Applied Intelligence (2021) 51:51–70
https://doi.org/10.1007/s10489-020-01668-6

Sentiment analysis: dynamic and temporal clustering of product reviews

Murtadha Talib AL-Sharuee[1] · Fei Liu[1] · Mahardhika Pratama[2]

Published online: 19 March 2020

Abstract

The increased availability of online reviews requires a relevant solution to draw chronological insights from review streams. This paper introduces temporal sentiment analysis by adopting the automatic contextual analysis and ensemble clustering (ACAEC) algorithm. ACAEC is a clustering algorithm which utilizes contextual analysis and a clustering ensemble learning. We propose chronological sentiment analysis using window sequential clustering (WSC) and segregated window clustering (SWC). WSC is a dynamic analysis, whereas SWC is solely based on the temporal characteristic of reviews. ACAEC is the base learning algorithm of WSC and SWC. ACAEC's ensemble approach is enhanced using an additional weight scheme and an additional learner to improve WSC's outcome. To understand the produced sentiment pattern, an unsupervised review selection is introduced which is based on review polarity. We also introduce *consistency*, a free-label measure to assess the algorithm's performance. For this study, new sets of reviews are introduced, these being four airlines and an Australian property agent. In terms of accuracy and stability, the proposed methods are effective in processing a review series. Experiments show that the average accuracy rates of SWC and WSC reach 87.54% and 83.87%, respectively. In addition, it is robust against the so-called imbalanced windows problem. The suggested solutions are unsupervised i. e. domain-independent and suitable for the analysis of a large review series.

Keywords Sentiment analysis · Temporal clustering · Unsupervised learning · Ensemble learning · Contextual analysis · k-means algorithm

1 Introduction

The number of online product reviews is enormous and is growing rapidly as a result of web development and the construction of a large number of user-friendly review platforms. Sentiment Analysis (SA) tools that can automatically analyze a large set of reviews is demanding because first, the abundance of online reviews is beyond human analysis; and second, public opinion is a significant consideration when governments, institutions, and individuals are making decisions and taking action. As SA plays a pivotal role in decision making, it has been the topic of over 7000 research works [12] in different areas such as politics [24], finance [6, 37, 43], medicine and also social issues [42], recommender systems [46] and cybersecurity [39]. It is important to consider the temporal order of product reviews which can reveal sentiment patterns.

SA can detect the public sentiment toward products, services, events, policies and regulations and also individuals such as politicians. Taking this information into consideration is significant to make the right decision. In relation to finance, for example, Yu, Duan and Cao [47] conclude that trends in the stock market and the sentiments expressed through text streams are correlated, and [43, 47] show that there is a strong relationship between the stock market and public sentiment. SA is utilized in recommender systems in recent work by Xing and Wang [46], where sentiment information from user reviews is incorporated to enhance point-of-interest recommender system. In [35], sentiment and implicit information improve the prediction of items that customers might like. Also, cybersecurity can benefit from SA as sentiments expressed on social media can be measured to detect a cyber attack [39].

Temporal SA can identify changes in public sentiment over time which is important for analyzing and identifying the

✉ Murtadha Talib AL-Sharuee
M.Al-sharuee@latrobe.edu.au

[1] Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia

[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

potential causes of a positive/negative sentiment trend at a particular point in time. It is useful for both the customer and the provider as it reveals current public sentiment and it can also change the impression that a customer may already have toward a particular brand or service by following changes in sentiment patterns. For providers, such as companies, sentiment patterns help to identify a product's drawbacks, information which companies can use to improve both their product and the customer experience.

A few SA research studies have considered the temporal characteristic of SA, for example, in [40], a supervised autoregressive framework is proposed for processing social media text and predicting the stock market. Another work by Nguyen et al. [32] uses supervised and semi-supervised methods to analyze social media text by tracking sentiment changes over time. In our work, we propose a temporal and an unsupervised framework to analyze a product review series over time using the ACAEC algorithm [2]. ACAEC comprises a contextual analysis method and an ensemble clustering learning technique which combines the results of clustering several data representations using a modified k-means algorithm.

This paper seeks to formalize a chronological sentiment analysis of product reviews. Two methods are suggested namely window sequential clustering (WSC) –dynamic learning and temporal analysis– and segregated window clustering (SWC) –temporal analysis. The results show that WSC and SWC are competitively accurate compared to supervised methods and significantly outperform unsupervised methods. In addition to producing a close-to-reality sentiment pattern, the main advantage of WSC and SWC is that they are domain-independent methods as they use ACAEC as a base clustering approach. ACAEC is a domain-independent method because it is unsupervised. This was proven, in our previous research work [2] by testing the algorithm on data from different domains. A domain-independent and inexpensive method is needed to analyze product reviews since an unlimited number of domains can benefit from SA in terms of decision making. Such a system also permits real-time SA because it does not require training or pre-processing nor does it require labeled data which is expensive and time-consuming to obtain. Therefore, we used a label-free unsupervised paradigm which can overcome the drawbacks of supervised methods.

Moreover, we introduce consistency, which is a new label-free measure of the stability and reliability of temporal analysis. When conducting experiments using experiment datasets, we can easily calculate the accuracy of the analysis, and hence judge the quality of the algorithm. However, accuracy cannot be calculated in real-world data analysis. This is because real-world review samples are not labeled. In this circumstance, a measure of the stability and reliability of the algorithm is important.

In addition to this, unsupervised review selection is introduced which will be jointly integrated with chronological sentiment analysis to better understand sentiment patterns. As a result, this combined framework delivers a comprehensive temporal understanding of public opinion regarding a particular product which is significant for both consumers and providers.

This study, therefore, contributes to the literature by providing a reliable and effective temporal SA which is domain-independent and is suitable for real-time analysis. It yields a comprehensive analysis and focuses on product reviews and their chronological characteristics. To evaluate the proposed framework, five new datasets are scraped, these being reviews on four airlines and reviews on one Australian property agent.

The main contributions of this paper are as follows:

- Introducing two efficient temporal clustering methods, WSC and SWC, to identify sentiment patterns which can reveal sentiment shifts over time.
- Combining a review selection method with the chronological analysis of a review series to assist in the understanding of sentiment patterns.
- Enhancing the ACAEC algorithm by using an additional weight scheme and a second sequential learner to improve the outcome of WSC.
- Addressing the so-called imbalanced windows issue, which is crucial as it is likely that imbalanced data will be encountered within a window series.
- Introducing *consistency* which is a new measurement to assess the readability and stability of temporal analysis without using labeled data.
- New sets of review series of four airlines and an Australian property agent are collected and are available online.[1]

The remainder of the article is organized as follows: section 2 gives a brief overview of the related work. In section 3, ACAEC is explained in detail. Section 4 describes the proposed framework in detail. Section 5 presents the data and provides a detailed analysis of the experiment results. In section 6, we draw a conclusion.

2 Related work

In the literature, the main research directions that have been taken to address SA are lexicon-based and machine learning methods. The earliest methods are lexicon-based which utilize a list of words as the most important indicators of sentiments [27]. These methods usually do not result in a high accuracy rate because they simply rely on the occurrence of a document's words in a lexicon. A lexicon is either manually

[1] The datasets can be accessed using this link http://homepage.cs.latrobe.edu.au/liufei/Supervision/index.html

generated such as MaxDiff [20] or automatically generated, such as SentiWordNet [4]. In our work, we employ ACAEC which utilizes SentiWordNet because it covers a large number of words and it is based on a consistent sentiment scouring, since it has been automatically built. In contrast, manually generated lexicons usually contain a fewer number of words and its sentiment scouring is subjected to the annotators' judgment which can be inconsistent.

For more effective performance, the focus has shifted to machine learning methods, particularly supervised machine learning. Initially, single classic data mining classifiers were used [33] such as SVM, ME and NB. In later work, more complex supervised learning algorithms were proposed to address natural language complexity, such as ensemble learning [44, 45] and deep learning [8, 36]. However, the main drawback of these methods is that they are domain-dependent as a result of using supervised learning, hence they usually cannot deal effectively with completely unseen data [48]. In addition, the unavailability of labeled data prevents real-time analysis when using supervised learning. In [19], to address the inaccessibility of sufficient tweets regarding an outbreak of an epidemic, a word vector model is trained on tweets and scholarly biomedical abstracts. However, this model is domain-specific. Thus, using an unsupervised approach to address SA can overcome domain-dependency and the training data unavailability issue.

An unsupervised clustering method for SA has been suggested in studies. In [25], an unsupervised learning method is proposed using an unweighted voting mechanism to combine multiple results of the k-means algorithm in order to determine a document's group membership. This unweighted voting is used as a solution for k-means instability. However, the approach depends on an arbitrary selection of centroids which can affect its performance and stability. The method also uses an experimentally chosen seeds for group identification which means that the seeds have to be determined each time a new dataset is processed. In ACAEC, the stability and the groups' identification are solved and it also produces a higher accuracy rate.

Currently, the need to develop a dynamic SA is becoming increasingly important due to the rapid increase in the availability of web reviews. Sentiment pattern which can be identified from the analysis of a stream of reviews is significant for decision making for both beneficiaries (customers) and providers (businesses/government institutions). Therefore, several research studies have considered dynamic analysis [22, 31, 36, 41] and temporal analysis [3, 13, 30]. The previous research papers use well-known supervised classifiers to conduct experiments. In [31], the authors analyze political tweets using SVM which needs to be regularly trained on new training sets in order to maintain a reliable system. The SVM classifier is trained using term frequency-inverse document frequency (TF-IDF) with 1–3 grams, polar hashtags and number of polar words and reaches about 0.65 recall, precision and F-measure. The authors, in [28], use test samples from years 2012 to 2014, and training data from 2001 to 2004 and from 2008 to 2011 to show that accuracy deteriorates when logistic regression models trained on the older data sample. Their experiments show that accuracy decreases because the polarity of features changes over time. To address this, predictive feature selection methods are proposed to train regression models. However, featuresare selected from data distributed over a long period of time.

In [41], active sentiment analysis experiments using Pegasos SVM are conducted to predict stock price movements of selected companies from the sentiment expressed in Twitter data. In [22], a dynamic learning method for SA is proposed using a cloud-based platform to classify Twitter data into positive and negative classes. The system is intended to handle changes in the data over time. However, these approaches [22, 41] are supervised which results in a domain-dependent model and they are also designed for microblogging data. Moreover, the method adopts a common practice in active learning in which a manual data labeling is conducted to update the classifier, thus manual annotation strategies are classifier-specific and result in more domain-dependent classifiers.

In [30], clustering and classification of users' comments on a presidential election campaign are conducting using standard machine learning methods. The clustering process is based on clustering words within a given day, therefore it is hard to interpret results of processing many days. Also, the pattern discovered from a small set of data which is probably not reflecting the public opinion. Alves et al. [3] propose a supervised and temporal classification method using Bayesian classifiers. In this work, the data considered are tweets from 4 months in 2013 and written in Portuguese about FIFA's Confederations Cup. Usually, supervised methods can result in an accuracy rate higher than what has been reported in this work, which could be a result of the language's complexity of the dataset.

Temporal clustering is considered in other research fields [1], especially in environmental applications [17, 21]. In recent work [1], a social recommender system is proposed, in which changes in a consumer's preferences over time are taken into consideration using graph-based clustering. The graph is constructed using user-item and user-user relations and the temporal information is obtained from the timestamp of the item rating.

Dynamic and temporal analysis is insufficiently explored, in the literature, and more research is required to address the drawbacks of the proposed methods. Therefore, to fill this gab, we propose WSC which applies an unsupervised temporal and dynamic learning concept. It is domain-independent due to it being label-free and using a clustering approach (ACAEC) which we tested in our previous paper [2] on review sets from multiple domains. In addition to WSC, segregated window clustering(SWC) is also proposed in which each set of reviews written during a given time window is clustered independently without using information from previous windows.

In this study, review selection is combined with temporal sentiment analysis. The focus is on selecting the most polar reviews in an unsupervised paradigm from a sequence of review windows. Some online platforms provide review selection services. For example, Amazon.com nominates reviews based on how many users have marked a particular review as useful. Many research studies consider the helpfulness of reviews to automatically select reviews [11, 16]. The findings of previous research support our method which is based on review polarity. A review's sentiment significantly influences the helpfulness of the review [38]. The selection is unsupervised and deploys ensemble learning to select reviews based on their polarity from a window series.

3 The automatic contextual analysis and ensemble clustering (ACAEC) algorithm

ACAEC [2] is an unsupervised and automatic method for SA, which incorporates contextual analysis and ensemble clustering. It is domain-independent as it is unsupervised. ACAEC (Fig. 1) utilizes SentiWordNet[2] in which the same word is associated with different sets of sentiment scores. Therefore, the average synset score for each term is calculated which is then used to determine the polar features and form two initial starting centroids of the k-means algorithm, positive S_{pos} and negative S_{neg}.

3.1 Text preparation and contextual analysis

ACAEC uses a detection language tool[3] to eliminate reviews written in languages other than English. The process also involves role-based tasks to remove duplicate reviews and separate non-separated text units which leads to a more accurate detection of sentence boundaries and tokens. Spelling correction is also considered using the Jazzy library.[4] ACAEC addresses three common linguistic forms, namely intensifying, negation and contrast.

- Intensifying: An intensifier is usually an adverb which amplifies the strength of an adjective. Let D be a dictionary of pairs of adjectives and adverbs with the same or similar sentiment scores $\left(d_j, d_j'\right)$, $D = \left\{\left(d_1, d_1'\right), \left(d_2, d_2'\right), \ldots, \left(d_m, d_m'\right)\right\}$ $(m > 0)$ and $I = \{I_1, I_2, \ldots, I_n\}$ $(n > 0)$ is a list of intensifiers. In sentence $S = \{w_1, w_2, \ldots, w_l\}$ $(l > 0)$ if there exists $k(k > 0)$ such that $w_k = I_i$ and $w_{k+1} = d_j(1 \le i \le n \text{ and } 1 \le j \le m)$, then I_i is an intensifier of d_j, and I_i can be replaced by d_j' which is a pair of d_j.
- Negation: Negation is handled by substituting adjectives and adverbs, which follow negation words, with their opposite sentiment words in an antonym dictionary. This dictionary contains pairs of adjectives and adverbs extracted from SentiWordNet. The antonym pairs are antonyms in terms of sentiment polarity, disregarding their semantic meaning.
- Contrast: When a contrast word appears in a sentence, part of the sentence will be a clause by which the overall author's opinion is expressed. Let $S = \{w_1, w_2, \ldots w_m\}$ be a sentence which contains a contrast word, and $C = \{c_1, c_2, \ldots c_n\}$ be a set of contrast words. If $c_i \in S$, all words w_j of the revoked clause will be removed and the words in the conclusion clause will be retained.

3.2 Ensemble clustering

The ensemble clustering of ACAEC operates a modified k-means algorithm on several vector space models (VSMs) of the processed data. The results of clustering VSMs are combined using an unweighted voting mechanism. VSMs are built using adjectives and adverbs as features, and each document is a vector in a VSM. To construct varying VSMs, presence and frequency matrices with different weight schemes are used.

ACAEC separates positive and negative features using their SentiWordNet score to create positive S_{pos} and negative S_{neg} sets. These two polar sets are then used as the initial starting centroids of the k-means algorithm. Using S_{pos} and S_{neg} enhances accuracy significantly and also stabilizes the performance of k-means. Moreover, the assignment of the two polar seeds, S_{pos} and S_{neg}, to the produced clusters is used to identify the polarity of the clusters. As these seeds are highly polar, each seed can be easily assigned to the corresponding group. ACAEC assembles the results of the base learners using unweighted voting to decide the membership of each review r_i in a set of reviews R (refer to algorithm 1).

4 Temporal analysis

The analysis of data within a sequence of time segments is suited to online SA. It can track sentiment shift of an already available set of reviews over time. It results in an informative sentiment pattern which can give a close-to-reality insight into the actual public sentiment.

[2] http://sentiwordnet.isti.cnr.it/

[3] http://labs.cybozu.co.jp/en/

[4] https://sourceforge.net/projects/jazzy

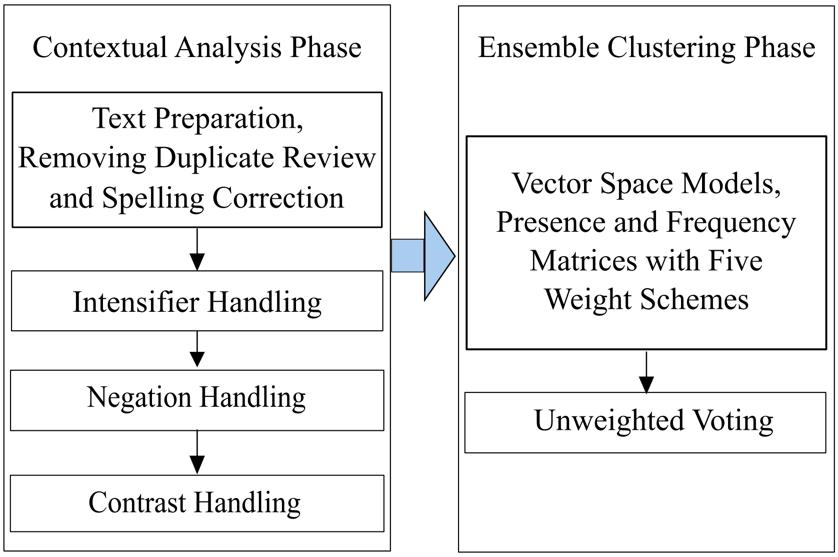
Contextual Analysis Phase
Text Preparation, Removing Duplicate Review and Spelling Correction
Intensifier Handling
Negation Handling
Contrast Handling
Ensemble Clustering Phase
Vector Space Models, Presence and Frequency Matrices with Five Weight Schemes
Unweighted Voting

Fig. 1 Automatic Contextual Analysis and Ensemble Clustering (ACAEC)

```
Algorithm 1 Ensemble voting
INPUT: A review set $R$ contains $m$ number of reviews $\{r_1, r_2, \ldots, r_m\}$
OUTPUT: Assign each review $r_i \in R$, $(i = 1, 2, \ldots, m)$, to the positive or negative cluster
for all $r_j \in R$ do
    for all result $t_i$ of $H_i$ do, where $H_i$ is a base learner
        if $H_i$ is accurate enough then
            if $\sum(r_j(t_i) = positive \geq \sum(r_j(t_i) = negative)$ then
                $r_j$ is assigned to the positive cluster
            else
                $r_j$ is assigned to the negative cluster
            end if
        end if
    end for
end for
```

A review window can be defined by a time period, for instance, a window of reviews that includes all reviews written in a period of 1 week.

Definition 1 *Review Series.* Review Series *RS* is a sequence of reviews $r_1, r_2, \ldots, r_m$. Each is associated with a time stamp.

$RS = \{(t_1, r_1), (t_2, r_2), \ldots, (t_m, r_m)\}$ where r_i is the review written at time t_i, and $t_1 \leq t_2 \leq \ldots \leq t_i \ldots \leq t_m$.

Definition 2 *Review Window.* A review window starts at time g and ends at time $g + l$, $W(g, g + l)$, is a subsequence of a review series *RS* with the earliest review written after or at time g and the latest review written before or at time $g + l$. The number of reviews is the size of the window, denoted as $|W(g, g + l)|$, and l is the length of the window, which is the period of time in which the reviews were written.

For $W(g, g + l) \subseteq RS$, and $W(g, g + l) = \{(t_1, r_1), (t_2, r_2), \ldots, (t_s, r_s)\}$ where $W(g, g + l)$ is review window, with each t_i satisfying $g \leq t_i \leq g + l$, and s is the size of the review series, $t_1 \leq t_2 \leq \ldots \leq t_s$.

Window length l is a parameter which can be decided after taking into consideration the density and the length of a review series. This is because the speed and quantity of a review series is likely to vary according to the entity being reviewed and the online platform on which it is published. The sliding process over a sequence of data can be defined by a period of time (refer to definition 3).

Definition 3 *Slide Time.* A slide time θ is the parameter of the time period between two consecutive windows of the same length, such that $W(g_1, g_1 + l)$ starts at time g_1 and ends at time $g_1 + l$, and $W(g_2, g_2 + l)$ starts at time g_2 and ends at time $g_2 + l$ where $(g_1 < g_2)$, then $\theta = |g_1 - g_2|$.

In accordance with definition 3, a user chooses to operate on non-overlapped windows, or on overlapped consecutive windows when the slide time satisfies $\theta < l$. It is assumed that the overlapping arrangement can result in a smooth representation of a small and gradual change in sentiment over time.

Definition 4 *Window Series.* A window series *WS* is a sequence of fixed length review windows *W* in a chronological order, such that.

$WS = \{(W(g_1, g_1 + l), W(g_2, g_2 + l), \ldots, W(g_m, g_m + l)\}$ where l is the window length and $W(g_i, g_i + l)$ is a review window which starts at g_i and ends at $g_i + l$, $g_1 < g_2 < \ldots < g_m$.

The window length l and the intensity of review series *RS* determines the number of windows. Since each window is of the same length, a sequence of windows can therefore be represented as $\{W_1, W_2, \ldots, W_n\}$.

Definition 5 *Sentiment orientation.* In temporal sentiment analysis *TSA*, let (t, r) be a review sample where review r was posted at time t. Let $WS = \{W_1, W_2, \ldots, W_n\}$ be the sequence of analysing windows where window W_i starts and ends at time g and $g + l$. The review sample (t, r) belongs to W_i, denoted as $(t, r) \in W_i$ if $g \le t \le g + l$. $SO((t, r), W_i)$ is denoted as the sentimental orientation of (t, r) in the window W_i analysis.

$$SO((t,r),W_i) = \begin{cases} \text{positive if } (t,r) \text{ is assigned} \\ \text{to the positive cluster} \\ \\ \text{negative if } (t,r) \text{ is assigned} \\ \text{to the negative cluster} \end{cases} \qquad (1)$$

The output of processing *WS* is a traced pattern of the sentiment polarity over time. Time-based window clustering can also assist in specifying the key factors contributing to sentiment change.

The ACAEC algorithm [2] is used to conduct temporal clustering using SWC and WSC. The algorithm is modified to perform sequential dynamic clustering (WSC). The performance of SWC and WSC is investigated using ACAEC and its enhanced version ACAEC+. To understand the produced pattern, we propose an unsupervised review selection based on a review's polarity which is combined with window series clustering.

4.1 Window sequential clustering (WSC)

WSC (Fig. 2) applies a dynamic concept to compute the proportion p of sentiment polarity using information from a previously clustered window to process the following window (refer to Algorithm 2). To implement WSC, ACACE is modified by using block sequential learning to perform the clustering ensemble instead of using non-sequential k-means to cluster a series of windows $WS = \{W_1, W_2, \ldots, W_n\}$. One of ACAEC's important settings is dividing the feature set into positiveS_{pos} and negative S_{neg} feature sets, then using these two sets as seeds and initial starting centroids for H, a set of base learners of ACAEC. For the first window W_1, the learners of ACAEC use these polar seeds as initial starting centroids. Then, the block sequential learners use the means $M_{(j,i-1)}$ of cluster j, $j = \{1, 2\}$ and previous window W_{i-1} to cluster W_i. To enhance accuracy, we combine the seed $S_{(j,i)}$, $j = \{1, 2\}$, that is S_{pos} or S_{neg}, of window W_i with the mean $M_{(j,i-1)}$ of window W_{i-1} (Eq. (2)).

$$CM_{(j,i)} = \frac{1}{2}\left(M_{(j,i-1)} + S_{(j,i)}\right) \qquad (2)$$

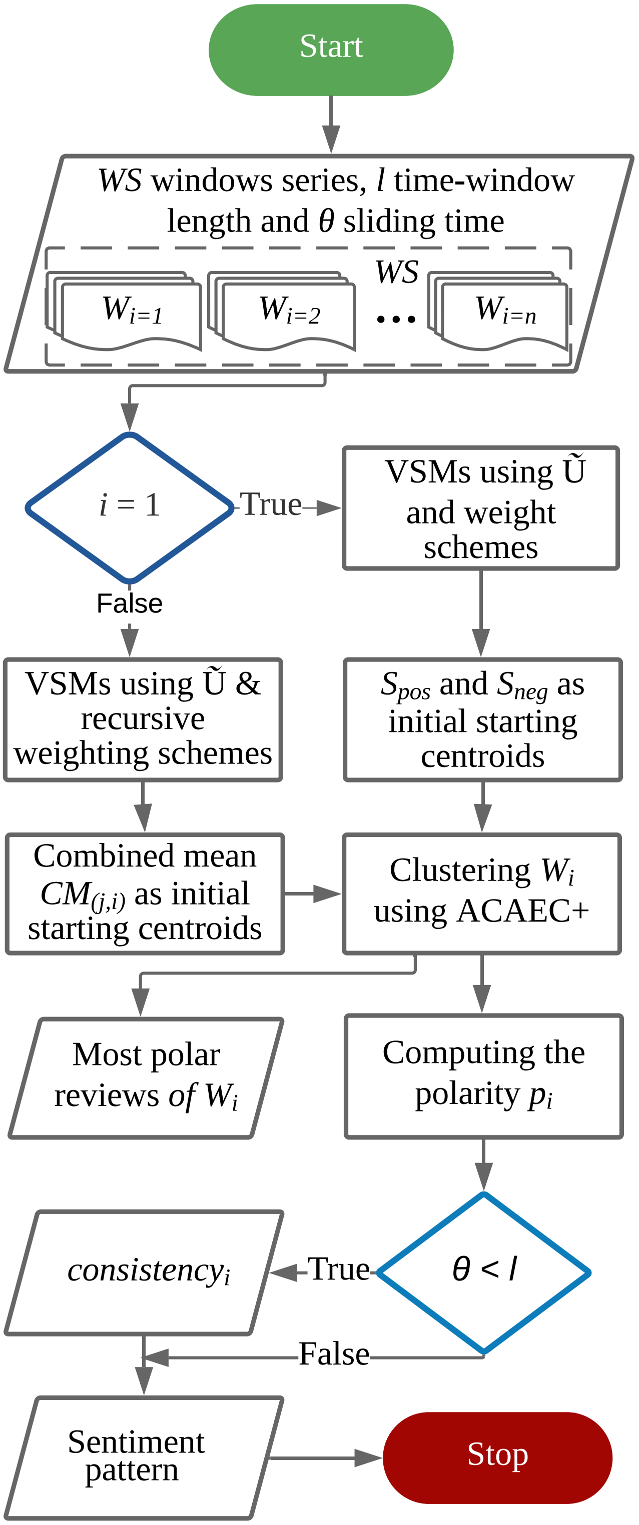
Start
WS windows series, l time-window length and θ sliding time
WS
$W_{i=1}$
$W_{i=2}$
...
$W_{i=n}$
$i = 1$
True
VSMs using Ũ and weight schemes
False
VSMs using Ũ & recursive weighting schemes
S_{pos} and S_{neg} as initial starting centroids
Combined mean $CM_{(j,i)}$ as initial starting centroids
Clustering W_i using ACAEC+
Most polar reviews of W_i
Computing the polarity p_i
$consistency_i$
True
$\theta < l$
False
Sentiment pattern
Stop

Fig. 2 Window Sequential Clustering (WSC)

where $CM_{(j,i)}$ is the combined mean, and $S_{(j,i)}$ is the seed (i. e. S_{pos} or S_{neg}).

However, to enable this dynamic analysis, some changes need to be made. Firstly, in ACAEC, the number of features is not fixed, which impedes the correlated processing of a window series. Therefore, all adjectives and adverbs contained in SentiWordNet are used as a feature set Ũ, instead of extracting the features from the processed set of reviews U. This means WSC's feature set is fixed and predefined for any window or dataset. Consequently, S_{pos} and S_{neg} are also fixed and predefined. Using a predefined feature set allows the correlated processing of means between windows by representing every document in all windows by the same set of features. Secondly, weighting schemes are computed recursively to construct the matrices using a look-up table which is a table that stores the numbers of documents and term frequency from every new window of review.

4.2 Segregated window clustering (SWC)

SWC (Fig. 3) is a separated clustering process where each review window is clustered without using any outcomes from the previously clustered windows (refer to Algorithm 3). SWC clusters each window W_i, in a series of windows $WS = \{W_1, W_2, \ldots, W_n\}$, using S_{pos} and S_{neg} as initial starting centroids.

In SWC, each window has a different set of features because the features are extracted from the processed review window. Thus, a correlation between the means of two windows in the window series is hard to find in contrast to WSC where the feature is unified for all windows. On the other hand, the computational complexity of SWC is lower than WSC as SWC's feature set usually contains less features. SWC and WSC can be implemented on non-overlapped windows, that is $W_i \cap W_{i+1} = \varnothing$, or overlapped windows i.e. $W_i \cap W_{i+1} \neq \varnothing$. As reviews are non-stationary data, overlapped windows can be useful if the processed data are small or the intent is to capture a small change in the sentiment.

Algorithm 2 Window Sequential Clustering (WSC)

INPUT: A series of windows $WS = \{W_1, W_2, \ldots, W_n\}$, θ and l

OUTPUT: A positive or negative pattern of the sentiment polarity over time

```
1:  for all $W_i \in WS$ do, where $W_i$ has $m$ number of reviews, $l$ time period and $\theta$ sliding time
2:      if $i = 1$ then
3:          Construct VSMs using Ũ and weighting schemes
4:          Set $S_i = \{S_{pos}, S_{neg}\}$, as initial starting centroids for $H$
5:          Cluster $W_i$ into $v$ number of positive reviews
s           and $g$ number of negative reviews
6:          Compute $p_i$, $p_i = y/m_i$, where $y = \{v, g\}$
7:          Select $r$ number of most polar reviews
8:          if $\theta < l$ then
9:              Compute $consistency$
10:         end if
11:     else
12:         Construct VSMs using Ũ and recursive weighting schemes
13:         Set $CM_{(j,i)}$ as initial starting centroids for $H$ of ACAEC+
                $CM_{(j,i)} = \frac{1}{2}(M_{(j,i-1)} + S_{(j,i)})$
14:         Do steps 5–10
15:     end if
16: end for
```

Algorithm 3 Segregated Window Clustering (SWC)

```
INPUT: A series of windows $WS = \{W_1, W_2, ..., W_n\}$, $\theta$ and $l$
OUTPUT: A positive or negative pattern of the sentiment polarity over time
for all $W_i \in WS$ do, where $W_i$ has $m$ number of reviews, $l$ time period and $\theta$ sliding time
    Construct VSMs using U and weight schemes
    Set $S_i = \{S_{pos}, S_{neg}\}$, as initial starting centroids for $H$
    Cluster $W_i$ into $v$ number of positive reviews and $g$ number of negative reviews
    Compute $p_i$, $p_i = y/m_i$, where $y = \{v, g\}$
    Select $r$ number of most polar reviews
    if $\theta < l$ then
        Compute $consistency$
    end if
end for
```

4.3 Consistency

In temporal sentiment analysis, consistency is an important indicator that reveals the stability of the analysis. Consistency is calculated based on the percentage of review samples that do not change their sentiment orientation in different windows throughout the analysis (Fig. 4).

Definition 6 *Consistent.* In temporal sentiment analysis *TSA*, let (t, r) be a review sample. Let $WS = \{W_1, W_2, \ldots, W_n\}$ be the sequence of analysing windows. Assume the sample (t, r) appears and only appears in h windows. These are $W_i, W_{i+1}, \ldots, W_{i+h}$. That is $(t, r) \in W_{i+j}\, j = 0, 1, \ldots, h$. The analysis *TSA* is consistent on the sample (t, r) if

$$SO((t,r), W_i) = SO((t,r), W_{i+1}) = \ldots = SO((t,r), W_{i+h})$$

The analysis *TSA* being consistent on review (t, r) means that (t, r) is always classified as a positive/negative review throughout the analysis. The sentiment orientation of the sample does not change. For example, point b shown in Fig. 5, is positive, the algorithm is consistent in clustering b if b is assigned to a positive cluster in windows W_1, W_2, W_3 and W_4.

Definition 7 *Consistent Set.* In temporal sentiment analysis *TSA*, let $RS = \{(t_1, r_1), (t_2, r_2), \ldots, (t_m, r_m)\}$ be the set of review samples. $t_i \le t_j$ if $i < j$. Let $WS = \{W_1, W_2, \ldots, W_n\}$ be the sequence of analysing windows. The consistent set of *TSA*

$$CRS = \left\{(t_i, r_i) \in RS | TSA \text{ is consistent on sample}(t_i, r_i)\right\}$$

Obviously, *CRS* is a subset of *RS*. That is $CRS \subseteq RS$. Consistency measures can be applied if *WS* is overlapped. Allowing overlapped windows in the analysis will fine-tune the outcome, and possibly enhance accuracy. When overlapped consecutive windows adjustment is applied, the slide time satisfies $\theta < l$ and $W(g, g + l) \cap W(g + \theta, g + \theta + l)$ is a non-empty set (definition 3).

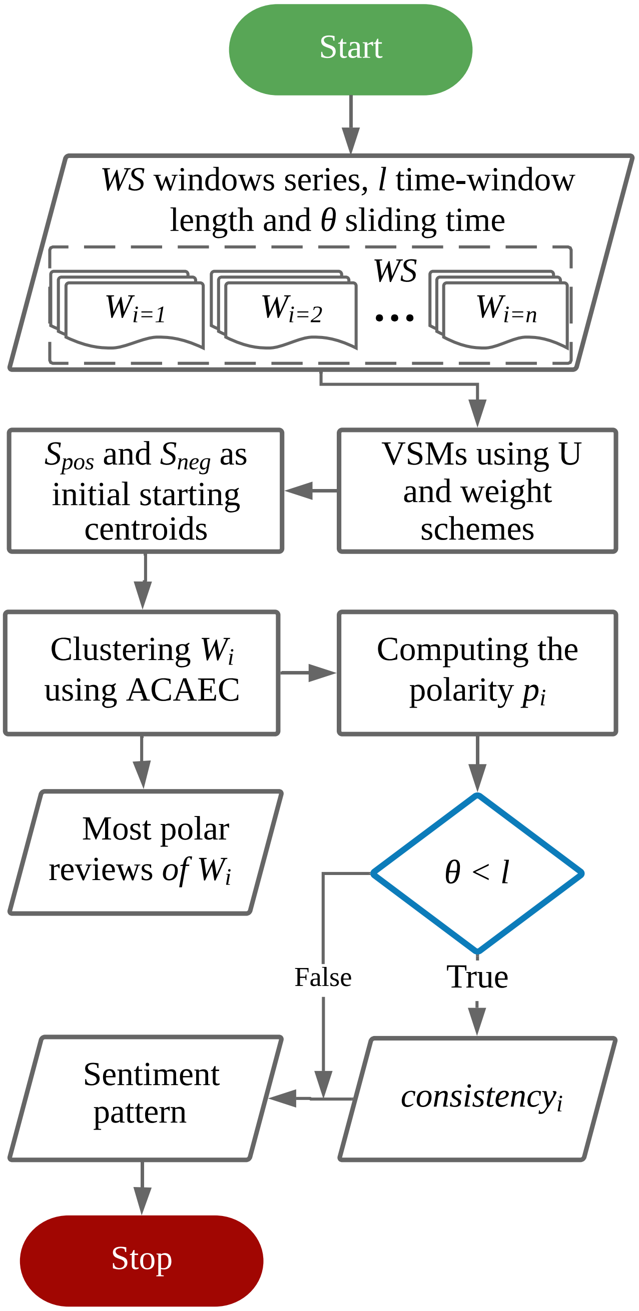
Start
WS windows series, *l* time-window length and θ sliding time
WS
$W_{i=1}$
$W_{i=2}$
…
$W_{i=n}$
VSMs using U and weight schemes
S_{pos} and S_{neg} as initial starting centroids
Clustering W_i using ACAEC
Computing the polarity p_i
Most polar reviews of W_i
θ < *l*
False
True
Sentiment pattern
$consistency_i$
Stop

Fig. 3 Segregated Window Clustering (SWC)

Definition 8 *Consistency.* In temporal sentiment analysis *TSA*, if *RS* is the set of review samples and *CRS* is the consistent set, the consistency of the analysis

$$consistency(TSA) = \frac{|CRS|}{|RS|} \tag{3}$$

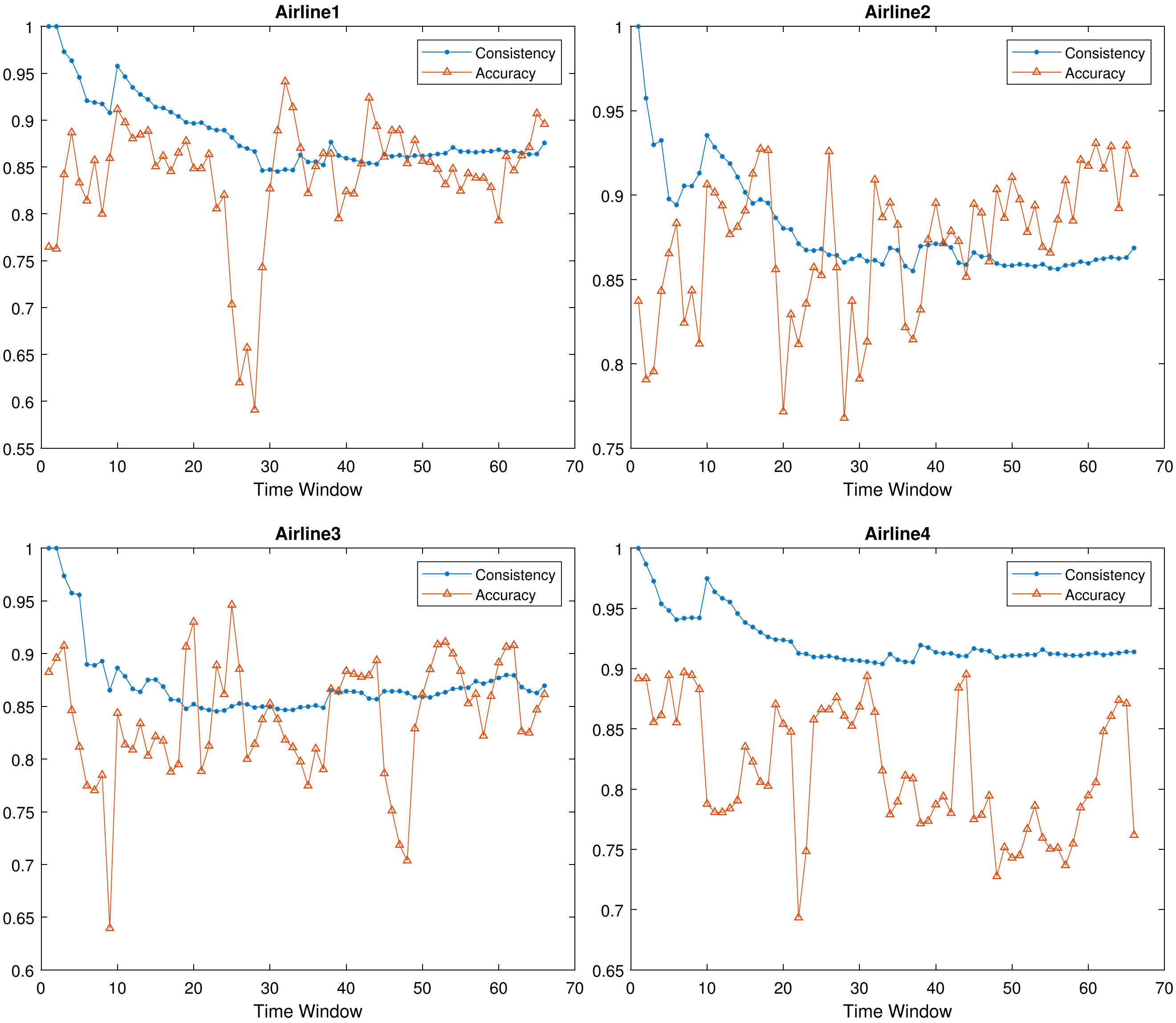

Fig. 4 Consistency and accuracy rates using SWC and overlapped windows

As indicated previously, consistency is an important measure of the stability of the analysis. The higher the consistency, the more stable the analysis, and hence the more reliable the outcome. In real-world applications where the accuracy of the

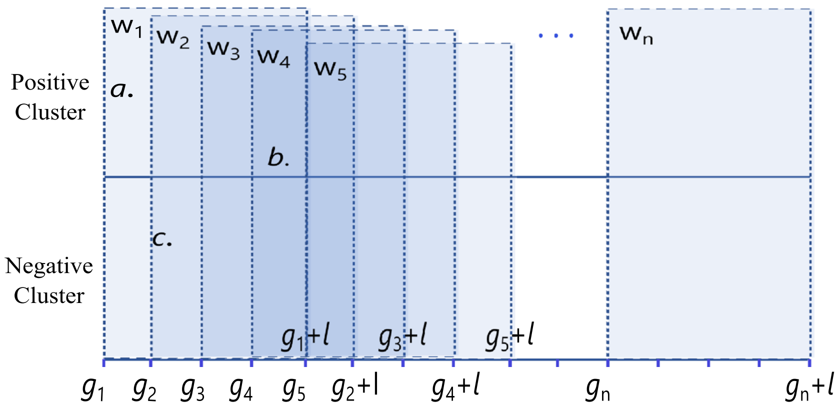

Fig. 5 Diagram of overlapped windows structure

analysis cannot be obtained, we can still calculate consistency. Hence, consistency can effectively reveal the quality of the analysis.

4.4 Review selection

Review selection, with respect to a specific time period for a review stream, provides a better understanding of sentiment patterns. It can effectively assist both providers and beneficiaries in decision making. For a provider, the selected reviews can provide information and give a useful insight into the potential reasons behind negative/positive sentiment change. This can assist in making the right decision to effectively address the drawbacks resulting in negative sentiments and enhance or maintain the factors that contribute to positive sentiments. For a customer, knowing the potential causes of sentiment shift can provide a clear picture that helps in deciding what may suit a customer's preferences and specifications more precisely.

We focus on selecting reviews that can be used to highlight the causes of sentiment change over time; therefore, these reviews can serve this purpose even if they are brief. This method is distinguished from the solutions suggested in the literature as it selects reviews in an unsupervised learning paradigm from a consecutive sequence of sliding time windows. Unsupervised review selection is suitable for the proposed framework as it uses unsupervised learning. It also selects reviews based on the strength of their polarity. To select the most positive reviews and the most negative reviews from each window, we conduct experiments using three measurements.

4.4.1 Distance from the seeds (DS)

DS measures distance $DS(X_i, S_i)$ of each point X_i, where $X_i \in C_i$, from the polar seed S_i that is S_{pos} or S_{neg} of cluster C_i. The review with the smallest distance from the seed will be selected from cluster C_i as most polar review in the group.

4.4.2 Seed silhouette coefficient (SSC)

SSC uses the polar seed S_i that is S_{pos} or S_{neg} of the clusters to measure the cohesion and separation of each point. The reviews with a maximum SSC, given in Eq. (4), are the most polar reviews.

$$SSC(X) = \frac{d\left(X_j, S_i\right) - d\left(X_j, S_j\right)}{max\left\{d\left(X_j, S_i\right), d\left(X_j, S_j\right)\right\}} \tag{4}$$

where $d(X_j, S_j)$ is the distance from point X_j, $(X_j \in C_j)$ to the seed S.

4.4.3 Silhouette coefficient [9] (SC)

The SC measures the cohesion and separation of each point. SC value varies between −1 and 1, indicating how well-clustered a review is. Therefore, reviews which have a maximum of SC are selected from each cluster produced by each learner.

$$a = \frac{1}{n} \sum_{i=1}^{n} d\left(X_{\left(X \in C_j\right)}, A_i\right)$$
$$b = \frac{1}{m} \sum_{j=1}^{m} d\left(X_{\left(X \in C_j\right)}, A_j\right)$$

$$SC(X) = \frac{a - b}{max\{a, b\}} \tag{5}$$

where cluster C_i contains n data points and cluster C_j contains m data points, and $d(X, A)$ is the distance between X and A.

When selecting reviews based on their polarity, the selection can be inaccurate if the polarity of a selected review is opposite to the polarity of its cluster. In other words, some of the selected reviews can be selected from wrongly clustered reviews. For a more accurate review selection, we use Kullback–Leibler [5] divergence, expressed by Eq. (6). This enables reviews to be selected only from distant clusters which indicates a relatively high-quality clustering.

$$D_{KL}(X \| Y) = \sum_i X(i) \log\left(\frac{X(i)}{Y(i)}\right) \tag{6}$$

where X and Y are the clusters' means.

After selecting a set of reviews from the clusters which are produced by each learner, the polar reviews from each window of review are selected. Using a majority vote with ensemble learning, the polar documents are those which are selected by most base clustering algorithms from the assembled learners. We assume that the selected reviews contain useful information, either positive or negative, that can assist in learning about the causes at a particular sentiment polarity of a window in time. If overlapped window processing is applied, the same review may be selected from different consecutive review windows. Therefore, instead, the next most polar review can be selected.

4.5 The enhancement of ACAEC (ACAEC+)

Adopting sequential learning for WSC negativity affects performance, hence ACAEC is enhanced by increasing the number of the ensemble components. This varies the base learners of the ensemble and improves its outcome even if the accuracy of some of the component learners is relatively low [23]. We refer to the enhanced version of ACAEC as ACAEC+. Ensemble learning can result in higher accuracy compared to a solo learner if the combined learners are accurate and diverse [15]. According to Dietterich [10], an accurate learner is a learner whose accuracy is better than random guessing. Therefore, the idea is to increase the number of diverse and accurate learners to contribute to the decision regarding each document membership. This adjustment leads to a satisfactory level of accuracy, reinforces the generalization power of the method and maintains stable performance. Moreover, it addresses the imbalanced data issue when using WSC.

4.5.1 Additional learner

To overcome the limitation of k-means and strengthen the method, another learner is added to operate on the same data representations then its results and k-means' results are assembled to decide each document's membership. The additional learner is a simple clustering method in which two positive and negative points are used in the process. In WSC, the points produced by Eq. (2) are used to assign each document to the positive group if it is closer to the positive point or to the negative group if it is closer to the negative point using cosine distance, except for the first window in the series, where the polar seeds (S_{pos} and S_{neg}) are used. When operating SWC using ACAEC+, the polar seeds (S_{pos} and S_{neg}) are used to form the additional learner (Fig. 7).

4.5.2 Additional weight (BM25) [29]

This additional weight scheme also aims to increase the number of ensemble components using additional data representations which are processed by the two clustering algorithms of the ensemble. BM25, mathematically expressed by Eq. (7), is a probabilistic model which takes into account the structure of the document and uses tuning parameters to calibrate the given document's quantities.

$$BM25 = \frac{IDF \times ((k+1)T)}{k\left((1-b) + b \times \left(\frac{|L|}{avgDl}\right)\right) + T} \tag{7}$$

where usually $k = 1.2$ to determine the impact of term frequency T, and $b = 0.75$, $(0 \leq b \leq 1)$ to modulate the influence of document length. $|L|$ is the document length and $avgDl$ is the average document length in the text. The described improvement has strengthened the overall performance of WSC as shown in subsequent sections.

4.6 Data imbalance

The window-based ensemble method deals with review windows which have a limited number of reviews. Therefore, the imbalanced data issue is likely to be encountered in some windows. Imbalanced data refers to a problem in machine learning when data contains a dominating class or classes to which a significant number of instances belong, compared to other classes. It negatively affects the learning process in both supervised and unsupervised paradigms [14]. Therefore, a considerable amount of research work has been done to address this problem. This issue is commonly addressed by redistributing data such as using oversampling which increases the computational complexity and undersampling which leads to a loss of important data.

As the dataset contains the actual labels, we are able to measure the data imbalance using Eq. (8). We experimentally find that the impact on accuracy is caused by the difference between the number of positive features and the number of negative features $\triangle$, $\triangle = (\mathcal{NF}/\mathcal{F}) - (\mathcal{PF}/\mathcal{F})$, where $\mathcal{NF}$ and $\mathcal{PF}$ are the number of negative and positive features, respectively and $\mathcal{F} = \mathcal{NF} + \mathcal{PF}$. Usually, $\mathcal{NF}$ is greater than $\mathcal{PF}$. As shown in Fig. 4, a gradual reduction of the number of positive reviews does not cause a noticeable change in $\triangle$ nor in accuracy rates. However, $\triangle$ drops and consequently the accuracy rate also drops when the number of negative reviews is reduced. It is concluded that an imbalance in data is an issue when the positive class is a dominating class. Experimentally, a high $\triangle$ results in a high difference between the distance averages (that is the average of distances from each point to the centroid) of clusters. This finding can be used in further research to handle the imbalance in data. For example, if $\triangle$ is small we can assume that a given review set is mainly positive, and the clustering can be biased to the positive class (Fig. 6).

$$data\_imbalance = \left|\frac{(a+c) - (b+d)}{a+b+c+d}\right| \tag{8}$$

5 Experiments and analysis

In this section, we evaluate the concepts and the introduced methods by conducting experiments and analysis on real-world sets of reviews. The accuracy rate can be calculated by using a confusion matrix (Table 1) and Eq. (9).

$$accuracy = \frac{a+d}{a+b+c+d} \tag{9}$$

where the summation of a, b, c, and d, the number of reviews in window W_t at time t. The clusters are identified as positive or negative by the allocation of S_{pos} and S_{neg} such that a positive cluster is a cluster which contains the positive seed S_{pos} and a negative cluster is a cluster which contains the negative seed S_{neg}.

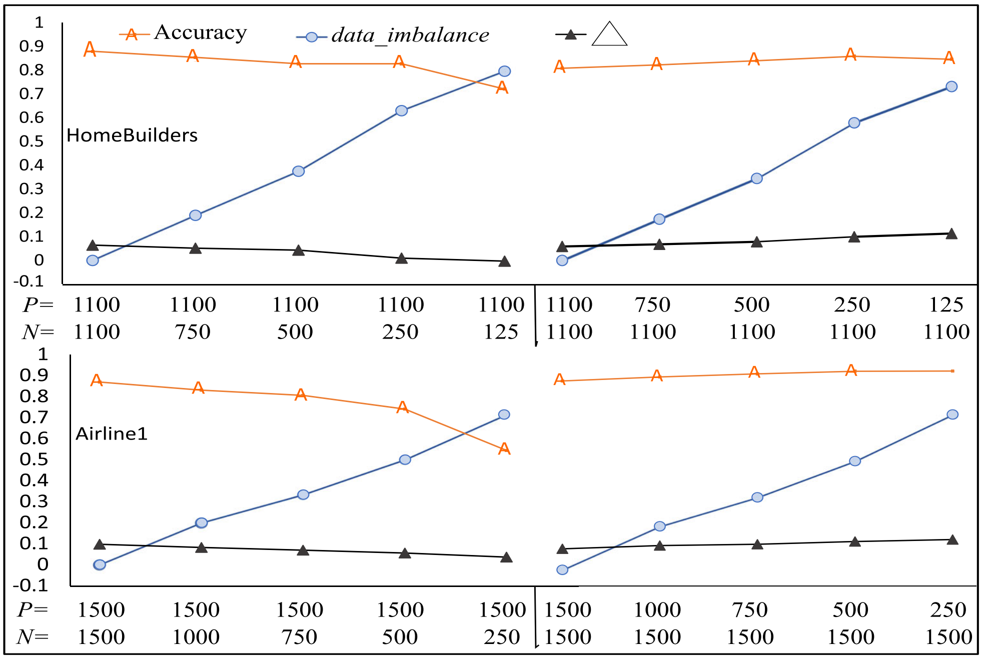

Fig. 6 Experiment are conducted using HomeBuilders [2] and Airline1 datasets, the base learner of ACAEC, which is k-means with S_{pos} and S_{neg} as initial starting centroids and using the presence matrix with TFIDF weight scheme. P and N are the number of positive and negative reviews, respectively

5.1 Datasets

The experiments are conducted on sets of reviews on airlines and an Australian property agent (refer to Table 2).[5] The airline datasets are collected from Tripadvisor.com. On this platform, the authors can associate their reviews with a star ranking where a one-star rank is the most negative and a five-star rank is the most positive. In our experiments, reviews ranked four or five stars are labeled positive and reviews ranked one or two stars are labeled negative, whereas reviews ranked three stars are considered neutral and are ignored. The Australian property agent review set is scraped from the productreview.com website. In this dataset, all reviews associated with excellent or good categorizations, are labeled positive and all reviews associated with bad or terrible categorizations, are labeled negative.

5.2 Experiment settings

In Fig. 7 and Tables 3, 4 and 5 we use a non-overlapped window series, where any window W_{t_1} at time t_1 does not share reviews with window W_{t_2}. Subsequently, $W_{t_1} \cap W_{t_2}$ is an empty set as $\theta = l$, where θ is the slide time and l is the window length. The experiment settings result in 16 non-overlapped review windows for airline datasets and 21 non-overlapped review windows for the property agent dataset. The window's length l is a period of 2 months for the three airline datasets and 3 months for the property agent dataset. To apply the *consistency* measure, we conduct temporal analysis through a sequence of overlapped windows using SWC. The experiment results shown in Fig. 4 are the outcome of analysing four datasets using a time window of 2 months length l and 2 weeks' slide time θ. This setting results in 66 overlapped review windows.

Table 1 Confusion matrix

	Actual positive	Actual negative
Predicted positive	a	b
Predicted negative	c	d

In Table 5, WSC and SWC are compared to unsupervised and supervised techniques. Two data representations, presence and frequency matrices with the TFIDF weight scheme, are used in the experiments. These representations are generated by using adjectives and adverbs as features and after applying ACAEC's contextual rules on the text. The accuracy rate of each unsupervised method is the average accuracy of 20 runs, whereas 10-fold cross-validation is used to evaluate the supervised methods XGBoost [7], support vector machine (SVM), naive base (NB), random forest (RF) and SPegasos. For incremental XGBoost and NADINE [34], a model is trained in a first window W_1 and in the following windows W_i, $(1 > i)$, it will be evaluated then updated. VADER [18] is a rule-based method, thus the evaluation is conducted using the confusion matrix and accuracy. For the method proposed by Lukeš and Søgaard [28] which uses a logistic regression (LR) classifier and a feature selection method, the feature set is selected from the Airline4 dataset. Selected features must pass the significance test $\mid \mu_2 - \mu_1/\mu_1 \mid > 0.05$ and pass the threshold 0.05 of the p value which is obtained using a linear regression trend line.

Multiple reviews can be selected from each window as the most polar reviews. Two reviews are selected from

Table 2 Four different real-world datasets are utilized to evaluate the proposed framework

Datasets	Number of reviews		Dates		Platform
	Positive	Negative	From	To	
Airline1	1509	1768	Feb-2016	Oct-2018	Tripadvisor.com
Airline2	1158	1415	Feb-2016	Oct-2018	
Airline3	1761	1620	Feb-2016	Oct-2018	
Airline4	19,418	8608	Feb-2016	Oct-2018	
Property agent	501	448	Mar-2012	Jul-2017	
HomeBuilders [2]	1100	1100	–	–	Productreview.com

each cluster which means there are four reviews from each window. The number of reviews to be selected from each cluster can be decided empirically. In our experiments, two reviews from the cluster can provide clues to sentiment change. Table 6 shows the error rates when applying the measurements illustrated in subsection 4.4. Error in review selection is a result of selecting reviews from wrongly clustered reviews. It was experimentally found that the best results are obtained when reviews are selected from clusters with Kullback–Leibler divergence higher than 20.

Figure 7 shows the accuracy rates of WSC and SWC using ACAEC+ and ACAEC. The figure also show the relativity between the accuracy rates and the data imbalance. Tables 3 and 4 present the average accuracy rates of the processed windows using ACAEC and ACAEC+. WSC performance is significantly improved when using ACAEC+ compared to using ACAEC (refer to Table 3), whereas the performance of SWC drops when using ACAEC+ (refer to Table 4).

5.3 Discussion

In this paper, we introduce two temporal SA methods which are time-window-based, namely window sequential clustering (WSC) which utilizes a dynamic clustering concept where every window uses information from the previously clustered windows and segregated window clustering (SWC) in which review windows are clustered with no previous information. In SWC, the set of features changes for every window and is of an unfixed size because the features are extracted from the processed data within a given window. This means it is comparatively a small set of features, hence it is computationally more efficient. However, WSC's feature set is unified and fixed for every window and dataset, and usually its size is larger compared to SWC's feature set. The predefined feature set of WSC is technically useful for more advanced development that requires an analysis of the correlation between the windows.

5.3.1 Comparison

A comparison is conducted between the proposed methods and the supervised incremental, supervised and unsupervised methods (refer to Table 5). Based on the average accuracy rates, WSC and SWC significantly outperform the clustering techniques, while achieving competitive performance compared to the supervised methods, such as XGBoost, SVM and LR [28] classifiers. It is a notable improvement to obtain better or more competitive results using the unsupervised SA method. Also, WSC has higher accuracy in general and a more stable performance in comparison to the incremental learning methods. The proposed methods provide an inexpensive solution by being label-free and they also allow real-time analysis. Our methods provide better performance compared to VADER, as WSC and SWC yield higher accuracy on three of the datasets.

5.3.2 Performance

The performance of both methods, which were successful in processing review series *RS*, is shown in Fig. 7. Experiments show that the average accuracy rates of SWC and WSC reach 87.54% and 83.87%, respectively. SWC yields better results compared to WSC, as can be seen in Tables 3 and 4, where the average of the accuracy rates of the windows is reported. This is mainly because the base learners of ACAEC and ACAEC+ of WSC are sequential and they use combined initial points, whereas SWC employs a non-sequential learner in which the polar seeds S_{pos} and S_{neg} are the initial starting centroids. Thus, the combination of the sequential learners is less effective when WSC operates ACAEC compared to using ACAEC+, which is mainly due to a drop in the recall rate (refer to Table 3). We address this by enhancing ACAEC using additional learner and additional weight scheme in the improved version ACAEC+.

SWC yields contrasting behaviour, as ACAEC outperforms ACAEC+ (refer to Table 4 and Fig. 7). The reason for this is that the weak base learners which result in the lower accuracy of ACAEC+ outnumber the learners with higher

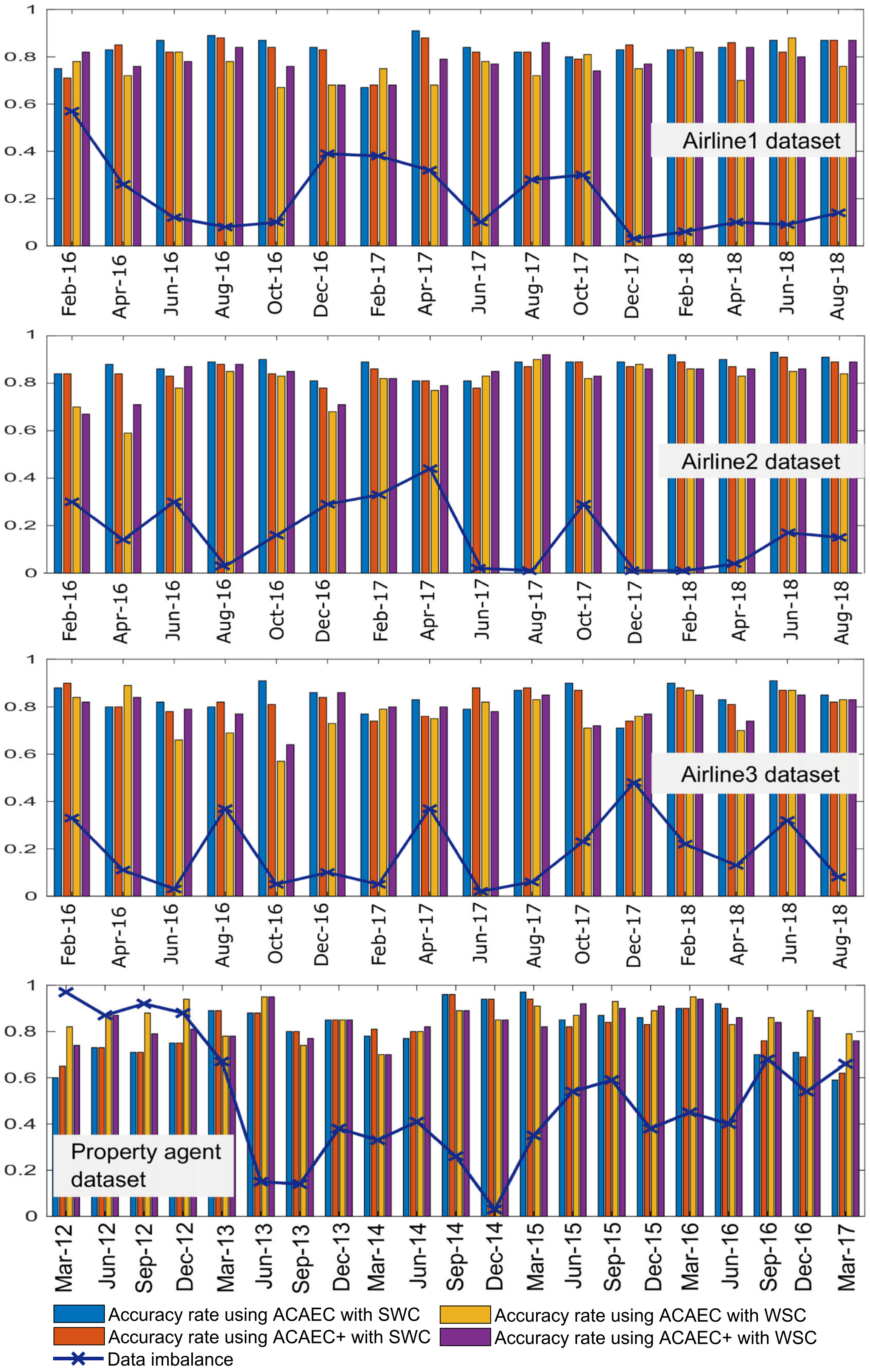

Fig. 7 Shows data imbalance and the accuracy rate of clustering each window using ACAEC and ACAEC+ with SWC and WSC

accuracy. In general, as non-sequential learner initially starts with the polar seeds as centroids, it is more effective compared to sequential learner which uses combined initial points.

5.3.3 Consistency measure

Performance can be assessed using the consistency measure which is a free-label measure and therefore, it is suitable for assessing the outcome of analysing real-world data. The consistency of the first window is always equal to 1, then decreases as more reviews are shared between time windows. In the experiment, θ is 2 weeks' long and window length l is 2 months' long. This gradual shift over time yields a more reliable *consistency* measure because it allows at least 90% of all reviews to be shared among windows which contribute in calculating *consistency*. In Fig. 4, it can be seen that *consistency* decreases slightly when the accuracy rate is comparatively low. Figure 4 reveals that the analysis of the four datasets is reliable which is mainly a result of ensemble clustering on which ACAEC is based.

5.3.4 Review selection

The temporal analysis is combined with an unsupervised review selection method to extract more information from a given series. We use three selection measurements which are DS, SSC and SC, with three distance measurements namely, cosine, Euclidean and Manhattan distances. Table 6 shows the error rates. One of the reasons behind the reported error is encountering imbalanced windows because when a window is extremely imbalanced, most if not all, of the reviews, are from one class; therefore, if four reviews are selected from this window, two of them will be selected from wrongly clustered reviews. DS results in low error rates for all datasets. Hence, in the following, we use DS with cosine distance to analyze the datasets.

Table 3 Shows the average accuracy rate of clustering window series using ACAEC+ and ACAEC with window sequential clustering (WSC)

Datasets		Accuracy	Precision	Recall	F_measure
Airline1	ACAEC+	78.66	76.89	62.6	68.64
	ACAEC	76.27	80.94	49.1	59.86
Airline2	ACAEC+	82.77	84.26	72.04	77.28
	ACAEC	80.2	88.25	58.29	68.97
Airline3	ACAEC+	79.48	85.85	69.1	76.03
	ACAEC	76.83	89.48	58.41	68.75
Property agent	ACAEC+	83.87	66.23	83.26	67.91
	ACAEC	85.54	69.59	82.66	70.13

Table 4 Shows the average accuracy rate of clustering window series using ACAEC+ and ACAEC with segregated window clustering (SWC)

Datasets		Accuracy	Precision	Recall	F_measure
Airline1	ACAEC+	82.11	76.56	77.95	76.82
	ACAEC	83.28	80.04	78.44	78.47
Airline2	ACAEC+	85.29	82.81	82.91	82.43
	ACAEC	87.54	87.34	83.36	84.85
Airline3	ACAEC+	82.43	84.25	80.46	81.88
	ACAEC	84.01	89.55	78.5	83.19
Property agent	ACAEC+	81.22	64.73	90.65	69.68
	ACAEC	81.12	64.89	91.33	69.74

5.3.5 Data imbalance

In addition to this, WSC and SWC can handle data imbalance to a certain degree. Imbalanced data negatively affects ACAEC's performance which can be observed when applying dynamic learning utilizing WSC. This is because WSC uses sequential k-means as a base learner of ACAEC's ensemble which is sensitive to imbalances in the data [26], since k-means assumes that the data contains clusters of approximately an even number of instances. Utilizing WSC with ACAEC+ addresses imbalanced review windows to a certain degree by using an additional sequential learner and an additional weight scheme. This improves the ensemble outcome and accuracy rate, whereas SWC can handle the imbalanced data issue due to the initialization of k-means and using polar seeds (Fig. 7).

5.3.6 Dataset analysis

Dataset analysis is conducted on new real-world datasets, these being three airline datasets and one Australian property agent dataset. Figures 8 and 9 show the predicted and the actual patterns of the positivity changes in review series RS. The sentiment patterns are produced by computing the proportion p of sentiment polarity for each review window $W_i \subseteq RS$, $p_i = y/m_i$, where m_i is the number of the reviews in W_i and y is the number of reviews of positive/negative cluster. p can be used to measure the predicted and the actual positivity and negativity of each window hence the actual label is available. We conduct experiments using non-overlapped windows for both WSC and SWC methods. In our experiments, the dataset analysis is conducted by computing the positive sentiment polarity (refer to Figs. 8 and 9). Based on the results obtained from using WSC and SWC, the produced sentiment pattern is close to the actual sentiment changes. Figures 8 and 9 also show the difference in the performance of ACAEC and ACAEC+. The results of SWC do not contradict WSC's results; thus, WSC operating ACAEC+ (refer to Fig. 8) is

Table 5 Average accuracy rate of all windows using WSC and SWC compared to the accuracy rate of the incremental methods and also the average accuracy rate of all windows using supervised and unsupervised methods

Method Type	Dataset		Airline1	Airline2	Airline3	Property agent
	WSC using ACAEC+		78.66	82.77	79.48	83.87
	SWC using ACAEC		83.28	87.54	84.01	81.12
Incremental	XGBoost [7]	FrequencyTFIDF	65.84	69.17	73.29	64.38
		PresenceTFIDF	64.63	70.87	73.56	62.59
	NADINE [34]	FrequencyTFIDF	54.35	50.85	46.12	60.25
		PresenceTFIDF	59.08	52.96	51.32	59.03
	HoeffdingTree	FrequencyTFIDF	65.46	72.08	64.56	66.53
		PresenceTFIDF	75.58	77.77	74.17	79.23
	NB	FrequencyTFIDF	62.59	73.96	58.13	66.53
		PresenceTFIDF	63.64	56.22	56.87	56.82
	SVM	FrequencyTFIDF	67.06	80	66.61	75.41
		PresenceTFIDF	82.77	67.43	73.56	86.26
	SPegasos	FrequencyTFIDF	63.58	58.33	58.98	61.47
		PresenceTFIDF	60.62	59.8	61.51	57.02
	VADER [18] (Rule-based)		77.18	82.61	81.12	82.3
Supervised	XGBoost [7]	FrequencyTFIDF	81.02	84.90	82.89	85.22
		PresenceTFIDF	81.36	84.05	83.51	83.88
	LR [28]	FrequencyTFIDF	70.34	74.21	73.96	–
		PresenceTFIDF	70.05	75.86	74.22	–
	SVM	FrequencyTFIDF	82.67	84.58	84.93	81.83
		PresenceTFIDF	83.81	85.19	84.88	84.94
	NB	FrequencyTFIDF	81.02	83.57	84.1	84.16
		PresenceTFIDF	82.62	84.37	84.04	86.28
	RF	FrequencyTFIDF	81.57	83.59	83.66	83.16
		PresenceTFIDF	82.5	84.2	83.77	84.39
Clustering	k-means	FrequencyTFIDF	61.54	67.03	62.84	67.41
		PresenceTFIDF	61.14	66.56	62.07	66.65
	k-means++	FrequencyTFIDF	61.05	67.52	63.06	67.31
		PresenceTFIDF	61.02	67.55	62.14	66.33
	k-means Uniform	FrequencyTFIDF	61.44	66.48	63.2	62.2
		PresenceTFIDF	61.39	66.44	61.99	62.4
	k-medoids	FrequencyTFIDF	66.86	68.83	65.62	65.89
		PresenceTFIDF	66.85	68.64	63.83	67.16
	Clara	FrequencyTFIDF	67.48	69.44	65.79	66.36
		PresenceTFIDF	66.54	69.03	64.63	68.13

utilized to analyze the sentiment pattern of one of the datasets as an example.

Airline1 dataset's sentiment pattern shown in Fig. 8 changes chronologically during a two-year period from February 2016 to August 2018. A negative sentiment is expressed during February 2016 and based on the two selected negative reviews, the reason for the drop in positivity seems to be poor service. Samples sentences from the selected reviews are as follows.

> *"I am glad I'm [sic] not 6 feet tall or my knees would have been by my chin."*
> *"The staff didn't even have the decency to make an announcement"*
> *"[..] there is no service on that flight unless you want to buy overpriced food"*

Positivity starts increasing from April 2016 to October 2016 which could be a result of enhancing some of the services.

Table 6 Error rate using the three measurements with three distances when applying window sequential clustering (WSC) and ACAEC+

Dataset	Distance	SC	SSC	DS
Airline1	Cosine distance	0.23	0.2	0.17
	Euclidean distance	0.36	0.15	0.14
	Manhattan distance	0.19	0.26	0.45
Airline2	Cosine distance	0.42	0.28	0.17
	Euclidean distance	0.5	0.19	0.2
	Manhattan distance	0.23	0.33	0.47
Airline3	Cosine distance	0.59	0.3	0.22
	Euclidean distance	0.55	0.31	0.34
	Manhattan distance	0.34	0.33	0.44
Property agent	Cosine distance	0.44	0.28	0.36
	Euclidean distance	0.44	0.34	0.28
	Manhattan distance	0.33	0.34	0.49

This is expressed in a customer's review. However, there is approximately an equally significant number of negative reviews compared to positive reviews written by customers as positivity is still lower than 50%. Selected segments from the positive review written between April and June 2016 are as follows.

> *"Great direct flight from Edinburgh to Toronto"*
> *"Return flight was early and just as comfortable"*

Positive and negative reviews assist both the customer and the provider of the service or product in their decision making and selecting the most polar review can give an overall insight on the causes of the sentiment pattern. The highest positivity can be seen in the last window, August 2018, which is slightly under 60%. The selected positive and negative reviews

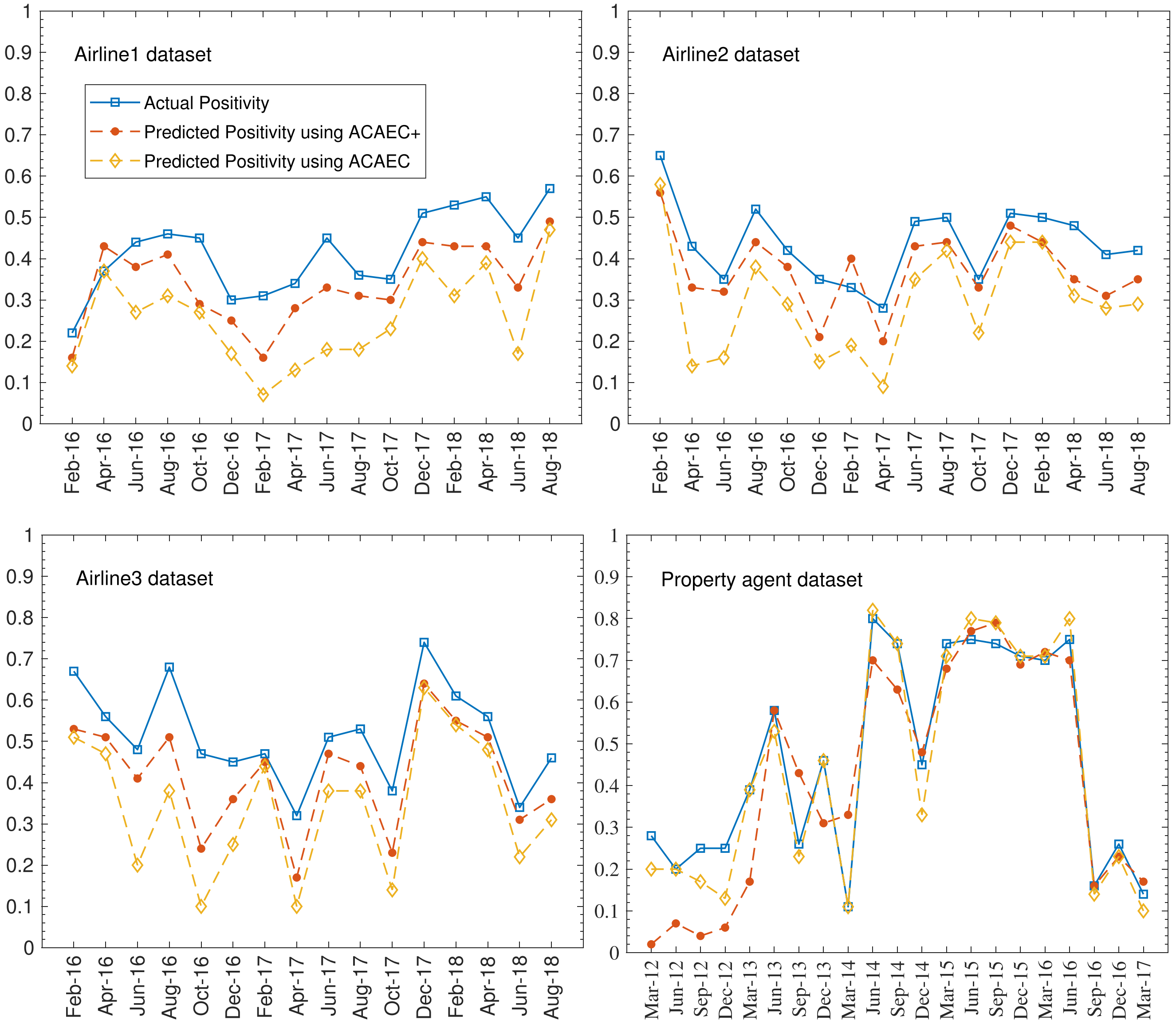

Fig. 8 Predicted positivity using window sequential clustering (WSC): non-overlapped windows

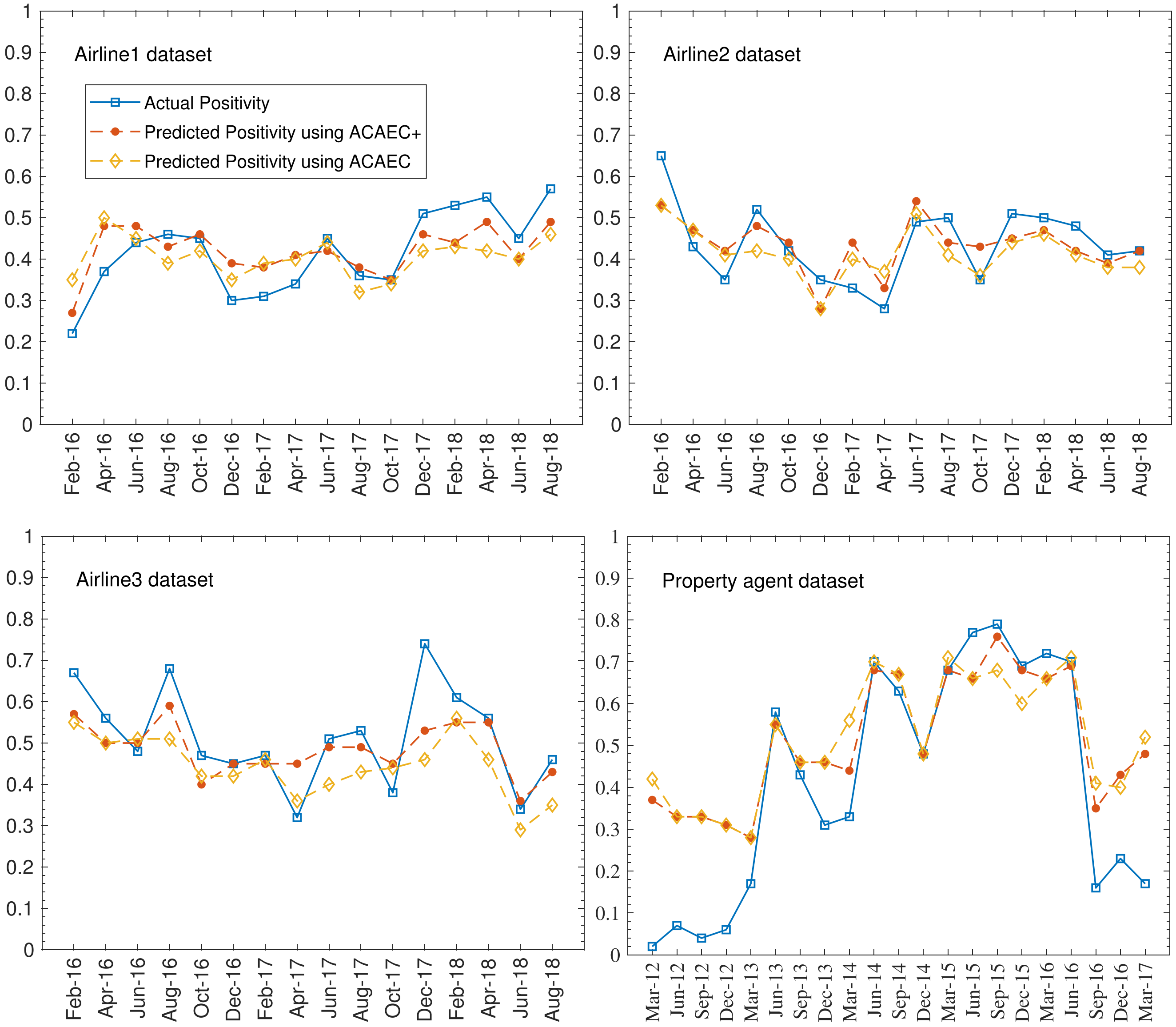

Fig. 9 Predicted positivity using window segregated window clustering (SWC): non-overlapped windows

discuss the cost and the quality of the services. Samples from the selected positive and negative reviews are as follows.

Positive reviews:

"Can't beat the value"

"Airline was easy to deal with and carrier fit easy under the sea"

"Check in was easy and staff friendly"

Negative reviews:

"We were flying with an infant and found the airplane uncomfortable, staff unaccommodating and unfriendly (except for one steward), with no understanding for a family traveling [sic] with a baby."

The analysis of the datasets shows that temporal sentiment clustering incorporated with review selection can be useful and informative. The proposed methods are effective and yield a close-to-reality sentiment pattern.

6 Conclusions

In this article, we introduced two machine learning methods for sentiment analysis (SA) these being automatic and unsupervised for processing review series. The window sequential clustering (WSC) and segregated window clustering (SWC) methods are introduced to adopt ACAEC which combines a contextual analysis and an ensemble of clustering algorithms. The outcome of WSC is improved by proposing ACAEC+ to overcome the drawbacks of k-means such as its sensitivity to imbalanced data. Using these methods to conduct chronological sentiment analysis can be effective and informative as a result of the key features that distinguish them from the other work in the literature. The methods are unsupervised which results in generic domain-independent methods that can process unseen data, and automatic as a minimal human intervention is required. In addition to this, a large volume of data can be processed effectively mainly because the ensemble's components are fast compared to the hierarchical clustering algorithms. The methods can also handle imbalanced review windows as was shown experimentally. The framework enables public sentiment polarity to be tracked by discovering a sentiment pattern using WSC or SWC. Furthermore, the incorporated review selection provides useful information that can assist in understanding sentiment shift. The performance of the algorithm in terms of reliability and stability can be judged using *consistency* which suits the analysis of real-world data.

As future work, the unified feature set of WSC can be utilized to enhance ensemble learning or develop a new algorithm using window correlation. An automatic window setting can be used to decide the length of the windows based on the data density.

Acknowledgements The authors would like to acknowledge the financial support from the Iraqi Ministry of Higher Education and Scientific Research (MoHESR).

References

1. Ahmadian S, Joorabloo N, Jalili M, Meghdadi M, Afsharchi M, Ren Y (2018) A temporal clustering approach for social recommender systems. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1139–1144
2. AL-Sharuee MT, Liu F, Pratama M (2018) Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison. Data Knowl Eng 115:194–213
3. Alves ALF, Baptista CdS, Firmino AA, de Oliveira MG, de Figueirêdo HF (2014) Temporal analysis of sentiment in tweets: a case study with FIFA confederations cup in Brazil. In: Decker H, Lhotsk'a L, Link S, Spies M, Wagner RR (eds) Database and expert systems applications. Springer International Publishing, Cham, pp 81–88
4. Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Paper presented at the 7th international conference on language resources and evaluation (LREC), Valletta, Malta, pp 2200–2204
5. Beyerer J, Richter M, Nagel M (2017) Pattern recognition: introduction, features, classifiers and principles. Walter de Gruyter GmbH & Co KG
6. Chan SW, Chong MW (2017) Sentiment analysis in financial texts. Decis Support Syst 94:53–64
7. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 785–794
8. Chen C, Wang Z, Lei Y, Li W (2016) Content-based influence modeling for opinion behavior prediction. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 2207–2216
9. Cichosz P (2015) Data mining algorithms: explained using R. Wiley, Chichester
10. Dietterich TG (2000a) Ensemble methods in machine learning. In: Multiple classifier systems. Springer Berlin Heidelberg, pp 1–15
11. Eslami SP, Ghasemaghaei M, Hassanein K (2018) Which online reviews do consumers find Most helpful? A multi-method investigation. Decis Support Syst 113:32–42
12. Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56:82–89
13. Fukuhara T, Nishida T (2007) Understanding sentiment of people from news articles: temporal sentiment analysis of social events. In: Proceedings of international conference on weblogs and social media (ICWSM)
14. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
15. Hansen LK, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12:993–1001
16. Hu YH, Chen YL, Chou HL (2017) Opinion mining from online hotel reviews-a text summarization approach. Inf Process Manag 53:436–449
17. Hüsch M, Schyska BU, von Bremen L (2018) CorClustST—correlation–based clustering of big spatio–temporal datasets. Futur Gener Comput Syst
18. Hutto CJ, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media (ICWSM-14)
19. Khatua A, Khatua A, Cambria E (2019) A tale of two epidemics: contextual Word2vec for classifying twitter streams during outbreaks. Inf Process Manag 56:247–257
20. Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762
21. Kisilevich S, Mansmann F, Nanni M, Rinzivillo S (2009) Spatio–temporal clustering. In: Data mining and knowledge discovery handbook. Springer, pp 855–874
22. Kranjc J, Smailović J, Podpečan V, Grčar M, Znidaršič M, Lavrač N (2015) Active learning for sentiment analysis on data streams: methodology and workflow implementation in the Clowdflows platform. Inf Process Manag 51:187–203
23. Kuncheva LI (2014) Combining pattern classifiers: methods and algorithms. Wiley
24. Kušen E, Strembeck M (2018) Politics, sentiments, and misinformation: an analysis of the twitter discussion on the 2016 Austrian presidential elections. Online Social Networks and Media 5:37–50
25. Li G, Liu F (2014) Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions. Appl Intell 40:441–452
26. Liang J, Bai L, Dang C, Cao F (2012) The k-means-type algorithms versus imbalanced data distributions. IEEE Trans Fuzzy Syst 20: 728–745
27. Liu B (2012) Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 5:1–167
28. Lukeš J, Søgaard A (2018) Sentiment analysis under temporal shift. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp 65–71
29. Manning CD, Raghavan P, Schütze H et al (2008) Introduction to information retrieval, vol 1. Cambridge University press, Cambridge
30. Medagoda N, Shanmuganathan S (2015) Keywords based temporal sentiment analysis. In: 2015 12th international conference on fuzzy

systems and knowledge discovery (FSKD), pp 1418–1425. https://doi.org/10.1109/FSKD.2015.7382152

31. Monireh E, Hossein YA, Amit S (2017) Challenges of sentiment analysis for dynamic events. IEEE Intell Syst 32:70–75
32. Nguyen LT, Wu P, Chan W, Peng W, Zhang Y (2012) Predicting collective sentiment dynamics from time-series social media. In: Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, Beijing, China. ACM, p 6
33. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the 2002 conference on empirical methods in natural language processing, (EMNLP 2002), Philadelphia, PA, USA. Association for Computational Linguistics, pp 79–86
34. Pratama M, Za'in C, Ashfahani A, Ong YS, Ding W (2019) Automatic construction of multi-layer perceptron network from streaming examples. In: Proceedings of the 28th ACM international conference on information and knowledge management. ACM, pp 1171–1180
35. Qian Y, Zhang Y, Ma X, Yu H, Peng L (2019) Ears: emotion-aware recommender system based on hybrid information fusion. Information Fusion 46:141–146
36. Rashkin H, Bell E, Choi Y, Volkova S (2017) Multilingual connotation frames: a case study on social Media for Targeted Sentiment Analysis and Forecast. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 2: short papers), pp 459–464
37. Ravi K, Ravi V, Prasad PSRK (2017) Fuzzy formal concept analysis based opinion mining for CRM in financial services. Appl Soft Comput 60:786–807
38. Salehan M, Kim DJ (2016) Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics. Decis Support Syst 81:30–40
39. Shu K, Sliva A, Sampson J, Liu H (2018) Understanding cyber attack behaviors with sentiment information on social media. In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, pp 377–388
40. Si J, Mukherjee A, Liu B, Li Q, Li H, Deng X (2013) Exploiting topic based twitter sentiment for stock prediction. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pp 24–29
41. Smailović J, Grčar M, Lavrač N, Znidaršič M (2014) Stream-based active learning for sentiment analysis in the financial domain. Inf Sci 285:181–203
42. Subramani S, Wang H, Islam MR, Ulhaq A, O'Connor M (2018) Child abuse and domestic abuse: content and feature analysis from social media disclosures. In: Proceedings of databases theory and applications, 29th Australasian database conference, ADC 2018, Gold Coast, QLD, Australia. Springer, pp 174–185
43. Vanstone BJ, Gepp A, Harris G (2019) Do news and sentiment play a role in stock Price prediction? Appl Intell:1–6
44. Wang G, Zhang Z, Sun J, Yang S, Larson CA (2015) Pos-rs: a random subspace method for sentiment classification based on part-of-speech analysis. Inf Process Manag 51:458–479
45. Xia R, Xu F, Yu J, Qi Y, Cambria E (2016) Polarity shift detection, elimination and ensemble: a three-stage model for document-level sentiment analysis. Inf Process Manag 52:36–45
46. Xing S, Wang Q, Zhao X, Li T et al (2019) Content-aware point-of-interest recommendation based on convolutional neural network. Appl Intell 49:858–871
47. Yu Y, Duan W, Cao Q (2013) The impact of social and conventional media on firm equity value: a sentiment analysis approach. Decis Support Syst 55:919–926
48. Zhou G, Zhou Y, Guo X, Tu X, He T (2015) Cross-domain sentiment classification via topical correspondence transfer. Neurocomputing 159:298–305

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Murtadha Talib AL-Sharuee received his PhD degree from the Department of Computer Science and Information Technology, La Trobe University, Australia. He received a B.S. degree in Computer Science from Al-Mustansiriya University, Iraq and a M.S. in Computer Science from BAMU University, India. His research interests include natural language processing, sentiment analysis, machine learning, deep learning, transfer learning, clustering analysis and ensemble learning.

Fei Liu received Bachelor of Science (in Applied Mathematics) degree from Zhejiang University, China and PhD in Computer Science from La Trobe University, Australia. She was awarded La Trobe University Postgraduate Research Scholarship for her PhD study.

Fei is currently a senior lecturer in the Department of Computer Science & Information Technology, La Trobe University. Her research area is Artificial Intelligence including Automated Reasoning, Semantic Web and Text Mining. She has authored/co-authored more than 80 conference and journal papers and has supervised and co-supervised 10 PhD students to completion.