Cross-Lingual Sentiment Classification from English to Arabic using Machine Translation

Adel Al-Shabi, Aisah Adel, Nazlia Omar, Tareq Al-Moslmi
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
University Kebangsaan Malaysia, Malaysia

***Abstract*—Cross-lingual sentiment learning is becoming increasingly important due to the multilingual nature of user-generated content on social media and the scarce resources for languages other than English. However, cross-lingual sentiment learning is a challenging task due to the different distribution between translated data and original data and due to the language gap, i.e. each language has its own ways to express sentiments. This work explores the adaptation of English resources for sentiment analysis to a new language, Arabic. The aim is to design a light model for cross-lingual sentiment classification from English to Arabic, without any manual annotation effort which, at the same time, is easy to build and does not require deep linguistic analysis. The ultimate goal is to find an optimal baseline model and to determine the relation between the noise in the translated data and the accuracy of sentiment classification. Different configurations of several factors are investigated including feature representation, feature reduction methods, and the learning algorithms to find the optimal baseline model. Experiments show that a good classification model can be obtained from translated data regardless of the artificial noise added by machine translation. The results also show a significant cost to automation, and thus the best path to future enhancement is through the inclusion of language-specific knowledge and resources.**

Keywords*—Cross-lingual sentiment classification; English to Arabic; machine translation

I. INTRODUCTION

Given the quantity and massive popularity of multilingual user-generated content on social media, the need for effective multilingual and cross-lingual sentiment analysis is becoming increasingly important. Typically, CLSC refers to the task of predicting the polarity of the opinion expressed in a text in a label-scarce target language using a classifier trained on the corpus from a label-rich source language. CLSC is popularly studied to reduce the expense of manual annotation efforts required in the target language domain [1]-[3]. To date, a variety of lexicon-based and corpus-based methods have been developed for sentiment classification. The lexicon-based methods rely heavily on a sentiment lexicon containing positive terms and negative terms. The corpus-based methods rely heavily on an annotated corpus for training a sentiment classifier. The sentiment lexicon and corpus are considered the most valuable resources for the sentiment classification task. However, such resources for the world's languages are rather unbalanced. Because most previous work focuses on English sentiment classification, many annotated sentiment lexica and corpora for English sentiment classification in various domains are freely available on the Web. The annotated resources for sentiment classification in many other languages are not abundant and it is time-consuming to manually label a rich and reliable sentiment lexicon or corpus in those languages [3]-[5]. In general, efforts towards building sentiment analysis methods for other languages have been hampered by the high cost involved in creating corpora and lexical resources for a new language. The present study investigates whether creating sentiment resources with machine translation is a viable alternative to labor-intensive manual annotation tasks. In particular, we focus on the problem of English-to-Arabic cross-lingual sentiment classification, leveraging only English sentiment resources for sentiment classification of Arabic product reviews, without using any Arabic sentiment resources.

Pilot studies have been performed to make use of machine translated English resources for sentiment analysis in other languages [2], [6]-[8]. However, adapting machine translated English resources to entirely new languages usually produces various challenges, as each language may be significantly different in terms of the characteristics and translation quality differs from language pair to language pair. Moreover, it is widely believed that aspects of sentiment may be lost in translation, especially in automatic translation [3]. The extent of this loss, in terms of drop in accuracy of automatic sentiment analysis remains undetermined [3]. In this work, one of the main objectives is to determine extent of this drop.

Keeping these thoughts in mind, we explore the ability of machine translation to generate reliable training data for scarce-resources languages such as Arabic. We employ machine translation to obtain training and test data for the Arabic language. In particular, the present work involved several experiments in order to perform extensive evaluation of the possible combination of different data preparation strategies (i.e., feature extraction, representation, and selection), as well as a variety of classification algorithms. Our goal here is two-fold. First, we are occupied with choosing the optimal model, which obtains the maximum performance for English-Arabic cross-lingual sentiment classification. Second, we seek to understand sentiment predictability of Arabic text using a classification model trained by using automatically translated data. The results obtained from these experiments will help users identify the methods best suited for their particular needs.

The rest of the paper is organized as follows. In Section II we summarized related work. Section III describes the problem formulation followed by Section IV which presents the proposed model. The evaluation criteria is discussed in Section V while the experimental results and discussion are presented in Section VI. Finally, Section VII concludes the paper.

II. Related Work

Cross-lingual sentiment classification is a popular topic in the sentiment analysis community. It aims to solve the sentiment classification task from a cross-language point of view. Previous research developed methods to map sentiment analysis and resources on English to other languages. Mihalcea et al. [9] proposed a method to learn multilingual subjective language via cross-language projections. Bautin et al [10] proposed cross-lingual sentiment analysis using machine translation. They use machine translation in order to convert all considered texts into English and subsequently perform sentiment analysis on the translated results. By doing so, the authors assume that the results of the analysis on both the original text and the translated text are comparable and that the errors made by the machine translation do not significantly influence the results of the sentiment analysis. Inui and Yamamoto [11] employed machine translation and, subsequently, sentence filtering to eliminate the noise obtained in the translation process. That work is based on the idea that sentences that are translations of each other should contain sentiment-bearing words that have the same polarity. Demirtas et al. [12] use machine translation to employ labelled instances in Turkish for expanding the training set in English considered as the target language for polarity detection. They also consider a co-training approach as a viable alternative to leveraged machine translated data. Wan [13] designed cross-lingual sentiment classification based on machine translation where the source language is English and the target language is Chinese. The available resources include both English sentiment lexicons and training corpuses.

More recently, Balahur and Trurchi, [4] and Becker et al. [14] investigated how a simple strategy can address the problem of sentiment analysis in multiple languages. Particularly, they analyze how the use of machine translation systems - such as Google Translate - can affect the performance of English Sentiment Analysis methods in non-English datasets. Their findings suggest that machine translation systems are mature enough to produce reliably translations to English that can be used for sentence-level sentiment analysis and obtain lower, but still competitive prediction performance results. They also show that some popular language specific methods do not have significant advantages over a machine translation approach. In these works, several commercial machine translation systems which can be publicly accessed are used to map English corpora and resources to other languages such as by Google Translate, Yahoo Babel Fish, Bing Translator, and Windows Live Translate.

Our proposal builds upon the above mentioned works to investigate the suitability of translation-based cross-lingual sentiment analysis for Arabic sentiment classification. To our knowledge, no published work has yet investigated this topic.

III. Problem Formulation

The problem of cross-lingual sentiment classification is to leverage available resources in a source language for sentiment classification in a target language. Here, the source language is English and the target language is Arabic. The aim of this study is to design a light model for cross-lingual sentiment classification from English to Arabic, without any manual annotation effort which is at the same time easy to build and does not require deep linguistic analysis. To do so, the following sections describe the problem formulation and the proposed model.

To identify the problem of cross-lingual sentiment classification in a formal manner, we adopt Balahur & Turchi's [4] formulation for sentiment classification. The profile of cross-lingual sentiment classification performance *CLSCP* can thus be defined as a function of five factors: the quality of the translated resources *tq*, the feature set, *fs*, the feature representation, *fr*, the learning algorithm, *l*, and the experimental design, *ed* (e.g. data split): *CLSCP = fn (tq, fs, fr, l, ed)*. To design an effective and optimal CLSC model, extensive evaluation of the different combination of these factors is needed. Error of translating sentiment expression leads to a much smaller sentiment expression intersection between translations and native expressions, as well as different semantic feature distributions between original language and target language contents. As a result, CLSC tasks cannot achieve performance comparable to that obtained for monolingual sentiment classification tasks. The maximum performance or the upper bound *scpmax* can be obtained by the perfect translations of the training data which are equivalent here to the monolingual sentiment classification. The lack of manually translated training data for the target language and the large cost of manually producing it do not allow us to compute the maximum sentiment classification performance, *scpmax*, in the target language using translated training and gold standard testing data.

Several evaluation metrics, methods and tools for machine translation (MT) are introduced. The BLEU evaluation metric is known to have good correlations with human evaluation. This work evaluates its suitability on measuring the sentiment predictability of the translated data.

IV. Proposed Model

As mentioned above, the aim of this study is to design a light model for cross-lingual sentiment classification from English to Arabic, without any manual annotation effort which is easy to build and does not require deep linguistic analysis. Based on the problem formulation, the proposed model illustrated in Fig. 1 consists of the following steps:

A. Data Acquisition

In order to overcome the language barrier, we must translate one language into another language. For this purpose, the present work adopted Google Translate (GT) to translate the corpora from English-to-Arabic as it offers API access and is considered the state-of-the-art machine translation system used today [4], [13], [14].

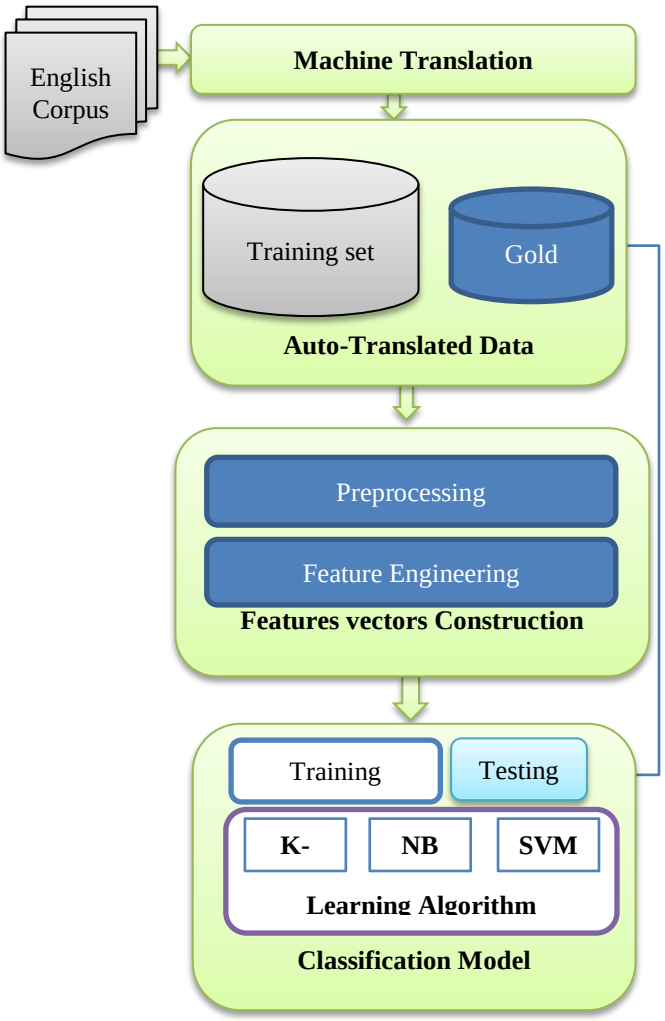

Fig. 1. Proposed framework.

With the aid of machine translation, we can translate the data in both directions to form target language to the source language (i.e., Arabic to English) or from source language to the target language (i.e. from English to Arabic). In this work, we choose the latter because it does not involve any translation at test time, i.e., there is no need to translate every new test dataset and hence has lesser test-time complexity and cost (it just has a fixed training time cost). Consequently, the Amazon Products Reviews data set was pushed through the machine translation to eliminate the gap and get an Arabic version of the data set.

B. Data split and Gold Standard

The auto-generated data set was split into training and testing. Then, a fraction of auto-translated data was selected randomly and manually corrected to serve as a gold standard test set. These correctly translated test sets allow obtaining a more precise measure of the impact of translation quality on the sentiment classification task. Although the upper bound for the proposed model would be possible to estimate using the Gold Standard for each of the training sets as well, at this point a scenario that is closer to real situations was selected as the issue is related to the non-existence of training data for a specific language.

C. Preprocessing

As with any sentiment classification system, the first step is pre-processing the plain texts. For Arabic texts, text pre-processing usually involves the following: removing punctuation marks, diacritics and non-Arabic letters, excluding the words with length less than three, and eliminating stop-words [15]. Arabic TREC-2002 Light Stemmer [16] have been employed to return the words to their stems by removing the most frequent suffixes and prefixes.

D. Feature Engineering

Choosing features is crucial in situations where no high-quality training data are available, as in our case. Sentiment analysis tasks require effective representations of textual inputs. These representations can arise from feature design and control the noise of data. In our case, the noise is likely to come from two sources, namely, incorrect translations or features that are not appropriate [4], [13], [14]. Thus it is crucial to distinguish between the drop in accuracy that caused by inappropriate feature representation from that might have occurred because of erroneous translation. By this method the extracted features have been represented in different ways on the one hand to determine the source of drop in accuracy and on the other hand we want to understand which feature representation/weighting is more robust with respect to the noise data and gives the best performance and under what conditions. The features used in this study include; unigrams, bigrams, trigram and the feature weightings used are term frequency and term frequency-inverse document frequency (TFIDF), Binary Occurrence (BO) and Term Occurrence (TO). Previous classification studies using n-gram modeling usually included some sort of feature reduction technique to reduce the dimensions space of the features vector and to extract the most important words or phrases. For this purpose, the Information Gain (IG) heuristic was used to conduct feature selection due to its reported effectiveness in previous text-classification research [17]. All the features with an information gain greater than 0.0025 were selected.

E. Classification Algorithm

Since there is no prior research on CLSC from English to Arabic, little guidance is available about which machine learning techniques work well for such a task. Therefore, several learning algorithms were explored and compared. In particular, SVM, NB and KNN classifiers were utilized. The choice of these classification algorithms is based on numerous experimental confirmations of its effectiveness for cross-lingual information retrieval tasks [18] and monolingual sentiment classification [19]. In the following, these learning algorithms are briefly described.

Support Vector Machine (SVM): A linear supervised algorithm, which works well both for regression and classification. An SVM model is a representation of the instances as points in space, mapped so that the instances of the separate categories are divided by a clear gap that is as wide as possible. An SVM algorithm constructs a hyperplane (or set of hyperplanes) that divides the space into dimensions

representing classes. The algorithm chooses the hyperplane(s) that maximizes the distance from it to the nearest data point of each class, the solution being handled as a quadratic programming (QP) optimization problem.

Naive Bayes (NB): This is one of the simplest probabilistic classification algorithms widely used for text and opinion mining due to its good results. It is based on the application of the Bayes Theorem, which assumes total independence of variables. The algorithm is fast, deals with high dimensionality (i.e., high number of features), and types of features.

K-nearest neighbour (KNN): The K-Nearest Neighbour (K-NN) is a well-known instance-based classifier. In this classification algorithm, a new input instance should belong to the same class as its k nearest neighbours in the training data set. Given a test review r, the system finds the K nearest neighbours among the training reviews. The similarity score of each nearest neighbour review to the test review is used as the weight of the classes of the neighbour review. The weighted sum in KNN classification and can be written as follows:

$$scored(\mathrm{r}, t_i) = \sum_{r_j = KNN(\mathrm{r})} sim(\mathrm{r}, r_j)\delta(\mathrm{r}_j, c_i) \quad (1)$$

Where *KNN(r)* indicates the set of *K* nearest neighbors of review *r*. If r_j belongs to $\mathrm{c_i}$, then $\delta(r_j, c_i)$ equals one; otherwise, it is zero. For test reviewd *r*, it should belong to the class that has the highest resulting weighted sum.

V. Evaluation

In this evaluation, we seek answers to the following questions: (1) With auto translated data, which feature representation/weighting leads to the best classification performance? (2) To what extent does the noise of translation in training data affect the accuracy of sentiment classification? (3) Which kind of classifier is most appropriate for sentiment classification under such conditions? In order to answer these questions and mainly to test the performance of Arabic sentiment classification when using translated data, different experimental settings of supervised learning were employed with different configurations.

A. Evaluation Setup

*Dataset***:** Standard evaluation benchmarks for cross-lingual sentiment classification from English to Arabic are not available. Therefore, we used the Amazon corpus [20] as a benchmark and developed our own gold standard. This dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics, and Kitchen appliances. Each review comes with the full text and the rating score by the reviewer. More details about this dataset are presented in Table I.

The Gold Standard was used to test the performance of sentiment classification using translated (noisy) versus correct data. Each review comes with the full text and the rating score by the reviewer.

TABLE I. Characteristics of the Data Set

Dataset/features	Books	DVDS	Electronics	Kitchen
No. of reviews	2000	2000	2000	2000
Positive	1000	1000	1000	1000
Negative	1000	1000	1000	1000
No of features	188050	179879	104027	89478
Average length/review	239	234	153	131

VI. Results and Analysis

Since we sought to study the ability of machine translation to generate reliable training data which can be employed to perform sentiment analysis for Arabic languages, several experiments were conducted to perform extensive evaluation of different configuration feature representation, feature weighting and classification algorithms.

Feature representation/weighting: To examine which feature representation/weighting lead to the best classification performance, we represented translated training datasets with three features representations in unigrams, bigrams, and trigrams, and four feature weighing methods, term frequency (TF), term frequency–inverse document frequency (TFIDF), Binary Term Occurrence (BTO) and Term Occurrence (TO). The effects of feature representation and feature weighing methods on sentiment analysis performance were examined. The classification accuracies that achieved using these different configurations on Books, DVDs, Electronics, and Kitchen datasets are shown in Tables II, III, IV, and V, respectively using the three mentioned classification methods. As shown in Tables II, III and IV, it is evident how bigrams representation with term frequency almost achieves the best results with naïve Bayes classifier compared to the unigram and trigram representation. On the other hand, trigram representation with Term Occurrence (TO) always achieve the best results with the SVM classifier compared to the unigram and bigram representation. The results show that each feature representation and weighting method acts differently with each classification model. As noted from the results obtained for SVM classifier in the four datasets, the Term Occurrence (TO) is more suitable than other weighting methods when the SVM classifier is used while the term frequency and TF-IDF respectively are the best weighting methods when naïve Bayes and KNN classifiers are used. The comparison between unigram, bigram, and trigram representation methods shows that the unigram is less suitable for the noisy data. This can be explained taking into account the nature of the task (sentiment analysis) where sentiment is usually expressed in phrases rather than a single word. For example negative words can shift the polarity of a specific word. So polarity analysis in phrase-level or expression level (bigrams and trigrams) is expected to give better results. Moreover, the unigram models do not consider how opinion is composed (e.g., intensifier, negation) and therefore fail to recognize many sophisticated opinion patterns. For Arabic, a morphological-complex

language, wrong translation also leads to an explosion of features, of which many are irrelevant for the learning process.

Classification methods: The main aim here is to answer the research question as to which type of classifier is most appropriate for sentiment classification under conditions of noisy translations in training data. Fig. 2, 3, 4 and 5 shows the performance of the three classifiers NB, KNN, and SVM.

Comparing the behaviors of the three classifiers results, the results show that the classification performances of the three classifiers with feature representation/weighting methods vary from dataset to dataset. In addition, there is no superior classifier for all feature representation/weighting methods. Table III and Fig. 2 show the experiments indicated that the SVM classifier produced superior results to other classification methods for almost all datasets. The experiments also indicated that the KNN classifier produced the worst results on all datasets. The highest performances are obtained by the SVM classifier on Books, DVDs, Electronics and Kitchen Appliances domains. However, given different experimental settings there is no classifier that is superior in overall.

D*omains and translation quality*: As it is known that, the quality of the machine translation differs from domain to domain. The aim here is to study effects on the quality of the machine translation and to determine to what extent the noise of translation in training data may affect the accuracy of sentiment classification. Table VI shows the best result obtained for each domain along with the translation quality measured by BLEU score. It is notable that there is a correlation between the BLEU score values and the classification performance of classifiers. The comparison results demonstrate that the different classification schemes rely heavily on the translation quality. In general, the bigger picture of the results obtained show that the existing machine translation reach a level of maturity to generate reliable training data for scarce-resource languages such as Arabic. However, the results still far of satisfactory comparing the results archived using the same dataset in the source language.

TABLE II. PERFORMANCE OF THE NB, SVM, AND KNN CLASSIFIERS ON BOOK DOMAIN DATASET

Feature Representation/Weighting	NB	KNN	SVM
Unigram_ TF	56.55	47.86	50.55
Unigram_Tf-Idf	54.58	52.62	57.94
Unigram_TO	52.99	48.23	55.23
Unigram_BTO	52.99	48.33	53.64
Bigram_ TF	60.32	52.45	48.2
Bigram_Tf-Idf	58.67	50.08	46.66
Bigram_TO	56.84	53.03	48.2
Bigram_BTO	57.08	52.78	48.2
Trigram_ TF	59.11	52.47	53.03
Trigram_Tf-Idf	57.15	48.83	61.26
Trigram_TO	55.53	51.42	62.89
Trigram_BTO	53.88	49.7	54.62

TABLE III. PERFORMANCE OF THE NB, KNN, AND SVM CLASSIFIERS ON DVD DOMAIN DATASET

Feature Representation/Weighting	NB	KNN	SVM
Unigram_ TF	58.66	48.51	50.65
Unigram_Tf-Idf	56.13	53.74	60.85
Unigram_TO	53.74	45.34	57.85
Unigram_BTO	54.77	46.5	58.97
Bigram_ TF	62.65	52.45	54.56
Bigram_Tf-Idf	59.88	57.71	62.87
Bigram_TO	57.85	49.16	64.52
Bigram_BTO	54.83	50.36	63.26
Trigram_ TF	61.23	51.8	53.41
Trigram_Tf-Idf	58.35	56.05	61.45
Trigram_TO	56.48	47.69	63.18
Trigram_BTO	55.66	47.35	59.89

TABLE IV. PERFORMANCE OF THE NB, SVM, AND KNN CLASSIFIERS ON ELECTRONICS DOMAIN DATASET

Feature Representation/Weighting	NB	KNN	SVM
Unigram_ TF	58.97	47.79	43.85
Unigram_Tf-Idf	56.63	54.24	44.22
Unigram_TO	52.83	47.09	58.32
Unigram_BTO	54.98	48.25	57.38
Bigram_ TF	62.53	52.08	47.71
Bigram_Tf-Idf	60.84	58.51	48
Bigram_TO	56.67	49.43	67.32
Bigram_BTO	59.14	52.18	61.61
Trigram_ TF	61.55	51.15	46.64
Trigram_Tf-Idf	59.01	57.61	46.93
Trigram_TO	55.55	50.05	**66.05**
Trigram_BTO	55.88	49.11	58.29

TABLE V. PERFORMANCE OF THE NB, SVM, AND KNN CLASSIFIERS ON KITCHEN DOMAIN DATASET

Feature Representation/Weighting	NB	KNN	SVM
Unigram_ TF	57.41	50.74	49.9
Unigram_Tf-Idf	56.85	54.12	55.98
Unigram_TO	55.63	48.51	57.29
Unigram_BTO	57.17	48.88	60.01
Bigram_ TF	60.43	53.51	54.37
Bigram_Tf-Idf	60.72	57.04	60.36
Bigram_TO	59.71	50.61	68.49
Bigram_BTO	61.4	52.83	64.32
Trigram_ TF	59.32	52.32	53.03
Trigram_Tf-Idf	59.71	55.39	58.96
Trigram_TO	58.51	50.03	68.84
Trigram_BTO	58.08	49.75	60.93

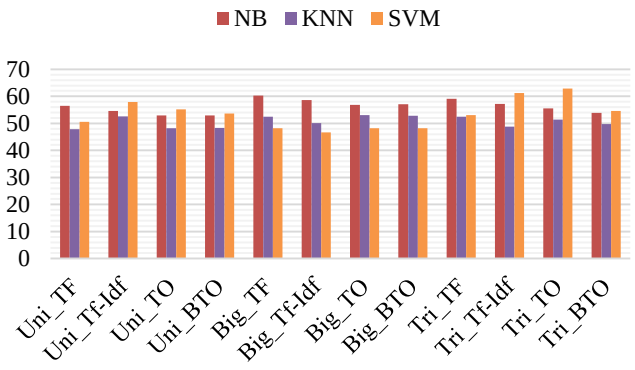

Fig. 2. Results of book dataset with different representations, weightings and classifiers.

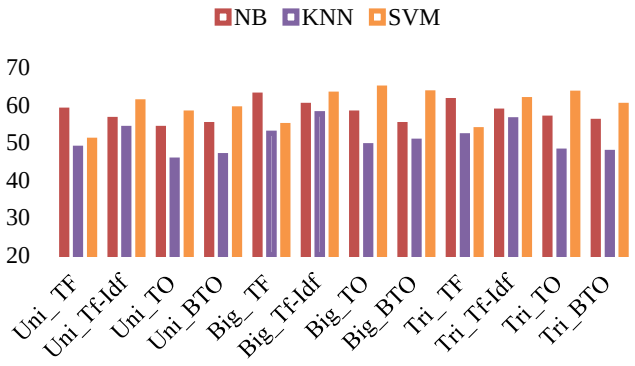

Fig. 3. Results of DVD dataset with different representations, weightings and classifiers.

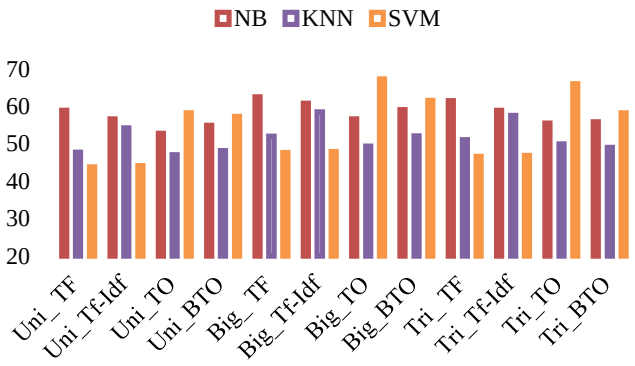

Fig. 4. Results of electronics dataset with different representations, weightings, and classifiers.

TABLE VI. THE BEST RESULT OBTAINED FOR EACH DOMAIN ALONG WITH THE TRANSLATION QUALITY MEASURED BY BLEU SCORE

Dataset	**SVM**	**BLEU score.**
Book	62.89	0.203
DVD	63.18	0.207
Electrics	66.05	0.209
Kitchen	68.84	0.212

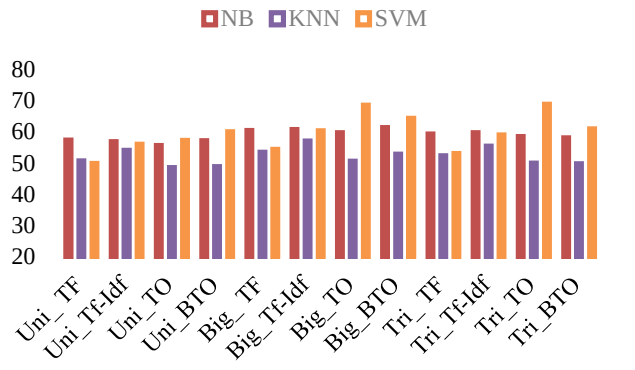

Fig. 5. Results of kitchen dataset with different representations, weightings and classifiers.

VII. CONCLUSION

In this work, we have proposed and pursued an extensive evaluation of the use of translated data in the context of Arabic sentiment analysis. Our findings show that translated data using state of the art statistical machine translation systems have reached a reasonable level of maturity to produce sufficiently reliable training data for scarce-resources languages. Different configurations of several factors have been investigated including feature representation, feature reduction methods, and the learning algorithms to find the optimal baseline model. To limit these problems, we tested three different classification approaches, using different types of features and feature weighting methods. The proposed approach clearly depends on the availability of the translation engines for the required languages.

In future work, we plan to investigate new data representation schemes. We believe that improvement of translation quality through a post processing module will lead to great improvements on results and can reduce the impact of the translation errors. Furthermore, future work should cope with semantic gap and distribution disparity by making use of target language resources and machine translation.

REFERENCES

[1] Bangalore, S., P. Lambert, and E. Montiel-Ponsoda, Introduction to the special issue on cross-language algorithms and applications. Journal of Artificial Intelligence Research, 2016. 55: p. 1-15.

[2] Hajmohammadi, M.S., et al., Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. Information sciences, 2015. 317: p. 67-77.

[3] Mohammad, S.M., M. Salameh, and S. Kiritchenko, How Translation Alters Sentiment. J. Artif. Intell. Res.(JAIR), 2016. 55: p. 95-130.

[4] Balahur, A. and M. Turchi, Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 2014. 28(1): p. 56-75.

[5] Omar, N., et al., Ensemble of Classification Algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews. International Journal of Advancements in Computing Technology, 2013. 14(5): p. 77-85.

[6] Chaturvedi, I., E. Cambria, and D. Vilares. Lyapunov filtering of objectivity for Spanish Sentiment Model. in Neural Networks (IJCNN), 2016 International Joint Conference on. 2016. IEEE.

[7] Korayem, M., K. Aljadda, and D. Crandall, Sentiment/subjectivity analysis survey for languages other than English. Social Network Analysis and Mining, 2016. 6(1): p. 75.

[8] Vilares, D., M.A. Alonso, and C. Gómez-Rodríguez, Supervised sentiment analysis in multilingual environments. Information Processing & Management, 2017. 53(3): p. 595-607.

[9] Mihalcea, R., C. Banea, and J.M. Wiebe, Learning multilingual subjective language via cross-lingual projections. 2007.

[10] Bautin, M., L. Vijayarenu, and S. Skiena. International Sentiment Analysis for News and Blogs. in ICWSM. 2008.

[11] Inui, T. and M. Yamamoto, Applying sentiment-oriented sentence filtering to multilingual review classification. Sentiment Analysis where AI meets Psychology (SAAIP), 2011: p. 51.

[12] Demirtas, E. and M. Pechenizkiy. Cross-lingual polarity detection with machine translation. in Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2013. ACM.

[13] Wan, Xiaojun. "Bilingual co-training for sentiment classification of Chinese product reviews." Computational Linguistics 37.3 (2011): 587-616.

[14] Becker, Karin, Viviane P. Moreira, and Aline GL dos Santos. "Multilingual emotion classification using supervised learning: Comparative experiments." Information Processing & Management 53.3 (2017): 684-704.

[15] Alabbas, W., H.M. Al-Khateeb, and A. Mansour. Arabic text classification methods: Systematic literature review of primary studies. in Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on. 2016. IEEE.

[16] Darwish, K. and D.W. Oard. Term selection for searching printed Arabic. in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002. ACM.

[17] Tang, B., et al., A Bayesian classification approach using class-specific features for text categorization. IEEE Transactions on Knowledge and Data Engineering, 2016. 28(6): p. 1602-1606.

[18] Xu, R., et al. Cross-lingual Text Classification via Model Translation with Limited Dictionaries. in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. ACM.

[19] Dey, L., et al., Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier. arXiv preprint arXiv:1610.09982, 2016.

[20] Blitzer, John, Mark Dredze, and Fernando Pereira. "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification." ACL. Vol. 7. 2007.