The effect of preprocessing techniques on Twitter Sentiment Analysis

Akrivi Krouska, Christos Troussas, Maria Virvou
Software Engineering Laboratory, Department of Informatics
University of Piraeus
Greece
{akrouska, ctrouss, mvirvou}@unipi.gr

Abstract—**As Twitter offers a fertile ground for expressing different thoughts and opinions, it can be seen as a valuable tool for sentiment analysis. Furthermore, properly identified reviews present a baseline of information as an input to different systems, such as e-learning systems, decision support systems etc. However, the data preprocessing is a crucial step in sentiment analysis, since selecting the appropriate preprocessing methods, the correctly classified instances can be increased. In view of the above, this research paper explains the necessary information to get preprocess the reviews in order to find sentiment and make analysis whether it is positive or negative. Extended comparison of sentiment polarity classification methods for Twitter text and the role of text preprocessing in sentiment analysis are discussed in depth. In the set of tests, possible combinations of methods and report on their efficiency were included, conducting experiments using manually annotated Twitter datasets. Finally, it is proved that feature selection and representation can affect the classification performance positively.**

Keywords—Preprocessing; N-grams; Attribute selection; Learning machines; Sentiment Analysis; Twitter

I. INTRODUCTION

Over the last decades, the rapid growth of the Information Technology has rendered information dissemination very important [1]. The advent of social media has offered the possibility to Internet users to express and share their thoughts and opinions on different topics and events. The impact of social networks like Facebook or Twitter has increased over the past few years and they continue to be the cornerstone in the dissemination of data and information given that people can have access to a large amount of information when using a social network [2]. In particular, Twitter provides a platform on which discussions on various topics can be detected sooner than other standard information channels.

In the scientific literature, Twitter can offer a more fertile ground for sentiment analysis than Facebook. The main reason is the fact that Twitter data and namely tweets can be easily extracted by emoticons (emotion icons). As such, positive, neutral and negative tweets can be extracted from a twitter API. Then, tweets can be downloaded using hashtags. Furthermore, Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brands by extracting and analyzing the sentiment of the Tweets posted by the public about them, their markets, and competitors [3]. Sentiment analysis over Twitter data and other similar micro-blogs faces several new challenges due to the typical short length and irregular structure of such content.

On the one hand, there is the challenge of the effort itself which often is a moving goal; application requirements differentiate (topics of interest, media, available resources, etc) and sentiment analysis inherently cannot cope automatically with these changes. On the other hand, there is the high demand, with a large number of organizations and individuals looking forward to lay their hands on mechanisms that will automatically harness the volume of data generated by users and assist them to evaluate public opinion regarding topics of interest (products, services, people, concepts, etc) [4]. Towards this direction, Twitter has comprised the most prominent playground for sentiment analysis solutions with businesses and scientists alike trying to tap into its users' enthusiasm for sharing opinions publicly online. It is not by chance that numerous works have suggested methods for implementing such mechanisms.

In view of the above, the data preprocessing is a crucial step in sentiment analysis, since selecting the appropriate preprocessing methods, the correctly classified instances can be increased [5]. This paper tackles with the extended comparison of sentiment polarity classification methods for Twitter text and with the role of text pre-processing in sentiment analysis, and with the report on experiment results demonstrating that with the feature selection and representation can affect the classification performance positively. The preprocessing methods evaluated by the current research are three different data representations: unigram, bigrams and 1-to-3 grams, and two feature extraction filters: one based on information gain and the other based on Random Forest algorithm. These settings applied on three different datasets: Obama-McCain Debate (OMD), Health Care Reform (HCR) and Stanford Twitter Sentiment Gold Standard (STS-Gold). Finally, four well-known machine learning algorithms were selected for tweets classification, namely Naïve Bayes (NB), Support Vector Machine (SVM), k- Nearest Neighbors (KNN) and Decision Tree (C4.5), using 10-fold cross-validation method.

The remainder of the paper is organized as follows. First, we present the related work in Twitter sentiment analysis, focus on preprocessing methods used. Following, we present the evaluation procedure of this research, describing the preprocessing techniques, datasets and algorithms used in

detail. Finally, we come up with a discussion on experiment results and we present our conclusions and future work.

II. Related Work

In [4], the authors conducted an extended comparison of sentiment polarity classification methods for Twitter text. Furthermore, they proceeded to the inclusion of combination of classifiers in the compared set and to the aggregation and use of a number of manually annotated tweets for the evaluation of the methods. Especially regarding the latter, they consider it to be a main contribution in the sense that from past experience the automated annotation of tweets based on the detection of features like the emoticons ("☺", "☹", etc) has been problematic since it does not always reflect the case about the overall sentiment expressed by the author, especially when one considers the expression of no-sentiment ("neutral") through the text.

In [6], the authors investigated a classification model for sentiment analysis in micro-blogging posts from Twitter. Using various preprocessing techniques, and applying various feature selection techniques to the Naïve Bayes classifier, they were able to achieve reasonably good performance for the training set used. Ultimately they also noticed that all the classifiers trained were performing slightly better for classifying the positive class compared to the negative class. They show that Naïve Bayes algorithm with application of Information Gain measured using Chi square with minimum value of 3 to select high information features, gives accuracy above 89%.

In [7], the authors described a system designed to detect political topics emerging in Twitter. The main focus lied on a fast detection based on a few tweets at an early stage of a discussion. Furthermore, they have extended their system by a sentiment analysis component to detect the polarity of topics marked by hashtags. Hence, they used special Twitter hashtags, called sentiment hashtags, which people used to tag their opinion about politicians or parties. Their idea was to build up relation graphs for emerging political topics enriched with information like context and polarity. These graphs can later be used to extend an existing web ontology or a semantic network by a new dimension. This will contribute to improve concept-level sentiment analysis methods that use such knowledge bases.

In [3], the authors introduced and implemented a hybrid approach for determining the sentiment carried by each tweet. Also, they demonstrated the value of pre-possessing data using detection, analyzed the slangs/abbreviations and lemmatization and correct and stop words removal. Furthermore, they tested the accuracy of sentiment identification on 6 Twitter datasets and produced an average harmonic mean of 83.3% and accuracy of 85.7%, with 85.3% precision and 82.2% recall. They resolved the data sparsity issue using domain independent techniques. Finally, they compared with other techniques to prove the effectiveness of the proposed hybrid approach.

In [8], the authors presented the two main approaches to build a framework for polarity analysis: based on lexical dictionaries (knowledge-based); or based on machine learning algorithms. The first approach when designed for Twitter text made the system vulnerable to the short messages, informal language with slangs and swear words, absence of explicit sentiments, presence of special characters, mixed language, among other features peculiar to tweets. Thus, they introduced a polarity analysis framework focused on Twitter messages, which combined both approaches. This framework used techniques specifically designed to deal with short messages, such as tweets. The classification was performed by a two-stage machine learning approach, which included an automatic knowledge-based classifier and the machine learning algorithms. This framework provided a modular framework in which each module had different approaches that can be configured according to the application domain.

In [9], the authors presented an analysis of various parameter settings for selected classifiers: Supported Vector Machines, Naïve Bayes and Decision Trees. They used ngrams of normalized words as features and observed the results of various combinations of positive, negative, neutral, and informative sets of classes. They made their experiments in Spanish language for the topic related to cell phones, and also partially used data from tweets related to the recent Mexican presidential elections (for checking the balanced vs. unbalanced corpus).

However, after a thorough investigation in the related scientific literature, we came up with the result that there has not been done extended comparison of sentiment polarity classification methods for Twitter text. Hence, we present the role of text preprocessing in sentiment analysis, and a report on experiment results demonstrating that feature selection and representation can affect the classification performance positively.

III. Evaluation Procedure

The goal of the current research is to evaluate the role of preprocessing techniques on classification problems. Hence, we examine the performance of several well-known learning-based classification algorithms using various preprocessing options on three different subject datasets. Fig. 1 illustrates the steps of evaluation followed in this study.

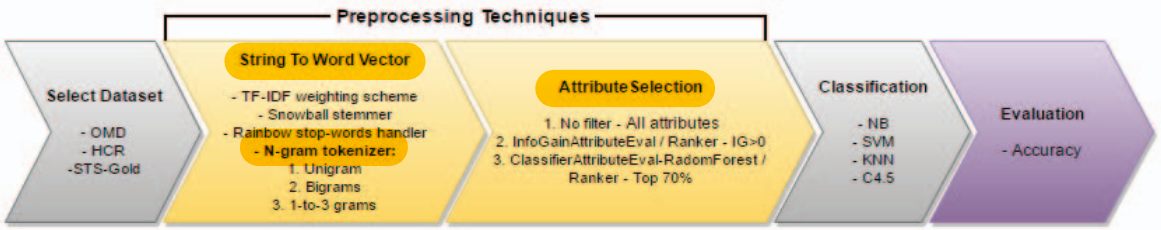

Fig. 1. Evaluation Procedure.

Firstly, we transform each tweet in a word vector form using the TF-IDF weighting model and applying the Snowball stemmer library and the Rainbow list for stop-words removal as fixed options, while we experiment with tokenization and feature selection. As tokenization setting, we choose word unigram, bigram and 1-to-3-gram to compare. After representing the tweets as term-weight vector, we apply feature extraction in order to estimate if the elimination of poorly characterizing attributes can be useful to get better classification accuracy. We examine two different attribute selection methods: one based on information gain and the other based on Random Forest classifier. For each vector model produced by the combination of tokenization and feature selection options, we run the selected classifiers using 10-fold cross-validation method. Afterwards, we evaluate the performance of classifiers according to their accuracy. Finally, the experiments' outcomes have been tabulated and a descriptive analysis has been conducted. All the preprocessing settings and the classification were employed in Weka data mining package[1].

A. *Preprocessing Techniques*

Preprocessing is a necessary data preparation step for sentiment classification. To perform the preprocessing in WEKA, we use the StringToWordVector filter. This filter allows the following configurations:

- **TF-IDF weighting scheme**: It is a standard approach to feature vector construction. TF-IDF stands for the "term frequency-inverse document frequency" and is a numerical statistic that reflects how important a word is to a document in a corpus.
- **Stemming**: Stemming algorithms work by removing the suffix of the word, according to some grammatical rules. In this study, we apply the Snowball stemmer library[2], which is the most popular and standard approach.
- **Stop-words removal:** It is a technique that eliminates the frequent usage words which are meaningless and useless for the text classification. This reduces the corpus size without losing important information. The Rainbow list[3] is used for our experiments.
- **Tokenization**: This setting splits the documents into words/terms, constructing a word vector, known as bag-of-words. We propose NGramTokenizer to compare word unigram, bigram and 1-to-3-gram.

The above preprocessing generates a huge number of attributes, many of them being not relevant with classification. Hence, we apply the following operation:

- **Feature selection**: It is a process by which the number of attributes is decreased into a better subset which can bring highest accuracy. The benefits of performing this option on the data are the limitation of overfitting, the improvement of accuracy and the reduction in training time. Feature Selection methods can be classified as Filters and Wrappers. Filters are based on statistical tests, such as Infogain, Chisquare and CFS, while Wrappers use a learning algorithm to report the optimal subset of features. For this task, WEKA provides the AttributeSelection filter which allows to choose an attribute evaluation method and a search strategy. In this paper, we examine three options:

  a. No filter applied. We use all attributes created by StringToWordVector filter.

  b. InfoGainAttributeEval, which evaluates the worth of an attribute by measuring the information gain with respect to the class and we set the Ranker search method to select attributes with IG>0, and

  c. ClassifierAttributeEval, which evaluates the worth of an attribute by using a user-specified classifier. We choose Random Forest as classifier and set the Ranker search method to select the top 70% attributes.

TABLE I. PREPROCESSING TECHNIQUES APPLIED

Preprocessing Technique	Applied option
Weighting scheme	TF-IDF
Stemming	Snowball stemmer
Stop-words removal	Rainbow list
Tokenization	1. Unigram 2. Bigram 3. 1-to-3-gram
Feature selection	1. All 2. InfoGainAttributeEval / Ranker - IG>0 3. ClassifierAttributeEval-RadomForest / Ranker - top 70%

B. *Data Collection*

For the experiments of current work, three freely-available in Internet datasets, collected from Twitter, were used; a dataset on the Obama-McCain Debate (OMD) [10], one on Health Care Reform (HCR) [11] and one with random subjects tweets, the Stanford Twitter Sentiment Gold Standard (STS-Gold) dataset [12]. These datasets have been created by reputable universities for academic scope and have been used in various researches [13]. Moreover, they consist of a remarkable number of tweets on either specific or no subjects. All these facts render these datasets proper for our research. Table II represents the statistics of above datasets.

TABLE II. STATISTICS OF THE THREE TWITTER DATASETS USED

Dataset	Tweets	Positive	Negative
Obama-McCain Debate (OMD)[4]	1904	709	1195
Health Care Reform (HCR) [20][1]	1922	541	1381
Stanford Twitter Sentiment Gold Standard (STS-Gold)[5]	2034	632	1402

C. *Classifiers*

Four well-known classifiers, namely Naïve Bayes, Support Vector Machine, k-Nearest Neighbor and C4.5, have been

[1] http://www.cs.waikato.ac.nz/ml/weka/index.html
[2] http://snowball.tartarus.org/
[3] http://www.cs.cmu.edu/~mccallum/bow/rainbow/
[4] https://github.com/utcompling/applied-nlp/wiki/Homework5
[5] http://tweenator.com/index.php?page_id=13

chosen in order to evaluate their performance depending on the preprocessing method applied on selected datasets. These classifiers comprise the most representative and state-of-the-art machine learning algorithms which are provided by Weka [14]. Table III shows the classifiers used.

TABLE III. EXPERIMENT CLASSIFIERS

Classifier	Approach
Naïve Bayes (NB)	Probabilistic learning algorithm
Support Vector Machines (SVM)	Supervised learning model
k- Nearest Neighbor (KNN)	Instance-based learning algorithm
C4.5	Decision tree

IV. EXPERIMENT RESULTS AND DISCUSSION

In this section, we applied several preprocessing techniques to three Twitter datasets and for each approach we classified their tweets using four machine learning algorithms. Afterwards, we performed a comparative analysis of the classification accuracy to estimate the effect of preprocessing on Twitter classification problems.

In order to specify the optimal settings of the preprocessing techniques and the classifiers, we conducted a variety of experiments testing the options that would return more accurate results. The chosen preprocessing methods has already described minutely in Section 3. Concerning the classifiers, the Naïve Bayes Multinomial (NBM) and nu-SVM type were chosen, while in KNN the optimal k was 19. For the validation phase, the 10-fold cross validation method was used.

According to n-gram and attribute selection options, a different number of attributes was created based on which the classification was performed. Table IV demonstrates that numbers. We observe that selecting the attributes with information gain upper than zero, the resultant number of them is decreased appreciably. For the second feature extraction, where the attributes are evaluated by Radom Forest algorithm, we choose approximately the 70% of attributes ranked as more worthy. An expected benefit of attribute selection is that algorithms train faster.

TABLE IV. NUMBER OF ATTRIBUTES CREATED BY ATTRIBUTE SELECTION OPTION

Dataset	Attribute Selection	N-gram		
		Unigram	Bigram	1-3gram
OMD	All	2150	7400	2430
	IG>0	264	519	1074
	Top 70%	1500	5180	1680
HCR	All	3000	1835	2945
	IG>0	281	645	1280
	Top 70%	2100	1280	2060
STS-Gold	All	2990	8420	2115
	IG>0	252	354	720
	Top 70%	2090	5890	1480

Table V demonstrates the performance of classifiers depending on the preprocessing methods applied. Regarding dataset representations, the behavior is not uniform. There is no representation that brings systematically better results in comparison with the others. In general, 1-to-3-grams perform better than the other representations, having a close competition with unigram.

Our evaluation results indicate that the attribute selection operation improves the performance of classification over selecting all attributes. This proceeds from the removal of redundant and irrelevant attributes from the datasets which can be misleading to modeling algorithms and result in overfitting. In Table VI, we observe that significant accuracy rates are obtained when applying the attribute selection based on information gain. Note particularly that in case of NB, the percentage of correctly classified instances is increased over 7 points. Moreover, the Random Forest algorithm as attribute selection classifier improves classification accuracy in comparison to using all attributes. Finally, in some experiment settings, there is no improvement in algorithms' performance by applying an attribute selection filter, but this is of insignificant value as the divergence is too low.

Evaluating the influence of dataset domain on preprocessing performance, we use three different datasets, one with no specific domain tweets and the others with a specific topic. The experiment results show that the effect of preprocessing techniques is the same regardless of the datasets.

TABLE V. CLASSIFIERS' ACCURACY RELATED WITH PREPROCESSING TECHNIQUES

		Dataset											
		OMD				HCR				STS-Gold			
N-gram	Attribute Selection	Classifiers				Classifiers				Classifiers			
		NB	SVM	KNN	C4.5	NB	SVM	KNN	C4.5	NB	SVM	KNN	C4.5
Unigram	All	79.46%	81.93%	73.32%	75.26%	77.59%	**83.15%**	72.02%	73.63%	82.10%	83.63%	68.93%	74.29%
	IG>0	86.61%	**83.61%**	**74.21%**	**75.89%**	85.34%	79.04%	**72.85%**	**74.57%**	88.35%	**83.92%**	**73.40%**	**74.29%**
	Top 70%	**88.08%**	83.25%	73.58%	74.79%	**88.25%**	82.53%	72.02%	73.27%	**91.40%**	83.68%	68.93%	74.04%
Bigrams	All	75.26%	85.82%	62.76%	**69.22%**	82.16%	78.00%	71.81%	71.66%	70.30%	**85.05%**	68.93%	70.50%
	IG>0	**89.02%**	82.88%	**63.39%**	68.28%	**89.13%**	**78.36%**	71.81%	71.92%	**86.77%**	80.63%	**68.98%**	70.21%
	Top 70%	83.04%	**87.08%**	62.76%	69.01%	87.52%	78.16%	**71.82%**	**72.07%**	80.53%	84.32%	68.93%	**70.55%**
1-3 grams	All	85.77%	82.51%	**68.86%**	76.05%	83.26%	**77.54%**	**72.13%**	73.95%	87.56%	81.91%	68.93%	74.39%
	IG>0	**92.59%**	**84.14%**	66.02%	**76.21%**	**91.94%**	76.76%	71.97%	**74.62%**	**92.67%**	**83.83%**	69.12%	74.73%
	Top 70%	88.34%	83.88%	66.44%	75.37%	87.94%	77.17%	71.92%	73.90%	90.81%	82.74%	**68.93%**	**74.93%**

TABLE VI. RELATIVE IMPROVEMENT IN ACCURACY OF CLASSIFIERS DEPENDING ON ATTRIBUTE SELECTION OPTIONS

Dataset		OMD		HCR		STS-Gold	
Attribute Selection		IG>0	Top 70%	IG>0	Top 70%	IG>0	Top 70%
N-gram	**Classifiers**						
Unigram	NB	**+7.15**	**+8.62**	**+7.75**	**+10.66**	**+6.25**	**+9.30**
	SVM	+1.68	+1.32	-4.11	-0.62	+0.26	+0.05
	KNN	+0.89	+0.26	+0.83	0	+4.47	0
	C4.5	+0.63	-0.47	+0.94	-0.36	0	-0.25
Bigram	NB	**+13.76**	**+7.78**	**+6.97**	**+5.36**	**+16.47**	**+10.23**
	SVM	-2.94	+1.26	+0.36	+0.16	-4.42	-0.73
	KNN	+0.63	0	0	+0.01	+0.05	0
	C4.5	-0.94	-0.21	+0.26	+0.41	-0.29	+0.05
1-3 gram	NB	**+6.82**	+2.57	**+8.68**	+4.68	**+5.11**	+3.25
	SVM	+1.63	+1.37	-0.78	-0.37	+1.92	+0.83
	KNN	-2.84	-2.42	-0.16	-0.21	+0.19	0
	C4.5	+0.16	-0.68	+0.67	-0.05	+0.34	+0.54

V. CONCLUSIONS AND FUTURE WORK

For the classification task to be done, a preliminary phase of text preprocessing and feature extraction is essential. The preprocessing operations affect the quality of the classification, for this reason we perform various experiments on different generated datasets.

In this paper, we experiment with a series of preprocessing methods applied on three different datasets of tweets, one with no specific domain and the others with certain topics, and evaluated the performance of four well-known classifiers. Our experiments' results illustrate that with appropriate feature selection and representation, sentiment analysis accuracies can be improved. In particular, unigram and 1-to-3-grams perform better than the other representations and feature extraction improves the classification accuracy in comparison with using all created attributes.

However, it is worthy to investigate further the available preprocessing options in order to find the optimal settings. Regarding the attribute selection strategies, a deeper study can focus on the choice of the best algorithm to assess the attributes or the evaluation of rankings methods such as Infogain, Chisquare, etc. Moreover, a future research should involve embedded methods which carry out feature selection and model tuning at the same time. Finally, the incorporation of such pre-processing techniques for sentiment analysis into e-learning systems could enhance the personalization of students and thus it consists of our future plan.

ACKNOWLEDGMENT

The authors of this paper would like to thank the University of Piraeus Research Center for the financial support of this research paper.

REFERENCES

[1] C. Troussas, M. Virvou and S. Mesaretzidis, "Comparative analysis of algorithms for student characteristics classification using a methodological framework", 6th International Conference on Information, Intelligence, Systems and Applications (IISA), doi: 10.1109/IISA.2015.7388038, pp. 1-5, 2015.

[2] C. Troussas, M. Virvou, and K. J. Espinosa, "Using visualization algorithms for discovering patterns in groups of users for tutoring multiple languages through Social Networking", Journal of Networks, vol. 10, no. 12, pp. 668-674, 2015.

[3] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme", Decision Support Systems, 57(1), pp. 245-257, 2014.

[4] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing methods for twitter sentiment analysis," KDIR 2014 - Proceedings of the Int. Conf. on Knowledge Discovery and Information Retrieval, pp. 225-232, 2014.

[5] E. Haddi, X. Liu, Y. Shi, "The role of text pre-processing in sentiment analysis", Proc. Comp. Sci. 17, pp. 26–32, 2013.

[6] S. Fouzia Sayeedunnissa, A. R. Hussain, and M. A. Hameed, "Supervised opinion mining of social network data using a bag-of-words approach on the cloud", 7th International Conference on Bio-Inspired Computing: Theories and Application, 2013.

[7] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," Knowledge-Based Systems, 69(1), pp. 24-33, 2014.

[8] A. C. E. S. Lima, L. N. De Castro, and J. M. Corchado, "A polarity analysis framework for twitter messages", Applied Mathematics and Computation, 270, pp. 756-767, 2015.

[9] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, .., and J. Gordon, "Empirical study of machine learning based approach for opinion mining in tweets", 11th Mexican International Conference on Artificial Intelligence, MICAI, 2012.

[10] D. Shamma, L. Kennedy, and E. Churchill, "Tweet the Debates: Understanding Community Annotation of Uncollected Sources," ACM Multimedia, ACM, 2009

[11] M. Speriosu, N. Sudan, N. Upadhyay, and J. Baldridge, "Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph," Proceedings of the First Workshop on Unsupervised Methods in NLP, Edinburgh, Scotland, 2011.

[12] H. Saif, M. Fernez, Y. He, and H. Alani, "Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold," 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), Turin, Italy, 2013.

[13] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," Information Sciences, 285(1), pp. 181-203, 2014.

[14] X. Wu, V. Kumar, J. Ross Quinlan, …, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2007.