords and stemming by applying the Snowball stemmer (Porter, 2001). We adopted a Bag-of-Words approach with single words (unigrams) to represent the reviews and TF-IDF as the weighting method (Manning et al., 2008). Table 2 characterizes the distribution of terms in the datasets after removing stopwords and stemming.

Table 2. Details of the datasets used in the experiments.

Domain	Number of distinct terms	Average number of terms per document
movies	25,456	665.6
GPS	6,880	171.5
books	10,422	189.9
cameras	5,996	122.6

RESULTS

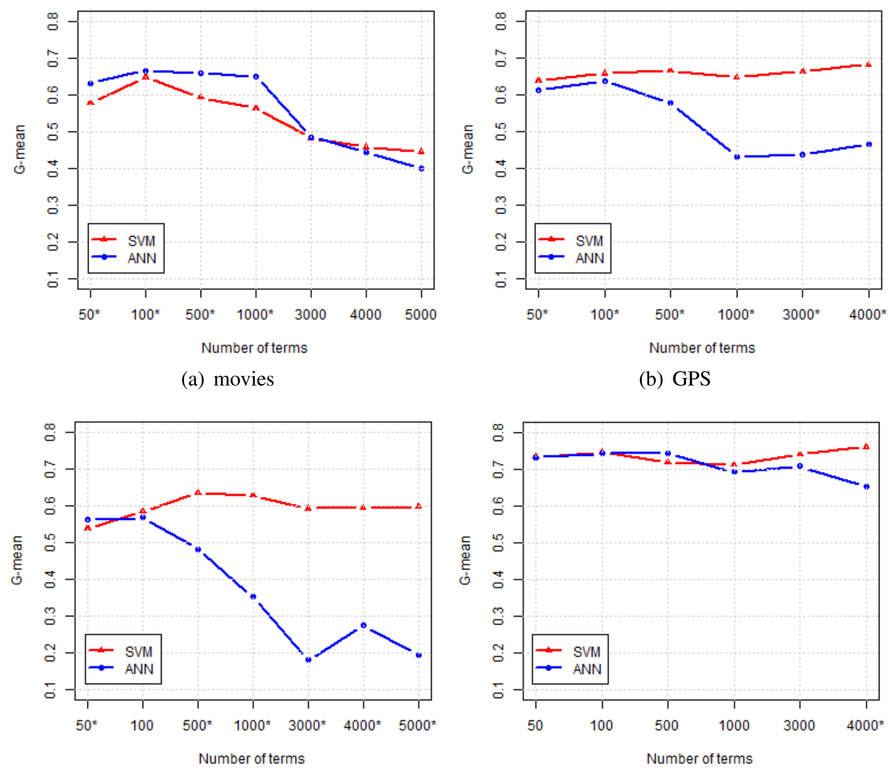
Our results are given as a function of vocabulary sizes since we aim to compare the behavior of classifiers and their requirements to achieve better classification performance. The vocabularies consisted of terms that were best ranked by the IG technique in the training stage. We arbitrarily chose seven quantities of terms between 50 and 5,000. Howerver, for some datasets, a low number of terms may result from the undersampling process and the resulting vocabulary size may not cover the entire range under analysis, and therefore no results are reported for some values of vocabulary sizes.

Figure 2 shows the average G-Mean for movies, GPS devices, books, and cameras datasets in the unbalanced context as a function of different number of selected terms, and Figure 3 shows the average G-Mean for the balanced scenario, which results from the undersampling approach. In the x-axes, the numbers of terms marked with "*" represent experiments in which the difference between the performance of ANN and SVM is statistically significant.

Tables 3 and 4 summarize the performance in terms of recall and precision for the unbalanced and undersampling contexts, respectively.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Figure 2. *Unbalanced datasets*: average G-Mean as a function of the number of selected terms.



(a) movies (b) GPS (c) books (d) cameras

Considering our results in the context of learning from unbalanced datasets (imbalance rate $\approx 80\%$), we observed the following:

- ANN outperformed SVM significantly in only 5 of 26 tests, while SVM outperformed significantly ANN in 12 tests.
- ANN tended to be more affected by noisy terms than SVM when the number of terms increases, as indicated by the decreasing G-Mean average for ANN in Figures 2(b)-(d). Since the selection of terms consisted of the top ranked terms according to IG score, it is reasonable assume that the larger is a set of selected terms, the higher is the chance of it containing less important (noisy) terms. Recall on the Negative class and precision on the Positive class (see Table 3) confirm the inferior performance of ANN.
- However, ANN was comparable with SVM in the Movies dataset, as shown in Figure 2(a). The reason for this may be due to the quality of terms in the dataset, since the reviews present characteristics that can result in a selection of terms with less noisy terms, like reviews with more terms (see Table 2) and terms that reach higher IG scores on average (Moraes et al., 2013).
- ANN was competitive with SVM when few terms (up to 100 terms) are selected to compose the vocabulary. Additionally, although the best performance of SVM has happened as a function of more than 100 terms, except for the Movies dataset (best performance at just 100 terms), it has not exceed 5% when compared with the performance achieved at 100 terms.

Considering our results in the context of undersampling the (majority) set of positive reviews, we observed the following:

- ANN outperformed SVM significantly in 6 of 22 tests, while SVM outperformed significantly ANN only twice.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Table 3. *Unbalanced datasets*: average recall and precision as a function of the number of selected terms. Best results for each classifier are in boldface.

Dataset	Metric	Classifier	Class	Number of terms						
				50	100	500	1,000	3,000	4,000	5,000
movies	Recall	ANN	Pos	0.967	0.963	0.972	0.971	0.987	0.984	0.988
			Neg	0.413	**0.462**	0.447	0.435	0.248	0.23	0.179
		SVM	Pos	0.953	0.936	0.94	0.937	0.931	0.925	0.942
			Neg	0.352	**0.451**	0.375	0.339	0.25	0.227	0.211
	Precision	ANN	Pos	0.622	**0.642**	0.637	0.632	0.569	0.564	0.548
			Neg	0.925	0.927	0.842	0.938	0.65	0.629	0.537
		SVM	Pos	0.596	**0.631**	0.601	0.586	0.553	0.544	0.544
			Neg	0.88	0.873	0.861	0.845	0.787	0.751	0.79
GPS	Recall	ANN	Pos	0.958	0.954	0.964	0.972	0.976	0.969	—
			Neg	0.393	**0.427**	0.355	0.206	0.21	0.24	—
		SVM	Pos	0.935	0.939	0.934	0.944	0.953	0.96	—
			Neg	0.436	0.46	0.475	0.445	0.462	**0.484**	—
	Precision	ANN	Pos	0.612	**0.625**	0.61	0.552	0.554	0.563	—
			Neg	0.904	0.903	0.914	0.694	0.697	0.59	—
		SVM	Pos	0.624	0.635	0.641	0.63	0.639	**0.65**	—
			Neg	0.874	0.884	0.879	0.89	0.909	0.926	—
books	Recall	ANN	Pos	0.957	0.966	0.976	0.992	0.995	0.991	0.996
			Neg	0.332	**0.334**	0.249	0.126	0.068	0.072	0.041
		SVM	Pos	0.94	0.935	0.933	0.941	0.944	0.946	0.944
			Neg	0.31	0.365	**0.432**	0.419	0.373	0.373	0.379
	Precision	ANN	Pos	0.589	**0.592**	0.566	0.531	0.516	0.517	0.509
			Neg	0.889	0.906	0.723	0.64	0.349	0.348	0.336
		SVM	Pos	0.577	0.595	**0.622**	0.619	0.601	0.601	0.603
			Neg	0.84	0.853	0.868	0.876	0.872	0.873	0.872
cameras	Recall	ANN	Pos	0.969	0.975	0.977	0.974	0.974	0.979	—
			Neg	0.554	0.565	**0.566**	0.493	0.509	0.438	—
		SVM	Pos	0.958	0.962	0.953	0.954	0.957	0.968	—
			Neg	0.562	0.577	0.543	0.532	0.571	**0.597**	—
	Precision	ANN	Pos	0.685	0.692	**0.693**	0.656	0.67	0.638	—
			Neg	0.947	0.957	0.86	0.85	0.753	0.851	—
		SVM	Pos	0.687	0.695	0.677	0.672	0.691	**0.707**	—
			Neg	0.93	0.939	0.919	0.92	0.93	0.948	—

- Again, but less expressively, the increase in the number of selected terms tended to affect negatively the ANN performance, as shown in Figures 3(b)-(d), except in the Movies dataset. ANN performed as stable as SVM on the movies dataset (see Fig. 3(a)) and the reason may be due to the quality of terms, as discussed above.

- In terms of recall and precision, both classifiers showed similar behaviors. Although the undersampling of positive reviews have significantly improved the performance of both classifier, in comparison with the unbalanced scenario, recall on the Positive class remained higher than recall on the Negative class as well as precision on the Negative class remained higher than precision on Positive class.

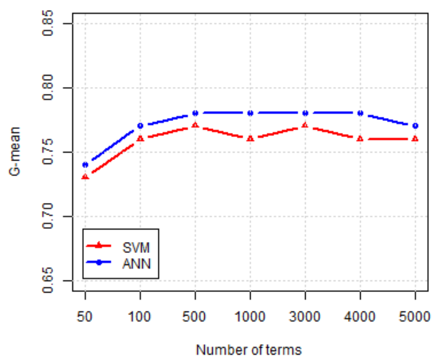
DISCUSSION

In accordance with Moraes et al. (2013), our results indicated that SVM tend to be more stable than ANN to deal with noisy terms in an unbalanced data context, since datasets of Books, GPS and Cameras have produced more noisy terms than the Movies reviews one (Moraes et al., 2013), and the behavior of G-Mean as a function of an increasing number of input (noisy) terms shown that the performance of ANN tend to decrease below the performance of SVM. Additionally, it is interesting to note that, although the number of reviews have been reduced considerably by the undersampling approach, ANN still tended to outperform SVM in a balanced context. It agrees with the results reported by Moraes et al. (2013), which were also produced in a context of balanced data, but with much more reviews since the experiments have not involved undersampling techniques.

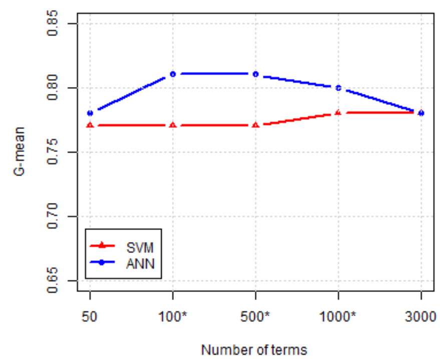
We adopted the classical neural classifier Multi-Layer Percepetron, but there are several kinds of neural networks that could be used, some of them perhaps more suitable for treating high dimensional, noise, and sparse data like textual information from the Internet. For example, a cost-sensitive neural network (Zhou and Liu, 2006) or convolutional neural networks (Severyn and Moschitti, 2015).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

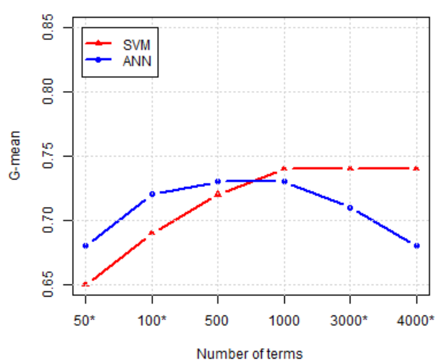
Figure 3. *Balanced datasets (via undersampling)*: average G-Mean as a function of the number of selected terms.



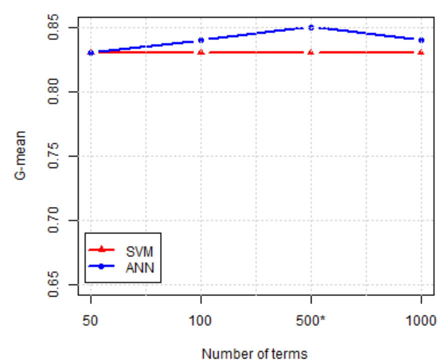
(a) movies



(b) GPS



(c) books



(d) cameras

We used terms (unigrams) as input features in our experiments. However, other features like n-grams (Vinodhini and Chandrasekaran, 2017), Part-of-Speech (Wang et al., 2015), Joint Sentiment Topic (He et al., 2011) or improvements in the quality of features (Xia et al., 2016) could also open new possibilities of investigation.

CONCLUSION

Considering the importance of negative reviews in purchasing decisions and the fact that such reviews are less common than positive reviews in e-commerce, this paper addressed the task of classifying positive versus negative-oriented reviews in data unbalanced scenarios, and focused on assessing the performance of ANN in the context of undersampling the (majority) set of positive reviews. Our experiments empirically compared ANN with SVM as a function of selected terms in a bag-of-words (unigrams) approach.

Results indicated that ANN is less stable than SVM in an unbalanced context, considering an increasing number of selected terms to represent the documents. As observed in Moraes et al. (2013), more terms may involve more noise to represent documents, showing that the neural network classifier is more sensitive to noise than the SVM classifier.

On the other hand, despite the negative aspects of random undersampling (Liu et al., 2009; Akbani et al., 2004), in all the experiments that it was employed, G-mean rates were higher than those unbalanced experiments. Even though the undersampling approach discards samples of the majority class, the performance improvements in the minority class seem to justify such a disadvantage. Although only one

Table 4. *Balanced datasets (via undersampling)*: average recall and precision as a function of the number of selected terms. Best results for each classifier are in boldface.

Dataset	Metric	Classifier	Class	Number of terms						
				50	100	500	1,000	3,000	4,000	5,000
movies	Recall	ANN	Pos	0.783	0.783	0.794	0.798	0.797	0.791	0.782
			Neg	0.702	0.759	**0.769**	0.762	0.767	0.762	0.753
		SVM	Pos	0.775	0.793	0.798	0.773	0.76	0.737	0.743
			Neg	0.696	0.727	0.744	0.758	0.775	**0.784**	0.776
	Precision	ANN	Pos	0.725	0.766	**0.775**	0.77	**0.775**	0.769	0.759
			Neg	0.766	0.779	0.79	0.793	0.794	0.786	0.776
		SVM	Pos	0.719	0.745	0.758	0.763	0.772	**0.775**	0.77
			Neg	0.755	0.778	0.787	0.771	0.764	0.749	0.751
GPS	Recall	ANN	Pos	0.853	0.843	0.846	0.82	0.814	—	—
			Neg	0.722	0.779	0.779	**0.786**	0.756	—	—
		SVM	Pos	0.796	0.793	0.804	0.828	0.852	—	—
			Neg	0.739	**0.752**	0.734	0.732	0.713	—	—
	Precision	ANN	Pos	0.756	0.792	0.793	**0.795**	0.77	—	—
			Neg	0.831	0.835	0.835	0.819	0.804	—	—
		SVM	Pos	0.753	**0.763**	0.752	0.756	0.748	—	—
			Neg	0.785	0.787	0.791	0.811	0.83	—	—
books	Recall	ANN	Pos	0.768	0.78	0.776	0.78	0.743	0.722	—
			Neg	0.617	0.661	**0.689**	0.688	0.675	0.654	—
		SVM	Pos	0.78	0.805	0.814	0.817	0.816	0.764	—
			Neg	0.558	0.597	0.644	0.677	0.668	**0.711**	—
	Precision	ANN	Pos	0.672	0.7	0.715	**0.716**	0.696	0.678	—
			Neg	0.736	0.753	0.755	0.76	0.727	0.702	—
		SVM	Pos	0.639	0.669	0.697	0.719	0.715	**0.727**	—
			Neg	0.731	0.757	0.779	0.788	0.786	0.753	—
cameras	Recall	ANN	Pos	0.874	0.86	0.866	0.865	—	—	—
			Neg	0.793	0.821	**0.831**	0.823	—	—	—
		SVM	Pos	0.855	0.847	0.856	0.866	—	—	—
			Neg	0.799	**0.811**	0.808	0.804	—	—	—
	Precision	ANN	Pos	0.81	0.829	**0.837**	0.831	—	—	—
			Neg	0.862	0.856	0.861	0.859	—	—	—
		SVM	Pos	0.811	**0.818**	**0.818**	0.816	—	—	—
			Neg	0.847	0.841	0.849	0.857	—	—	—

aleatory sample of positive opinions has been made, our experiments indicated that the undersampling strategy considerably benefits the ANN classifier. In situations where it is possible to select the best attributes, the ANN classifier could be competitive with or better than the SVM classifier.

Future work can extend this research to apply other approaches to treat the imbalance. As discussed by López et al. (2012), some studies have indicated that the drop in classifier's performance may be not solely caused by class imbalance, but it may be also related to intrinsic data characteristics like the degree of data overlapping among the classes. In this manner, a future contribution may be in analysing the influence of the imbalance ratio over the classification process as a function of a class overlapping metric on sentiment data.

REFERENCES

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26:12:1–12:34.

Abbasi, A., France, S., Zhang, Z., and Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering*, 23(3):447 –462.

Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D., editors, *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 39–50. Springer Berlin Heidelberg.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Anand, R., Mehrotra, K. G., Mohan, C. K., and Ranka, S. (1993). An improved algorithm for neural network classification of imbalanced training sets. *Neural Networks, IEEE Transactions on*, 4(6):962–969.

Bai, X. (2011). Predicting consumer sentiments from online text. *Decis. Support Syst.*, 50(4):732–742.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Barranquero, J., Díez, J., and del Coz, J. J. (2015). Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591 – 604.
Berry, M. W. and Kogan, J., editors (2010). *Text Mining: Applications and Theory*. Wiley, Chichester, UK.
Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. In *CIKM'11*, pages 375–382.
Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics*, pages 440–447.
Burns, N., Bi, Y., Wang, H., and Anderson, T. (2011). Sentiment analysis of customer reviews: Balanced versus unbalanced datasets. In *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6881, pages 161–170. Springer Berlin / Heidelberg.
Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
Chen, L.-S., Liu, C.-H., and Chiu, H.-J. (2011). A neural network based approach for sentiment classification in the blogosphere. *Journal of Informetrics*, 5(2):313 – 322.
Cheung, C. M. and Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems*, 53(1):218 – 225.
Cheung, C. M. K. and Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decis. Support Syst.*, 54(1):461–470.
Chevalier, J. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354.
Claster, W. B., Hung, D. Q., and Shanmuganathan, S. (2010). Unsupervised artificial neural nets for modeling movie sentiment. *Computational Intelligence, Communication Systems and Networks, International Conference on*, pages 349–354.
Dang, Y., Zhang, Y., and Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25:46–53.
Fersini, E., Messina, E., and Pozzi, F. (2014). Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, 68(0):26 – 38.
Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.
Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266 – 6282.
Guo, H. and Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: The databoost-im approach. *SIGKDD Explor. Newsl.*, 6(1):30–39.
Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag.
Haykin, S. (2008). *Neural Networks and Learning Machines (3rd Edition)*. Prentice Hall.
He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284.
He, Y., Lin, C., and Alani, H. (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
Horváth, G. (2003). Neural networks in measurement systems (an engineering view). In Suykens, J., Horváth, G., Basu, S., Micchelli, C., and Vandewalle, J., editors, *Advances in Learning Theory: Methods, Models and Applications*, volume 190 of *NATO Science Series. Series III: Computer and Systems Sciences*, chapter 18, pages 375–402. IOS Press, Amsterdam.
Hu, N., Pavlou, P. A., and Zhang, J. (2006). Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, pages 324–330, New York, NY, USA. ACM.
Huang, T. M., Kecman, V., and Kopriva, I. (2006). *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning*, volume 17 of *Studies in Computational Intelligence*. Springer, Secaucus, NJ, USA.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192.

Kim, C., Park, S., Kwon, K., and Chang, W. (2012). How to select search keywords for online advertising depending on consumer involvement: An empirical investigation. *Expert Systems with Applications*, 39(1):594 – 610.

Kubat, M., Holte, R., and Matwin, S. (1997). Learning when negative examples abound. In *In ECML-97, Lecture Notes in Artificial Intelligence*, pages 146–153. Springer Verlag.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

Lane, P. C., Clarke, D., and Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems*, 53(4):712–718.

Lee, J., Park, D.-H., and Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electron. Commer. Rec. Appl.*, 7(3):341–352.

Li, S., Ju, S., Zhou, G., and Li, X. (2012). Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, S., Wang, Z., Zhou, G., and Lee, S. Y. M. (2011a). Semi-supervised learning for imbalanced sentiment classification. In *Proc. of Int. Joint Conf. on Artificial Intelligence*, pages 1826–1831.

Li, S., Xia, R., Zong, C., and Huang, C.-R. (2009). A framework of feature selection methods for text categorization. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 692–700.

Li, S., Zhou, G., Wang, Z., Lee, S. Y. M., and Wang, R. (2011b). Imbalanced sentiment classification. In *Proc. of ACM Int. Conf. on Information and Knowledge Management*, pages 2469–2472.

Liu, B. (2011). *Web data mining : exploring hyperlinks, contents, and usage data*. Springer, New York, 2nd ed. edition.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *Trans. Sys. Man Cyber. Part B*, 39(2):539–550.

Liu, Y. (2006). Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3):74–89.

López, V., Fernández, A., Moreno-Torres, J. G., and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistic*, pages 22–31.

Luger, G. F. (2008). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison-Wesley Publishing Company, USA, 6th edition.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Markey, R. G. and Hopton, C. (2000). E-customer loyalty applying the traditional rules of business for online success. *European Business Journal*, 12(4):173–79.

Moraes, R., Valiati, J. F., and Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Syst. Appl.*, 40(2):621–633.

Mountassir, A., Benbrahim, H., and Berrada, I. (2012). An empirical study to address the problem of unbalanced data sets in sentiment classification. In *SMC*. IEEE.

Müller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525 – 533.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653 – 7670.

Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395, Stroudsburg, PA, USA. Association for Computational Linguistics.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics*, pages 271–278.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Park, D.-H., Lee, J., and Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *Int. J. Electron. Commerce*, 11(4):125–148.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Porter, M. (2001). *Snowball: A language for stemming algorithms*.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Riloff, E., Patwardhan, S., and Wiebe, J. (2006). Feature subsumption for opinion analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 440–448.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., and Ventura, S. (2013). Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146.

Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition.

Schlosser, A. (2011). Can including pros and cons increase the helpfulness and persuasiveness of online reviews? the interactive effects of ratings and arguments. *Journal of Consumer Psychology*, 21(3):226 – 239.

Sen, S. and Lerman, D. (2007). Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of Interactive Marketing*, 21(4):76–94.

Severyn, A. and Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 464–469.

Su, C.-T. and Hsiao, Y.-H. (2007). An evaluation of the robustness of mts for imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 19(10):1321–1332.

Sun, A., Lim, E.-P., and Liu, Y. (2009a). On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009b). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719.

Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Application*, 36(7):10760–10773.

Tian, F., Wu, F., Chao, K.-M., Zheng, Q., Shah, N., Lan, T., and Yue, J. (2016). A topic sentence-based instance transfer method for imbalanced sentiment classification of chinese product reviews. *Electronic Commerce Research and Applications*, 16:66 – 76.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

Turney, P. D. (2002). 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*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Verhagen, T., Nauta, A., and Feldberg, F. (2013). Negative online word-of-mouth: Behavioral indicator or emotional release? *Comput. Hum. Behav.*, 29(4):1430–1440.

Vinodhini, G. and Chandrasekaran, R. (2017). A sampling based sentiment mining approach for e-commerce applications. *Information Processing & Management*, 53(1):223–236.

Wang, B. X. and Japkowicz, N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, 25(1):1–20.

Wang, G., Zhang, Z., Sun, J., Yang, S., and Larson, C. A. (2015). Pos-rs: A random subspace method for sentiment classification based on part-of-speech analysis. *Information Processing & Management*, 51(4):458 – 479.

Wang, S., Li, D., Zhao, L., and Zhang, J. (2013). Sample cutting method for imbalanced text sentiment classification based on brc. *Know.-Based Syst.*, 37:451–461.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2004). *Text Mining. Predictive Methods for Analyzing Unstructured Information*. Springer, Berlin, 1 edition.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2010). *Fundamentals of predictive text mining*. Springer-Verlag, London; New York.

Woong Yun, G. and Park, S.-Y. (2011). Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication*, 16(2):201–227.

Wu, G. and Chang, E. (2005). Kba: kernel boundary alignment considering imbalanced data distribution. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):786–795.

Wu, G. and Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning. In *In ICML 2003 Workshop on Learning from Imbalanced Data Sets*, pages 49–56.

Xia, R., Xu, F., Yu, J., Qi, Y., and Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1):36–45.

Xia, R. and Zong, C. (2010). Exploring the use of word relation features for sentiment classification. In *Proc. of the International Conference on Computational Linguistics*, pages 1336–1344.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, R. and Tran, T. (2011). A helpfulness modeling framework for electronic word-of-mouth on consumer opinion platforms. *ACM Trans. Intell. Syst. Technol.*, 2(3):23:1–23:18.

Zhou, Z.-H. and Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. on Knowl. and Data Eng.*, 18(1):63–77.

Zhu, J., Xu, C., and Wang, H.-s. (2010). Sentiment classification using the theory of anns. *The Journal of China Universities of Posts and Telecommunications*, 17:58–62.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26618v1 | CC BY 4.0 Open Access | rec: 5 Mar 2018, publ: 5 Mar 2018