2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI) | 978-1-7281-7397-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IIAI-AAI50415.2020.00025

Multilingual Sentiment Analysis for Web Text Based on Word to Word Translation

Keita Fujihira
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
keita.fujihira@jaist.ac.jp

Noriko Horibe
Depertment of Computer and Information Science
Sojo University
Kumamoto, Japan
horibe@cis.sojo-u.ac.jp

Abstract— **People's sentiments are known to have a large impact on changes in stock prices, products sales, and trends. Since web users generally state their opinion in various languages, it is important to develop a method of multilingual sentiment analysis for web texts. In this research, we design a multilingual sentiment analysis method based on word to word translation using a sentiment dictionary in arbitrary native language. This method consists of three phases: morphological analysis of text, a sentiment extraction of each word with sentiment dictionary, and a sentiment extraction of text based on words sentiments. We conduct a sentiment classification experiment for tweets in English, German, French, and Spanish. In the experiment, we evaluate our classifier's performance by comparing the classifier with the other previous classifiers based on the evaluation standards "Accuracy", "Precision", "Recall", and "F1 score". The experimental results show that our classifier has an applicability to sentiment analysis for multilingual, because our classifier's performance is independent of the differences languages.**

Keywords—Multilingual sentiment analysis, Opinion mining, Sentiment dictionary, Machine translation, Natural Language Processing.

I. Previous Researches on Use and Extract of Sentiment Information in Web Text

A wide variety of web services such as micro blog and commercial website have been rapidly spread all over the world. Rapid growth of web services has been inducing an increase of web users and a diversification of data on website. Contents of web texts can influence the movement of real society, and advance of market research and social analysis using web texts are becoming more important on Economics, Politics, and Sociology. In particular, people's sentiments are known to have a large impact on the analysis with respect to changes in stock prices, products sales, and trends. Therefore, many researchers are studying in this field. Bollen *et al.* proposed a prediction method for the stock market using sentiment information from tweet texts on Twitter [1]. In the research, sentiment information was extracted with "OpinionFinder" and "Google-Profile of Mood States (GPOMS)". Moreover, they built a Self-Organizing Fuzzy Neural Network model using the sentiment information as a feature for prediction and showed that prediction accuracy of the model is higher than accuracy of conventional prediction model. Tumasjan *et al.* analyzed tweets regarding political parties or politicians to investigate whether Twitter is used as a place for political discussion and whether user's sentiments reflect the results of election [2]. The results showed that election results have correlation with the number of political tweets. Moreover, they reported that political sentiments contained in the tweets coincide with political reputation for parties and politicians. In this way, since analysis using sentiments on web texts is useful for solutions to social problems, researchers in the field of Natural Language Processing are interested in studying regarding sentiment analysis methods.

However, in most sentiment analysis systems, usable languages are restricted to only one kind of language because of the difficulty of natural language processing. In general, sentiment classification performance can be low when a user uses such systems for other languages. Low performance hinders accurate analysis on research and commerce using sentiment information. A development of multilingual sentiment analysis systems is indispensable for realizing a sentiment analysis of texts written in various languages. Although various resources (such as corpora and lexicons) are required for a development of sentiment analysis systems, it is difficult to obtain these resources sufficiently for some languages [3, 4]. Reuse of the resources has been taken in multilingual sentiment analysis researches because major language (generally, English) has many resources. Araújo *et al.* investigated whether English sentiment analysis tools are effective for translated texts in nine languages and compared the performance of the tools [5]. Comparison results showed that machine translation leads to lower sentiment classification performance on nonEnglish texts compared to performance on native English texts. Balahur and Turchi translated English texts into French, German, and Spanish [6]. Although they classified sentiments using Support Vector Machine that is trained with translated texts, classification accuracy was not sufficient in their experimental results. In the results of these papers, one of the most serious causes on the performance degradation is considered as an error of translation for each text. Can *et al.* initially built a machine learning model using recurrent neural networks for texts included in English reviews [7]. Furthermore, the learning model applied for texts on other languages. In the classification experimental results, they claimed that machine translation does not affect the performance because translation errors are not noticeable and necessary information for sentiment analysis are not lost. Other researches also reported that sentiment analysis performance is not significantly affected by machine translation results [8, 9].

Various researchers in the field of multilingual sentiment analysis have been using a machine translation as one of the functions of their analyzing processes. A performance of the

978-1-7281-7397-9/20/$31.00 ©2020 IEEE
DOI 10.1109/IIAI-AAI50415.2020.00025

machine translation is considered to have a big influence on the results of multilingual sentiment analysis using machine translation. Therefore, the difficulties of machine translation for each language is a bottleneck for multilingual sentiment analysis using machine translation. As an example, it is difficult to maintain sentimental phrases, nuances and expressions peculiar to a language on simple machine translation of a text, and this becomes a factor that deteriorates the performance of sentiment analysis. For this reason, there are various opinions about using machine translation for multilingual sentiment analysis. In order to analyze sentiment while maintain sentimental phrases, we consider that word to word translation can be a valid method instead of machine translation for a whole text.

In this paper, we propose a dictionary based multilingual sentiment analysis method based on word to word translation. This method uses only one sentiment dictionary on user's native language and has a word translation processes instead of translation for whole a text. In the method, text is divided into words by morphological analysis and then each word is translated into a word on native language. A native language sentiment dictionary calculates sentiment of each word and sentiment of text is classified based on each word's sentiments. A classifier using the method reduces a risk of losing sentimental phrases, and extract sentiment of text. The method is applicable for many languages, and reduces performance variability due to language difference because contextual information is not used. Another feature of the method is that each word in text is represented in a multiple word on native language that have similar meanings, that is, in the analysis results, users can easily understand which word cause sentiment. To evaluate the performance of our classifier, we conduct an experiment using tweets on Twitter in English, French, German, and Spanish, and compare with other multilingual sentiment classifiers. Contributions of this study are: 1) proposal of a multilingual sentiment analysis approach based on word to word translation, 2) reducing variability of accuracy associated with language difference and mistranslation.

II. Proposal of Sentiment Analysis Method Based on Word to Word Translation

In order to extract sentiment information from texts in a nonnative language, a translation on each language generally becomes an important process. Actually, some sentiment analysis methods were constructed using translation processes as a part of their functions. However, in general, translation process is a high cost, and it is difficult to compute translations considering appropriate expressions on each language. Therefore, in this research, we propose an efficient sentiment analysis method which is applicable into various languages without translation processes for whole a text. Our classifier deals with multilingual and uses only the translation process on words. In this section, we introduce a sentiment analysis method and tools used in the method. Figure 1 shows an outline of our sentiment analysis method from a web text. The method consists of three phases: a morphological analysis of a text, a sentiment extraction from each word with native language sentiment dictionary, and a sentiment extraction from a text based on sentiments of words.

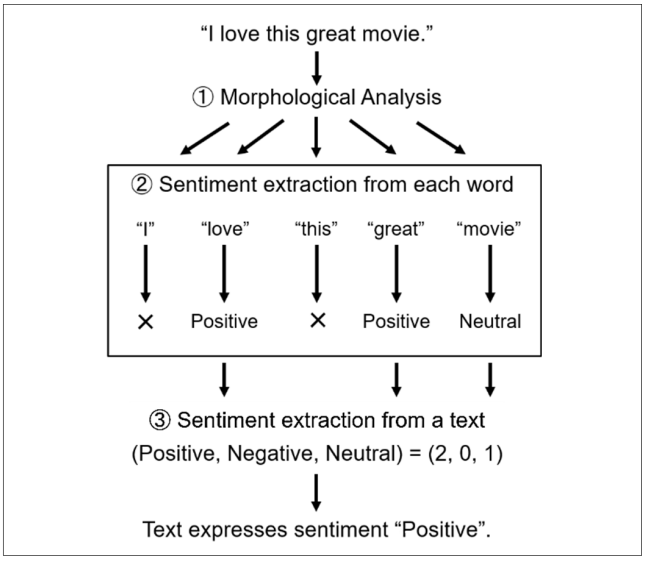

Fig. 1. Outline of a multilingual sentiment analysis method.

A morphological analysis is a process that divides a sentence into the smallest meaningful unit called " morphemes" for each language. We use "TreeTagger" as a morphological analysis tool [10]. The morphological tool "TreeTagger" deals with 25 languages including English and German. Furthermore this tool can be extended to be applicable for other languages if a lexicon and a tagged training corpus are available. Note that our sentiment analysis method uses only the words which aren't a part of speech such as article and interrogative. In the preprocess, we remove useless words and use only adjective, adverb, common noun, and verb. It is known that if useless words aren't removed sufficiently and a part of them remain, then the results of a sentiment classification becomes worse.

The second phase of our method is a sentiments extraction of words. Figure 2 shows the process of the second phase by exemplifying a word "Love" which is obtained as a part of outputs from the morphological analysis. At first, all words which have a similarity to the original word "Love" are detected by using a corpus. In order to calculate word similarity, we use an open-source library "fastText" and word vectors models trained on corpora such as "Common Crawl" and "Wikipedia" [12]. The library "fastText" is used for learning of text classification and word embedding. The word vectors model is adopted as an expression of the meaning of each word. Secondly, the similar words are translated into native language by a translation system "Google Translate". In our method, the corpus and the word-to-word dictionary are the only two necessary knowledges about the language of the input text. Thirdly, some words are extracted from the set of words obtained by the translation. As shown in Figure 2, the original word "Love" and each word obtained by translating the word are paired up, and an average value of vectors is calculated from the vector of the pair. Words are extracted as the high similarity words in descending order of the average values. Fourthly, sentiments of each pair are classified by a sentiment dictionary

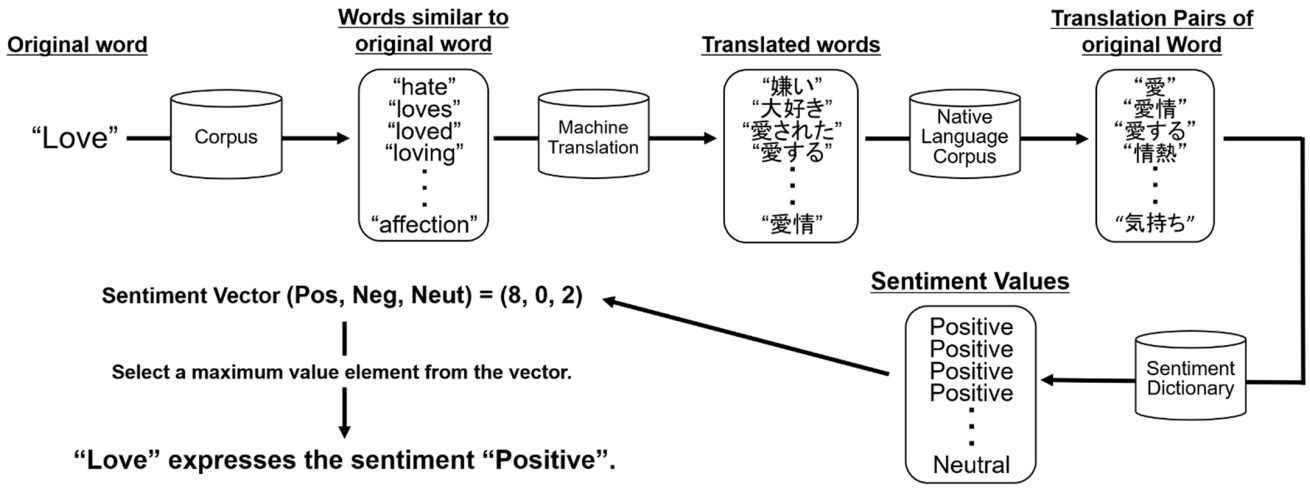

Fig. 2. Procedure to calculate the sentiment for each word.

in native language. The sentiment dictionary used in this research is released by Takamura *et al.* [13]. In this dictionary, sentiment values for each words are real numbers in the range –1 to +1 automatically assigned using a lexical network. The words which have value close to –1 represent "Negative", and the words which have value close to +1 represent "Positive". Let $w_1, w_2, \ldots, w_n$ be words in sentiment dictionary D, and $F(w_i)$ be the sentiment value of a word w_i. We define three sets "Negative", "Neutral", and "Positive" as

$$Negative = \{w_i \in D \mid -1.0 \leq F(w_i) \leq -0.4\},$$

$$Neutral = \{w_i \in D \mid -0.4 < F(w_i) < -0.37\},$$

$$\text{Positive} = \{w_i \in D \mid -0.37 \leq F(w_i) \leq 1.0\},$$

respectively. These borderlines are used in our experiment and defined, considering a balance of the proportion of the number of words in a sentiment dictionary, and results obtained from our preliminary experiments for English texts.

The last phase of our method is a computation of a sentiment value for a whole text. From the result of the sentiment extraction of a word in the second phase, a sentiment value of a text is obtained by three-dimensional vectors, in which the first, the second, and the third elements represent the number of words assigned by "Positive", "Negative", and "Neutral", respectively. We represent the three-dimensional vector by

$$V = (v_{pos}, v_{neg}, v_{neu}).$$

An algorithm to determine a final sentiment value for each vector is illustrated in Figure 3. For example, we consider the case of vector = (4, 2, 4). Since the maximum value of the elements is 4 and it is in the "Positive" and "Neutral" positions, the sentiment value of the word represented by the vector (4, 2, 4) is "Positive". By using this decision rule, the classifier can extract sentiment value as one of "Positive", "Negative", or "Neutral" from sentiment values of words in a text.

Our method have an advantage that even if a user does not understand a language subject to analysis, the user can extract the information by using our method.

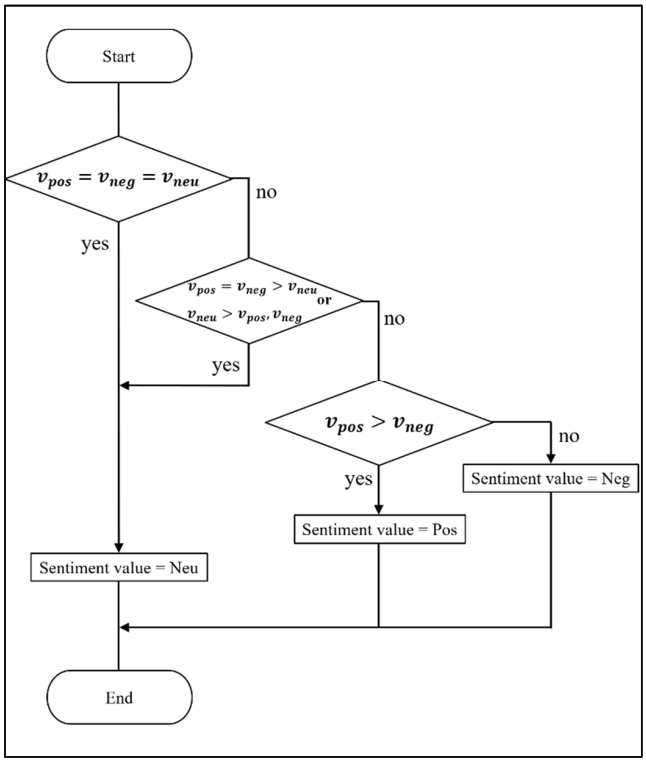

Fig. 3. Classification rule on the sentiment analysis method.

III. Sentiment Classification Experiment

In order to evaluate the performance of our sentiment classifier, we conduct a classification experiment of sentiment information from tweet texts in four languages, English, German, French, and Spanish. In this section, two previous sentiment classifiers "Valence Aware Dictionary for Sentiment Reasoning" ("VADER", for short) and "Google Cloud Platform" ("GCP", for short) are described to compare the performance with them.

A. Previous Sentiment Classifiers

Two sentiment classifiers are used in an experiment for making comparisons with our classifier. The one is called "VADER", which is a simple rule-based model for general sentiment analysis from "Natural Language Toolkit" ("NLTK", for short) [14]. "VADER" is well-known to be an especially suitable sentiment analysis method for microblog-like context (*e.g.* Tweets). Before using "VADER" in our experiment, we need to translate tweets into English texts. The performance of "VADER" can be considered to become lower if translation process is omitted in the preprocess, because "VADER" on "NLTK" focuses on sentiment classification for English texts. "VADER" returns a real value named "compound" in the range from –1 to +1 instead of sentiment categories such as "Positive" and "Negative". Typically, the value close to –1 is supposed to be negative sentiment, and the value close to +1 is supposed to be positive sentiment. Let $s_1, s_2, \ldots, s_m$ be texts in datasets, and $V(s_i)$ be the sentiment value of a sentence s_i calculated by "VADER". We define three sets "Negative", "Neutral", "Positive" as

$$Negative = \{s_i \mid -1.0 \leq V(s_i) \leq -0.05\},$$
$$Neutral = \{s_i \mid -0.05 < V(s_i) < 0.05\},$$
$$Positive = \{s_i \mid 0.05 \leq V(s_i) \leq 1.0\},$$

respectively. Each s_i is classified to negative, neutral, or positive, if it is in the sets "Negative", "Neutral", or "Positive", respectively.

The other is a "Natural Language API" on "GCP". A sentiment classifier based on machine learning is available through this API. Same as "VADER", "GCP" returns a real value named "score" in the range –1 to +1 instead of sentiment categories. Let $G(s_i)$ be the sentiment value of a sentence s_i calculated by "GCP". We define three sets "Negative", "Neutral", "Positive" as

$$Negative = \{s_i \mid -1.0 \leq G(s_i) \leq -0.5\},$$
$$Neutral = \{s_i \mid -0.5 < G(s_i) < 0.5\},$$
$$Positive = \{s_i \mid 0.5 \leq G(s_i) \leq 1.0\},$$

respectively. As with results of "VADER", Each s_i is classified to negative, neutral, or positive, if it is in the sets "Negative", "Neutral", or "Positive", respectively. These borderlines are determined by considering preliminary experimental results for English texts. In the preliminary experiments, we assumed several candidate values for borderlines of each classifier, and selected values which have the best score on the average evaluation metrics values.

B. Experimental Datasets

In the experiment to evaluate performance of our method, we use free open datasets of tweet texts [15, 16]. The datasets consist of texts with annotation of sentiment category as "Positive", "Negative", or "Neutral". Table □ shows the number of texts in each category.

TABLE I. DATASETS USED IN CLASSIFICATION EXPERIMENT

Language	Positive	Negative	Neutral	Total
English	375	338	197	910
German	124	144	86	354
French	293	337	207	837
Spanish	272	309	211	792

C. Functions for Evaluating Classifier's Performance

In the experiment, we adopt "Accuracy", "Precision", "Recall", and "F1 score" as evaluation metrics. "Accuracy" is the ratio of the number of texts to which correct sentiment categories are assigned, to the number of all texts. Let C be a classifier, and s be a sentence in the given set S of texts. Then, we denote $C(s) = "Positive"$, $C(s) = "Negative"$, or $C(s) = "Neutral"$ if the judgement regarding sentiment information of s by C is positive, negative, or neutral, respectively. Let C_t be a target classifier which outputs the correct sentiment value for each given sentence. For each classifier C, "Accuracy" of C is denoted by $Accuracy(C)$, and defined as follows:

$$Accuracy\ (C) = \frac{|\{s \in S \mid C(s) = C_t(s)\}|}{|S|}. \quad (1)$$

"Precision" is the ratio of the number of texts to which a correct sentiment category is assigned, to the number of texts classified as the category. "Recall" is the ratio of the number of texts to which a correct sentiment category is assigned, to the number of texts which have an annotation of the category. "F1 score" is harmonic mean of "Precision" and "Recall". Furthermore, for each classifier C and $Value$ in {Positive, Negative, Neutral}, "Precision", "Recall", and "F1 score" are also defined as follows:

$$Precision(C, Value) = \frac{|\{s \in S \mid C(s) = C_t(s) = Value\}|}{|\{s \in S \mid C(s) = Value\}|}, \quad (2)$$

$$Recall(C, Value) = \frac{|\{s \in S \mid C(s) = C_t(s) = Value\}|}{|\{s \in S \mid C_t(s) = Value\}|}, \quad (3)$$

$$F1\ score(C, Value) = \frac{2 \times Precision(C, Value) \times Recall(C, Value)}{Precision(C, Value) + Recall(C, Value)}. \quad (4)$$

These functions (1) to (4) are used to estimate performance of each classifiers explained in the following sections.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we report the results obtained from the experiment explained in Section Ⅲ. At first, we conduct an experiment to measure the performance of our classifier. For each language, the sets of texts are divided according to the labeled sentiment values. Table Ⅱ shows values of "F1 score" for each sets of texts by our classifier. From the results, we can see that our classifier has an advantage for extracting from "Negative" texts than "Positive" or "Neutral" texts, because the scores for "Negative" are larger than scores for "Positive" or

TABLE II. VALUES OF F1 SCORE BY OUR CLASSIFIER

Language	Positive	Negative	Neutral
English	0.53	0.61	0.23
French	0.56	0.66	0.18
German	0.54	0.56	0.49
Spanish	0.58	0.69	0.42

"Neutral". These results are not enough to show outstanding advantage of our classifier, because is not especially good at the classification for "Neutral" texts.

Figure 4, Figure 5, Figure 6, and Figure 7 show average values of "Accuracy", "Precision", "Recall", and "F1 score" by three classifiers for tweets in English, German, French, and Spanish. Note that the sets contain texts labeled by any sentiment value in these experiments. We can see that our classifier has the same capabilities at the other classifiers from several experimental results. The scores of "Accuracy" and "F1 score" of our classifier exceed on them of "GCP" in three languages, and "VADER" in one language. In particular, for the set of texts in Spanish, since there is a difference of 16 points on the scores of "Accuracy", this result represents superiority of our classifier.

Table shows the maximum values of differences on the scores of "Accuracy", "Precision", "Recall", and "F1 score" by each classifier. "VADER" has the largest value of differences (23% in accuracy, 18% in F1 score) and "GCP" has the smallest value of differences (5% in accuracy, 7% in F1 score). The difference values of scores of our classifier are 10% in "Accuracy" and 10% in "F1 score". These results indicate that classification performance of "VADER" fluctuates depending on type of languages and translation processes for a whole text. In contrast, the differences of performances of "GCP" and our classifier for four languages are not significant. Therefore, we can say that "GCP" and our classifier have a flexibility for the variety on the kinds of languages.

Results of sentiment classification experiment show that our classifier has the performance similar to some existing classifiers, and the stability with regard to the difference of languages. Totally, the performance of our classifier is enough to be applicability for some situations in which sentiment analysis is needed for texts in various languages. However, the classifier needs to be improved on the performance for practical use. Main reason of no significant performance is lack of consideration for contextual weight of sentimental words. For example, "hate" is strong sentimental degree than "like". In this research, the sentiment weights have not been considered in our classification. In order to improve the score of the performance of the classifier, we have to introduce syntactic parser to the classification process.

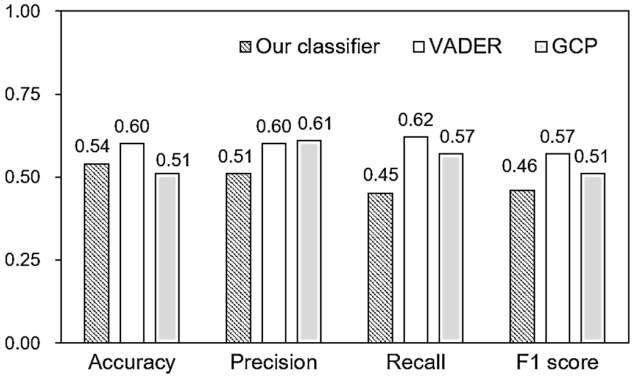

Fig. 4. Performance comparison by three classifiers about English tweets.

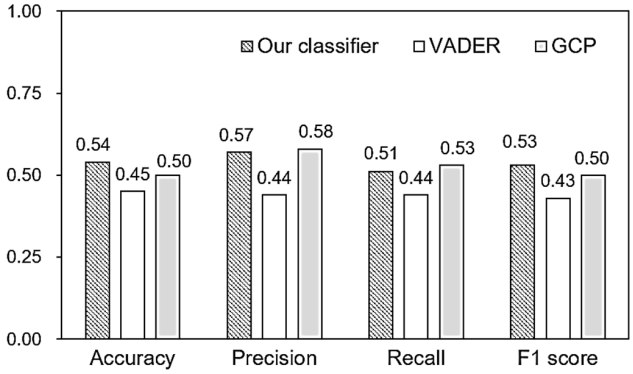

Fig. 5. Performance comparison by three classifiers about German tweets.

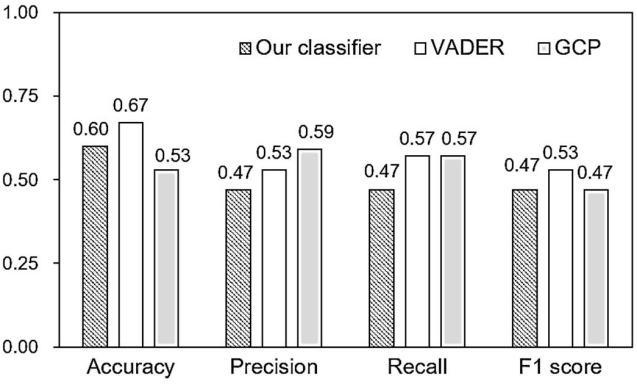

Fig. 6. Performance comparison by three classifiers about French tweets.

V. CONCLUSION AND FUTURE RESEARCH

We proposed a multilingual sentiment analysis method based on word to word translation using a sentiment dictionary, and evaluated the performance by comparing it with the previous classifiers "VADER" and "GCP". It was shown that our classifier has an advantage of low costs in the process of language translations, because it estimates sentiment values for each sentence by using only the word to word translation without translating process for whole a text. In other words, since our classifier can be applied even if the syntax of input texts is unknown, the method has a potential to be used for various unknown language. This advantage of our method

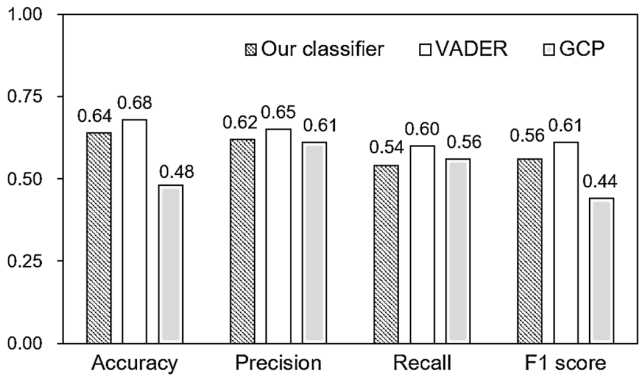

Fig. 7. Performance comparison by three classifiers about Spanish tweets.

TABLE III. MAXIMUM VALUES OF DIFFERENCES AMONG SCORES BY THREE CLASSIFIERS FOR FOUR LANGUAGES

Method	Accuracy	Precision	Recall	F1 score
Our classifier	10%	15%	9%	10%
VADER	23%	21%	18%	18%
GCP	5%	3%	4%	7%

derives that each user in various nationalities can analyze sentiment information for various unknown languages based on each native language. The evaluation experiment of the classifiers was conducted for languages, English, German, French, and Spanish. Any significant fluctuation of values on each evaluation standards cannot be observed in the difference of each language in the results of experiment. The results show that the classifier is acceptable in terms of applicability for multilingual. However, overall accuracy of the classifier is not sufficient for practical use. Syntactic analysis with multilingual versatility is essential for enhancement of the performance. For future work, we would like to investigate the effectiveness of the method for a different type of texts such as customer review and formal document, and other languages that have diverse structures from English including Chinese and Arabic.

REFERENCES

[1] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[2] A. Tumasjan, Timm. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178-185, 2010.

[3] S. L. Lo, E. Cambria, and R. Chiong, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," Artificial Intelligence Review 48(4), pp. 499-527, 2017.

[4] K. Dashtipour *et al.*, "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognitive Computation, vol. 8, pp. 757-771, 2016.

[5] M. Araújo, J. Reis, A. Pereira, and F. Benevenuto, "An evaluation of machine translation for multilingual sentence-level sentiment analysis," Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 1140-1145, 2016.

[6] A. Balahur and M. Turchi, "Multilingual Sentiment Analysis using Machine Translation?," Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 52-60, 2012.

[7] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data," arXiv preprint arXiv:1806.04511, 2018.

[8] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and translation for multilingual sentiment analysis," Computer Speech and Language, vol. 28, no. 1, January 2014.

[9] E. Demirtas and M. Pechenizkiy, "Cross-lingual Polarity Detection with Machine Translation," Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, 2013.

[10] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

[11] T. Nasukawa, D. Andrade, Y. Umino, Y. Muramatsu, and K. Yamamoto, "Finding translation pairs for cross-lingual text mining," Proceedings of the Annual Meeting of the Association for Natural Language Processing, vol. 15, pp. 108-111, 2009 (in Japanese).

[12] G. Edouard, B. Piotrm, G. Prakhar, J. Armand, and M. Tomas, "Learning Word Vectors for 157 Languages," Proceedings of the International Conference on Language Resources and Evalustion, 2018.

[13] H. Takamura, T. Inui, and M. Okumura, "Extracting Semantic Orientations of Words using Spin Model," Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 133-140, 2005.

[14] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Proceedings of the English International AAAI Conference on Weblogs and Social Media, 2014.

[15] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," CS224N Project Report, pp. 1-12, 2009.

[16] C. Malafosse, "Open datasets for sentiment analysis," 2019, https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis (accessed 2019-12-13).