IOP Conf. Series: Journal of Physics: Conf. Series **870** (2017) 012005 doi:10.1088/1742-6596/870/1/012005

Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques

Redouane Karsi[1], Mounia Zaim[1] and Jamila El Alami[1]

[1] Laboratory of System Analysis, Information Processing and Integrated Management, Mohammadia School of Engineers, Mohammed V University, Rabat, Morocco

E-mail: rdkarsi@yahoo.fr, zaim.mounia@yahoo.fr, alamijamila1@gmail.com

Abstract. Thanks to the development of the internet, a large community now has the possibility to communicate and express its opinions and preferences through multiple media such as blogs, forums, social networks and e-commerce sites. Today, it becomes clearer that opinions published on the web are a very valuable source for decision-making, so a rapidly growing field of research called "sentiment analysis" is born to address the problem of automatically determining the polarity (Positive, negative, neutral,…) of textual opinions. People expressing themselves in a particular domain often use specific domain language expressions, thus, building a classifier, which performs well in different domains is a challenging problem. The purpose of this paper is to evaluate the impact of domain for sentiment classification when using machine learning techniques. In our study three popular machine learning techniques: Support Vector Machines (SVM), Naive Bayes and K nearest neighbors(KNN) were applied on datasets collected from different domains. Experimental results show that Support Vector Machines outperforms other classifiers in all domains, since it achieved at least 74.75% accuracy with a standard deviation of 4,08.

1. Introduction

Nowadays, companies and organizations that provide public services increasingly need to extract strategic and decision information from data that has long been produced only through structured information systems, Thus data mining tools have been widely used for data analysis. With the emergence and evolution of social networks, blogs, discussion forums and e-commerce sites, today there is an explosion in the creation of massive data through user-generated web content, these data particularly reflect the preferences and opinions of users in various fields such as politics, health, education, movies, finance, and many research studies have found that the analysis of opinions generated by users is very useful For the prediction of stock price evolution and elections results. Thus, sentiment analysis becomes a very hot field of research [1].

Sentiment analysis is the natural language processing techniques for determining polarity of a text data. The main approaches for sentiment classification are machine learning and lexicon-based approaches. Users express their opinions towards different topics and domains. Thus the challenge is to propose classification methods witch guaranties good performance when applied to different domains, tackling this challenge appears tedious since two different domains use different expressions, for example “honest candidate” expression is frequently used in political domain, however it rarely appears in hotel

domain. Thus, a classifier trained in hotel domain will not be accurate in political domain [2], to overcome this problem many searchers have modelled cross-domain sentiment classification systems which enable training a classifier from one or many domains and applying the trained classifier on a different domain. However, in reality it is not practicable to ensure relatedness between features of the source domain and a large number of target domains, therefore sentiment classification methods are performing well when applied to the same domain as the training data [3].
In this paper supervised machine learning performance was investigated through different domains, choosing to explore machine learning approach is motivated by the fact that searches found that it is more accurate than lexicon based approach when training and testing data are from the same domain [4].
Our contribution is to determine the performance of different machine learning methods applied to different domains, and to explore the effect of varying domain on the accuracy of classifiers by performing experiments using three supervised machine learning: Support Vector machines (SVM), Naive Bayes (NB) and k-nearest neighbors (KNN) applied to distinct datasets covering four different domains: Book, Electronic, Movie and Automobile. Related work on the use of supervised machine techniques for sentiment classification is presented in section 2, our proposed approach is described in section 3, results are commented and discussed in section 4, finally, a conclusion of the paper and an overview of future work are given in section 5.

2. Related Work

Many researchers have investigated supervised machine learning techniques for sentiment classification. From movie reviews dataset, Pang, Lee and Vaithyanathan [5] experimented three machine learning methods (Naive Bayes, maximum entropy and SVM) over n-gram features, they concluded that Naive Bayes shows poor results, whereas SVM algorithm performs good performance. Tripathy, Agrawal and Rath [6] chose to classify movie reviews using Naive Bayes (NB) and Support Vector Machines (SVM), then they compared their performance with results obtained in literature, they observed that SVM yields the best results. Musabah Alkalbani, Mohamed Ghamry and Khadeer Hussain [7] applied five models that are based on five algorithms, the Support Vector Machines algorithm, Naive Bayes algorithm, Naive Bayes (Kernel) algorithm, k-nearest neighbors algorithm, and the decision tree algorithm to classify SaaS (Software as a service) online reviews, they tested all models with unigrams, Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme, using 3-fold, 5- fold and 10-fold cross-validation, the obtained results shows that SVM gives best accuracy. Travel domain was explored by Ye, Zhang and Law [8], their study consist on mining reviews from travel blogs, for this purpose they compared three supervised machine learning algorithms, namely Naive Bayes, SVM and the character based N-gram model for sentiment classification of the reviews on travel blogs for seven popular travel destinations in the US and Europe, the experimental study indicates that SVM and n-gram achieve better performance than the Naive Bayes Method, in addition, When training data sets had 500 or more reviews, all three approaches could reach accuracies of more than 80%. Shi and Li [9] conducted their study on hotel domain, they experimented a supervised machine learning approach using unigram features with two types of weighting scheme (frequency and TF-IDF) to realize opinion classification of documents, they found that using TF-IDF provides more accurate results. Tripathy, Agrawal and Rath [10] Classified movie reviews with four machine learning algorithms: Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM), they found from empirical tests that as the value of "n" in n-gram increases the classification accuracy decreases.

3. Methodology

The steps of our classification approach are shown in figure 1. In our study, three machine learning algorithms were applied, namely Support Vector Machines (SVM), Naive Bayes and K-nearest neighbors (KNN) for sentiment classification of reviews from datasets in different domains.

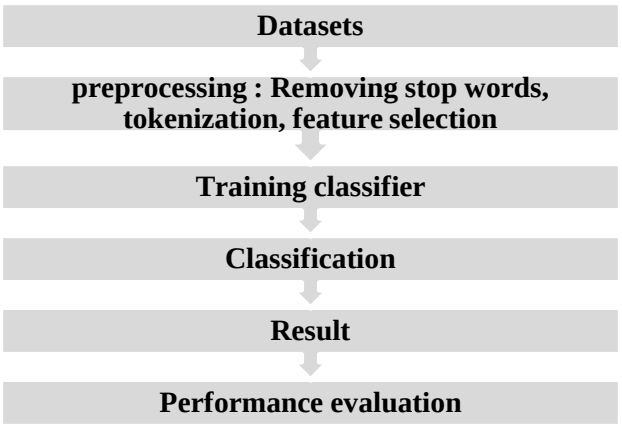

Figure 1. Sentiment classification steps with supervised machine learning.

3.1. Dataset Collection

For our study, four publicly available datasets for research purposes were used, these datasets are described as follows:

Multi-Domain Sentiment Dataset: it is a dataset constructed by Blitzer, Dredze and Pereira [11], they collected Amazon product reviews representing different domains, in our experiment, two domains: Book and Electronic were selected, each domain was built over a collection of 2000 reviews (1000 positive, 1000 negative).

Epinions dataset is built through automobile reviews downloaded from epinions.com web site, the dataset contains 1382 posts with 691 positive and 691 negative reviews.

Polarity dataset: this dataset was created by Pang and Lee [12], it contains 1000 positive and 1000 negative movie reviews all written before 2002.

3.2. Supervised machine learning techniques

In this paper, supervised learning techniques towards different domains were applied by selecting commonly used algorithms for addressing classification problems based on supervised learning methods, in our case, Support vector machines (SVM), Naive Bayes, k-nearest neighbors (KNN) are experimented.

- Support vector machines (SVM): Support Vector Machines (SVM) are new discriminating techniques in statistical learning theory. They have been proposed in 1995 by V. Vapnik in his book "The Nature of Statistical Learning». They allowed tackling several different problems such as regression, Classification… SVM consists in projecting the data of the input space (belonging to two different classes) non-linearly separable in a space of a larger dimension called a feature space so that the data becomes linearly separable. In this space, an optimal hyperplane separating classes such that the two sides of the hyperplane separate the vectors belonging to the different classes and the smallest distance between the vectors and the hyperplane, (the margin) is Maximum [13].

- Naive Bayes: Naive Bayes classifiers are linear and based on Bayes' theorem [10], and the naive term refers to the assumption that words in a dataset are independent. Given an document represented by a dimensional vector $D_i = (X_1,X_2,\ldots,X_n)$, where X_k is the term k of the document D_i, and given a set of m classes,$C_1,C_2,\ldots,C_m$. Using Bayes theorem, the naive Bayesian classifier calculates the posterior probability of each class conditioned on D_i. Di is assigned the class label of the class with the maximum posterior probability conditioned on D_i. Therefore, the most likely class for a document D_i is computed as

$$C^* = argmax_j\, p(Cj)p(^{Di}/_{Cj}) \quad (1)$$

- $P(C_j)$ is the probability of class C_j.
- $P(D_i/C_j)$ is the probability of document D_i conditioned on C_j.

- K-nearest neighbors (KNN): the idea of KNN algorithm is that the class of an observation X is determined according to a majority vote of the k nearest neighbors of the observation [7].

3.3. Proposed approach

Before applying machine learning algorithms, datasets collected for our experiment are preprocessed by following several steps as schematized in figure 1.

- Stop words removal: Unnecessary and insignificant words are removed from datasets.

- Stemming: each word is transformed into its stem.

- Feature selection: Only unigrams from text was extracted, each selected feature was assigned a weight, in this study a popular weighting scheme, TF-IDF was used, where TF is the number of times a term occurs within a document and TF-IDF witch measure the importance of a feature in the document and the whole corpus is calculated according to the following formula:

$$\mathrm{TFIDF} = \mathrm{TF} * \log(\frac{\mathrm{N}}{\mathrm{df}}) \quad (2)$$

• N is the total number of documents.

• df is the number of documents in which the term appears.

-Training classifiers: 3-fold, 5- fold and 10-fold cross-validations in training and test processes [7] were used.

- Performance evaluation: To evaluate the performance of the algorithms, evaluation metrics (precision, recall and accuracy) were computed.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

- TP (True Positive): correctly classified as positive.
- FP (False Positive): incorrectly classified as positive.
- TN (True Negative): correctly classified as negative.
- FN (False Negative): incorrectly classified as negative.

4. Experimental Results

To evaluate the performance of our selected supervised machine learning techniques, multiple performance indicators were computed, namely: Accuracy, recall and precision, for training and testing data, 3, 5, 10 fold cross validation models were used as shown in tables 2 to 5.

After calculating the average accuracy of each algorithm on all domains, SVM shows the best accuracy with an average of 78.92%, followed by NB with 75.95% accuracy, and then KNN achieves a low average of 58.07%.

To evaluate the impact of domain on classification accuracy, the results of our experiment revealed that for SVM the accuracy varies between 74.75% in electronic domain and 85.25% in the movie domain, while NB shows a Minimum accuracy in the book domain with 72.3% and a maximum accuracy of 81.4% in the movie domain, 5NN achieves classification accuracy ranging from 53.5% in the book domain to 64.32% in the automobile domain, 20NN gives accuracies varying between 51.45% in the book domain and 67.22% in the automobile domain, finally 45NN offers an accuracy of 50.3% in the electronic domain and 67.87% in the Automobile domain.

Table 1. Standard deviation of accuracy for applied machine learning algorithms.

Algorithm	Standard deviation
SVM	4.28
NB	4.04
5NN	5.15
20NN	6.74
45NN	8.42

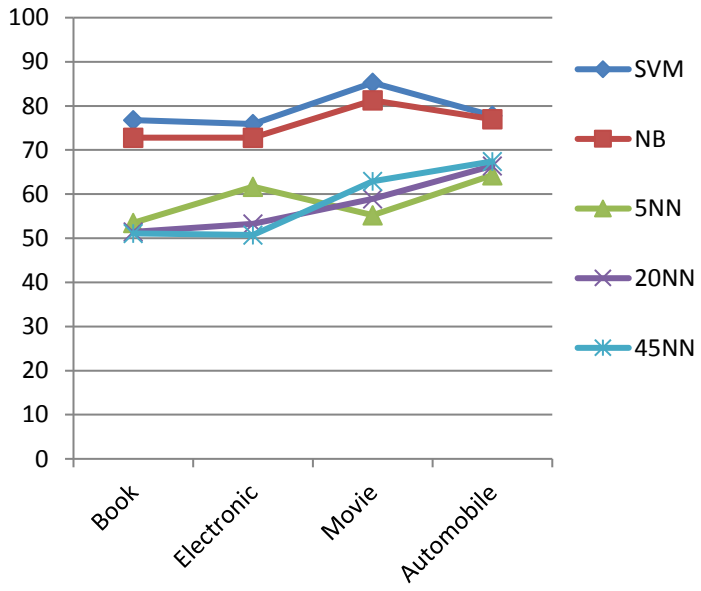

Figure 2. Classification accuracy variation within different domains.

With values of standard deviation less than 5 as shown in table 1, NB and SVM appear to be the most suitable classifiers performing almost similar performances over several domains. Combining low precision and high dispersion, KNN is not good to be used as classifier in different domains.
According to the results of our experiment as shown in figure 2, SVM and NB are more efficient when applied to movie domain, otherwise, KNN behaves well in the automobile domain.
The results also show that k-fold cross validation model affect randomly the classification performance.

Table 2. Performance of different supervised machine learning algorithms on Book domain.

	Fold-3			Fold-5			Fold-10		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
SVM	75.05	75.10	75.90	77.20	77.20	78.00	76.75	76.80	77.40
NB	72.30	72.30	72.30	73.40	73.40	73.40	72.80	72.80	72.80
5NN	54.55	54.60	61.20	53.75	53.80	60.20	53.50	53.50	59.40
20NN	51.85	51.90	58.10	51.70	51.70	63.30	51.45	51,50	62.70
45NN	53.15	53.20	61.90	51.85	51.90	70.20	51.10	51.10	68.60

Table 3. Performance of different supervised machine learning algorithms on Electronic domain.

	Fold 3			Fold 5			Fold 10		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
SVM	74.75	74.80	77.60	77.05	77.10	79.10	75.90	75.90	76.00
NB	75.15	75.20	75.30	75.75	75.80	75.80	72.80	72.80	72.80
5NN	59.00	59.00	59.60	61.10	61.10	61.60	61.70	61.70	62.10
20NN	52.90	52.90	64.40	54.10	54.10	65.50	53.25	53.30	63.90
45NN	50.30	50.30	65.10	50.55	50.60	75.10	50.75	50.80	72.20

Table 4. Performance of different supervised machine learning algorithms on Movie domain.

	Fold 3			Fold 5			Fold 10		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
SVM	83.85	83.90	83.90	84.20	84.20	84.20	85.25	85.30	85.30
NB	81.25	81.30	81.50	81.40	81.40	81.70	81.25	81.30	81.50
5NN	57.15	57.20	57.20	54.45	54.50	54.60	55.20	55.20	55.50
20NN	57.65	57.70	57.80	58.40	58.40	58.50	58.95	59.00	59.30
45NN	61.40	61.40	62.00	60.75	60.80	60.90	62.90	62.90	62.90

Table 5. Performance of different supervised machine learning algorithms on Automobile domain.

	Fold 3			Fold 5			Fold 10		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
SVM	75.54	75.50	76.80	76.91	76.90	77.80	77.78	77.80	78.30
NB	76.84	76.80	76.90	76.41	76.40	76.40	76.98	77.00	77.10
5NN	64.25	64.30	67.00	63.02	63.00	65.50	64.32	64.30	66.70
20NN	67.22	67.20	71.10	65.48	65.50	69.10	66.42	66.40	70.60
45NN	67.80	67.80	72.10	67.87	67.90	73.10	67.43	67.40	72.90

5. Conclusion

The aim of this paper was to investigate the impact of domain on classification performance by proceeding with a machine learning-based approach tested on publicly available datasets covering Book, Electronic, Movie and Automobile domains, According to the results of our experimental study, it was found that Support Vector Machines (SVM) is the most appropriate classifier to be used on different domains.

For further work, it will be examined how the size of the dataset affects classification performance, for that, several datasets of different sizes collected from different domains will be tested.

References

[1] W. Fangzhao, H. Yongfeng and Y. Zhigang, "Domain-Specific Sentiment Classification via Fusing Sentiment," *Information Fusion*, 2016.

[2] B. Heredia, T. M. Khoshgoftaar, J. Prusa and M. Crawford, "Cross-Domain Sentiment Analysis: An Empirical Investigation," in *IEEE 17th International Conference on Information Reuse and Integration*, 2016.

[3] J. Carroll, D. Bollegala and D. Weir, "Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus," *IEEE Transactions on Knowledge and Data Engineering*, vol. **25**, pp. 1719-1731, 2013.

[4] D. Shuyuan, P. S. Atish and Z. Huimin, "Adapting sentiment lexicons to domain-specific social media texts," *Decision Support Systems*, vol. **94**, p. 65–76, 2017.

[5] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," 2002.

[6] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning," *Procedia Computer Science*, vol. **57**, pp. 821-829, 2015.

[7] A. Musabah Alkalbani, A. Mohamed Ghamry and F. Khadeer Hussain, "Predicting the sentiment of SaaS online reviews using supervised machine learning techniques," in *International Joint Conference on Neural Networks (IJCNN)*, 2016.

[8] Q. Ye, Z. Zhang and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications*, vol. **36**, p. 6527–6535.

[9] H.-X. Shi and X.-J. Li, "A sentiment analysis model for hotel reviews based on supervised learning," in *2011 International Conference on Machine Learning and Cybernetics*, 2011.

[10] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. **57**, p. 117–126, 2016.

[11] J. Blitzer, M. Dredze and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," in *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[12] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proceedings of the 42nd ACL*, pp. 271-278, 2004.

[13] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Conference on Empirical Methods in Natural Language* , 2004.