SVM-based sentiment classification: a comparative study against state-of-the-art classifiers

Dionisios N. Sotiropoulos*,
Demitrios E. Pournarakis and
George M. Giaglis

Dept. of Management Science & Technology,
Athens University of Economics and Business,
Evelpidon 47a & Lefkados St, Greece
Email: dsotirop@gmail.com
Email: pournadi@aueb.gr
Email: giaglis@aueb.gr
*Corresponding author

Abstract: Transforming the unstructured textual information contained in various social media streams into useful business knowledge is an extremely difficult computational task, mainly, due to the underlying hard pattern classification problem of sentiment analysis, especially within the context of the Greek language. In this paper, we address the pattern classification problem of sentiment analysis through the utilisation of support vector machines (SVMs). In particular, we conducted an extensive experimental comparison where we tested the aforementioned classifier against a set of state-of-the-art machine learning classifiers on a benchmark dataset originating from the Greek bank sector by collecting data from the streaming API of Twitter that were explicitly referring to the major banks of Greece. Our results present classification accuracy and execution time metrics for each classifier, revealing the superiority of the SVM learning paradigm in assigning patterns to the correct sentiment class.

Keywords: sentiment analysis; support vector machines; SVMs; classification; Twitter.

Reference to this paper should be made as follows: Sotiropoulos, D.N., Pournarakis, D.E. and Giaglis, G.M. (2017) 'SVM-based sentiment classification: a comparative study against state-of-the-art classifiers', *Int. J. Computational Intelligence Studies*, Vol. 6, No. 1, pp.52–67.

Biographical notes: Dionisios N. Sotiropoulos received his PhD in Computer Science from the Department of Informatics at the University of Piraeus, Greece in 2011. He holds a BSc degree in Computer Science since 2003. He is currently working as a Post-Doctoral Researcher at Athens University of Economics and Business in the Department of Management Science and Technology as a member of SocioMine. He is also a Visiting Researcher at Norwich Business School, University of East Anglia. His primary research interests are in the areas of machine learning, data mining, evolutionary computing and signal processing, and applications in user modelling, social networks analysis.

Demitrios E. Pournarakis holds a BSc in Business Administration from Athens University of Economics and Business (AUEB), an MSc in Information Systems from City University London and an MBA from AUEB. He is currently pursuing his PhD at the Athens University of Economics and Business, focusing on Online Social Networks Analysis, under the supervision of Professor George M. Giaglis. He is also a Research Associate at the Athens University of Economics and Business, ISTLab. His research interests include web design (HTML, CSS, Inkscape), data analysis (SQL,R), digital marketing, product branding and social network science.

George M. Giaglis is Vice Rector of Finance and Development and Professor of e-Business at the Athens Universityof Economics and Business, Greece. He has previously worked with the University of the Aegean (Greece) and Brunel University (UK), while he has held visiting posts in universities in the UK, Australia, USA, Finland and Denmark. In 2001, George founded the ISTLab Wireless Research Center, the first research centre in Greece with a focus on mobile business, applications and services, while since 2009 he is the Director of Sociomine, a newly-founded research centre with a focus on social network analytics.

1 Introduction

The popularity and increased usage of online social networks (e.g., Twitter, LinkedIn, Facebook, Instagram, and many others) has generated an unprecedented wealth of content that may be leveraged for social media analysis. Social media analytics (SMA) may be defined as the application of data mining and associated information retrieval techniques to extract patterns of enacted knowledge from complex sets of relationships between members of social systems – such as online social networks and microblogging applications. From a business perspective, social media data represent a rich information source to capture public sentiment and analyse it for actionable decision making in several application contexts as marketing (Cambria et al., 2012), politics (Hong and Nadler, 2012), finance (Bollen and Mao, 2011), and civil security (Cheong and Lee, 2011). Recently, scholars advocated that social media metrics may yield better prediction of corporate performance than conventional metrics (Luo et al., 2013).

Sentiment analysis is an important research field of SMA, which concentrates on detecting the emotional or opinionated polarity of online social media text segments, commonly referred to as social sentiment. Extant research has focused on perfecting the prediction accuracy of social sentiment by developing approaches based on machine learning algorithms or dictionary-based sentiment classification (Thelwall et al., 2011; Paltoglou and Thelwall, 2012). Interestingly, although academic scholars and practitioners may select from a plethora of tools to capture social sentiment, the domain of explaining what influences the formulation of a given social sentiment state remains largely understudied.

This paper addresses the extremely hard pattern classification task that underlies the problem of sentiment classification through the utilisation of the machine learning paradigm of support vector machines (SVMs). Their classification superiority is demonstrated through the utilisation of an extensive experimentation session were their classification accuracy is tested against a benchmark set of state-of-the-art classifiers. Our

results provide significant insights that justify the efficiency of the SVM classifier in addressing the extremely sparse and imbalanced problem of sentiment analysis.

The structure of the paper is as follows: Section 2 reviews the relevant literature of sentiment analysis while Sections 3 and 4 summarise the data collection and corpus vectorisation procedures. Section 5 discusses the problem of sentiment analysis within the context of pattern classification emphasising on its extremely sparse and imbalanced nature. Section 6 inspects the fundamental notions behind the machine learning paradigms of SVMs and Section 7 provides an extensive presentation of the acquired comparative classification results. Finally, Section 8 concludes the paper and highlights avenues of future research.

2 Literature review

Performing sentiment analysis and opinion mining through Twitter is an area that has drawn the interest of many researchers. The challenge to accurately predict social mood based on text mined from Twitter, still remains a big challenge and is currently being explored in various market and academic segments. O'Connor et al. (2010) connected measures of public opinion measured from polls with sentiment measured from text and found that opinions measured from polls correlate to sentiment word frequencies in contemporaneous Twitter messages. The study concludes with the potential of the use of text streams as a substitute and supplement for traditional polling. Jansen et al. (2009) investigated the overall structure of micro-blog postings, types of expressions, and sentiment fluctuations discussing the implications for organisations in using micro-blogging as part of their overall marketing strategy and branding campaigns. Mishne and Glance (2006) in their study, show that, in the domain of movies, there is good correlation between references to movies in weblog posts – both before and after their release – and the movies' financial success. Furthermore, they demonstrate that shallow usage of sentiment analysis in weblogs can improve this correlation. Tumasjan et al. (2010) used the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment. In more detail, the study found that the mere number of messages reflects the election result and even comes close to traditional election polls. Bollen et al. (2011) argue that Twitter mood predicts the stock market. In their study, they conclude that changes in public mood state can indeed be tracked from the content of large-scale Twitter feeds, by means of rather simple text processing techniques and that such change responds to a variety of socio-cultural drivers in a highly differentiated manner, which in turn is correlated or even predictive of DJIA values.

Sentiment analysis of online text content is now in a mature state and a big part of market business analytics software such as Radian6 or IBM Cognos Consumer Insight. L.A. Times, IBM and the University of Southern California Annenberg Innovation Lab (L.A. Times et al., http://graphics.latimes.com/sentimeter/) have used sentiment analysis in twitter feeds to predict the Oscars, in the 2012 ceremony. IBM along with USC Annenberg Innovation Lab (SuperBowl, http://asmarterplanet.com/blog/2012/02/superbowl-analysis-takes-us-beyond-the-tweets.html) performed sentiment analysis on Super Bowl XLVI analysing fan sentiment across 600,000 tweets to determine which players and teams have the most support. Although research in the field of sentiment analytics includes many studies that estimate the polarity of social sentiment with a

variety of approaches (e.g., based on machine learning algorithms or lexicon classification methods) there is limited knowledge regarding the causes leading to a particular social sentiment state. Indeed, social sentiment may be viewed as a multi-dimensional phenomenon; collective opinions on key discussion topics may positively or negatively influence the polarity of social sentiment. This explanatory investigation of social sentiment may yield critical insights for the performance of the object under study. Social media data analysts would know the decomposing factors of social sentiment in the form of semantically defined properties. Our study adopts this investigation perspective and aims at developing a framework that explains the causes of social sentiment rather than simply capturing it. The following section highlights the activities we undertook to develop and empirically assess the proposed framework.

3 Data collection

We randomly collected and analysed a set of over 9,000 tweets during the time period between 2013 and 2015, by utilising the streaming API of Twitter. The data collection process was focused on gathering tweets that were explicitly referring to the four leading banks of Greece, namely National Bank of Greece (NBG), Alpha Bank, Piraeus Bank and Eurobank. This task was accomplished by parsing the official streaming API of Twitter through keyword filtering on the terms 'National Bank', 'NBG', 'Alpha Bank', 'Piraeus Bank' and 'Eurobank'. The resulting dataset was subsequently submitted to a series of data clearing and pre-processing operations. The data preparation process, in particular, involved text tokenisation into words, elimination of Greek stop-words and words with less than three characters, and stem extraction from each word. Therefore, the final version of our corpus was formed by a collection of 9,552 purified documents where each document contained the text from a single tweet. The number of documents pertaining to the class of positive sentiment was 3,132 while the number of documents pertaining to the negative class of sentiment was 6,460. Sentiment class labels were manually obtained by assigning each document with the majority label of polarity provided by a group of post-graduate students. Each student, in particular, undertook the task to associate each document with a distinct sentiment category according to his/hers individual opinion. Therefore, majority voting was utilised in order to eliminate personal biases.

4 Corpus vectorisation

A fundamental prerequisite in order to perform sentiment analysis through the exploitation of any machine learning algorithm is to obtain a mathematical representation of the corpus, so that each document can be treated as a point in a multi-dimensional vector space. A natural approach towards this end was the employment of the standard vector space model (VSM) for our corpus, which was originally introduced by Salton et al. (1975). The main idea behind VSM is to transform each document d into a vector containing only the words that belong to the document and their frequency by utilising the so called 'bag of words' representation. According to VSM, each document is represented exclusively by the words it contains by tokenising sentences into elementary term (word) elements losing the associated punctuation, order and grammar information.

The underlying mathematical abstraction imposed by VSM entails a mapping which transforms the original purified document to its corresponding bag of terms representation. This transformation can be formulated by the following equation:

$$\varphi : d \to \varphi(d) = \left[tf\left(t_1, d\right), ..., tf\left(t_M, d\right) \right] \in \mathbb{R}^M \tag{1}$$

where $tf(t_i, d_j)$ is the normalised frequency of term t_i in document d_j given by the following equation:

$$tf\left(t_i, d_j\right) = \frac{f\left(t_i, d_j\right)}{\max\left\{ f\left(t, d_j\right) : t \in d_j \right\}} \tag{2}$$

given that $f(t_i, d_j)$ is the absolute frequency term t_i in document d_j.

Based on the adopted mathematical formulation for the fundamental notions of corpus and dictionary, such that a corpus D of n documents and a dictionary T of M terms may be represented according to equations (3) and (4).

$$D = \left\{ d_1, ..., d_n \right\} \tag{3}$$

and

$$T = \left\{ t_1, ..., t_M \right\} \tag{4}$$

Having in mind, equation (1) and the formal definitions for the notions of corpus and dictionary, the mathematical representation for corpus in the context of VSM can be done through the utilisation of the document-term matrix given by the following equation:

$$D = \begin{bmatrix} tf\left(t_1, d_1\right) & \cdots & tf\left(t_M, d_1\right) \\ \vdots & \ddots & \vdots \\ tf\left(t_1, d_n\right) & \cdots & tf\left(t_M, d_n\right) \end{bmatrix} \tag{5}$$

where N is typically, quite large resulting in a sparse VSM representation such that a few matrix entries are non-zero. In our approach, in order to mitigate the effect relating to the compete loss of context information around a term, we incorporated the term-frequency inverse document frequency (TF-IDF) weighting scheme according to which each term t_i is assigned a weight of the form:

$$w_i = idf\left(t_i, D\right) = \log \frac{|D|}{\left|\left\{ d \in D : t_i \in d \right\}\right|} \tag{6}$$

so that the relative importance of each term for the given corpus is taken into consideration.

5 Sentiment analysis as a pattern classification problem

Sentiment analysis may be regarded as the computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views and emotions that are exclusively expressed in textual form. In the context of machine learning, however,

sentiment analysis constitutes an extremely hard pattern classification task which in turn can be formally defined as the problem of estimating a mapping of the following form:

$$F : D \to C \tag{7}$$

where $C = \{C_{pos}, C_{neg}\}$, C_{pos} such that C_{pos} indicates the class of positive sentiment and C_{neg} indicates the class of negative sentiment. Having in mind that the original unstructured textual information contained in a given corpus D will be mapped onto an M-dimensional vector space according to VSM, equation (7) could be rewritten in the following form:

$$F : \mathbb{R}^M \to C \tag{8}$$

Letting $I = \{1, \cdots, n\}$ be set of indices that span the purified collection D of documents that pertain to our dataset, the subsets of tweets that are associated with positive and negative sentiment evaluations may be designated as I_{pos} and I_{neg}, respectively such that:

$$I_{pos} \cup I_{neg} = I \tag{9}$$

and

$$I_{pos} \cap I_{neg} = \varnothing \tag{10}$$

In this setting, the ideal functionality provided by the discrimination function defined in equation (8) can be reduced to the following operations:

$$\forall i \in I_{pos}, F\left(\varphi\left(d_i\right)\right) = C_{pos} \tag{11.1}$$

$$\forall i \in I_{neg}, F\left(\varphi\left(d_i\right)\right) = C_{neg} \tag{11.2}$$

Figure 1 Three-dimensional corpus representation (see online version for colours)

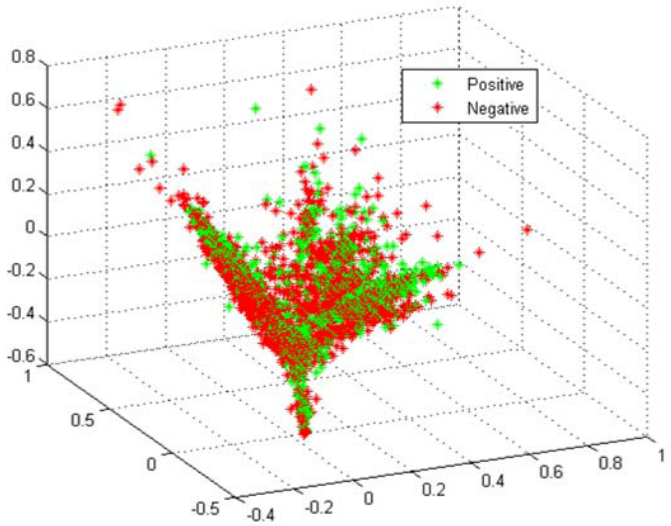


Figure 1 presents a three-dimensional representation of our dataset where the green dots correspond to the positive class patters while the red dots correspond to the negative class patterns. The original dimensionality of our corpus was significantly larger, however, since the number of (TF-IDF)-based feature was experimentally selected to be $M = 400$. Therefore, acquiring a three-dimensional representation of our dataset could only be possible through the utilisation of a dimensionality reduction technique such as principal components analysis. The resulting spatial distribution of the sampled tweets which is illustrated in Figure 1 justifies the severe complexity of the underlying pattern classification problem as well as the highly nonlinear nature of the mapping F that is to be estimated.

The severity of the pattern classification problem that underlies the task of sentiment analysis, however, relates to the extreme degree of sparsity which is associated with the (TF-IDF)-based representation of our corpus. In order to estimate the overall sparsity ratio of the dataset along with the partial sparsity ratio values associated with each distinct class of patterns we need to adapt the following formulation. Having in mind that $\Phi \in \mathbb{R}^{n \times m}$ is the matrix storing the particular (TF-IDF)-based feature values for each document in the given corpus, we may consider its column-wise expansion as:

$$\Phi = \left[\Phi_1, \cdots, \Phi_j, \cdots, \Phi_n\right] \tag{12}$$

where $\Phi_j = [\varphi_{1j}, \cdots, \varphi_{nj}]^T \in \mathbb{R}^n, \ \forall j \in [M]$. The sparsity ratio ($0 \leq \lambda_j \leq 1$) associated with the term $t_j \in T$ may then be defined as:

$$\lambda_j = \frac{\left|\sup\left(\Phi_j\right)\right|}{n}, \forall j \in [M] \tag{13}$$

having in mind that $\sup(\Phi_j)$ corresponds to the support of the j^{th} textual feature according to following equation:

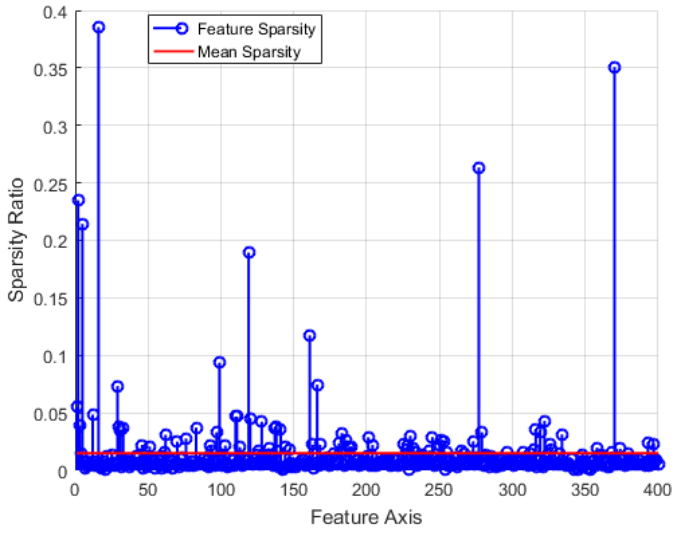
$$sip\left(\Phi_j\right) = \left\{i \in I : \varphi_{ij} \neq 0\right\} \tag{14}$$

Thus, the overall sparsity ratio for the complete set of terms T and therefore for the whole dataset stored in matrix Φ will be given by the following equation:

$$\lambda_0 = \frac{\sum_{j \in [M]} \left|\sup\left(\Phi_j\right)\right|}{nM} = \frac{\sum_{j \in [M]} \lambda_j}{M} \tag{15}$$

The stem plot appearing in Figure 2 illustrates the sparsity ratio per term $t_j \in T$ where the red line parallel to the x-axis depicts the overall sparsity ratio which was estimated to be $\lambda_0 = 0.0149$. It is clearly evident that the majority of the derived textual features are associated with sparsity ratios that are close to zero. This fact is indicative of the complexity of the discrimination function that has to be learned given that the available training instances lie in a high dimensional and extremely sparse vector space.

Figure 2 Sparsity ratio per feature (see online version for colours)



Additional insights concerning the critical role of sparsity when faced with the pattern classification task associated with sentiment analysis could be derived by measuring the partial sparsity ratio per textual feature $t_j \in T$ over the acquired corpus D. Such a measurement may be conducted by imposing a slight modification on equations (11.1) and (11.2) so that the relative computation of the support is constrained within each distinct sentiment category as designated by the following set of equations:

$$S_j^{pos} = \left\{ i \in I_{pos} : \varphi_{ij} \neq 0 \right\} \tag{16.1}$$

$$S_j^{neg} = \left\{ i \in I_{neg} : \varphi_{ij} \neq 0 \right\} \tag{16.2}$$

Equations (16.1) and (16.2) may be subsequently utilised in order to derive the exact formulation for the partial sparsity ratio per term as follows:

$$\lambda_j^{pos} = \frac{\left| S_j^{pos} \right|}{\left| I_{pos} \right|}, \forall j \in [M] \tag{17.1}$$

$$\lambda_j^{neg} = \frac{\left| S_j^{neg} \right|}{\left| I_{neg} \right|}, \forall j \in [M] \tag{17.2}$$

Therefore, the overall sparsity ratio per sentiment class can be easily estimated as:

$$\lambda_0^{pos} = \frac{\sum_{j \in [M]} \left| S_j^{pos} \right|}{\left| I_{pos} \right| M} \tag{18.1}$$

$$\lambda_0^{neg} = \frac{\sum_{j \in [M]} \left| S_j^{neg} \right|}{\left| I_{neg} \right| M} \tag{18.2}$$

The stem plots presented in Figures 3 and 4 depict the partial sparsity values for the positive and the negative class respectively. The red lines parallel to the x-axis correspond to the overall sparsity ratios for each class and they are associated with the particular values $\lambda_0^{pos} = 0.0130$ and $\lambda_0^{neg} = 0.0158$. Once again, only a insignificant fraction of textual features are assigned partial sparsity ratios that considerably exceed zero. This is indicative of the fact that the majority of the extracted textual features cannot be associated with a particular sentiment class rendering the problem of sentiment classification as extremely hard.

Figure 3 Positive sentiment sparsity ratio per feature (see online version for colours)

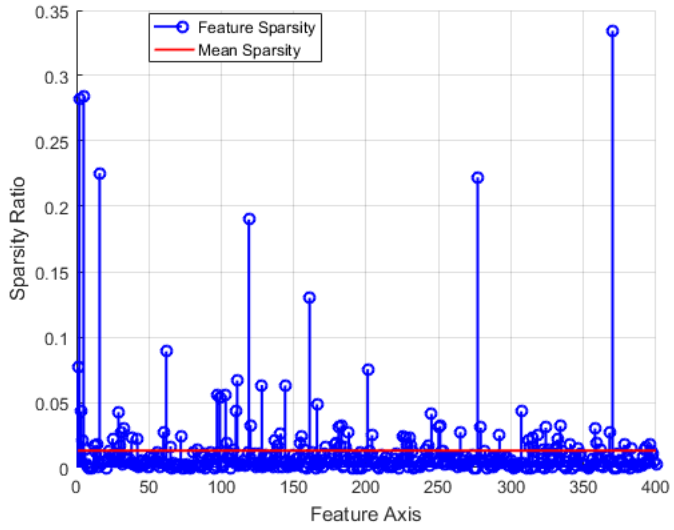
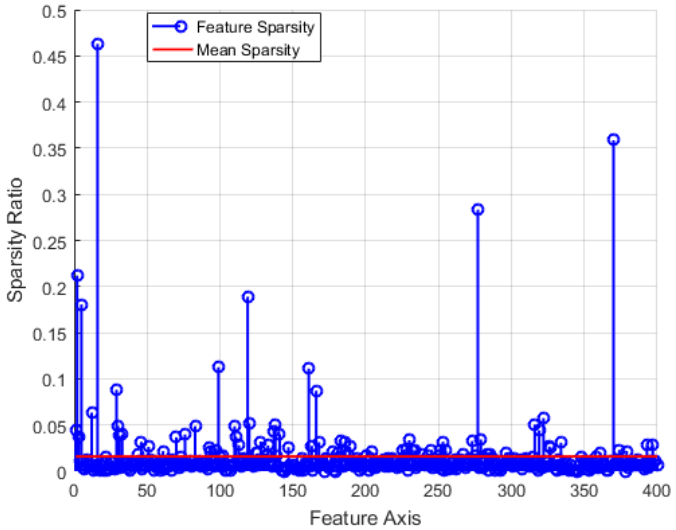


Figure 4 Negative sentiment sparsity ratio per feature (see online version for colours)



6 Support vector machines

Sentiment analysis was conducted through the utilisation of a state-of-the-art classifier, namely SVMs. SVMs are nonlinear classifiers that were initially formulated by Vapnik (1995), operating in higher-dimensional vector spaces than the original feature space of the given dataset. Letting $S = \{(\overrightarrow{x_1}, y_i) \in \mathbb{R}^n \times \{-1, +1\},\ \forall i \in [m]\}$ be the set of m training patterns with associated binary labels, such that –1 denotes the class of negative sentiment and +1 the class of positive sentiment, the learning phase of the SVMs involved solving the following quadratic optimisation problem:

$$\min_{\vec{w},\vec{\xi},b} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{19.1}$$

$$\text{s.t. } y_i\left(\left\langle \vec{w}, \overrightarrow{x_1}\right\rangle + b\right) \geq 1 - \xi_i, \forall i \in [m] \tag{19.2}$$

and

$$\xi_i \geq 0, \forall i \in [m] \tag{19.3}$$

Equations (19.1), (19.2) and (19.3) define the primal optimisation problem whose corresponding dual gives rise to a discrimination function of the form:

$$g(\vec{x}) = \sum_{i \in SV}^{m} \alpha_i^* y_i \left\langle \vec{x}, \overrightarrow{x_1}\right\rangle + b^* \tag{20}$$

where $\{\alpha_i^*, i \in [m]\}$ and b^* denote the optimal solutions for the corresponding optimisation variables and SV is the subset of training patterns associated with positive Lagrange multipliers. Given that the training patterns appear only in dot product terms of the form $\langle \overrightarrow{x_1}, \vec{x} \rangle$, a positive definite kernel function such as $K(\vec{u}, \vec{v}) = \Phi(\vec{u}) f(\vec{v})$ can be employed in order to implicitly map the input feature space into a higher-dimensional vector space and compute the dot product. In this paper, we utilised the Gaussian kernel function defined by the following equation:

$$K\left(\vec{u}, \vec{v}\right) = -\gamma \exp\left(\|\vec{u} - \vec{v}\|^2\right) \tag{21}$$

7 Experimental results

In order to demonstrate the validity of the SVM algorithm for the sentiment classification problem, we adopted the standard ten-fold cross-validation process on a set of 9,552 previously labelled Tweets and measured the corresponding training and testing sentiment classification accuracy. Each fold involved splitting the complete set of pre-labelled samples into a 90% training data – 10% testing data ratio, where the fist subset of data instances was utilised to build the classifier and the latter for assessing its ability to infer the sentiment polarity of unseen data patterns.

To justify the superiority of SVMs in order to address the problem of sentiment classification we measured its classification accuracy against a series of state-of-the-art classifiers which is presented in the following list:

- linear SVMs
- radial basis function neural networks (RBFNs)
- random forests
- multi-layer neural networks (MLPs)
- Bayesian networks
- naïve Bayes
- classification via clustering.

It has to be mentioned that the linear SVM classifier implements a different kernel function than the Gaussian one defined in equation (21) which is given by the inner product of the corresponding input vectors according to the following equation:

$$K(\vec{u}, \vec{v}) = \langle \vec{u}, \vec{v} \rangle \tag{22}$$

RBFN classifier is parameterised by the number of clusters to be estimated which was set equal to the number of sentiment classes pertaining to our dataset. The number of clusters to be estimated was also set to 2 for the classification via clustering approach since the natural number of clusters should coincide with the number of distinct categories of patterns in the dataset. Accordingly, random Forrest classifier is parameterised by the number of trees to be estimated which was allowed to vary in the discrete range {10, 15}. The structure of the MLP classifier is parameterised by the number of hidden layers that was experimentally set to vary in the discrete range {1, 3, 5}.

Table 1 Overall classification accuracy per classifier

Classifier	*Correct (%) classification rate*	*Incorrect (%) classification rate*
SVM RBF	91.7118	8.2882
Random Forrest (trees number = 15)	91.1906	8.8094
Random Forrest (trees number = 10)	91.055	8.945
Multi-layer network (hidden layers = 5)	90.0855	9.9145
Multi-layer network (hidden layers = 3)	89.7519	10.2481
Multi-layer network (hidden layers = 1)	89.1993	10.8007
SVM linear	88.2923	11.7077
Bayesian network	83.5905	16.4095
RBF network	79.5246	20.4754
Naïve Bayes	65.6902	34.3098
Classification via clustering	64.9708	34.9562

Table 1 presents a comparative evaluation of the classification accuracy for each of the utilised machine learning paradigms by averaging the acquired correct and incorrect classification percentages over all folds. SVM classifier with RBF kernel exhibits the

highest classification accuracy by obtaining a correct classification rate of over 91.7%. Comparable classification results were achieved by the random Forrest classifier which exceeded the 91% of accuracy for both number of trees. The machine learning paradigm of multi-layer networks occupies the third ranking position in decreasing order of correct classification accuracy by approaching a value that is marginally greater than 90%.

Table 2 Positive sentiment classification accuracy per classifier

Classifier	*Precision*	*Recall*
SVM RBF	0.901	0.838
Random Forrest (trees number = 15)	0.884	0.841
Random Forrest (trees number = 10)	0.87	0.853
Multi-layer network (hidden layers = 5)	0.855	0.839
Multi-layer network (hidden layers = 3)	0.848	0.836
Multi-layer network (hidden layers = 1)	0.843	0.823
SVM linear	0.862	0.764
Bayesian network	0.708	0.847
RBF network	0.634	0.882
Naïve Bayes	0.486	0.906
Classification via clustering	0.399	0.139

Note: Precision/recall: metrics

The linear SVM classifier despite its simplicity dominates the machine learning approaches provided by the Bayesian network, the RBF network, naïve Bayes and classification via clustering by ranking in the fourth position. Bayesian network is the fifth classifier that exceeds the correct classification rate boundary of 80% with RBF network, naïve Bayes and classification via clustering occupying the sixth, seventh and eighth positions, respectively.

An issue worth discussing relates to the imbalanced nature of the acquired dataset. That is, the majority of the available training patterns pertain to the negative sentiment category with the ratio $\boldsymbol{R}$ of positive to negative training instances be given as:

$$R = \frac{|I_{pos}|}{|I_{neg}|} = 0.4848 \tag{23}$$

This entails that the a priori probability of a document to convey a negative sentiment is at least twice as much as the corresponding probability of expressing a positive sentiment. This fact could in principle obscure the correct interpretation of the obtained classification results since a trivial majority classifier that would completely ignore the positive class of patterns could achieve a minimum correct classification accuracy given by the following equation:

$$P_{\min} = \frac{|I_{neg}|}{n} = 0.6763 \tag{24}$$

Therefore, it is possible for a classifier to score a 67.63% of correct classification rate but at the same time being totally ignorant of the presence of positive patterns in the dataset.

Table 3 Negative sentiment classification accuracy per classifier

Classifier	*Precision*	*Recall*
SVM RBF	0.924	0.955
Random Forrest (trees number = 15)	0.925	0.946
Random Forrest (trees number = 10)	0.93	0.938
Multi-layer network (hidden layers = 5)	0.923	0.931
Multi-layer network (hidden layers = 3)	0.921	0.927
Multi-layer network (hidden layers = 1)	0.915	0.926
SVM linear	0.892	0.941
Bayesian network	0.918	0.83
RBF network	0.929	0.753
Naïve Bayes	0.922	0.536
Classification via clustering	0.682	0.898

Note: Precision/recall: metrics

Table 4 Positive sentiment classification accuracy per classifier

Classifier	*F-measure*	*ROC*
SVM RBF	0.869	0.897
Random Forrest (trees number = 15)	0.862	0.967
Random Forrest (trees number = 10)	0.862	0.962
Multi-layer network (hidden layers = 5)	0.847	0.935
Multi-layer network (hidden layers = 3)	0.842	0.93
Multi-layer network (hidden layers = 1)	0.833	0.903
SVM linear	0.81	0.852
Bayesian network	0.771	0.919
RBF network	0.738	0.842
Naïve Bayes	0.633	0.732
Classification via clustering	0.206	0.519

Note: (F-measure/ROC: metrics

According to the previous analysis, the bottom ranked classifiers, namely, naïve Bayes and classification via clustering could in fact be characterised as worse than majority classifiers since they achieve a worse classification accuracy than $P_{\min}$. Therefore, one should provide additional classification metrics such as precision and recall focusing on each distinct sentiment class. These particular classification measures for each sentiment class over the complete set of the utilised machine learning paradigms are summarised in Tables 3 and 4. It has to be mentioned that the presentation order reflects the initial ranking of the classifiers according to their overall accuracy scores. Surprisingly, there exists no classifier that completely ignores the minority class of positive sentiment. Moreover, our top ranked classifier, SVM with RBF kernel, is able to recognise both sentiment classes with an accuracy level that exceeds the 90% over all folds. Similar classification performance on both classes is also exhibited by the random Forrest.

Table 5 Negative sentiment classification accuracy per classifier

Classifier	*F-measure*	*ROC*
SVM RBF	0.955	0.939
Random Forrest (trees number = 15)	0.935	0.967
Random Forrest (trees number = 10)	0.934	0.962
Multi-layer network (hidden layers = 5)	0.927	0.935
Multi-layer network (hidden layers = 3)	0.924	0.93
Multi-layer network (hidden layers = 1)	0.92	0.903
SVM linear	0.915	0.852
Bayesian network	0.872	0.919
RBF network	0.832	0.842
Naïve Bayes	0.678	0.85
Classification via clustering	0.776	0.518

Note: F-measure/ROC: metrics

The effect of class imbalance becomes apparent for the machine learning paradigms that obtained the worst overall classification results, namely, Bayesian network, RBF networks and naïve Bayes. The classification accuracy for each one of the aforementioned classifiers on the majority (negative) class is over 90% while the corresponding correct classification percentage on the minority class is significantly lower. The ability of our top scored classifiers to cope with the imbalanced sentiment classification problem is also evident by taking into consideration the F-measure and ROC metrics which are summarised in Tables 4 and 5.

Finally, besides reporting classification related metrics we also provide an execution time-based ranking of the utilised classifiers. Such an effort is critical in order to assess the trade-off between computational complexity and classification accuracy. The training time for each classifier is summarised in Table 6 according to which a positive correlation may be revealed between computational complexity and classification accuracy.

Table 6 Execution time per classifier

Classifier	*Training time (secs)*
SVM RBF	472.45
Random Forrest (trees number = 15)	22.26
Random Forrest (trees number = 10)	16.84
Multi-layer network (hidden layers = 5)	426.02
Multi-layer network (hidden layers = 3)	258.38
Multi-layer network (hidden layers = 1)	160.88
SVM LINEAR	63.46
Bayesian network	2.85
RBF network	22.73
Naïve Bayes	1.27
Classification via clustering	8.54

8 Conclusions and future work

This paper addressed the extremely hard pattern classification task that underlies the problem of sentiment classification through the utilisation of the machine learning paradigm of SVMs. Their classification superiority was demonstrated through the utilisation of an extensive experimentation session were their classification accuracy was tested against a benchmark set of state-of-the-art classifiers. Our results justify the efficiency of the SVM classifier in addressing the extremely sparse and imbalanced problem of sentiment analysis. Specifically, SVMs acquired the highest overall classification accuracy in assigning patterns to the correct sentiment category. Comparable classification results were exhibited by random Forests and multi-layers networks. SVMs, however, were proved to be extremely successful in detecting patterns from both sentiment categories despite the fact that the majority of available patterns originated from the negative class. This particular behaviour was not common amongst the rest of the employed classification mechanisms which were significantly biased towards the majority sentiment class. In fact, lower overall classification accuracy was found to be positively correlated with a higher degree of bias towards the negative sentiment category.

Future research will focus on the utilisation of alternative machine learning approaches such as the biologically inspired classification paradigm of artificial immune systems (AIS). AIS-based classification has been proven to be highly efficient in addressing classification tasks under sever imbalance conditions. Sentiment classification provides an ideal experimentation framework in order to extend the existing research that investigates the effects of the class imbalance problem on AIS-based classification mechanisms. This is true, since the vast majority of the publicly announced opinions over the various social media streams are biased towards the negative sentiment class. Moreover, the problem of sentiment analysis reduces to a hard text classification problem over extremely sparse datasets. AIS-based classification algorithms rely on sophisticated subspace sampling operations that aim to generate a minimal set of representative training patterns that reproduce the spatial distribution of the original dataset. Textual feature spaces, however, tend to be extremely sparse rendering the utilisation of AIS-based classifier as an open research problem.

References

Bollen, J. and Mao, H. (2011) 'Twitter mood as a stock market predictor', *IEEE Computer*, Vol. 44, No. 10, pp.91–94.

Bollen, J., Mao, H. and Zeng, X. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, Vol. 2, No. 1, pp.1–8.

Cambria, E., Grassi, M., Hussain, A. and Havasi, C. (2012) 'Sentic computing for social media marketing', *Multimedia Tools and Applications*, Vol. 59, No. 2, pp.557–577.

Cheong, M. and 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*, Vol. 13, No. 1, pp.45–59.

Hong, S. and Nadler, D. (2012) 'Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience', *Government Information Quarterly*, Vol. 29, No. 4, pp.455–461.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009) 'Micro-blogging as online word of mouth branding', in *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, ACM, April, pp.3859–3864.

L.A. Times, IBM and the University of Southern California Annenberg Innovation Lab predict the Oscars, *Senti-Meter* [online] http://graphics.latimes.com/sentimeter/ (accessed 1 February 2013).

Luo, X., Zhang, J. and Duan, W. (2013) 'Social media and firm equity value', *Information Systems Research*, Vol. 24, No. 1, pp.146–163.

Mishne, G. and Glance, N. (2006) 'Predicting movie sales from blogger sentiment', in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, March, Vol. 30, No. 2, pp.301–304.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010) 'From tweets to polls: linking text sentiment to public opinion time series', in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, May, pp.122–129.

Paltoglou, G. and Thelwall, M. (2012) 'Twitter, Myspace, Digg: unsupervised sentiment analysis in social media', *ACM Trans. Intell. Syst. Technol*., Vol. 3, No. 4, pp.1–19.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Commun. Journal*, November, Vol. 18, No. 11, pp.613–620, ACM.

SuperBowl, *SuperBowl Analysis Takes us beyond the Tweets* [online] http://asmarterplanet.com/blog/2012/02/superbowl-analysis-takes-us-beyond-the-tweets.html (accessed 1 February 2013).

Thelwall, M., Buckley, K. and Paltoglou, G. (2011) 'Sentiment in Twitter events', *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 2, pp.406–418.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2010) 'Predicting elections with twitter: what 140 characters reveal about political sentiment', in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, May, pp.178–185.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.