3) Support Vector Machine: SVM is one of the classifiers that utilizes the vector space technique in which the features are represented in 2-D via X and Y axes [19]. The values of the features that will be used for representation in the 2-D space are considered to be the occurrence of each term in accordance to the dataset. Once the features are depicted in the vector space, a hyperplane, which is a margin that separates the data into two classes, will be implemented. Accurate acquisition of the hyperplane will lead to accurate classification results. The hyperplane can be calculated based on Equation (3):

$$f(\vec{x}) = \begin{cases} +1: & (\vec{x} \times \vec{w}) + b > 0 \\ -1: & Otherwise \end{cases} \qquad (3)$$

The SVM model adjusts to the most accurate hyperplane that has the greatest margin. One example of this is the chemical and non-chemical data instances that are divided by a hyperplane in which the shortest path is between the nearest chemical instance and nearest non-chemical instance [20].

D. Feature Selection

This phase applies the feature selection, whereby the most appropriate features will be identified. Hence, the Wrapper Subset Selection (WSS) approach was adopted. This approach is based on a wrapping mechanism in which a search will be performed to find the most robust subset within the featured space [21]. WSS employs a classification method to assess the effectiveness of each feature. Therefore, this study will integrate both SVM and NB with WSS in order to measure the accuracy of each combination of features.

To describe the problem of dimensionality in the chemical compound extraction task, a chemical data D is considered, which consists of sequences $D = \{t_1, t_2, t_3, \dots, t_m\}$, where every token denotes a term within the data. The term is either a regular term or a chemical compound. Evidently, for every token there are different features that correlate with it $f = \{f_1, f_2, f_3, \dots, f_n\}$. In this manner, every feature should be assessed separately to obtain the best combination. However, evaluating each feature separately may lead to numerous possibilities. The single evaluation required to specify the number of combination of features are bi-combination (e.g. the combination of f_1 and f_2 or the combination of f_1 andf_3), tri-combination (e.g. the combination of f_1, f_2 andf_3) or even any number of possible combinations ranging from 1 to n, where n represents the number of features. In this manner, the problem can be formulated based on Equation (4):

$$\sum_{n=13} \frac{n!}{(n-r)! \times r!} \qquad (4)$$

Where *n* is the number of features and *r* is the number of combinations. The number of utilized features in the detailed representation is 13, which seems to be small. However, examining every possibility of each possible combination would be tedious. Table 4 shows the number of possibilities for each combination.

TABLE IV
NUMBER OF POSSIBILITIES FOR EACH COMBINATION

Number of combinations	Number of possibilities
r = 1	13
r = 2	78
r = 3	286
r = 4	715
r = 5	1287
r = 6	1716
r = 7	1716
r = 8	1287
r = 9	715
r = 10	286
r = 11	78
r = 12	13
r = 13	1
Total	8191

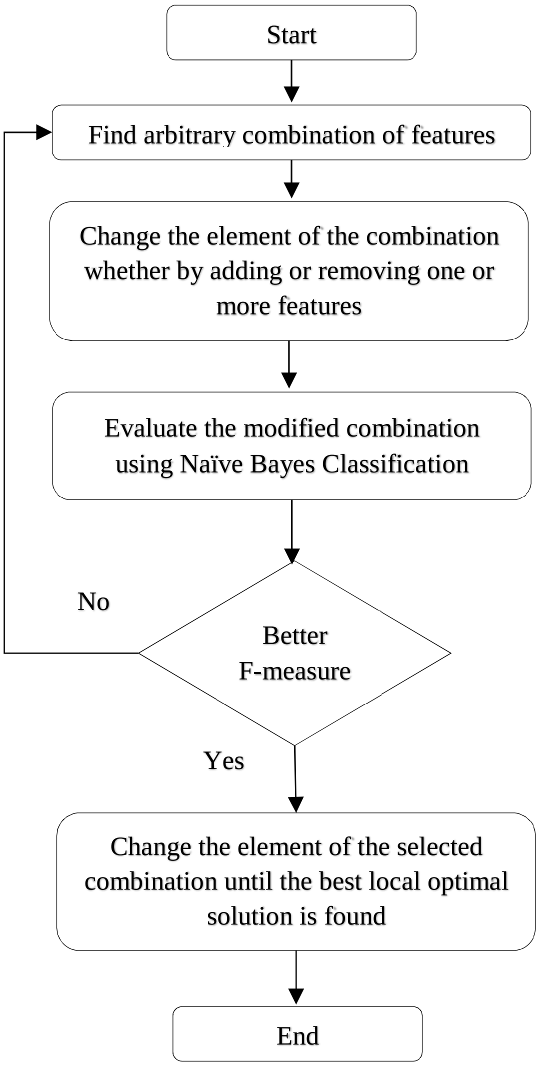
Start
Find arbitrary combination of features
Change the element of the combination whether by adding or removing one or more features
Evaluate the modified combination using Naïve Bayes Classification
No
Better F-measure
Yes
Change the element of the selected combination until the best local optimal solution is found
End

Fig. 3. HC algorithm flowchart

As shown in Table 4, the total number of possibilities for individual combinations is 8191. Examining each possibility separately would prove tedious, especially when the computation for an individual run is time-consuming. In the same manner, examining the possibilities for the N-gram, which contains 465 features, would increase the problem of dimensionality. Therefore, it is necessary to apply feature reduction.

It is important to note that the search algorithm used in our study is Hill Climbing. Hill Climbing (HC) is a heuristic search algorithm that seeks to find nearly optimized solutions [22]. HC is a local search algorithm that has been used on hard optimization problems. A key characteristic of the local search algorithm is that it can be applied on problems that require finding a solution with the maximized criterion among a number of candidate solutions [23]. Local search algorithms work by moving from one solution to another in the search space through making some local changes until the optimal solution is found.

Similarly, HC begins with an arbitrary solution, then tries to figure out a better solution by incrementally changing the elements of the solution [24]. The flowchart for the HC algorithm is depicted in Figure 3.

III. Results and Discussion

As with any machine-learning task, the evaluation will be conducted using precision, recall, and f-measure. Also, the evaluation will be based on two paradigms; the detailed-attribute using NB and the N-gram using SVM. The following sub-sections show the results obtained in this study.

1) Results of Detailed-attribute using NB: As mentioned earlier, this section shows the results of applying the NB classifier with the detailed-attribute paradigm. The features are evaluated separately and with the total combination of features. Table 5 shows the results.

TABLE V
Results of NB with detailed-attribute

Feature	Precision	Recall	F-measure
Length	0.4755	0.5	0.48747
IsCapital	0.4755	0.5	0.48747
ContainsDigit	0.4755	0.5	0.48747
ContainsPunctuation	0.4755	0.5	0.48747
ContainsRoman	0.4755	0.5	0.48747
Prefixes	0.7422	0.6186	**0.6561**
Suffixes	0.5694	0.5949	**0.5792**
POS tagging	0.6563	0.6202	**0.6354**
Modifier	0.6377	0.5188	0.5241
Abbreviation	0.7256	0.5020	0.4916
Trivial	0.9756	0.5021	0.4917
Sum	0.4755	0.5	0.48747
Family	0.7411	0.5181	0.5229
Total	0.6488	0.6606	0.6544

Table 5 shows that prefix, POS and suffix obtained the greatest f-measure values. This denotes the importance of these features in extracting chemical compounds.

On the other hand, even though morphological features (i.e., F1 to F5) and dictionary features (i.e., F9 to F13) have yielded lower performance, different studies have suggested that these features be combined with other features to yield reasonable performance [5]. The total combination of all features has shown similar performance to that of the independent use of the prefix (i.e., around 0.65).

2) Results of N-gram using SVM: Also, the results of applying SVM with the N-gram are depicted in Table 6.

TABLE VI
Results of SVM with N-gram

Features	Precision	Recall	F-measure
465 terms	0.716	0.694	0.704

As shown in Table 6, the results of applying SVM are 0.716 for precision, 0.694 for recall, and 0.704 for the f-measure. It is evident that the results of applying the SVM with N-gram have outperformed the results of applying NB with detailed-attributes. This is proven by the 0.704 f-measure achieved by SVM and 0.654 achieved by NB with all features. This outperformance can be justified from the numerous features used in the SVM paradigm (i.e., 465 features) compared to the 13 features used by NB.

3) Results of applying the WSS feature selection: This section highlights the results of applying the WSS feature selection for both paradigms—SVM with N-gram and NB with the detailed-attribute. Table 7 depicts the results.

As shown in Table 7, the results of applying the feature selection on SVM with N-gram has led to the selection of 100 features with an f-measure of 0.718. In contrast, the results of applying the feature selection on NB with detailed-attributes have led to 4 features with an f-measure of 0.722. It is clear that the detailed-attribute representation has outperformed the N-gram representation regarding classification accuracy. Also, the selected features of the N-gram representation can be depicted as meaningless terms. Comparatively, the selected features of the detailed representation tend to be more generalized. This can facilitate the process of applying the selected features to new datasets to achieve higher accuracy.

As shown in Table 7, the results of applying the feature selection on SVM with N-gram has led to the selection of 100 features with an f-measure of 0.718. In contrast, the results of applying the feature selection on NB with detailed-attribute has led to 4 features with an f-measure of 0.722. It is clear that the detailed-attribute representation has outperformed the N-gram representation regarding classification accuracy. Also, the selected features of the N-gram representation can be depicted as meaningless terms. Comparatively, the selected features of the detailed representation tend to be more generalized. This can facilitate the process of applying the selected features to new datasets to achieve higher accuracy. To compare the acquired results with the related works, it is evident that NB with detailed-attribute showed superior performance by acquiring a 72.2% f-measure compared to the work of Rocktaschel et al. [6], which used the SCAI dataset, acquiring a 63% f-measure, and Usie et al. [9], which used the same dataset, and acquired a 68% f-measure.

TABLE VII
RESULTS OF APPLYING WSS FEATURE SELECTION

Paradigm	Selected Features	No. of features	Precision	Recall	F-measure
SVM with N-gram	hyd, py, carb, dime, nitr, trypto, but, meth, acy, sulf, dipep, dihydro, benz, mono, trii, iso, palm, iod, naph, etho, niso, testo, tyr, threo, cyclo, chol, prop, deox, uri, flu, adria, alka, glu, trig, ethy, nucl, xyl, phth, oxo, pip, brom, thio, acid, aden, dini, hetero, tamox, lact, cefo, tazo, allop, augus, yl, xy, ones, one, in, cin, mino, cetate, lic, yla, ic, phene, ium, sium, ine, chlor, ene, ide, ate, pril, lix, cid, rile, am, MD, VBZ, JJR, RP, CD, NNPS, PRP, WDT, NNS, JJ, qutation, EX, CC, VBG, POS, :, -RRB-, VBN, VB, NNP, DT, JJS, fullstop, QotationItalic	100	0.7545	0.698	0.718
NB with detailed-attributes	Contains-Digit, Prefix, POS tagging, Trivial	4	0.703	0.745	0.722

IV. CONCLUSION

This paper conducted a comparative study between two data representations—N-gram and detailed-attribute. N-gram was used with a SVM classifier, while the detailed-attribute was used with a NB classifier. Both data representations underwent a feature selection using the WSS approach. The results show that the detailed-attribute with NB yielded superior performance by achieving a 72.2% f-measure. For future researches, it is highly recommended that new data representations such as word embedding be applied and the results examined.

ACKNOWLEDGMENTS

Universiti Kebangsaan Malaysia supported this research.

REFERENCES

[1] Basel Alshaikhdeeb and Kamsuriah Ahmad, "Integrating correlation clustering and agglomerative hierarchical clustering For holistic schema matching," Journal of Computer Science, vol. 11, p. 484, 2015.

[2] B. Alshaikhdeeb and K. Ahmad, "Feature selection for chemical compound extraction using wrapper approach with Naive Bayes classifier," in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), 2017, pp. 1-6.doi:10.1109/ICEEI.2017.8312421.

[3] Yaoyun Zhang, Jun Xu, Hui Chen, Jingqi Wang, Yonghui Wu, Manu Prakasam, and Hua Xu, "Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning," Database, vol. 2016, p. baw049, 2016.

[4] Baydaa Hashim and Nazlia Omar, "A Back Propagation Neural Network for Identifying Multi-Word Biomedical Named Entities," 2016, vol. 11, 2016.doi:682-690 http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&amp;page=article&amp;op=view&amp;path%5B%5D=19206.

[5] Basel Alshaikhdeeb and Kamsuriah Ahmad, "Biomedical Named Entity Recognition: A Review," International Journal on Advanced Science, Engineering and Information Technology, vol. 6, 2016.

[6] Tim Rocktäschel, Michael Weidlich, and Ulf Leser, "ChemSpot: a hybrid system for chemical named entity recognition," Bioinformatics, vol. 28, pp. 1633-1640, 2012.

[7] Andre Lamurias, Tiago Grego, and Francisco M Couto, "Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI," in BioCreative Challenge Evaluation Workshop, 2013, p. 75.doi.

[8] Riza Batista-Navarro, Rafal Rak, and Sophia Ananiadou, "Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features, and heuristics," J Chem Inf, vol. 7, p. S6, 2015.

[9] Anabel Usié, Joaquim Cruz, Jorge Comas, F Solson, and Rui Alves, "CheNER: a tool for the identification of chemical entities and their classes in biomedical literature," J Cheminform, vol. 7, p. S15, 2015.

[10] Haider Banka and Suresh Dara, "A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification, and validation," Pattern Recognition Letters, vol. 52, pp. 94-100, 2015/01/15/ 2015.doi:https://doi.org/10.1016/j.patrec.2014.10.007 http://www.sciencedirect.com/science/article/pii/S0167865514003146

[11] Iñaki Inza, Pedro Larrañaga, Rosa Blanco, and Antonio J Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," Artificial intelligence in medicine, vol. 31, pp. 91-103, 2004.

[12] Robert Leaman, "Advancing biomedical named entity recognition with multivariate feature selection and semantically motivated features," Arizona State University, 2013Retrieved from.

[13] Corinna Kolárik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius, and Juliane Fluck, "Chemical names: terminological resources and corpora annotation," in Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference), 2008.

[14] E Alharbi and S Tiun, "A Hybrid Method of Linguistic Features and Clustering Approach for Identifying Biomedical Named Entities," Asian Journal of Applied Sciences, vol. 8, pp. 210-216, 2015.

[15] Stanford, "Part-of-Speech Tagger," ed, 2014.

[16] Peter Willett, "The Porter stemming algorithm: then and now," Program, vol. 40, pp. 219-223, 2006.

[17] Bo Tang, Steven Kay, and Haibo He, "Toward optimal feature selection in naive Bayes for text categorization," 2016.

[18] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu, "Adapting naive bayes to domain adaptation for sentiment analysis," in Advances in Information Retrieval, ed: Springer, 2009, pp. 337-349.

[19] Ahmed Almusawi and Haleh Amintoosi, "DNS Tunneling Detection Method Based on Multilabel Support Vector Machine," Security and Communication Networks, vol. 2018, 2018.

[20] Samaneh Moghaddam and Martin Ester, "AQA: aspect-based opinion question answering," in Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011, pp. 89-96.doi.

[21] Suzanne Little, Ovidio Salvetti, and Petra Perner, "Evaluation of feature subset selection, feature weighting, and prototype selection for biomedical applications," Advances in Case-Based Reasoning, pp. 312-324, 2008.

[22] Yuri Bykov and Sanja Petrovic, "A Step Counting Hill Climbing Algorithm applied to University Examination Timetabling," Journal of Scheduling, vol. 19, pp. 479-492, 2016.

[23] Ruizhi Li, Shuli Hu, Yiyuan Wang, and Minghao Yin, "A local search algorithm with tabu strategy and perturbation mechanism for generalized vertex cover problem," Neural Computing and Applications, vol. 28, pp. 1775-1785, 2017.

[24] Ivan Piza-Davila, Guillermo Sanchez-Diaz, Manuel S Lazo-Cortes, and Luis Rizo-Dominguez, "A CUDA-based Hill-climbing Algorithm to Find Irreducible Testors from a Training Matrix," Pattern Recognition Letter*s*, 2017.