Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 14, Issue. 1, January 2025, pg.78 – 87

A Vertical Mining Approach for Interesting Sentiment Associative Patterns Discovery

Zainab A. Al-Sayyady[1]; Basheer M. Al-Maqaleh[2]

[1]Faculty of Computers and Informatics, Thamar University, Yemen
[2]Faculty of Computers and Informatics, Thamar University, Yemen
[1] zainabalsyyadi@gmail.com; [2] basheer.almaqaleh@tu.edu.ye

DOI: https://doi.org/10.47760/ijcsmc.2025.v14i01.008

Abstract: Social media platforms have become an important part of our daily lives due to the widespread use of the Internet. They contains a great wealth of valuable information which provide opportunities for us to explore hidden patterns or unknown correlation relationships. The most existing Association Rule Mining (ARM) algorithms focus on mining associative patterns from large sentiment datasets based on traditional support-confidence framework. Those algorithms generate a large number of redundant patterns, the majority of which are uninteresting to the user or do not imply a true correlation relationship between related items. In this paper, an effective approach that integrates Natural Languages Processing (NLP) techniques, and ARM concepts in one vertical mining approach in order to discover interesting sentiment associative patterns is proposed. The proposed approach uses common NLP techniques as a pre-processing phase to generate stemmed wordsets from large sentiment datasets. In the second phase, the proposed approach pushes the support-all-confidence interestingness measures as a new framework deep during the mining process in order to generate a reduced and complete set of All-Confident Frequent Sentiment Wordsets (ACFSWs) directly from large vertical sentiment datasets. Furthermore, the generated ACFSWs are used as an underlying knowledge representation for Interesting Sentiment Association Rules (ISARs) discovery. The obtained results show the usefulness and effectiveness of the proposed approach.

Keywords: All-Confident Frequent Sentiment Wordsets, Interesting Sentiment Association Rules, Natural Languages Processing, Sentiment Analysis, Interestingness Measures

I. INTRODUCTION

Social media platforms such as Twitter, Instagram, and Facebook are frequently utilized by users to access and share information about important events, news, and other topics [1]. Sentiment analysis is also known as opining mining refers to the task of extracting the sentiment of people from textual data [2]. It has received popularity across various domains due to the significant increase in user-generated content in social media platforms as they contain a great wealth of valuable information [3], [4]. The field of Knowledge Discovery in Databases (KDD) is the automated process that attempts to make sense of the information explosion embedded in this big volume of data [5]. Data mining is a core component in the KDD analysis process that involves using data analysis tools to find valid patterns in massive datasets [6]. Frequent Itemset Mining (FIM) is a prominent

area of research in the field of KDD and data mining as it used to discover association relationships among items in large transactional datasets [7]. The mining goal is to identify all itemsets whose Frequent Itemsets (FIs), called support in the datasets exceeds a user- defined threshold called minimum support (*minsup*). The transactional datasets can be organized horizontally or vertically [6]. Horizontally means storing for each transaction the items contained, whereas vertically means storing for each item the transactions containing it. Association Rule Mining (ARM) is a data mining technique that aims to uncover hidden relationships and associations among items in datasets [6], [8]. Association rules are implication of the form[6]: $X \rightarrow Y$, where X (antecedent part) and Y (consequent part) are FIs, and $X \cap Y = \Phi$.

The strength of the association is measured using the confidence of the rule, which is the probability that item Y is present given that item X is present. Any association rule has a confidence value exceeds a user-defined threshold called minimum confidence (*minconf*) is called strong association rule. The support – confidence framework is commonly used as a basis to discover such association rules [6]-[9]. Sentiment mining is one of the important aspects of data mining where useful information can be mined from the collected data as it plays an essential role in the decision-making process [2], [3].

In this paper, the ARM notations and terminologies are customized and employed to fit the sentiment mining domain. We substitute transactional dataset with sentiment dataset, transactions with sentiment reviews or tweets, items with words, itemsets with wordsets (i.e. sets of words), and association rules with Sentiment Association Rules (SARs). The problem of support – confidence framework is generating a huge number of associative sentiment patterns, which is considered as computationally challenging, time consuming and makes difficulty for a user to use and understand the generated patterns [7]. Furthermore, because this framework lacks a test for capturing the correlation of generated associative sentiment patterns, it does not reflect the true correlation relationship among wordsets in the generated associative sentiment patterns [10].

To overcome the shortcomings of the support-confidence framework, the proposed approach extends this framework to deal with correlation by augmenting this framework with all-confidence measure [11]. The all-confidence is used as a correlation measure to filter out uncorrelated patterns in support dimension. The proposed approach is an attempt that combines and integrates the Natural Languages' Processing (NLP) techniques, ARM concepts in one mining approach in order to discover a reduced set of interesting sentiment associative patterns in terms of All-Confident Frequent Sentiment Wordsets (ACFSWs) and Interesting Sentiment Association Rules (ISARs) directly from large sentiment datasets.

The remainder of this paper is organized as follows:- Section II presents related work. Section III presents the interestingness measures used. The description of the proposed approach is introduced in Section IV. Section V reports the experimental results. Conclusion and future works are given in Section VI.

II. RELATED WORKS

Several surveys on opinion mining and sentiment analysis including their methods, applications and challenges were presented in [1], [12]-[13]. They explored the challenges of sentiment analysis and opinion classification in large, irregular online data, discussed extensively a wide range of methods used to solve the problems of mining various types of big data, and explained the latest applications, developments and challenges in sentiment analysis and opinion mining. In [14], a comparison of sentiment analysis techniques: polarizing movie blogs was presented. This study used IMDB dataset and incorporated WordNet lexicon resource to extract opinion from review. Various machine learning classifiers such as Support Vector Machine (SVM), Naive Bayse (NB) and alternating decision tree are used to classify the dataset with more than 75% accuracy.

In [15], a sentiment analysis method to clarify the relationship between music and human feelings based on data mining was proposed. The main idea of this method is to detect the singer's emotions such as happiness, hope, sadness, and anger while the song is playing. The authors used data-mining algorithms such as multi-label *k*-nearest neighbours and random *k*-label to build the model. In [16], a mining text patterns over fake and real tweets approach was suggested. This approach suggested a text mining solution to find patterns related to fake news and true news in tweets and it tested and validated the system using a pre-labelled dataset of fake and real tweets during the U.S. election. In [17], discovering long COVID-19 symptom patterns using association rule method was suggested. This method aimed to understand the patterns and behaviour of long COVID-19 symptoms reported by patients on Twitter by using ARM techniques to identify frequent symptoms and establish relationships among them, with a high confidence level, *minconf* of 10%, and *minsup* of 0.01%.

In [18], mining twitter data on COVID-19 for sentiment analysis and frequent patterns discovery method was conducted. This method adapted the FP-Growth algorithm to discover frequent patterns and association rules in the tweets, providing insights into the tweeters' perspectives on COVID-19. This method used support – confidence framework to evaluate the discovered knowledge. In [19], an efficient framework for sentiment classification using Apriori based feature reduction was proposed. This framework converted every sentence into binary form using horizontal data format and applied the Apriori algorithm to reduce the dataset. It then implemented four machine learning algorithms, with the proposed framework showing accuracy improvements

of up to 5.9% compared to other feature selection methods across various domain datasets. In [10], discovering sentiment patterns from online customer reviews in vertical mining was proposed. This method focused on sentiment patterns in online product evaluation, utilizing techniques for mining customer opinions. It used traditional support- confidence framework to evaluate the discovered patterns, so these patterns lack the correlation relationship among their items.

All related works mentioned in this study use support-confidence framework as a feature selection method to discover patterns from textual datasets in horizontal format. As a results, huge number of discovered patterns are found, they ignored the true correlation among those discovered patterns, and they may took long execution time. Therefore, the proposed approach is an attempt that integrates the NLP techniques and ARM concepts in one mining approach in order to discover interesting sentiment associative patterns from large vertical sentiment datasets. The discovered interesting sentiment association patterns are evaluated using support, all-confidence and confidence objective interestingness measures. As a result, a reduced set and concise knowledge in terms of ACFSWs and ISARs are discovered. By using the vertical format, the proposed approach reduces the sentiment dataset scans which also reduces the amount of time required to discover such knowledge.

III. INTERESTINGNESS MEASURES

Interestingness measures select, filter, prune and order patterns based on their potential interest to the end users [6], [20]. There are two approaches to classify a measure of interestingness of discovered patterns, objective and subjective [6], [8]. The objective measures such as support, all-confidence, confidence, lift, and coverage are based on the structure of discovered patterns and the underlying statistics [21]. They can obtain a quantitative value by the algorithm, and they are easy to implement. In contrast, the subjective measures are based on user belief in the data like novelty, actionability, unexpectedness , etc. [22]. The proposed approach follows the direction of objective interestingness measures such as support, all-confidence and confidence to extract the interesting sentiment associative patterns directly from large vertical sentiment datasets as described below.

A. Support Measure

The association rule X → Y is supported in the percentage of transactions that contain both itemsets X and Y in set of transactions T exceeds a certain threshold, called *minsup* threshold. The support of the association rule X → Y is defined as [6], [9]:-

$$support(X \rightarrow Y) = \frac{|X \cap Y|}{|D|} \tag{1}$$

where |D| is the total number of transactions, and |X ∩ Y| is the number of transactions containing both X and Y. The *minsup* threshold controls the minimum number of transactions that a itemset must cover in a transactional dataset.

B. Confidence Measure

The confidence for the association rule X → Y is defined by the percentage of the transactions that contain itemset Y among transactions containing itemset X. The confidence of the association rule X → Y is defined as [6], [9]:-

$$confidence(X \rightarrow Y) = \frac{|X \cap Y|}{|X|} \tag{2}$$

where |X| is the number of transactions containing X.

The pitfall of confidence can be traced to the fact that the measure ignores the support of the itemset in the rule consequent [8]. Confidence indicates, that the appearance of some itemsets will promote to appearance of other itemsets without considering the relationship between "X" and "Y" when does not occur [6].

C. All-Confidence Measure

The all-confidence measure for an association rule (X → Y) is defined as under [11] :-

$$All-Confidence(X \rightarrow Y) = \frac{support(X \cap Y)}{max[support(X), support(Y)]} \tag{3}$$

Any itemset passes a *minAllconf* threshold is called all-confident or correlated itemset, where *minAllConf* is the user-defined minimum all-confidence threshold value. This definition simply addresses that all-confidence is the smallest confidence of any rule for the set of items of X. That is, all rules produced from this itemset would have a confidence greater than or equal to its all-confidence value. The all-confidence measure is appealing since it decreases the amount of mined patterns and only generates correlated patterns [23].

IV. THE PROPOSED APPROACH DESCRIPTION

The important concepts related to the proposed approach are explained in the following subsections.

A. *Frequent Sentiment Wordsets (FSWs)*

FSWs are sentiment wordset that appear in a sentiment dataset frequently. For example, a set of sentiments, such as movie and funny, that appear frequently together in a dataset is a FSW. A wordset (i.e. sets of words) can be defined as follows:-

Let W = $\{w_1,w_2,...,w_m\}$ to be a set of words(items), and D be a sentiment dataset consists of a set of sentiments (reviewers or tweets) in vertical format. Each sentiment T consists of a set of words such that T ⊆ W and it is associated with an identifier, called TID. Let X be a set of words, referenced to as a wordset. A wordset that contains *k* words is a *k*-wordset. The set {movie , funny} is a 2-wordsets.

The support of a wordset X in D defined as support (X), is the number of sentiments in D containing X. The formal definition of FSW as follows:-

A sentiment wordset X is a FSW if it occurs no less frequently than a minsup threshold such that : X ϵ FSWs if support (X) $\geqslant$ minsup.

All FSWs are called Supported Sentiment Wordsets (SSWs),

B. *All- Confident Frequent Sentiment Wordsets*

The formal definition of ACFSWs is as follows:-

Given a sentiment dataset D, a minsup and a minAllConf thresholds, a wordset X is said to be all-confident or correlated if support (X) $\geqslant$ minsup and all-confidence (X) $\geqslant$ minAllConf.

The concise representation of ACFSW is a subset of FSW (ACFSWs ⊆ FSWs). These ACFSWs represent the most correlated and frequent wordsets that exist in the sentiment dataset and they would only participate in the further mining processes. More precisely, all infrequent sentiment wordsets, whose support values are less than *minsup*, would not be participated in the further mining processes, so the search space would be reduced greatly. As a result, the expected execution time would be also reduced.

According to all-confidence measure, a ACFSW is interesting if all Sentiment Association Rules(SARs) that can be generated by partitioning that ACFSW have all-confidence greater than or equal to *minAllConf* threshold.

C. *Interesting Sentiment Association Rule*

In the context of sentiment analysis, a SAR based on the notion of FSWs is defined as under :

R: X $\rightarrow$Y is called SAR if X $\cup$Y is a FSW, support(R) $\geqslant$ minsup, confidence(R) $\geqslant$ minconf, and (X $\cap$Y)= Φ.

Mining strong correlation from large datasets often generates more valuable results as it reduces the number of discovered rules than mining SARs. So, in this work, the correlation constrained in terms of all-confidence measure is integrated with support to form a support-all-confidence framework in order to evaluate the generated candidate sentiment wordsets. Generally, support and confidence measures are used to evaluate the discovered association rules [9], [24], [25]. Similarly, in the context of sentiment analysis the formal definition of ISAR is defined as under:-,

R: X $\rightarrow$Y is called ISAR if X $\cup$Y is a ACFSW, support(R) $\geqslant$ minsup, confidence(R) $\geqslant$ minconf, and (X $\cap$ Y)= Φ.

By the definition the union of the antecedent and the consequent parts of such ISARs form ACFSWs for that rule. As that, the number of ACFSWs are usually less than the number of FSWs, therefore the number of ISARs is also much less than the number of SARs. Thus, the discovered ISARs are special subset of SARs such that ISARs ⊆ SARs.

D. *The proposed approach phases*

The proposed approach consists of two phases, the first phase is the pre-processing phase, where standard NLP text pre-processing techniques like punctuation marks removal, tokenization, stop words removal and stemming are applied on a raw dataset [26], [27]. This pre-processing phase reduces the size of dataset as well as eliminates the information that does not contain sentiment or emotional meaning to it. It is to be noted that, the output of the pre-processing phase is the processed dataset in the horizontal format. This format consists of a list of wordsets (stemmed words) associated with an identifier (TID).These steps are necessary to make the sentiment texts ready to be applied in the second phase. Mining phase is the second and core phase in the proposed approach and it consists of the following steps:-

1) Vertical format conversion step

Data representation is a crucial in the proposed approach as the dataset format and the searching strategy involved are contribute to the performance of mining each wordset. The processed dataset in horizontal format

is used as input to the mining phase. In this step, the dataset in horizontal format is converted into the vertical format, where sentiment dataset consists of a list of wordsets, each wordset followed by the list of Tid (also called Tidset). The vertical format helps the proposed approach to scan the dataset only once, hence the expected total execution time would be decreased.

2) ACFSWs generation step

In this step, the proposed approach performs the following procedure:

- Generating all ACFSWs directly from the vertical sentiment dataset using support-all-confidence framework , i.e., generates all ACFSWs that have support and all-confidence greater than, or equal to user specified *minsup* and *minAllConf* thresholds respectively.

That is, generating a complete and reduced set of ACFSWs. More precisely, any candidate wordset has support and all-confidences values greater than or equal to *minsup* and *minAllConf* respectively are considered as ACFSW and they would only participate in the next process. Pruning by support and all-confidence measures can be applied at each iteration of the proposed approach to help greatly reducing the search space and then improve the effectiveness of whole generation process by generating only the most interesting patterns in terms of ACFSWs.

3) ISARs discovery step

In this step, the proposed approach performs the following procedure:

- Discovering all ISARs from the generated ACFSWs that have *minconf* threshold in the following simple way:- For each ACFSW Y, generate all nonempty subsets of Y. For every nonempty subset S of Y, output the rule S → Y − S if its confidence value is greater than, or equal to, the *minconf* , then this rule is considered as a ISAR.

The workflow of the proposed approach is shown in Fig. 1.

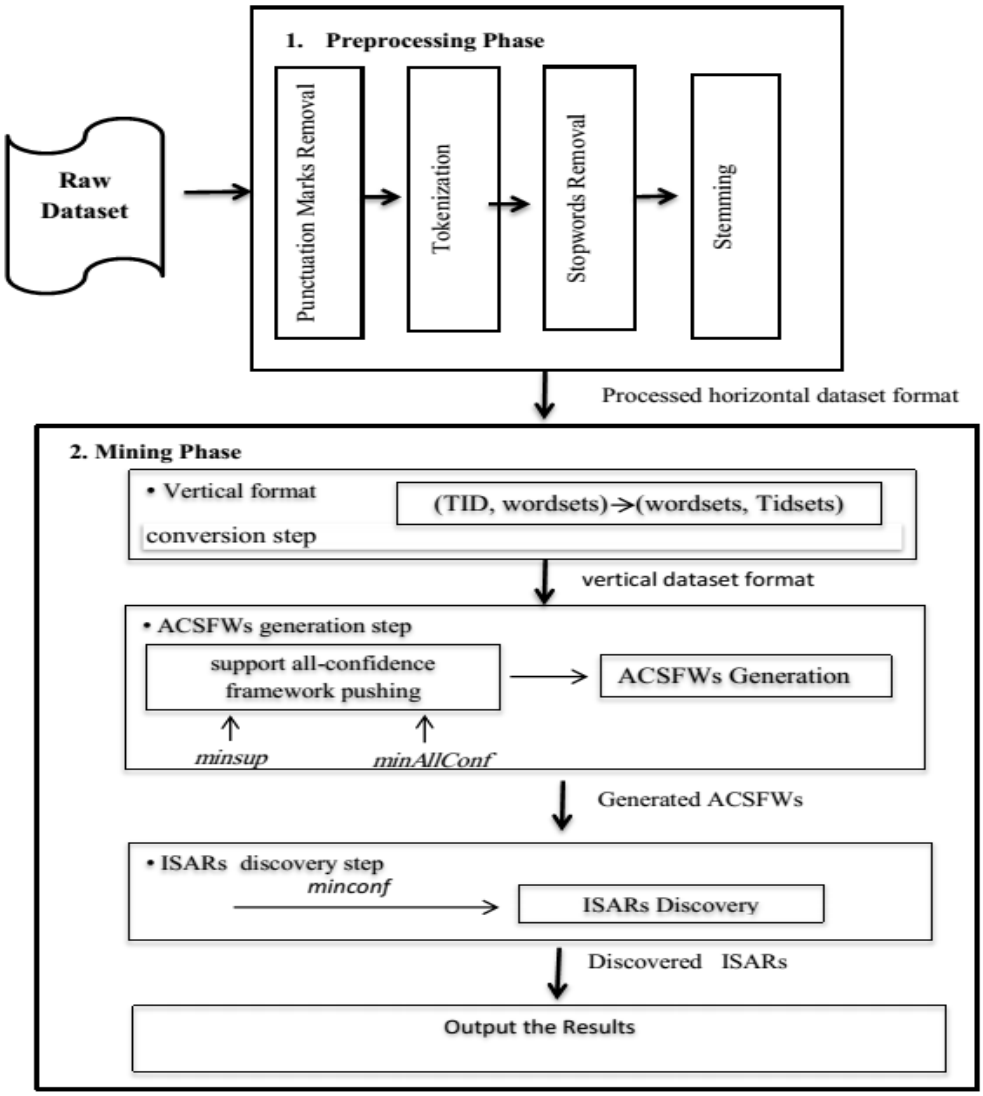

Fig. 1. Workflow of the Proposed Approach

V. COMPUTATIONAL RESULTS

The performance of the proposed approach is validated on benchmark real- world sentiment datasets. Those datasets are obtained from UCI Machine Learning Respiratory which is a collection of widely used benchmark and real-world datasets for KDD community and data mining [29]. The performance of the proposed approach is evaluated and compared with the well-known Apriori [9] and ECLAT [28] algorithms. The Apriori algorithm is the widely referenced techniques in FIM, ARM [6], [7] and the ECLAT is the base algorithm of the proposed approach. All experiments are performed on a laptop with 2.60GHz core i7-66004 processer, 4GB RAM and windows 10.0. The performance of the proposed approach on different datasets is demonstrated below.

A. Experiment one

This experiment was carried out on the SMILE Twitter Emotion Dataset. It contains nominal-value attributes and contains 3 attributes with 1000 instances. Table 1 below shows the number of generated ACFSWs by the proposed approach with different *minsup* values such as 0.01, 0.02, 0.06 and 0.07 and with different *minAllConf* values such as 0.03, 0.04 ,0.10 and 0.20 respectively.

TABLE 1 : NUMBER OF GENERATED ACFSWS FROM SMILE TWITTER EMOTION DATASET

minsup	***minAllConf***	**#ACFSWs**
0.01	0.03	2330
0.02	0.04	441
0.06	0.10	5
0.07	0.20	3

It can be observed that the number of generated ACFSWs decreases as the values of *minsup* and *minAllConf* increase as shown in Table 1. For an illustration, the proposed approach would generate five ACFSWs when *minsup* = 0.06 and *minAllConf* = 0.10 as shown in Table 2.

TABLE 2 : GENERATED ACFSWS FROM SMILE TWITTER EMOTION DATASET

No.	**Generated ACFSWs**	**support**	**all-confidence**
1	happy, britishmuseum	0.14	0.35
2	britishmuseum, nocode	0.16	0.40
3	happy, nationalgalleri	0.07	0.18
4	nationalgalleri, nocode	0.09	0.22
5	not-relevant, nationalgalleri	0.06	0.16

The proposed approach uses the generated ACFSWs as input to the ISARs discovery step in the mining phase. As a result, the discovered ISARs with *minsup* = 0.06 and *minconf* = 0.30 are shown in Table 3.

TABLE 3 : DISCOVERED ISARS FROM SMILE TWITTER EMOTION DATASET

No.	**Discovered ISARs**	**support**	**confidence**
R1	happy → britishmuseum	0.14	0.39
R2	britishmuseum →happy	0.14	0.35
R3	britishmuseum → nocode	0.16	0.40
R4	nocode →britishmuseum	0.16	0.44
R5	nationalgalleri → nocode	0.09	0.33
R6	not-relevant → nationalgalleri	0.06	0.41

The results in Table 3 proves that the support measure is a symmetric measure which means X →Y and Y→ X have the same support value, whereas confidence measure is not.

B. Experiment two

The sts_gold_tweet dataset was used for this experiment. This dataset is designed for evaluating sentence similarity, specifically focusing on tweets from the Twitter platform. It is primarily used in NLP tasks, consists of 1000 tweets, their corresponding Ids, and polarity. This dataset has 1000 instance, and 3 attributes. the proposed approach would generate the following ACFSWs with *minsup* = 0.03 and *minAllConf* = 0.05 as shown in Table 4.

TABLE 4 : GENERATED ACFSWS FROM STS_GOLD_TWEET DATASET

No.	**Generated ACFSWs**	**support**	**all-confidence**
1	cant, negative	0.043	0.063
2	negative, get	0.049	0.071
3	negative, go	0.064	0.093

4	headach, negative	0.043	0.063
5	miss, negative	0.037	0.053
6	negative, sad	0.039	0.056
7	negative, want	0.035	0.051
8	negative, work	0.038	0.055
9	love, positive	0.041	0.059

The proposed approach would generate the most confident ISARs from the generated ACFSWs that shown in Table 4, by introducing the confidence measure to the mining process. Table 5 shows the discovered ISARs from this dataset with *minsup* = 0.03 and *minconf* = 0.70.

TABLE 5: DISCOVERED ISARS FROM STS_GOLD_TWEET DATASET.

No.	Discovered ISARs	support	confidence
R1	cant →negative	0.043	0.781
R2	get →negative	0.049	0.763
R3	headach →negative	0.043	1.000
R3	sad→negative	0.039	1.000
R5	miss →negative	0.037	0.949
R6	want→negative	0.035	0.814
R7	work →negative	0.038	0.826
R8	love → positive	0.041	0.854
R9	positive → love	0.041	0.729

C. Experiment Three

Internet Movie Database (IMDB) was used for this experiment. This dataset has 1000 instances, 2 attributes. An example, the proposed approach would generate five ACFSWs when *minsup* = 0.20 and *minAllConf* = 0.50 as shown in Table 6.

TABLE 6 : GENERATED ACFSWS FROM IMDB DATASET.

No.	Generated ACFSWs	support	all-confidence
1	movie, film	0.370	0.564
2	film, one	0.364	0.554
3	like, movie	0.356	0.543
4	movie, negative	0.347	0.529
5	movie, one	0.398	0.607
6	movie, love, positive	0.378	0.565

It is to noted that, the more wordsets of generated ACFSWs are dependent or strongly correlated to each other, the higher value of all-confidence measure is, since the support of the generated ACFSWs would be closer to the maximum support of the wordsets within the generated ACFSWs itself as shown in Table 6. The proposed approach would discover nine ISARs from the generated ACFSWs that shown in Table 6 with *minsup* =0.20 and *minconf* = 0.60 as shown in Table 7.

TABLE 7 : DISCOVERED ISARS FROM IMDB DATASET.

No.	Discovered ISARs	support	confidence
R1	film → movie	0.370	0.615
R2	film → one	0.364	0.605
R3	One → film	0.364	0.620
R4	like → movie	0.356	0.711
R5	negative → movie	0.347	0.694
R6	movie →one	0.398	0.607
R7	one → movie	0.398	0.678
R8	movie → love, positive	0.378	0.723
R9	movie, love → positive	0.378	0.712

D. Comparative Study

The performance of the proposed approach is compared with Apriori and ECLAT algorithms in two directions as follows:-

1) Number of discovered sentiment patterns

In order to evaluate the performance of the proposed approach over original Apriori and ECLAT algorithms, experiments have been conducted several times with different *minsup* values. It is to be noted that, the values of *minAllConf* are used only during the generation of ACFSWs by the proposed approach. The obtained results from the proposed approach, Apriori and ECLAT algorithms based on the number of generated ACFSWs and FSWs are shown in Table 8.

TABLE 8 : COMPARISON RESULTS OF THE PROPOSED APPROACH, APRIORI AND ECLAT ALGORITHMS.

Dataset Name	***minsup***	***minAllConf***	**#FSWs/ #ACFSWs**		
			Apriori	**ECLAT**	**The proposed approach**
SMILE Twitter Emotion	0.01	0.03	4336	4336	2330
	0.02	0.04	494	494	441
	0.06	0.10	18	18	5
	0.07	0.20	11	11	3
sts_gold_ tweet	0.008	0.01	378	378	194
	0.010	0.03	299	299	36
	0.020	0.04	119	119	20
	0.030	0.05	51	51	9
IMDB	0.04	0.07	27115	27115	15928
	0.06	0.09	6380	6380	6087
	0.09	0.20	1542	1542	324
	0.20	0.50	103	103	6

In general, Apriori and ECLAT algorithms may generate large and equal numbers of FSWs in all datasets, depending on how the *minsup* is set. Discovering too many FSWs makes it difficult for a user to analyze them. The proposed approach reduces the number of generated ACFSWs greatly as it uses two ways of reducing the number of ACFSWs found and presents more meaningful wordsets to the user. The first way, it involves using the support measure, and the second way, it pushes correlation measure called all-confidence, to assess how correlated each FSW is, as an another constraint deep in the mining process, in order to filter less interesting wordsets and discover only interesting wordset called ACFSWs. So, the correlation has been adopted as an interesting measure since the most users are interested in not only association-like co-occurrences but also the possible strong correlations implied by such co-occurrences. So, the proposed approach would generate less number of sentiment associative patterns in terms of ACFSWs as shown in Table 8.

2) Execution Time

Execution time is a measure of time taken by the proposed approach to generate only ACFSWs and then ISARs. A comparative analysis of the execution time between the proposed approach and Apriori algorithm (in seconds) is shown in Table 9.

TABLE 9 : COMPARISON RESULTS OF THE PROPOSED APPROACH AND APRIORI ALGORITHM BASED ON EXECUTION TIME

Dataset Name	***minsup***	***minconf***	**Apriori algorithm**	**The proposed approach**
SMILE Twitter Emotion	0.25	0.80	40	10
sts_gold_ tweet	0.15	0.70	90	20
IMDB	0.05	0.60	120	55

As shown in Table 9, the execution time increases as the values of *minsup* and *minconf* decrease. Furthermore, Apriori algorithm consumes long time as it generates patterns by scanning the dataset many times, such as if the length of wordsets is *k* then the Apriori algorithm should scan the datasets *k-1*. But the proposed approach takes less time as it uses the vertical format and it scans the dataset only once to get the wordsets. For the ACFSWs generation from the *2*- wordsets, it only needs to refer the previous wordsets. This eliminates the need to scan through the dataset each time to count the frequency(support) of wordsets for each iteration. So, The proposed approach outperforms the Apriori algorithm in terms of execution time.

Fig. 2 depicts the comparative performance based on execution time taken to discover SARs and ISARs by Apriori algorithm and the proposed approach respectively.

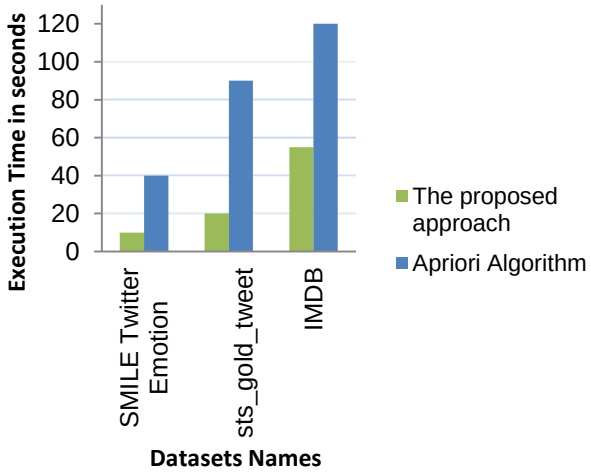

Fig. 2. Comparison Performance based on Execution Time.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, the proposed approach integrates NLP techniques and ARM concepts in one mining approach in order to discover the truly correlated relationship among sentiment associative patterns in terms of ACFSWs and ISARS from large vertical sentiment datasets. In the first phase of the proposed approach, common NLP techniques are applied on a raw sentiment dataset in order to generate a processed dataset in horizontal format. In the second phase, the proposed approach converts the processed dataset into the vertical format and it pushes the support-all-confidence framework into the mining process to prune the search space and then reduced the generated ACFSWs using both tests simultaneously. Also, those ACFSWs are used as a new basis to mine a reduced set of ISARs by using confidence interestingness measure. So, the discovered knowledge in terms of ACFSWs and ISARs are supported, correlated and confident. The obtained results show that mining of such knowledge generates a much smaller set but truly correlated sentiment patterns with less execution time comparing to the related algorithms.

The proposed approach can be extended to use some of supervised machine learning techniques that automatically classify the generated correlated sentiment patterns to the proper class value. In this case, the improved approach should have a dynamic extraction features and real-time classification of sentiment datasets streams. Also, the proposed approach can be extended to extract sentiment associative patterns from Arabic language.

REFERENCES

[1]. F. Aftab, S. U. Bazai, S. Marjan, L. Baloch, S. Aslam, A. Amphawan, and T-K. Neo, " A comprehensive survey on sentiment analysis techniques," *International Journal of Technology*, vol. 14, no. 6, pp. 1288-1298, 2023.

[2]. A. G. Katsafados, S. Nikoloutsopoulos, and G. N. Leledakis "Twitter sentiment and stock market: A COVID-19 analysis," *Journal of Economic Studies*, vol. 50, no.8, pp.1866-1888, 2023.

[3]. T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, pp. 6-9, 2023.

[4]. A. H. Shapiro, M. Sudhof, and D. J. Wilson, "Measuring news sentiment," *Journal of Econometrics*, vol. 228, no. 2, pp. 21-243, 2022.

[5]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27-34, 1996.

[6]. J. Han, and M. Kamber. *Data Mining Concepts and Techniques*, 4th Edition, Morgan Kaufmann Publishers, Waltham, 2022.

[7]. J. A. D. Garcia, M. D. Ruiz, and M. J. Martin Bautista, 'A survey on the use of association rules mining techniques in textual social media," *Artificial Intelligence Review*, vol. 5, pp.1175–1200, 2023.

[8]. P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, 2nd Edition, Pearson Education India, New Delhi, 2021.

[9]. R. Agrawal, and R. Srikanth, "Fast algorithms for mining association rules in large databases," *In Proceedings of International Conference on Very Large Databases*, pp. 487–499, 1994.

[10]. X. Li, and S. Zhang, "Discovering sentiment patterns in online customer reviews using vertical mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no.8, pp.1567-1580, 2019.

[11]. E. R. Omiecinski, "Alternative interest measures for mining associations in databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no.1, pp. 57-69, 2003.

[12]. M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, pp. 25-33, 2021.

[13]. M. Wankhade, A.C.S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, 2022.

[14]. S. S. Vavilapalli, P. ReddyKorepu, S. Saggam, M. Pentyala, and S. A. Devi, "Summarizing and sentiment analysis on movie critics data," *In 2021 6th International Conference on Inventive Computation Technologies (ICICT)*, vol. 3 no. 1, pp.1-5, 2021.

[15]. R. L. Rosa, D. Z. Rodriguez, and G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 3, pp. 359-367, 2015.

[16]. J. A. Díaz-García, C. Fernandez-Basso, M. D. Ruiz, and M. J. Martin-Bautista, "Mining text patterns over fake and real tweets," *In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Cham: Springer International Publishing, pp. 648-660, 2020.

[17]. M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina,"Discovering symptom patterns of COVID-19 patients using association rule mining," *Computers in Biology and Medicine*, vol. 131, pp. 16-19, 2021.

[18]. H. Drias, and Y. Drias, "Mining twitter data on COVID-19 for sentiment analysis and frequent patterns discovery," *Medial Research Archive*, pp. 20-33, 2020.

[19]. A. Jain, and V. Jain, "Efficient framework for sentiment classification using Apriori based feature reduction," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, no. 31, pp. 11-16, 2021.

[20]. J. Chen, S. Yang, W. Ding, P. Li, A. Liu, A. H. Zhang, and T. Li, "Incremental high average-utility itemset mining: survey and challenge," *Scientific Reports*, vol. 14, no. 1, pp.8-13, 2024.

[21]. S. D. Chandraveer, S. Arora, and Z. Makani, "Comparison of interestingness measures: support-confidence framework versus lift-irule framework," *International Journal of Engineering Research and Applications(IJERA)*, vol. 3, no. 2, 208-215, 2013.

[22]. Sethi, and B. Shekar, "Subjective interestingness in association rule mining: A theoretical analysis," *Digital Business: Business Algorithms, Cloud Computing and Data Engineering* , pp. 375-387, 2019.

[23]. F. M. Al-Kebsi, K. S. Al-Wagih, and B. M. Al-Maqaleh., An effective algorithm for mining interesting maximal association rules," *In 2021 IEEE International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI)*, pp.1-6, 2021.

[24]. B. M. Al-Maqaleh, and S. K. Shaab, "An efficient algorithm for mining association rules using confident frequent itemsets,' *3rd IEEE International Conference on Advanced Computing & Communication Technologies*, pp. 90-94, 2013.

[25]. A. M. Al-Badani, and B. M. Al-Maqaleh, "Efficient mining of frequent itemsets using improved FP-Growth algorithm," *International Journal of Applied Information Systems (IJAIS), Published by Foundation of Computer Science*, New York, USA, vol. 12, no. 14, pp.15-20, 2018.

[26]. A. Farkiya, P. Saini, S. Sinha, and S. Desai, "Natural language processing using NLTK and WordNet," *International Journal of Computer Science and Information Technology*, vol. 6, no.6, pp. 5465-5469, 2015.

[27]. M. E. Porter, "Snowball: A language for stemming algorithms," *Computer Science, Linguistics*. http://snowball.tartarus.org/texts/introduction.html.

[28]. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," *In KDD*, vol. 97, pp. 283-286, 1997.

[29]. C.L Blake ,and M.J. Merz ,UCI repository of Machine Learning Database [http://www.ics.uci.edu/mlearn/ ML Repository .html], Irvine, CA: University of California, Department of information and computer Science.

