International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2022, Vol. 6, No. 2, IJSKM664

Stability of a Multilingual Sentiment Analysis based on Word-to-Word Translations

Keita Fujihira [*], Noriko Horibe [†]

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 study, we design a multilingual sentiment analysis method based on word-to-word translation. The method classifies sentences by using a sentiment dictionary in a native language. The method consists of three phases: morphological analysis of a sentence, sentiment extraction of each word with the sentiment dictionary, and sentiment extraction of a sentence based on words sentiments. We conduct sentiment classification experiments for sentences in English, German, French, and Spanish. In the experiments, we compare our method with three previous methods by the evaluation metrics “Accuracy,” “Precision,” “Recall,” and “F1-score.” The experimental results show that our method has an advantage on the stability for variations of languages.

Keywords: machine translation, multilingual, sentiment analysis, sentiment dictionary.

1 Introduction

Social media have rapidly spread all over the world. In the media, web users can post their opinions about all products, events, and social problems. Sentences written on the services can influence decision making of web users. Researchers are interested in utilization of web texts for marketing, politics, health, and social studies [1][2][3][4]. Bollen *et al.* proposed a prediction method for the stock market using sentiment information from tweets on Twitter[1] [5]. In this study, sentiment information was extracted with “OpinionFinder” and “Google Profile of Mood States.” They built a Self-Organizing Fuzzy Neural Network model using the sentiment information as one of features. Experimental results showed that prediction accuracy of the model is higher than accuracy of a conventional model. Tumasjan *et al.* analyzed tweets regarding political parties and politicians [6]. Their purpose was to investigate usefulness of Twitter as a forum for political discussion and correlations between web user’s sentiments and results of election. Results of the investigation showed that the election results have the correlation with the number of political tweets and that sentiments on Twitter reflect political reputation for parties and politicians. Since

[*] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[†] Sojo University, Kumamoto, Japan

[1] https://twitter.com/

analysis using sentiments on web texts is useful for solutions to social problems, researchers in the field of Natural Language Processing are interested in sentiment analysis, and sentiment analysis systems have been actively developed [7].

Most systems support only one language because of technical difficulties and lack of resources such as corpora and lexicons [8][9][10]. Using the systems for sentences written in unsupported languages can decrease analysis performance. Web users write their opinions in various languages and a multilingual sentiment analysis method is indispensable to utilize sentiment information on the web. As with other academic fields, machine learning is a popular approach in sentiment analysis research [11][12][13][14]. However, these methods are not flexible because the methods require large training datasets in each language. It is difficult to support different domains and languages with small resources. In the recent research of multilingual sentiment analysis, many researchers tend to adopt an approach using machine translation. A feature of the approach is the reuse of resources and sentiment classifiers in a major language (generally, English) through machine translation [15][16][17]. Araújo *et al.* investigated whether English sentiment analysis methods are effective for sentences in other languages translated into English [18]. Can *et al.* initially built a Recurrent Neural Networks model using English reviews as a training dataset and they applied the model for sentences in different languages translated into English [19]. Balahur and Turchi proposed machine learning algorithms using sentences translated from English into French, German, and Spanish as a training dataset [20]. Several studies reported that one of the most serious causes of low sentiment classification performance is machine translation errors and noises. On the other hand, other studies reported that machine translation does not affect the performance because the errors are negligible. The difficulties of machine translation for each language are a bottleneck for sentiment analysis. It is difficult to maintain sentimental phrases, nuances, and expressions peculiar to a language on simple machine translation for a sentence. The difficulties become a factor that deteriorates the sentiment analysis performance. In this way, although various researchers in the field of multilingual sentiment analysis have been using machine translation as one of their analyzing processes, the translation performance depends on the kinds of languages, and there are different opinions about evaluation of machine translation.

In order to maintain sentimental phrases and reduce accuracy variability due to language difference, we consider that word-to-word translation can be a valid method instead of translation for a whole sentence. This paper proposes a multilingual sentiment analysis method based on word-to-word translation. The method has a word translation process instead of translation for a whole sentence and uses only one sentiment dictionary in a native language. The method initially divides a sentence into words by morphological analysis. The divided words are mapped into sets of similar words of them through word embeddings, and the similar words are translated into the native language. We define sentiment values of translated words by using the sentiment dictionary, and a sentence is classified based on the sentiment values. The method is applicable for many languages because contextual information is not used. The method also reduces the risk of losing sentimental phrases and performance variability due to language differences. Another feature of the method is that each word in a sentence is represented as multiple words in the native language with similar meanings. That is, in sentiment analysis results, users can easily understand which words cause sentiment even if they do not get used to foreign languages. To evaluate the method, we conduct sentiment classification experiments for sentences in English, French, German, and Spanish. In the experiments, we compared our method with three previous methods and our method has stability for multiple languages. Contributions of this paper are: (1) proposal of a multilingual sentiment analysis method based on word-to-word translation, (2) reducing the variability of classification accuracy associated with language differences and mistranslation. The method can be expected to realize the analysis of sentiment information on web texts without specifying languages and to expand a scope of marketing or research into regions that are difficult to enter because of language barriers.

The remainder of this paper is organized as follows. In Section 2, we describe our multilingual sentiment analysis method. In Section 3, we describe settings of sentiment classification experiments. In Section 4, we describe experimental results and discussion. Finally, we give conclusion in Section 5.

2 Proposal of Multilingual Sentiment Analysis Method

We propose a multilingual sentiment analysis method based on word-to-word translation called SAWW. In this section, we explain details of SAWW.

2.1 The Flow of SAWW

Figure 1 shows an outline of SAWW. SAWW consists of three phases: morphological analysis of a sentence, sentiment extraction of each word with a sentiment dictionary, and sentiment extraction of the sentence based on each sentiment from the words. The first phase morphological analysis is a process that divides a sentence into the smallest meaningful unit called morpheme in each language. We use "TreeTagger" as a morphological analysis tool [21]. "TreeTagger" deals with 25 languages including English and German. Furthermore, "TreeTagger" can be extended to be applicable for other languages if a lexicon and a tagged training corpus are available. After dividing the sentence, we remove some parts of speech such as articles and interrogatives and use only adjectives, adverbs, common nouns, and verbs. Sentiment classification performance of SAWW becomes worse if useless words are not removed sufficiently.

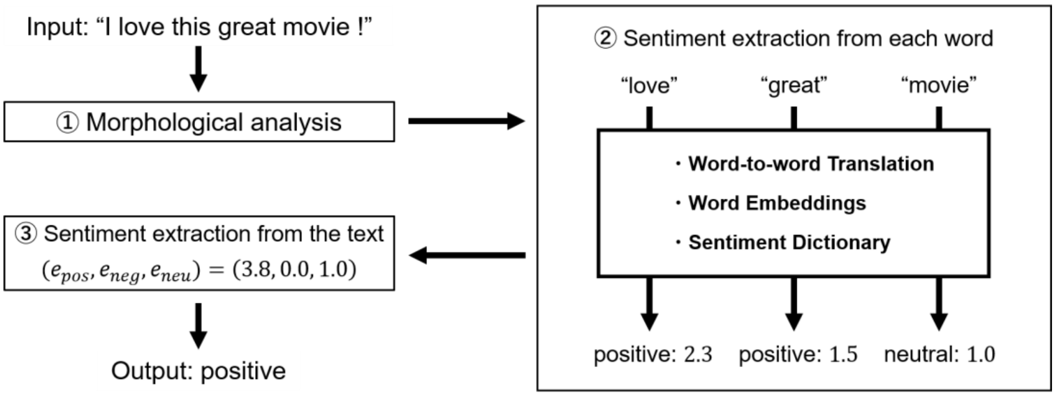

Input: "I love this great movie !"
① Morphological analysis
③ Sentiment extraction from the text
$(e_{pos}, e_{neg}, e_{neu}) = (3.8, 0.0, 1.0)$
Output: positive
② Sentiment extraction from each word
"love"
"great"
"movie"
· Word-to-word Translation
· Word Embeddings
· Sentiment Dictionary
positive: 2.3
positive: 1.5
neutral: 1.0

Figure 1: Outline of a multilingual sentiment analysis method

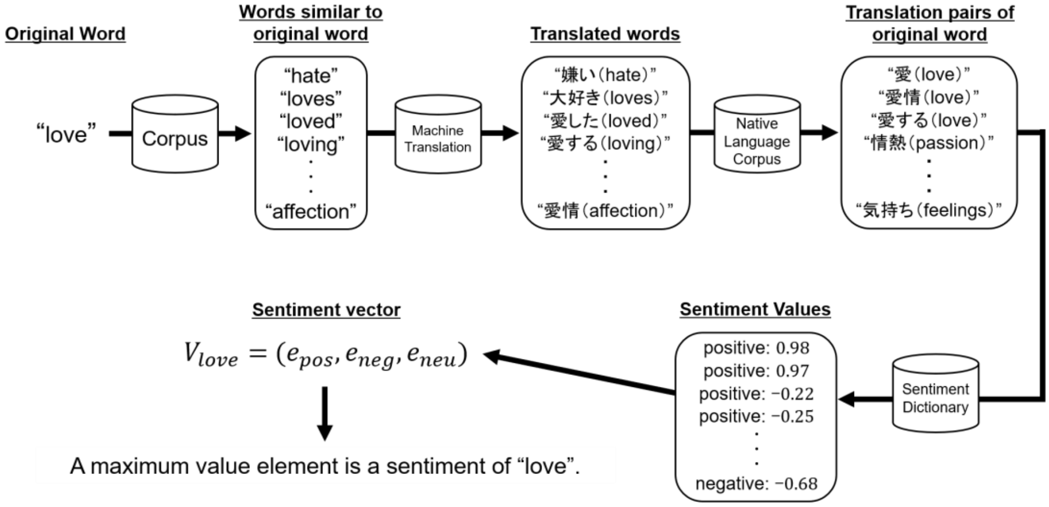


Figure 2: Procedure to calculate a sentiment for each word

The second phase is sentiment extraction of each word. Figure 2 shows the procedure by exemplifying the word "love" obtained from the first phase outputs. Assuming that a user's native language is Japanese. In SAWW, translation pairs are obtained based on a process proposed by Nasukawa *et al.* [22]. At first, a word embedding model detects N words similar to the original word "love", where N is set as 5, 10, or 15 in this paper. We use an open-source library "fastText" and word embedding models trained on corpora such as "Common Crawl" and "Wikipedia" to calculate word similarity [23][24]. "fastText" is used for learning text classification and word embedding. Secondly, the similar words are translated into the native language using a machine translation system, where "Google Translate[2] " is used in this paper. Thirdly, N words are extracted as translation pairs of "love" from the set of words obtained by translation. The word embedding model calculates an average vector of translated words and extracts translation pairs as the high similarity words to the average vector. Fourthly, sentiment values of each pair are calculated by a sentiment dictionary. A sentiment dictionary used in SAWW is designed by Takamura *et al.* [25]. In the dictionary, sentiment values of each word are real numbers in the range −1 to +1 automatically assigned using a lexical network. The word which has a value close to −1 represents negative sentiment, and the word which has a value close to +1 represents positive sentiment. Let $w_1, w_2, \ldots, w_n$ be words in sentiment dictionary D, and $F(w_i)$ be a sentiment value of a word w_i. We define three sets *Positive*, *Neutral*, and *Negative* as

$$Positive = \{w_i \in D \mid -0.37 \leq F(w_i) \leq 1.0\}, \tag{1}$$

$$Neutral = \{w_i \in D \mid -0.4 < F(w_i) < -0.37\}, \tag{2}$$

$$Negative = \{w_i \in D \mid -1.0 \leq F(w_i) \leq -0.4\}, \tag{3}$$

respectively. These thresholds are defined considering results of our preliminary experiment for English sentences, and a sentiment of a word not included in the dictionary is set to neutral sentiment. A sentiment value of "love" is the maximum value in a three-dimensional vector, in which the first, the second, and the third elements represent a value of positive, negative, and neutral, respectively. We represent the three-dimensional vector of "love" as an example by

[2] https://translate.google.com/

$$V_{love} = (e_{pos}, e_{neg}, e_{neu}), \tag{4}$$

and consider the case of vector = (4.0, 1.3, 0.8). Since the maximum value in the vector is 4.0 and it is in the positive position, the sentiment value of the original word “love” is positive: 4.0.

The last phase is a sentiment extraction of a sentence based on each sentiment from the words. From the second phase results, each word has a sentiment value and represents positive, negative, or neutral. A sentiment of sentence is also judged from a three-dimensional vector, in which the first, the second, and the third elements represent the sum of positive, negative, and neutral, respectively. In the vector, a sentiment represented by an element with the maximum value be a sentiment of sentence.

We still need to discuss the number of similar words and translation pairs for an original word and variations of calculation for sentiment values in the vector. We will describe three variations of calculation in Section 2.2 and show experimental results about the variations in Section 4.1. In SAWW, each word in a sentence is represented as multiple words in a native language with similar meanings. This is one of the advantages that even if a user does not understand a language subject to analysis, the user can easily understand which words cause sentiment.

2.2 Variations of Calculation for Sentiment

As explained in Section 2.1, SAWW has some variations to calculate values in the sentiment vector of words. Figure 3 shows three variations that we implemented. In Figure 3, sentiment values are assumed 1.0, 0.8, 0.5, −0.6, −0.38. The first is a way called SAWW-ADD that calculates the total of an absolute sentiment value categorized as *Positive*, *Negative*, and *Neutral*. In the first way, the sentiment vector is (2.3, 0.6, 0.38), and the sentiment value of a word is positive: 2.3. The second is a way called SAWW-COUNT that counts the number of words categorized as

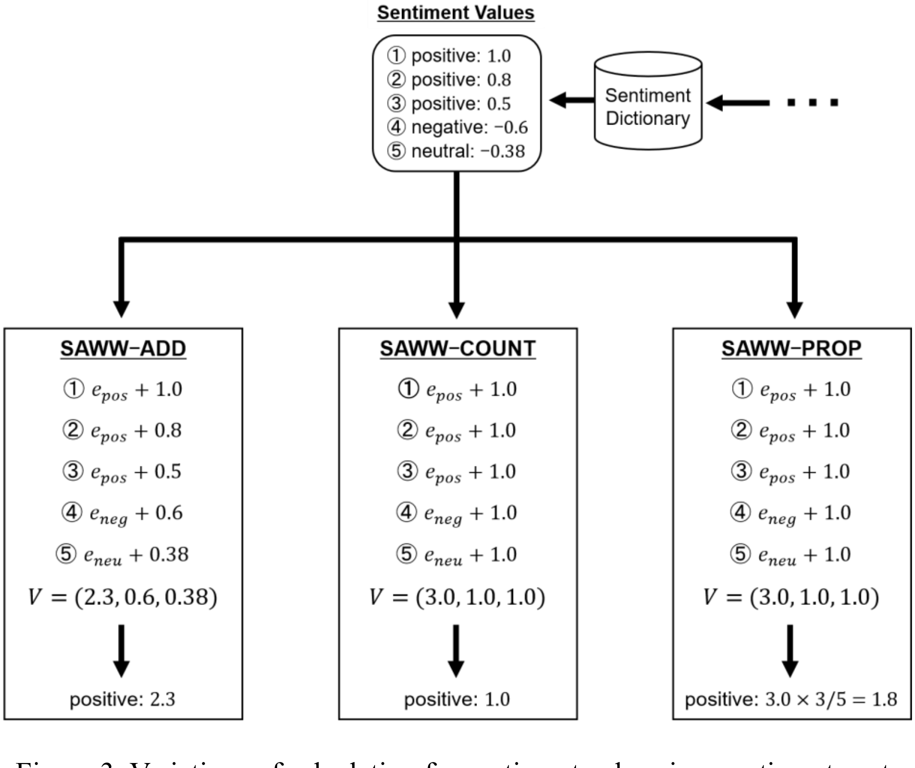


Figure 3: Variations of calculation for sentiment values in a sentiment vector

each sentiment. In the second way, the sentiment vector is (3.0, 1.0, 1.0), and the sentiment value of a word is positive: 1.0. The third is a way called SAWW-PROP that multiplies the number of words categorized as each sentiment and the proportion. In the third way, the sentiment vector is (3.0, 1.0, 1.0), and the sentiment value of a word is positive: 1.8.

3 Experimental Settings

We conduct sentiment classification experiments for web texts in four languages, English, German, French, and Spanish. In this section, we describe details of test data and evaluation metrics, and previous sentiment analysis models to make a comparison with SAWW.

3.1 Experimental Datasets

The kind of test data used in experiments is a tweet on Twitter and we use free open datasets [26][27]. Datasets consist of sentences with sentiment annotation as positive, negative, or neutral. Table 1 shows the number of sentences in each annotation. Three-class experiments are described in Section 4.1 and 4.2. In addition to three-class experiments, we also conduct two-class experiments that target only positive and negative sentiment. Table 2 shows the number of sentences in each annotation for two-class experiments. Two-class experiments are described in Section 4.3.

3.2 Evaluation Metrics

We adopt "Accuracy," "Precision," "Recall," and "F1-score" as evaluation metrics in classification experiments. "Accuracy" is the ratio of the number of sentences to which correct sentiments are assigned to the number of all sentences. Let C be a classifier, and s be a sentence in the given set S of sentences. Then, we denote $C(s) = Pos$, $C(s) = Neg$, or $C(s) = Neu$ if the prediction of s by C is positive, negative, or neutral, respectively. Let C_t be a target classifier that outputs the

Table 1: Test datasets for three-class experiments

Language	Positive	Negative	Neutral	Total
English	2417	2914	1293	6624
German	3804	3690	2372	9866
French	3552	3899	2841	10292
Spanish	3216	3974	2233	9423

Table 2: Test datasets for two-class experiments

Language	Positive	Negative	Total
English	1208	1457	2665
German	1902	1845	3747
French	1776	1949	3725
Spanish	1608	1987	3595

correct sentiment for each given sentence. For each classifier C, “Accuracy” of C is denoted by $Accuracy(C)$, and defined as follows:

$$Accuracy(C) = \frac{|\{s \in S \mid C(s) = C_t(s)\}|}{|S|}. \tag{5}$$

“Precision” is the ratio of the number of sentences to which correct sentiments are assigned to the number of sentences classified as the sentiment. “Recall” is the ratio of the number of sentences to which correct sentiments are assigned to the number of sentences which have the sentiment annotation. “F1-score” is a harmonic mean of “Precision” and “Recall.” For each classifier C and $Value$ in $\{Pos, Neg, Neu\}$, “Precision,” “Recall,” and “F1-score” are also defined as follows:

$$Precision(C, Value) = \frac{|\{s \in S \mid C(s) = C_t(s) = Value\}|}{|\{s \in S \mid C(s) = Value\}|}, \tag{6}$$

$$Recall(C, Value) = \frac{|\{s \in S \mid C(s) = C_t(s) = Value\}|}{|\{s \in S \mid C_t(s) = Value\}|}, \tag{7}$$

$$F1\text{-}score(C, Value) = \frac{2 \times Precision(C, Value) \times Recall(C, Value)}{Precision(C, Value) + Recall(C, Value)}. \tag{8}$$

3.3 Previous Sentiment Classification Models

We compare SAWW with three sentiment classification models in evaluation experiments. The first model is “Valence Aware Dictionary for Sentiment Reasoning (VADER),” which is a simple rule-based model for general sentiment analysis [28]. “VADER” is well-known to be an especially suitable model for microblog-like context. We use “VADER” through “Natural Language Toolkit (NLTK).” Note that we translate test data into English before using “VADER” because “VADER” on “NLTK” specialize in English sentences. The classification performance of “VADER” can be considered to become lower if a translation process is omitted. The second model is a machine learning model available through “Natural Language API” on “Google Cloud Platform (GCP)[3].” The third model is based on Recursive Neural Network that builds on top of grammatical structures and available through “CoreNLP” [29]. “CoreNLP” model is suitable for longer phrases and maintains the order of words and syntactic information.

We do not execute preprocessing such as converting from an uppercase to a lowercase because it is unclear how some comparison models process them, and our purpose is to verify the performance of each model under fair conditions.

4 Experimental Results and Discussion

In this section, we report preliminary experiment results about SAWW followed by comparison results of classification performance for sentiment analysis methods.

4.1 Preliminary Experiments

The purpose of preliminary experiments is to verify (1) differences between SAWW-ADD, SAWW-COUNT, and SAWW-PROP and (2) differences between performance of SAWW, where the number of similar words N is set as 5, 10, or 15.

[3] https://cloud.google.com/natural-language/

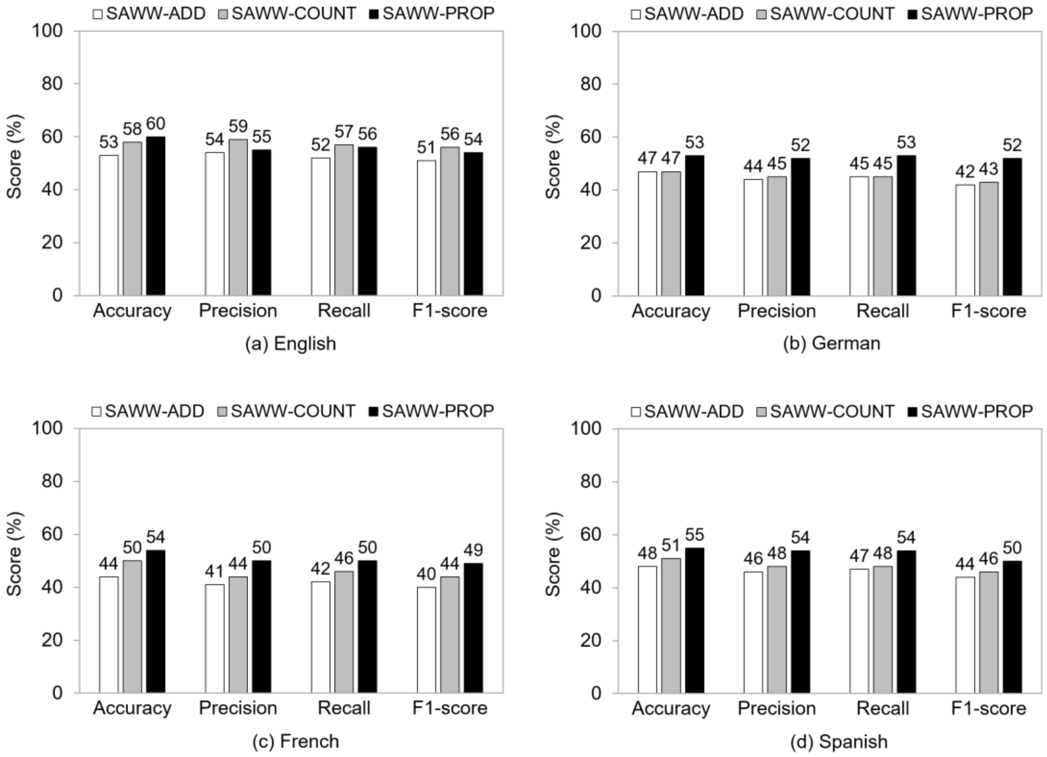


Figure 4: Average scores of each evaluation metric on three-class experiments for (a) English, (b) German, (c) French, and (d) Spanish by SAWW-ADD, SAWW-COUNT, and SAWW-PROP

First, we describe experimental results about differences of each method. Figure 4 shows average scores of each evaluation metric by SAWW-ADD, SAWW-COUNT, and SAWW-PROP ($N = 10$). SAWW-PROP has the best "F1-score" in three languages and the best "Accuracy" in all languages. SAWW-COUNT achieves the best "F1-score" in one language. SAWW-ADD has the worst scores of all metrics in all languages. The poor performance of SAWW-ADD would cause by a large difference between each value in the sentiment vector of words. Translation pairs are classified into each sentiment category using the sentiment dictionary and Equation 1 to 3. In SAWW-ADD, the sum of neutral position value is prone to smaller than positive and negative position values. Therefore, misclassification of neutral sentences has increased, and it leads to lower performance than other calculation way. These results indicate that classification performance can be fluctuated by calculation way for sentiment value and that SAWW-PROP is the most effective in our methods.

Figure 5 shows average scores of each evaluation metric by SAWW-PROP, where the number of similar words N is set as 5, 10, or 15. The best "Accuracy" scores are achieved by $N = 10$ and 15 in three and one languages, respectively. The method with $N = 10$ achieves the best "F1-score" in all languages. The method with $N = 5$ has the worst scores of all metrics in all languages. In the case of $N = 5$, mistranslation and misclassification of words would have a considerable influence for calculation results because the total number of words is small. Conversely, in the case of $N = 15$, the performance is prone to low because the relation between translation pairs and the original word can be weak. Examples such words include "love" and "kekkon ganbo" in Japanese (the meaning of yearning for marriage), "hate" and "heiki" in Japanese (the meaning

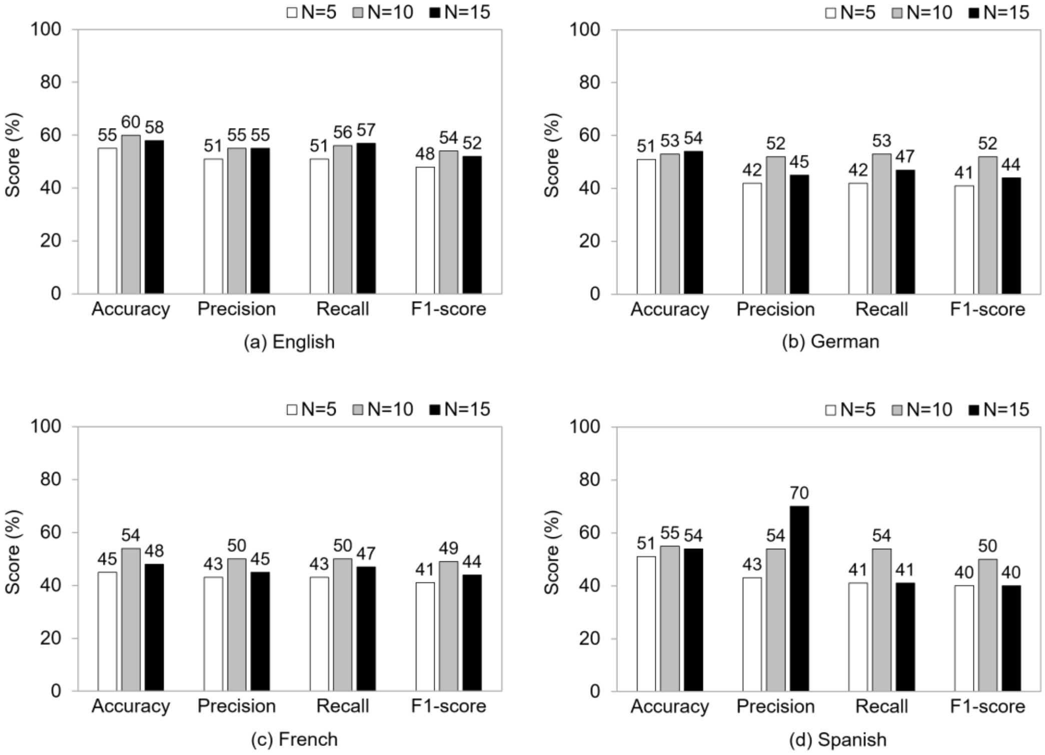


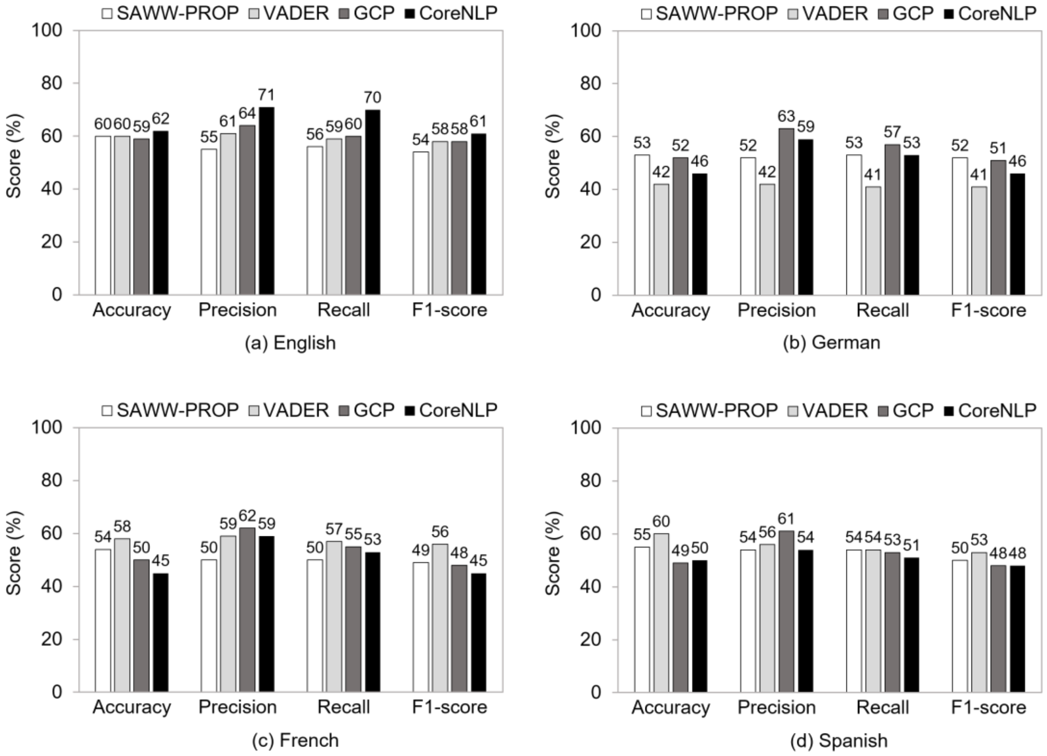
Figure 5: Average scores of each evaluation metric on three-class experiments for (a) English, (b) German, (c) French, and (d) Spanish by SAWW-PROP with N = 5, 10, and 15

of no problem). These results indicate that the suitable number of similar words and translation pairs is approximately 10.

4.2 Three-class Experiments

We compared SAWW-PROP ($N = 10$) with "VADER," "GCP," and "CoreNLP" for three classes. Figure 6 shows average scores of evaluation metrics by each method. SAWW-PROP achieves the best "Accuracy" and "F1-score" (53% and 52%, respectively) in German. "VADER" gives the best "Accuracy" and "F1-score" (58% and 56%, respectively) in French, and the best "Accuracy" and "F1-score" (60% and 53%, respectively) in Spanish. "CoreNLP" achieves the best "Accuracy" and "F1-score" (62% and 61%, respectively) in English. We can see that SAWW has the same capabilities at the other sentiment analysis methods from these results.

Table 3 shows the differences between the maximum and the minimum score of each method in the experimental results. The differences are performance variability due to language differences. That is a smaller value means a better versatility for multiple languages. SAWW-PROP achieves "Accuracy" difference of 7% (from 53% in German to 60% in English) and "F1-score" difference of 5% (from 49% in French to 54% in English). In other methods, the only "GCP" gives the "Accuracy" and "F1-score" difference of less than 10%. These results indicate that "VADER" does not have sufficient versatility for multiple languages although it is acceptable for specific languages and that SAWW-PROP has the equivalent performance to other sentiment methods and stability for multiple languages.



(a) English

(b) German

(c) French

(d) Spanish

Figure 6: Average scores of each evaluation metric on three-class for (a) English, (b) German, (c) French, and (d) Spanish by SAWW-PROP ($N = 10$), VADER, GCP, and CoreNLP

Table 3: The difference between the maximum score and the minimum score of each method on three-class

Method	Accuracy	Precision	Recall	F1-score
SAWW-PROP	**7**	7	**6**	**5**
VADER	18	19	18	17
GCP	10	**3**	7	10
CoreNLP	17	17	19	16

4.3 Two-class Experiments

We also conducted two-class experiments that target only positive and negative sentiment. Figure 7 shows the average scores of evaluation metrics by each method. SAWW-PROP achieves “Accuracy” of over 60% in all languages and “F1-score” of 60% in three languages. “VADER” gives the best “Accuracy” of 77% in French and “F1-score” in three languages (72% in English, 76% in French, and 65% in Spanish). “GCP” gives the best “Accuracy” and “F1-score” (74% and 72%, respectively) in English. “CoreNLP” achieves the best “Accuracy” in two languages (75% in German and 66% in Spanish) and “F1-score” in three languages (72% in English, 74% in German, and 65% in Spanish). We can see that every method increases the evaluation scores due to changing the number of classes from three to two.

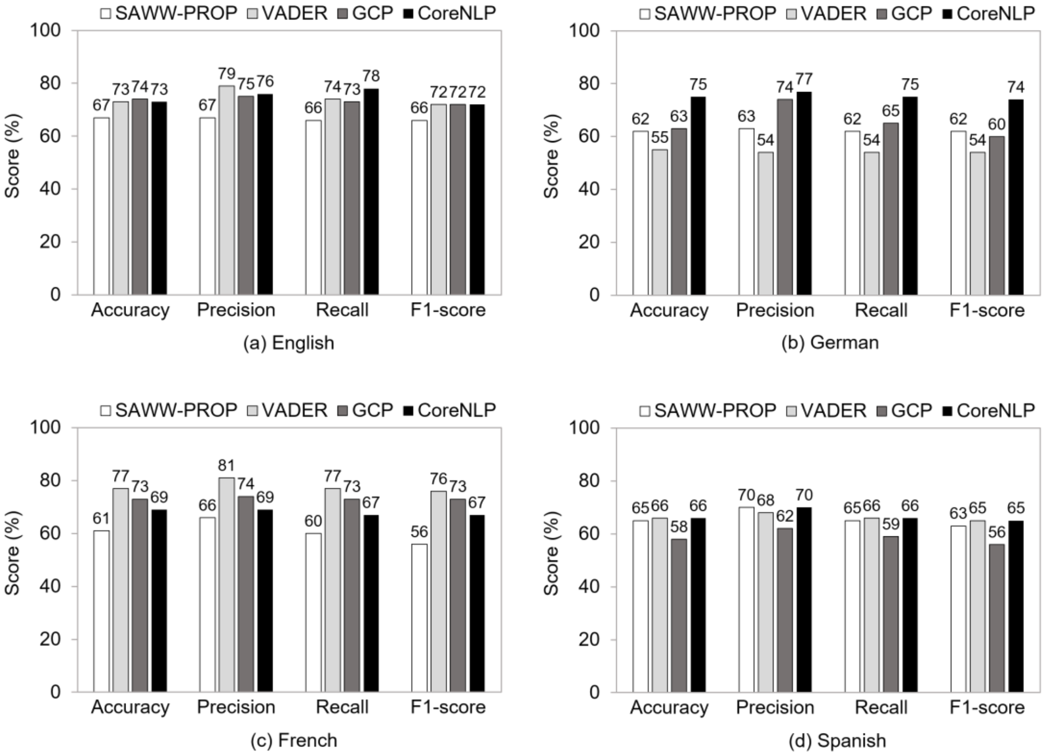


Figure 7: Average scores of each evaluation metric on two class classification in (a) English, (b) German, (c) French, and (d) Spanish by SAWW-PROP ($N = 10$), VADER, GCP, and CoreNLP

Table 4: The difference between the maximum score and the minimum score of each method on two classes

Method	Accuracy	Precision	Recall	F1-score
SAWW-PROP	**6**	**7**	**6**	10
VADER	22	27	23	22
GCP	16	13	14	17
CoreNLP	9	8	12	**9**

Table 4 shows the difference between the maximum and the minimum score of each method on the experimental results. As with three-class experiments, SAWW-PROP gives differences of less than 10% for all metrics. The best "Accuracy" difference of 6% (from 61% in French to 67% in English) and "F1-score" difference of 10% (from 56% in French to 66% in English). "CoreNLP" also gives differences of less than 10% for three metrics. These results suggest that SAWW-PROP and "CoreNLP" have versatility for multiple languages. Moreover, "CoreNLP" gives practical classification performance on two classes.

4.4 Discussion

SAWW has two superior features in addition to stability for multiple languages. The one is flexibility with thresholds of sentiment values in the sentiment dictionary. For example, when users want to collect negative opinions as many as possible, by changing the range of Equation (1) to (3), they will make it easier to collect sentences containing negative opinions. The other is a comprehensibility of the analysis results for users who analyze sentences in non-native languages. In the process of SAWW, each word in a non-native language sentence is represented in multiple words in the native language with similar meanings. That is, users can understand the meaning of each word in sentences to some extent.

Although SAWW has these advantages, "Accuracy" and "F1-score" of SAWW are approximately 60% and SAWW still has room to improve toward practical application. For example, translation pairs are extracted using word embedding models and extracted pairs sometimes contained spelling distortions or typographical errors. The sentiment dictionary does not usually support many spelling distortions or typographical errors and a sentiment of such word is set to neutral as described in Section 2.1. The errors cause misjudges and decrement of classification accuracy. For this reason, we plan to reduce spelling distortions and typographical errors during extraction of translation pairs by word embedding models or increase the number of words in the sentiment dictionary as one of future works. Another future work is development of a syntactic analysis tool with multilingual versatility. Although the multilingual syntactic analysis is a challenging topic in the field of natural language processing and is not introduced into SAWW, we expect that it improves the classification performance of SAWW.

5 Conclusion

We proposed a multilingual sentiment analysis method based on word-to-word translation. Our method was shown to have the advantage of low costs in language translation because the method classifies sentences by using only a process of not translations for a whole sentence but correspondences on word-to-word. In other words, since the method can be applied even if the syntax of an input sentence is unknown, the method has the potential to be used for various unknown languages. This advantage derives that each user in various nationalities can analyze sentiment information for various unknown languages based on each user's native language. We conducted sentiment classification experiments for sentences in English, German, French, and Spanish. For all languages used in the experiments, score differences of the evaluation metrics were shown to be within 10%. The results show that the method is acceptable in terms of applicability for multilingual. The overall accuracy of the method still has room to improve for practical application. Syntactic analysis with multilingual versatility can help for performance enhancement. For future work, we would like to investigate the effectiveness of the method for different types of sentences such as customer review or formal document and other languages that have diverse structures including Chinese and Arabic.

References

[1] T. Araujo, P. Neijens, and R. Vliegenthart, "Getting the Word Out on Twitter: The Role of Influentials, Information Brokers and Strong Ties in Building Word-of-mouth for Brands," International Journal of Advertising, Vol. 36, No. 3, 2017, pp. 496-513.

[2] N. Anstead and B. O'Loughlin, "Social Media Analysis and Public Opinion: The 2010 UK General Election," Journal of Computer-Mediated Communication, Vol. 20, No. 2, 2015, pp. 204-220.

[3] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study," Journal of Medical Internet Research, Vol. 22, No. 4, 2020, e19016.

[4] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter Events," Journal of the American Society for Information Science and Technology, Vol. 6, No. 2, 2011, pp. 406-418.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," Journal of Computational Science, Vol. 2, No. 1, 2011, pp.1-8.

[6] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proc. International AAAI Conference on Weblogs and Social Media, 2010, pp.178-185.

[7] K. Ravi and V. Ravi, "A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications," Knowledge-Based System, Vol. 89, 2015, pp. 14-46.

[8] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual Sentiment Analysis: From Formal to Informal and Scarce Resources Languages," Artificial Intelligence Review, Vol. 48, No. 4, 2017, pp.499-527.

[9] M. Kaity and V. Balakrishnan, "Sentiment Lexicons and non-English Languages: A Survey," Knowledge and Information Systems, Vol. 62, 2020, pp. 4445-4480.

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

[11] E. Boiy and M. F. Moens, "A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts," Information Retrieval, Vol. 12, No. 5, 2009, pp. 526-558.

[12] A. Cheng and O. Zhulyn, "A System for Multilingual Sentiment Learning on Large Data Sets," Proc. International Conference on Computational Linguistics, 2012, pp. 577-592.

[13] M. Attia, Y. Samih, A. Elkahky, and L. Kallmeyer, "Multilingual Multi-class Sentiment Classification Using Convolutional Neural Networks," Proc. International Conference on Language Resources and Evaluation, 2018, pp. 635-640.

[14] T. Kincl, M. Novák, and J. Přibil, "Improving Sentiment Analysis Performance on Morphological Rich Languages: Language and Domain Independent Approach," Computer Speech and Language, Vol. 56, 2019, pp. 36-51.

[15] X. Wan, "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis," Proc. Conference on Empirical Methods in Natural Language Processing, 2008, pp. 553-561.

[16] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-linguistic Sentiment Analysis: From English to Spanish," Proc. International Conference on Recent Advances in Natural Language Processing, 2009, pp. 50-54.

[17] K. Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis," Proc. IEEE International Conference on Data Engineering Workshop, 2008, pp. 507-512.

[18] M. Araújo, A. Pereira, and F. Benevenuto, "A Comparative Study of Machine Translation for Multilingual Sentence-level Sentiment Analysis," Information Sciences, Vol. 512, 2020, pp. 1078-1102.

[19] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data," Proc. International Workshop on Learning from Limited/Noisy Data for IR, 2018, arXiv:1806.04511.

[20] A. Balahur and M. Turchi, "Comparative Experiments Using Supervised Learning and Translation for Multilingual Sentiment Analysis," Computer Speech and Language, Vol. 28, No. 1, 2014, pp. 56-75.

[21] H. Schmid, "Probabilistic Part-of-speech Tagging Using Decision Trees," Proc. International Conference on New Methods in Language Processing, 1994.

[22] T. Nasukawa, D. Andrade, Y. Umino, Y. Muramatsu, and K. Yamamoto, "Finding Translation Pairs for Cross-lingual Text Mining," Proc. Annual Meeting of the Association for Natural Language Processing, 2009, pp. 108-111. (In Japanese)

[23] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics, Vol. 5, 2017, pp. 135-146.

[24] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning Word Vectors for 157 Languages," Proc. International Conference on Language Resources and Evaluation, 2018, pp. 3483-3487.

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

[26] I. Mozetič, M. Grčar, J. Smailović, "Twitter Sentiment for 15 European Languages," Slovenian language resource repository CLARNI.SI, http://hdl.handle.net/11356/1054. (Last access: March 2021)

[27] T. Blard, "French Sentiment Analysis with BERT," GitHub repository, https://github.com/TheophileBlard/french-sentiment-analysis-with-bert. (Last access: March 2021)

[28] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Proc. Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp. 216-225.

[29] R. Socher *et. al.*, "Recursive Deep Models for Semantic Compositionality Over A Sentiment Treebank," Proc. Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631-1642.