Expert Systems With Applications 169 (2021) 114318

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Analysis of concept drift in fake reviews detection

Rami Mohawesh *, Son Tran, Robert Ollington, Shuxiang Xu

School of Technology, Environments and Design, University of Tasmania, Tasmania, Australia

ARTICLE INFO

Keywords:
Fake reviews
Concept drift detectors
Fake reviews detection techniques
Time-changing data

ABSTRACT

Online reviews have a substantial impact on decision making in various areas of society, predominantly in the arena of buying and selling of goods. As such, the truthfulness of internet reviews is critical for both consumers and vendors. Fake reviews not only mislead innocent clients and influence customers' choice, leading to inaccurate descriptions and sales. This raises the need for efficient fake review detection models and tools that can address these issues. Analysing a text data stream of fake reviews in concept drift appears to reduce the effectiveness of the detection models. Despite several efforts to develop algorithms for detecting fake reviews, one crucial aspect that has not been addressed is finding a real correlation between the concept drift score and the classification of performance over-time in the real-world data stream. Consequently, we have introduced a comprehensive analysis to investigate the concept drift problem within fake review detection. There are two methods to achieve this goal: benchmarking concept drift detection method and content-based classification methods. We conducted our experiment using four real-world datasets from Yelp.com. The results demonstrated that there is a strong negative correlation between concept drift and the performance of fake review detection/prediction models, which indicates the difficulty of building more efficient models.

1. Introduction

Opinions are an intrinsic attribute of humanity. Through the popularisation of the internet, reviews have become more readily accessible in different places around the world. We use websites like Yelp and TripAdvisor to exchange feedback about facilities, locations, and establishments (Jindal & Liu, 2008). Similarly, e-commerce websites, including Amazon, allow their users to post reviews on products and services. Such electronic platforms popularise the exchange of comments and raise general trust in electronic opinions (Barbado, Araque, & Iglesias, 2019). They also improve competition, and some businesses have regrettably recruited people to post fake reviews to defame the goods and services of their competitors (Zhu & Zhang, 2010).

Fake reviews, also known as spammers, are often classified as spam opinion, deceptive opinions or spam reviews. They can cause financial problems for service providers, and product manufactures, due to the adverse effects they have on the reputation of their brand (Ho-Dac, Carson, & Moore, 2013). Businesses may also lose clients because fake reviews give undue advantage to their rivals (Ho-Dac et al., 2013). There are several news articles regarding the fraudulent use of reviews. In 2015, a chef was fired after posting a deceptive negative review on TripAdvisor about rivals' restaurants.[1] In 2013, Samsung hired spammers to publish a deceptive negative review about HTC smartphones.[2] The distribution of fake reviews on the internet is a severe problem. For example, as of 2016, 16% of Yelp reviews are estimated as fake reviews by Luca and Zervas (2016).

In recent years, the number of customer reviews on the internet, generated for the promotion of goods and services across various website, has increased dramatically. There are considered to be two kinds of fake reviews:

- A deceptive positive review, which is a review giving positive feedback, although it is a false expression of the product.
- A deceptive negative review, which is a review that gives negative feedback, although it is a false expression of the product.

Fake reviews by unlawful users can cause consumers to make poor decisions. Therefore, detecting fake reviews has become a significant

* Corresponding author.
E-mail addresses: rami.mohawesh@utas.edu.au (R. Mohawesh), sn.tran@utas.edu.au (S. Tran), robert.ollington@utas.edu.au (R. Ollington), shuxiang.xu@utas.edu.au (S. Xu).

[1] ' Chef sacked after putting negative reviews about rival on TripAdvisor. Available at https://goo.gl/QcR3EZ, accessed on April 20, 2020

[2] Samsung Fined for Paying People to Criticize HTC Products. Available at https://goo.gl/tmFwYk, accessed on April 20, 2020.

https://doi.org/10.1016/j.eswa.2020.114318
Received 17 June 2020; Received in revised form 12 November 2020; Accepted 12 November 2020
Available online 18 November 2020

area of study (Karumanchi, Fu, & Deng, 2018). Most of the existing methods found in the literature of content-based fake review detection ignore the chronological order of the reviews (Harris, 2012; Ott, Choi, Cardie, & Hancock, 2011; Al Najada & Zhu, 2014; Li, Chen, Mukherjee, Liu, & Shao, 2015; Jindal & Liu, 2007, 2008; Mukherjee, Venkataraman, Liu, & Glance, 2013b) which is extremely important in real-world data since the spammers try to avoid the spam filter (Silva, Alberto, Almeida, & Yamakami, 2017; Xiao et al., 2015). These methods may also not be appropriate for fake review detection in real-world applications, where, due to their time sensitive nature, the features of the reviews usually change-over-time. Despite research already carried out in this area, none of it has investigated the existing concept drift phenomena in fake review detection.

Concept drift in text streams indicates that the features of the subject variable, estimated by the model, shifts in unexpected ways over time (Widmer & Kubat, 1996). When the concept is drifting, the form and structure of previous data induced might not be significant for new data, resulting in weak speculation and decision results. The concept drift has been established in many systems as a way of reducing efficiency. In a dynamic data world, it has become critical to develop more accurate evidence-driven forecasts and decision-making tools (Widmer & Kubat, 1996). In-text data streams which appear commonly in real-world applications like online reviews (Widmer & Kubat, 1996), social networks (Scott, 1988), and online document collections (Drzadzewski & Tompa, 2016), concept drift is common and regular as it results from data distribution (Widmer & Kubat, 1996; Nguyen, Woon, & Ng, 2015). For example, the topics concerned in Twitter are changing over time (Bifet & Frank, 2010), and the same is consistent with online news (Šilić & Bašić, 2012) as well as shopping reviews (Zhang, Chu, Li, Hu, & Wu, 2017). Thus, concept drifts can appear in text data streams of fake reviews, affecting the detection/prediction model performance.

For example, consider an online reviews stream of a hotel on real estate and the task of classifying incoming reviews into fake and genuine reviews. Supposing we have two posted reviews “The hotel is good” posted in 2007 and “The hotel is awesome” posted in 2015. The model trained on the first review is no longer able to detect the second review that has been posted more recently because the vocabulary used to express positive and negative sentiments may have changed over time. Since the collection of reviews is not static, the feature space representing the current collection is dynamic and may require specific updates of the models. This process is referred to as virtual drift, which is the change that occurs in the non-class attribute. This change may come from the spammer behaviour or could be the change of genuine user habits or opinions.

This paper aims to investigate and detect the concept drift problem in fake reviews. We conduct a comprehensive analysis that benchmarks concept drift detection methods and content-based classification methods of fake reviews detection. The experiments were conducted using four fake review real-world datasets (Yelp CHI, Yelp NYC, Yelp ZIP and Yelp consumer electronic) and three machine learning algorithms as base learners (SVM, Logistic Regression and Perceptron). It is worth mentioning that as of yet, the concept drift problem in fake reviews detection has not been studied, and this study is the first to detect this problem and investigate its correlation and impact on fake review detection models.

The contribution of this paper can be summarised as it:

1. Provides the performance of the classification techniques after sorting the reviews in chronological order over-time in terms of accuracy while using different machine learning classifiers that can be used as a baseline for future studies.
2. Investigates the concept drift problem in terms of predictive accuracy, number of drifts and evaluation time by using the benchmarks drift detection algorithms.
3. Analyses the impact of the concept drift problem on fake review detection/prediction model’s performance.

The paper is organised as follows:

- Section 2 describes the related work for fake reviews classification performance and benchmarking concept drift detection methods.
- Section 3 describes problem formulation and methodology and provides detailed descriptions of the real-world datasets and datasets pre-processing.
- Section 4 presents the setting of the experiments undertaken for the average accuracy of classification performance and predictive accuracy for all the benchmark drift detection methods.
- Section 5 presents the results and discusses the average accuracy of classification performance and evaluates them statistically. This involves an analysis of the predictive accuracy for all the benchmark drift detection methods and evaluates them statistically. This section also presents the evaluation time for the drift detection methods for both classification performance and concept drift measuring.
- Section 6 presents the conclusion and provides context for future work.

2. Related work

This section is divided into two sub-sections. The first provides an overview of fake reviews, followed by a summarisation of the content-based models for classifying the fake reviews. The second outlines the definition of concept drift, along with the benchmark concept drift detection.

2.1. Fake review overview

In the real-world web, users can automatically post comments or reviews on websites. These user reviews are valuable for both cooperation and consumers (Pang & Lee, 2009; Fitzpatrick, Bachenko, & Fornaciari, 2015; Liu, 2012). For example, consumers read reviews about a product or service before deciding to make a purchase and individual businesses (e.g. restaurants, and hotels) rely on reviews coming from their consumers to enhance the quality of their services and products as well as improve their businesses models (Jindal & Liu, 2007).

The benefits associated with reviews have been corrupted with the posting of fake reviews, where fake opinions (positive or negative) are written to be genuine (Ott et al., 2011). As such, businesses must have a tool which may detect fake reviews to not only provide benefits for consumers but also for business and industry which rely immensely on these reviews for their business models. There is a plethora of research around discovering fake news and email spam, but fake review detection has not been seen to be a top priority of spam research due to limited reviews online. However, as online reviews influence how users express their opinions about services and products, it is essential to ensure the credibility and accuracy of those reviews.

To explain what a fake review is, we need to consider the following two examples of reviews from publicly available real-life Yelp CHI dataset (Mukherjee, Venkataraman, Liu, & Glance, 2013a). The reviews are categorised as genuine and fake Review 1, below, is an example of a genuine review, while Review 2 shows how a typical fake review looks in comparison. By analysing these two reviews, we conclude that it is difficult for a human to establish the distinction between the genuine and the fake. Current research in this field of study, as evidenced in literature, human manually annotates the reviews and have achieved limited performance with an accuracy of 60% (Ott et al., 2011). So, developing effective models that can identify fake reviews automatically is essential (Crawford, Khoshgoftaar, Prusa, Richter, & Al Najada, 2015).

- Review 1: “Nice location to stay. It is close to everything – plenty of restaurants, shopping places, the lake, the city, and Wrigley! It is a small place with good management. It is not cheap and not so much because of the excellent amenities, but for the excellent location”.

- Review 2: "Sutton Place is a very elegant hotel. The room and staff were excellent. Room service did a spectacular job cleaning the room while I was away. The bed was a little harder than expected, but still comfortable. If I'm ever in the area again, this would be the hotel I would stay at".

2.1.1. Fake reviews detection

Most existing studies adopted classical machine learning techniques for fake review detection. For example, various models were built based on supervised learning techniques where the labelled data is used to learn the classifier to predict whether the review was fake or not.

Jindal and Liu (2008) utilised ensemble models for detecting fake review based on unigram and bigram features. The models achieved the best performance in terms of accuracy compared with Naïve Bayes, Random Forest and Support Vector Machines (SVM). However, considering the famous assertion that duplicate reviews are fake and unreliable, Lin, Zhu, Wang, Zhang, and Zhou (2014) introduced a cross-domain fake review detector based on Sparse Additive Generative Model (SAGE). The results showed that SAGE model achieved the best accuracy compared with SVM in one-domain. However, it achieved less performance for intradomain and cross-domain classification, which indicates that LIWC feature is not suitable for intradomain and cross-domain. Recently, Hernández-Castañeda, Calvo, Gelbukh, and Flores (2017) discovered the efficiency of using a Support Vector Network (SVN) in detecting fake reviews in one domain, mixed domain and cross-domain based on combined features. Furthermore, using Latent Dirichlet Allocation (LDA) and Word Space Model (WSM) as feature extraction, the proposed model achieved good accuracy in one and mixed fake domains. However, the proposed model focuses on a specific domain only.

Recent fake review detection models adopted advanced techniques, such as deep learning (Collobert et al., 2011; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Jing, 2014; Ren & Ji, 2017; Ren, Wang, & Ji, 2016). Compared to traditional machine learning, deep learning techniques can easily extract latent data representation which can contribute to improving fake review detection performance. Notably, this type of learning is highly appropriate for text data, as it can capture the semantic meaning of the text using a word embedding method. Li, Ren, Qin, and Liu (2015) proposed deep learning-based fake review detectors using the Convolutional Neural Network (CNN). The authors utilised the document representation concept, where each review was converted to a word vector for training and testing the model. The experimental results showed that the proposed model was effective in detecting cross-domain fake reviews. Furthermore, their proposed model outperformed the Long Short-Term- Memory (LSTM) in mixed-domain fake reviews. Similarly, Zhao, Xu, Liu, and Guo (2017) utilised CNN for detecting fake reviews, though, the authors used the word order reserve pooling method, instead of the original max-pooling to construct the CNN architecture. Comparing Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN), the optimized CNN-based fake review detector achieved the best performance with regards to short text reviews due to the word order embedding, while RNN performed better for long-text reviews.

Liu, Jing, and Li (2019) utilised the Bidirectional LSTM algorithm to build their fake review detector. The authors also suggested a combination of features that were Part of Speech (POS), and First-Person Pronoun (I) to obtain better performance. The results showed that the proposed model outperformed the state-of-the-art methods such as paragraph average (Le & Mikolov, 2014), sentence weight neural network (SWNN) (Li, Qin, Ren, & Liu, 2017), SWNN&POS&I (Li et al., 2017), CNN-LSTM, Bidirectional LSTM and basic CNN&POS&I in one domain, mix-domain and cross-domain. However, the proposed model is dependent on the review text and ignores behavioural features.

The main problem that is associated with fake review detection is the changing of data over-time. Some models were evaluated using a dataset that fails to detect fake reviews with training data in a specific time and testing data at different times. Fake reviews, like spam email and twitter that are written by a person, may change over time, due to the dynamic nature of human behaviour. Furthermore, in-stream mining, the nature of the velocity, unbound size and time of arriving reviews can pose a problem for classification data in real-time.

2.2. Concept drift

In this subsection, we first outline the definition of the concept drift problem, then summarise the benchmark concept drift detection for supervised classification streams.

2.2.1. The definitions of concept drift

Concept drift is a phenomenon where the characteristics of the output arbitrarily change over time (Lu, Zhang, & Lu, 2014). It was first proposed by (Schlimmer & Granger, 1986) that noisy data could be used to transmit non-noisy information at specific times. These types of changes may be caused by occurrences that cannot be measured directly within hidden variables (Liu, Song, Zhang, & Lu, 2017).

The representations of concept drifts can be distinctive, as such different definitions by some researcher can be categorised into various groups (Nguyen et al., 2015; Widmer & Kubat, 1996; Zhang, Zhu, & Shi, 2008). Lu et al. (2018) summed up these opinions and emphasised that changes to the distribution of data induced by the different implicate contexts, can be categorised into four types:

- Incremental drift which is an old concept that changes incrementally to a new concept during a period of time
- Sudden drift which is a new concept that occurs within a short period of time
- Gradual drift which is a new concept and is gradually being swapped with an older one over a period of time.
- Reoccurring drift that is after a while, an old concept could reoccur.

Concept drift detection in text data streams is essential and can be categorised into three types; the occurrence of concept drift between features and labels, the occurrence of concept in the change of the distribution of features and occurrence of concept when the relationship between features and labels have changed (Zhang et al., 2017). Concept drift within text data streams of fake reviews can be challenging to detect as the text data streams can be high-dimensional, making the detection of concept drift in text data essential (Joachims, 1998).

2.2.2. Benchmarking concept drift detection methods

Concept drift detection refers to algorithms that detect concept drift through the change points which it occurs. Extensive literature exists for concept drift detection for classification of data streams in different domains (Tsymbal, 2004; Wang, Fan, Yu, & Han, 2003; Gama, Medas, Castillo, & Rodrigues, 2004; Aggarwal, 2005; Bifet & Gavalda, 2007; Baena-Garcıa et al., 2006; Page, 1954). Concept drift detection is critical for the binary classification of text data streams where it can affect the classification performance. The concept drift detection for text classification can be divided into two groups; unsupervised learning methods which are used for unlabelled streams that measure the concept drift between two clusters based on distance and radius (Aggarwal, 2005; Bouchachia & Vanaret, 2013). Supervised learning methods have been developed for labelled text data for concept drift detection that measure the concept drift based on the error rate (Gama et al., 2004; Baena-Garcıa et al., 2006), tracking the distribution between two windows (Bifet & Gavalda, 2007), and based on sequential analysis (Page, 1954).

Gama, Žliobaitė, Bifet, Pechenizkiy, and Bouchachia (2014) divided concept drift detection methods into three categories:

1- **Methods based on sequential analysis:** these methods evaluate predictive results as appropriate. Page Hinkley and The Cumulative Sum (CUSUM) methods are members of this category (Page, 1954).
2- **Methods based on statistical analysis:** these methods analyse the standard deviation and mean connected with the predicted results to detect concept drift. The Drift Detection Method (DDM) (Gama et al., 2004), Early Drift Detection Method (EDDM) (Baena-Garcıa et al., 2006), Reactive Drift Detection Method (RDDM) (Barros, Cabral, Gonçalves, & Santos, 2017), and Exponentially Weighted Moving Average (EWMA) (Ross, Adams, Tasoulis, & Hand, 2012) are representatives of this group.
3- **Methods based on windows:** These methods restore the knowledge of the past and provide a sliding window to restore more current knowledge. A significant difference between these two sub-windows indicates the presence of concept drift. The Adaptive Windowing (ADWIN) (Bifet & Gavalda, 2007), the Drift Detection Methods based on Hoeffding's Bound (HDDMA-test and HDDMW-test) (Frías-Blanco et al., 2014), the SeqDrift detectors (Pears, Sakthithasan, & Koh, 2014; Sakthithasan, Pears, & Koh, 2013) and fast hoeffding Drift Detection Method (Pesaranghader, Viktor, & Paquet, 2018a; Pesaranghader & Viktor, 2016) are members of this category.

DDM, ADWIN, EDDM and Page Hinkley methods have been considered as benchmarks in previous literature (Frías-Blanco et al., 2014; Pesaranghader & Viktor, 2016; Baena-Garcıa et al., 2006; Bifet & Gavalda, 2007; Huang, Koh, Dobbie, & Bifet, 2015). Accordingly, we chose the benchmark method from each category and evaluated them. These methods are explained further below.

Gama et al. (2004) introduced the Drift Detection Method (DDM) to detect the problem of concept drift based on the classification error rates. Increasing the error rate indicates that there is a change in the data distribution while the current base learner becomes ineffective due to the Probability Approximately Correct method (PAC) (Mitchell, 1997). DDM used binomial distribution which provided the common types of probability for the random variables, which presents the number of errors in n instance samples. For every instance i in the sequence, the probability of miss-classification is represented by the error rate *(pi)* and the standard deviation is calculated by $si = pi\ (1 - pi)/i$. Therefore, they store the values of si and pi when $si + pi$ achieved its minimum value through the process (*obtaining smin and pmin*), then it monitors the following trigger conditions; for the warning level, $pi + si \geq pmin + 2 \cdot smin$. This level indicates that there is a possibility in the change of context. For the drift level, $pi + si \geq pmin + 3 \cdot smin$. When a concept drift is detected, a new base learner is built using the instances stored, since the warning level and the value for *smin* and *pmin* are reset as well. The DDM parameters for the warning and drift levels and the minimum number of instances before the concept drifts are detected 2.0, 3.0 and 30, respectively.

Similarly, Baena-Garcıa et al. (2006) introduced an Early Drift Detection Method (EDDM) to detect the concept drift problem which depends on the value of the distance between two classification errors instead of the amount of errors rate. They considered two threshold values for the warning level and drift level; for the warning level, $(pi + 2 \cdot si)/(pmax + 2 \cdot smax) < \alpha$. For the drift level, $(pi + 2 \cdot si)/(pmax + 2 \cdot smax) < \beta$. If a new concept drift is detected, a new base learner is built using the instances stored since the warning level, and the value of *smax* and *pmax* is reset as well. The values for α and β have been set to 0.99 and 0.90.

Bifet and Gavalda (2007) introduced an Adaptive Windowing (ADWIN) algorithm that detects concept drifts based on a sliding window of instances. The two sub-windows that changed dynamically are stored, which represents the recent and old data. When two sub-windows' mean values are significantly larger than a given threshold, ADWIN detects a drift and removes the oldest data from the adjustable pane. This process is known as cutting detection because it decides when the adaptive window should delete the old data. The cut detection repeats removing old tuples until the output of the adaptive window no longer suggests that there is a concept drift. The default delta value parameters for ADWIN by the author is 0.002.

The Page Hinkley Test (Page, 1954) is a sequential analytical technique that is used to detect concept drift. It calculates the values observed, which is the base learner accuracy and their mean. When concept drift is detected, the base learner does not identify incoming instances correctly, thereby reducing the mean accuracy. The total distinction between the two values (UT) and the minimal distinction between the two (mT) is calculated. Higher UT values indicate that the values measured vary substantially from previous values. If the distinction from UT to MT is higher than the defined threshold that matches the size of the permissible changes (т), then a concept drift is detected. Higher μ values result in fewer false alarms, but some adjustments can be skipped or delayed. The default parameters for the delta value, threshold value, alpha value and a minimum number of instances before a drift is detected with Page Hinkley, are 0.005, 50, 0.0001 and 30 respectively.

3. Problem formulation and methodology

In this section, we provide a brief description of problem formulation, Concept drift in fake review detection, datasets used for evaluation and pre-processing methods.

3.1. Problem formulation

The primary objective of data stream classification is to predict the label of unseen classification. In fake review data stream classification, there is a sequence of reviews: $(X_1, y_1), (X_2, y_2), (X_3, y_3), \ldots (X_t, y_t)$ Where X_t is a vector containing S attributes: $X_t = (x_1, x_2, x_3, \ldots x_s)$, while y_t is a class label which belongs to a {0,1}. For each incoming review t, the classifier predicts its class label y_t. A concept drift refers to the joint distribution $P_t(X,y)$ that the performance of the model decreased over time due to the dynamic change of reviews features over the years (Gama, Žliobaitė, et al., 2014; Widmer & Kubat, 1996; Krawczyk, Minku, Gama, Stefanowski, & Woźniak, 2017). This change may come from the spammer behaviour that changes their behaviour in posting fake reviews to avoid the spam filter, and another reason could be the change of positive user habits or opinions.

To understand more about the concept drift in fake review, consider the hotel reviews as seen in Fig. 1. These were collected over a period of time (for example, from X_1 *to* X_t) and we trained our model using this data. Later on, we used the trained model to detect fake hotel reviews that were posted more recently, and here we assume it X_{t+1}. We can say that a model trained on older historical data (from X_1 *to* X_t) is no longer able to efficiently detect fake reviews (X_{t+1}) due to the change in features of reviews over the years.

The objective of this paper is to study concept drift phenomena and its impact on fake reviews detection. Given that these reviews change over time, due to velocity, unbound size and time nature of reviews in streaming data, we performed a comprehensive analysis on the fake review datasets along with the evaluation method used in this experiment to study the following scenarios:

- Study the change of characteristics in data over time.
- Measure the concept drift problem in fake review detection.
- Study the correlation between concept drift and classification performance of reviews which are sorted chronologically by posting time.

3.2. Concept drift in fake review detection

Scenario 1: Studying the changing characteristics of data over-time and determining its effects on fake review detection with regards to the chronological order of the reviews.

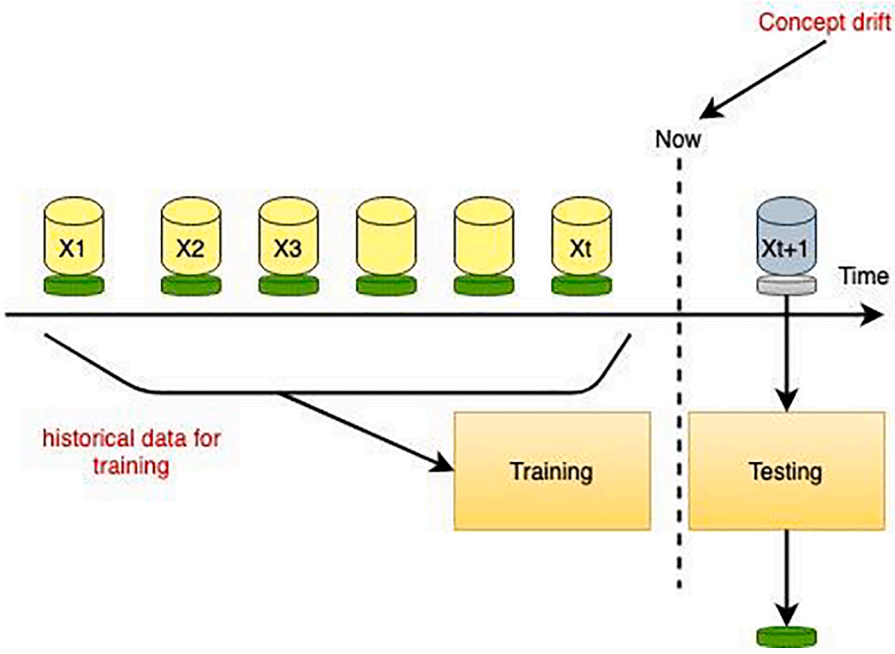


Fig. 1. Concept Drift in Fake Review.

We propose the alternative hypotheses that classification performance is decreased significantly over-time in fake reviews detection, due to concept drift impact. Subsequently, the null hypothesis is that the classification performance is not significantly affected by concept drift.

To evaluate the classification performance over-time, we arrange the data in chronological order, based on the given date and using several machine learning methods: Support Vector Machine (SVM), Logistic Regression (LR) and Perceptron (PNN). We chose these methods due to their robustness and accuracy (Mukherjee, et al., 2013a; Khurshid, Zhu, Xu, Ahmad, & Ahmad, 2018).

Scenario 2: Measuring the concept drift in fake review detection.

We propose an alternative hypothesis that the concept drift can be detected in fake reviews, where the variance of the base learner is different. The null hypothesis proposed indicates that the variance of the base learner's performance is the same.

To measure the concept drift problem, we used the benchmark drift detection methods including DDM, EDMM, ADWIN and Page Hinkley Test, with different incremental base learners such as, Perceptron, Stochastic gradient descent for SVM and logistic regression which are broadly used for classification (Bifet, Gavaldà, Holmes, & Pfahringer, 2018; Rajaraman, Leskovec, & Ullmann, 2014).

In fake review text data streams, where the data are collected over time, we assume a sequence of reviews which contain pairs (Xi, Yi), where Xi is a sequence of reviews represented as a vector and Yi is the label of review which can be fake or genuine. As shown in Fig. 2, concept drift detection methods classify each incoming review. For each text review, the output prediction is compared to the true class labels indicated as (1) to correct the classification and (0) representing misclassification. Depending on the classification results that are passed to the drift detection methods, these methods can determine if the concept drift has occurred or not. Lastly, the classifier is trained on the review, and the process is repeated for all the reviews.

Scenario 3: To study the correlation between the concept drift scores and the classification performance.

We propose the alternative hypothesis that the coefficient correlation between concept drift and performance is a significantly strong negative, where the null hypothesis has no significant correlation between concept drift and classification performance.

3.3. Datasets

We employed four large real-world datasets (Yelp CHI, Yelp NYC, Yelp ZIP and Yelp consumer electronic) from Yelp.com:

- Yelp CHI dataset contains 67,365 restaurant and hotel reviews in Chicago city from 2004 to 2012. The reviews include user information, product information, rating, timestamp and text review (Mukherjee, et al., 2013a).
- Yelp NYC and Yelp ZIP datasets contain restaurant reviews from 2004 to 2015 in NYC, NJ, VT and PA, where Yelp NYC contains 322,167 reviews and Yelp ZIP contains 608,598 reviews. The reviews include user information, product information, rating, timestamp and text review (Rayana & Akoglu, 2015).
- Yelp consumer electronic contains 9456 genuine reviews and 9456 fake reviews collected from four US cities from 2004 to 2017. The reviews include user information, product information, rating, timestamp and text review (Barbado et al., 2019).

These datasets were built based on an unknown filtering algorithm and web-scraper techniques to label each review as fake or genuine. Consequently, these datasets have been extensively used in the literature review, due to the lack of fake reviews datasets. These datasets also represent real-life data (Ren & Ji, 2019). Evaluating fake reviews detection models based on real-life data is preferred as this helps build more robust models that can work efficiently in the real-world environments (Li, Ott, Cardie, & Hovy, 2014).

3.4. Data pre-processing

In our experiments, we applied non-alphanumeric and lowercase conversion. Stop words were not removed. We did not apply stemming and lemmatisation, as these things might remove essential features required for the classification performance (Méndez, Iglesias, Fdez-Riverola, Díaz, & Corchado, 2005). Features were extracted using

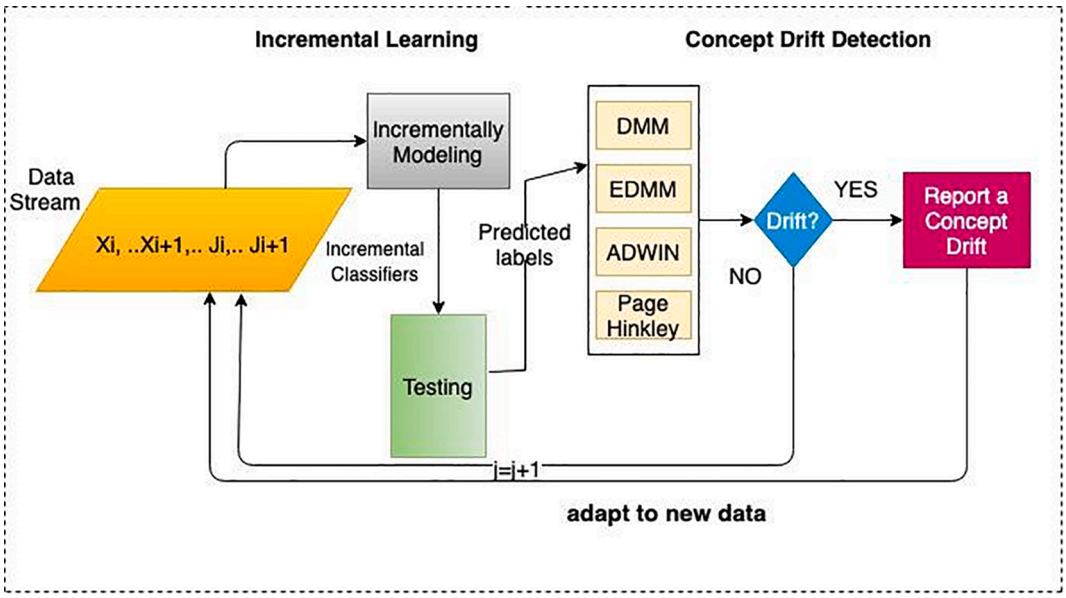


Fig. 2. Detailed process of analysing concept drift in fake reviews detection.

unigram, with term frequency and inverse document frequency (TFIDF) for text representations to achieve excellent results in text classifications (Das & Chakraborty, 2018). TF-IDF represented each review by a weighted vector by calculating the number of times a particular word appeared in the document and inversely the number of documents containing the word. Then we made the data partitions balanced to provide accurate results.

To address the open question in this paper, we divided the Yelp CHI dataset into three different parts based on the given years. The first includes reviews from 2004 to 2008 (D1), the second from 2009 to 2010 (D2) and the third from 2011 to 2012 (D3). For Yelp NYC and Yelp ZIP, each dataset is divided into four parts based on the given years. The first includes reviews from 2004 to 2009 (D1), the second from 2010 to 2011 (D2), the third from 2012 to 2013 (D3) and the fourth from 2014 to 2015 (D4).

For the Yelp consumer electronics dataset, we divided it into five parts based on different periods. The first includes reviews from 2004 to 2009 (D1), the second from 2010 to 2011 (D2), the third from 2012 to 2013 (D3), the fourth from 2014 to 2015 (D4), and the fifth from 2016 to 2017 (D5). After that, we balanced the dataset partitions.

The primary purpose for dividing the datasets was to make the first part of split data for training the classifiers and the remaining datasets partitions for testing purpose. As we need a large number of reviews for training, and due to the insufficient number of posted reviews each year, we divided the datasets as they presented in our work.

As these text datasets are high dimensionally, they may affect the concept drift detection methods that detect concept drift based on the base learners' performance. This makes the classification of the error rate variability appear high and thus inaccurate (Joachims, 1998). We used Principal Component Analysis (PCA) (Pearson, 1901) to reduce the data dimension and transform the feature vectors to new data representations (components). We chose the most significant number of components (10,000). PCA is the most common dimension reduction technique in text data processing tasks and has less sensitivity for noise data, low computational complexity and low memory capacity (Taloba, Eisa, & Ismail, 2018).

4. Experiments

4.1. Fake review – content-based performance based on machine learning classifiers

To study how the performance of the fake review detection changes over time, and test whether the null hypothesis, is significantly affected by the concept drift problem or not, we utilised three Stochastic Gradient Descent (SGD) classifiers as machine learning classifiers (Zhang, 2004). They were Support Vector Machine (SVM), Logistic Regression (LR) and Perceptron Neural Network (PNN) (Rosenblatt, 1958). If the null hypothesis is rejected, then the classification performance is significantly decreased over time, and the concept drift problem will be measured. We conducted our experiments using python[3]Language (scikit-learn library[4]) and employed cross-validation for the first dataset collection, then tested each classifier on the other datasets. The evaluation procedure was repeated ten times to calculate the average accuracy with 95% confidential interval. Here, we adopted only the accuracy as we focussed on the performance of methods in terms of correctly classified reviews. Further, it is a good enough metric for balanced datasets.

4.2. Fake review -concept drift detection based on benchmarking algorithms

For testing drift detection methods, described in Section 2.3.2., we utilised the benchmark drift detection methods including DDM, EDMM, ADWIN and Page Hinkley with the incremental machine learning classifiers (SVM, LR, and PNN). We used the default parameters of DDM (n $= 30$, $\alpha_w = 2$, $\alpha_d = 3$), EDDM (n $= 30$, $\alpha = 0.99$, $\beta = 0.90$), ADWIN ($\delta = 0.002$) and Page Hinkley (n $= 30$, $\delta = 0.005$, $\lambda = 50$, $\alpha = 0.0001$) according to the original work to avoid an accurate results.

The null hypothesis for concept drift detection algorithms stipulates whether the variance of the base learner's performance is the same. If the null hypothesis is rejected, then the concept drift can then be detected. To evaluate the concept drift methods performance, we used the Prequential methodology with a sliding window that it is forgetting

[3] . Available at http://www.python.org, accessed on August 20, 2019.
[4] Scikit learn. Available at http://scikit-learn.org/stable/, accessed on October 16, 2019.

mechanism (Dawid, 1984) used with default parameters in scikit-multiflow libraries to predict streaming data instances that have not been seen yet.

We calculated accuracy over the most recent instances as this method is more suitable and can enhance the drift detection results instead of using the entire stream in data evolving (Prasad & Agarwal, 2017). The model was updated based on the most recent data represented by the window size. The accuracy evaluation was calculated using prequential with a sliding window of size 200 instance that it is the default in scikit-multi-flow library. In this method, each incoming review is used for testing and subsequently for training to calculate the accuracy based on the cumulative sum of the sequential errors over time.

In our experiments, we used different review sizes to train the base learner before starting the evaluation. The evaluation procedure was repeated five times to calculate the average predictive accuracy with 95% confidential interval. To perform these experiments, we used the python language (scikit-learn and scikit-multi-flow libraries[5]).

5. Results and discussion

This section includes all the experimental results mentioned in this paper, which can be divided into two different scenarios.

5.1. Classifiers performance

Table 1–4 show the average accuracy for each classifier. In this scenario, we randomly partitioned the first collection of datasets (D1) into K subsets with (k = 5), where each subset had the same number of reviews in each class. We then tested the model on the last fold, and other dataset partitions D2 (2010–2011), D3 (2012–2013), D4 (2014–2015), D5 (2016–2017). This evaluation procedure was repeated ten times to calculate the average accuracy with 95% confidential interval.

As shown in the table below, all the classifiers suffered in their detection process and achieved a poor performance where the average accuracy was within the range [57.45%, 66.36%] for Yelp NYC dataset. The same observations for the Yelp ZIP dataset showed an average accuracy within the range [58.36%, 68.54%]. Observations for Yelp CHI dataset showed an average accuracy within the range [57.17%, 64.64%]. Lastly, the same observations for the Yelp consumer electronic dataset showed an average accuracy within the range [53.38%, 60.92%]. This is because training methods with real-world datasets at a specific time and testing at other times are complicated due to the time-ordered nature of reviews.

To emphasise that these results were not achieved by chance, we used a non-parametric statistical analysis called the Friedman test (Demšar, 2006). The Friedman test uses the average ranking to test the null hypothesis, indicating that all the approaches with the same performance can be dismissed. Figs. 3– 6 shows the average ranking of each approach based on the accuracy of the four datasets Yelp NYC, Yelp ZIP, Yelp CHI and Yelp consumer electronic.

We also used the post-hoc Nemenyi test to compare the performance of the classifiers. This test works if the difference between their ranks is more significant or equal to a critical difference (CD) (Demšar, 2006). We calculated the CD using the Nemenyi test with confidence interval α = 0.05, which was 0.74, 0.74, 0.85 and 0.66 on Yelp NYC, Yelp ZIP, Yelp CHI and Yelp consumer electronic datasets, respectively. We also found that the performance of LR, with the lowest average ranking, is significantly better than SVM and PNN for all Yelp NYC, Yelp ZIP, and Yelp CHI datasets partitions due to accuracy. However, the performance of SVM and PNN is significantly better than LR for Yelp consumer electronic dataset in term of accuracy.

[5] Scikit-multi-flow. Available at https://scikit-multiflow.github.io/scikit-multiflow/index.html. It was accessed on January 10, 2020.

We can conclude that as each classifier, SVM, LR and PNN, demonstrated a significant drop in its performance in terms of accuracy, as most recent real-world reviews contain features which are not reflected in the methods, since the spammers try to avoid the spam filter. However, these results are not sufficient to emphasise that there is a concept drift in fake review datasets. It could be that the significant drop in classifiers performance comes from other factors, such as noise. In this regard, we use the benchmark concept drift algorithms, as this method can detect the concept drift accurately. Based on these observations, the null hypothesis is thus rejected.

5.2. Measuring the concept drift problem

In this section, we present the results of the benchmark concept drift detection algorithms experiments for four real-world datasets from Yelp. The presence of concept drift in the real-world stream is unknown due to unbound size and velocity of streaming data. We investigated the performance of these algorithms (DMM, EDDM, ADWIN and Page Hinkley Test) in terms of predictive accuracy, the number of drifts and evaluation time.

5.2.1. Predictive accuracy

The predictive accuracy is computed as an Area Under Curve (AUC) metric with 95% confidential interval. The results for Yelp NYC, Yelp ZIP, Yelp CHI and Yelp consumer electronic real-world datasets are presented in Tables 5–7, respectively, using different base learners. The best results are written in bold due to the highest predictive accuracy.

First, we used the incremental classifiers without drift detection, also known as blind detection (Pesaranghader, Viktor, & Paquet, 2018b). This method adopts the base learner to the new data by training it on the most recent data without using any drift detection method. To measure the concept drift over time, we merged the first partition for each dataset with other partitions. For example:

- D1-D1 – using the first partition for training and testing incrementally.
- D1-D2 – merging the first partition and second for training and testing incrementally.
- D1-D3 – merging the first partition and third for training and testing incrementally.
- D1-D4 – merging the first partition and fourth for training and testing incrementally.
- D1-D5 – merging the first partition and fifth for training and testing incrementally.

Fig. 7(a.1, a.2, a.3 and a.4) shows the predictive accuracy curves for the Yelp NYC dataset. Based on the results on Yelp NYC, using LR as a base learner, ADWIN had the best predictive accuracy on all partitions when concept drift occurs. The predictive accuracy of ADWIN, EDDM and Page Hinkley decreased over time steps in all the merged partitions until the base learner adopted the new context. DDM had the worst performance since the DDM method could not detect any drift, using LR as a base learner. Fig. 7(b.1, b.2, b.3 and b.4) using SVM as a base learner on the same dataset showed that ADWIN had the best predictive accuracy. Page Hinkley had the worst performance since it failed to detect any drift. Fig. 7(c.1, c.2, c.3 and c.4) using PNN as a base learner, showed that ADWIN had the best predictive accuracy. Page Hinkley and DDM had the worst performance since these methods could not detect any drift.

Similarly, Fig. 8(a.1, a.2, a.3 and a.4) shows the experiment results for Yelp ZIP dataset using LR as a base learner. It can be noticed that EDDM performed best in D1-D1 partition. Though, using D1-D2, D1-D3, and D1-D4 dataset partitions, ADWIN achieved the best predictive accuracy. However, Page Hinkley and DDM performance were worst and these methods were not able to detect any drift.

Based on the results of Yelp Zip using SVM as a base learner (shown

Table 1
The main statistics and classification performance with 95% confidential interval of various classifiers over-time for Yelp NYC real world dataset parts - D1 (2004–2009), D2 (2010–2011), D3 (2012–2013), D4 (2014–2015) (F: Fake G: Genuine).

Dataset part	Training size		Testing size		Average accuracy SVM	Average accuracy LR	Average accuracy PNN
	F	G	F	G			
D1	2495	2471	621	621	63.65% ± 0.026	**66.36% ± 0.026**	59.62% ± 0.027
D2			621	621	62.86% ± 0.027	**65.65% ± 0.026**	58.43% ± 0.027
D3			621	621	62.31% ± **0.027**	**63.59% ± 0.026**	57.49% ± 0.027
D4			621	621	60.51% ± 0.027	**62.31% ± 0.027**	57.45% ± 0.027

Table 2
The main statistics and classification performance with 95% confidential interval of various classifiers over-time for Yelp ZIP real world dataset parts based on dates (years) D1 (2004–2009), D2 (2010–2011), D3 (2012–2013), D4 (2014–2015) (F: Fake G: Genuine).

Dataset part	Training size		Testing size		Average accuracy SVM	Average accuracy LR	Average accuracy PNN
	F	G	F	G			
D1	4736	4681	1150	1205	65.93% ± 0.019	**68.54% ± 0.018**	61.79% ± 0.019
D2			1178	1178	64.28% ± 0.026	**67.16% ± 0.025**	61.21% ± 0.027
D3			1178	1178	63.01% ± 0.027	**65.45% ± 0.026**	59.36% ± 0.027
D4			1178	1178	61.41% ± 0.027	**63.44% ± 0.026**	58.36% ± 0.027

Table 3
The main statistics and classification performance with 95% confidential interval of various classifiers over-time for Yelp Chi real world dataset parts D1 (2004–2008), D2 (2009–2010), D3 (2011–2012) (F: Fake G: Genuine).

Dataset part	Training size		Testing size		Average accuracy SVM	Average accuracy LR	Average accuracy PNN
	F	G	F	G			
D1	1021	1004	262	245	62.08% ± 0.042	**64.64% ± 0.043**	61.05% ± 0.043
D2			254	254	61.45% ± 0.042	**64.13% ± 0.041**	60.14% ± 0.042
D3			254	254	59.22% ± 0.042	**60.86% ± 0.042**	57.17% ± 0.042

Table 4
The main statistics and classification performance with 95% confidential interval of various classifiers over-time for Yelp consumer electronic real-world dataset parts D1 (2004–2009), D2 (2010–2011), D3 (2012–2013), D4 (2014–2015), D5 (2016–2017) (F: Fake G: Genuine).

Dataset part	Training size		Testing size		Average accuracy SVM	Average accuracy LR	Average accuracy PNN
	F	G	F	G			
D1	443	443	110	112	60.82% ± 0.092	**60.92% ± 0.096**	59.95% ± 0.097
D2			519	519	**56.19% ± 0.030**	55.71% ± 0.030	55.37% ± 0.030
D3			519	519	**55.18% ± 0.030**	54.83% ± 0.030	54.83% ± 0.030
D4			519	519	**54.49% ± 0.030**	54.29% ± 0.030	53.75% ± 0.030
D5			519	519	53.61% ± 0.030	**54.12% ± 0.030**	53.38% ± 0.030

in Fig. 8(b.1, b.2, b.3, and b.4)), ADWIN obtained the best predictive accuracy while Page Hinkley and DDM performance achieved the worst accuracy. This is because Page Hinkley and DDM were both not able to identify any drift in the data partitions. Furthermore, we obtained the same results using PNN as a base learner on Yelp ZIP dataset (see Fig. 8 (c.1, c.2, c.3 and c.4)). ADWIN functioned well and achieved the best predictive accuracy while Page Hinkley and DDM obtained the worst performance.

For Yelp CHI, results can be seen from Fig. 9 using different base learners. Using Yelp CHI with LR as a base learner, as shown in Fig. 9 (a.1, a.2, and a.3), ADWIN performed better and achieved the best predictive accuracy. On the other hand, Page Hinkley and DDM had a poor performance and they could not detect any drift. Fig. 9(b.1, b.2, and b.3) shows results for SVM as a base learner; in the given scenario, EDDM achieved the best performance and DMM performed the worst in terms of predictive accuracy. DDM failed to identify any drift in Yelp CHI dataset partitions. As in Fig. 9(c.1, c.2, and c.3), using Perceptron as a base learner, EDDM achieved a remarkable accuracy. Meanwhile, Page Hinkley and DDM had the worst performance.

Fig. 10 illustrates the predictive accuracy of the Yelp consumer electronic dataset using different base learners. Fig. 10(a.1, a.2, and a.3) shows the results of LR as the base learner. It can be seen that EDDM achieved the best performance in D1-D3, D1-D4 dataset partitions, while ADWIN was the best in (D1-D5) dataset partition. However, Page Hinkley and DDM performed the worst because drift could not be identified by these methods. Further, Fig. 10(b.1, b.2, and b.3) shows the best performance achieved by EDDM for this dataset partitions (D1-D3, D1-D4) using SVM as a base learner while ADWIN performed the best in (D1-D5) dataset partition. However, DDM had the worst results using SVM as a base learner. Furthermore, Fig. 10(c.1, c.2, and c.3) indicates

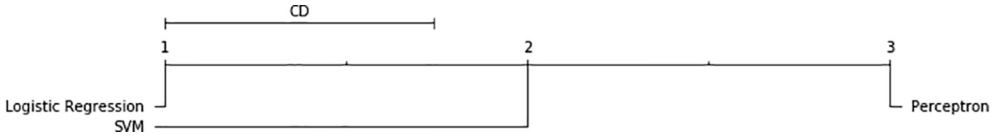


Fig. 3. Comparison of all classifiers against each other using Nemenyi test on Yelp NYC that showed a significant difference between various classifiers at (p-value = 0.05). where the classifiers are not connected.

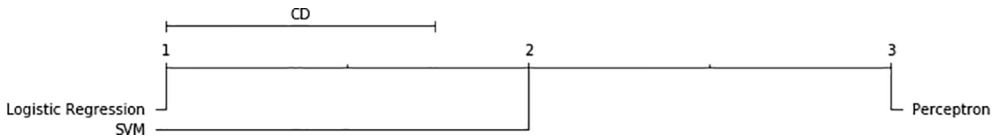


Fig. 4. Comparison of all classifiers against each other using Nemenyi test on Yelp ZIP that showed a significant difference between various classifiers at (p-value = 0.05) where the classifiers are not connected.

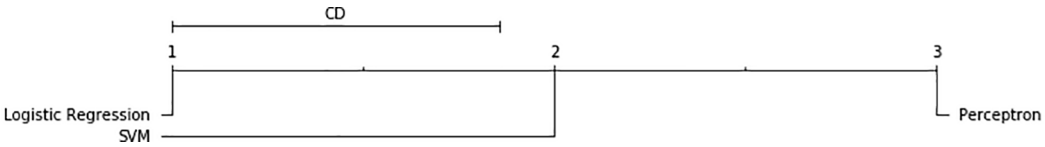


Fig. 5. Comparison of all classifiers against each other using Nemenyi test on Yelp CHI that showed a significant difference between various classifiers at (p-value = 0.05) where the classifiers are not connected.

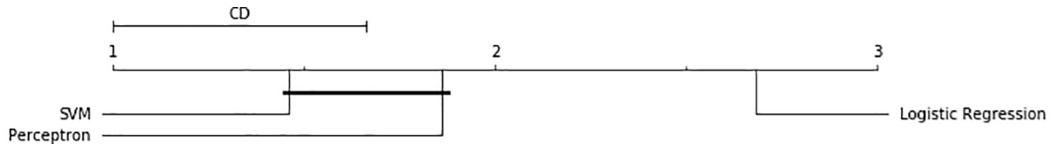


Fig. 6. Comparison of all classifiers against each other using Nemenyi test on Yelp Consumer Electronic that showed a significant difference between various classifiers at (p-value = 0.05) where the classifiers are not connected.

Table 5
Predictive accuracy using PNN Classifier with 95% confidential interval that showed the highest performance score for each drift detection method and without drift detection methods (Only classifier).

Dataset Name	Dataset parts	PNN Classifier	DDM	EDDM	ADWIN	Page Hinkley
Yelp NYC	D1-D1	60.87% ± 0.012	60.87% ± 0.012	60.57% ± 0.011	**61.42**% ± 0.012	60.87% ± 0.012
	D1-D2	69.37% ± 0.005	69.37% ± 0.005	69.63% ± 0.005	**77.37**% ± 0.005	69.37% ± 0.005
	D1-D3	74.81% ± 0.004	74.81% ± 0.004	75.02% ± 0.004	**82.73**% ± 0.004	74.81% ± 0.004
	D1-D4	74.90% ± 0.005	74.90% ± 0.005	75.14% ± 0.005	**84.89**% ± 0.005	74.90% ± 0.005
Yelp ZIP	D1-D1	65.88% ± 0.008	65.88% ± 0.008	67.67% ± 0.008	**71.63**% ± 0.007	65.88% ± 0.008
	D1-D2	77.12% ± 0.003	77.12% ± 0.003	78.21% ± 0.003	**86.04**% ± 0.003	77.12% ± 0.003
	D1-D3	82.59% ± 0.02	82.59% ± −0.02	83.53% ± −0.02	**90.56**% ± −0.02	82.59% ± −0.02
	D1-D4	81.56% ± 0.003	81.56% ± 0.003	82.65% ± 0.003	**91.02**% ± 0.003	81.56% ± 0.003
Yelp CHI	D1-D1	56.68% ± 0.019	56.68% ± 0.019	55.66% ± 0.019	56.68% ± 0.019	56.68% ± 0.019
	D1-D2	61.80% ± 0.009	61.80% ± 0.009	61.13% ± 0.009	**65.43**% ± 0.009	61.80% ± 0.009
	D1-D3	63.58% ± 0.009	63.58% ± 0.009	63.08% ± 0.009	**66.84**% ± 0.008	63.58% ± 0.009
Yelp Consumer Electronic	D1-D1	54.98% ± 0.044	54.98% ± 0.044	54.98% ± 0.044	54.98% ± 0.044	54.98% ± 0.044
	D1-D2	55.94% ± 0.022	55.94% ± 0.022	55.94% ± 0.022	55.94% ± 0.022	55.94% ± 0.022
	D1-D3	57.48% ± 0.015	57.48% ± 0.015	57.19% ± 0.015	57.48% ± 0.015	57.48% ± 0.015
	D1-D4	66.90% ± 0.010	66.90% ± 0.010	**69.31%** ± 0.010	66.07% ± 0.010	66.90% ± 0.010
	D1-D5	71.62% ± 0.010	71.62% ± 0.010	**74.28%** ± 0.010	74.19% ± 0.010	71.62% ± 0.010

that EDDM had the highest predictive performance in (D1-D4, D1-D5) dataset partitions using Perceptron as a base learner while Hinkley and DDM were not able to detect any drift in the dataset partitions.

Finally, the use of drift detection allowed the base learners to achieve higher classification performance only using a base learner without drift detection. However, we cannot make a strong statement that drift detectors outperformed other methods without drift method. The predictive accuracy for all base learners would fall once a concept drift had occurred until they adapt to the most recent reviews. The results showed that there is a significant concept drift problem in fake reviews detection that indicates the data changes over time, due to the spammers' behaviour.

We conclude that these drift detection methods are beneficial in real-world data streams. Nevertheless, we cannot make a strong statement depending on the predictive accuracy since the position of concept drift is unknown. ADWIN achieved the best performance in Yelp Zip, Yelp NYC and Yelp CHI real-world datasets partitions based on the predictive accuracy while EDDM achieved the best performance in most of Yelp consumer electronic real-world dataset partitions based on the predictive accuracy.

Similarly, to emphasise that the experiment's results were not achieved by chance, we used the non-parametric statistical analysis tool, the Friedman test, as described in section 5.1. The primary null hypothesis for the Friedman test is that all the methods which have the same performance can be missed. Figs. 11– 13 show the average ranking of the drift detection based on predictive accuracy using different base learners. We conclude that the behaviour of different methods depends on the datasets.

The Friedman test displayed a more significant difference in the approach scored with a confidence interval of $\alpha = 0.05$. The post-hoc Nemenyi test was used to conduct a paired comparison to show that the performance between two methods differed significantly when the difference between their results is more significant or equal to a critical difference (CD) (Demšar, 2006). The CD was calculated using the

Table 6
Predictive accuracy using SGD (LR) Classifier with 95% confidential interval showed the highest performance score for each drift detection method and without drift detection methods (Only classifier).

Dataset Name	Dataset parts	SGD (LR) Classifier	DDM	EDDM	ADWIN	Page Hinkley
Yelp NYC	D1-D1	62.21% ± 0.012	62.21% ± 0.012	60.74% ± 0.012	62.21% ± 0.012	62.21% ± 0.012
	D1-D2	67.17% ± 0.005	67.17% ± 0.005	67.32% ± 0.005	**81.11**% ± 0.005	67.17% ± 0.005
	D1-D3	70.20% ± 0.004	70.20% ± 0.004	70.52% ± 0.004	**86.58**% ± 0.004	70.93% ± 0.004
	D1-D4	72.05% ± 0.005	72.05% ± 0.005	72.43% ± 0.005	**86.75**% ± 0.005	72.05% ± 0.005
Yelp ZIP	D1-D1	67.05% ± 0.008	67.05% ± 0.008	**76.37**% ± 0.008	75.12% ± 0.008	67.05% ± 0.008
	D1-D2	71.41% ± 0.004	71.41% ± 0.004	76.65% ± 0.004	**88.22**% ± 0.004	71.41% ± 0.004
	D1-D3	74.67% ± 0.003	80.61% ± 0.003	82.30% ± 0.003	**92.14**% ± 0.003	85.52% ± 0.003
	D1-D4	76.39% ± 0.003	76.14% ± 0.003	81.66% ± 0.003	**91.72**% ± 0.003	79.99% ± 0.003
Yelp CHI	D1-D1	58.57% ± 0.019	58.57% ± 0.019	57.97% ± 0.019	58.57% ± 0.019	58.57% ± 0.019
	D1-D2	62.20% ± 0.009	62.20% ± 0.009	62.25% ± 0.009	**69.35**% ± 0.009	62.20% ± 0.009
	D1-D3	63.73% ± 0.009	63.73% ± 0.009	63.73% ± 0.009	**73.52**% ± 0.009	63.73% ± 0.009
Yelp Consumer Electronic	D1-D1	59.49% ± 0.043	59.49% ± 0.043	59.49% ± 0.043	59.49% ± 0.043	59.49% ± 0.043
	D1-D2	57.75% ± 0.022	57.75% ± 0.022	55.67% ± 0.022	57.75% ± 0.022	57.75% ± 0.022
	D1-D3	61.01% ± 0.015	61.01% ± 0.015	**62.93%** ± 0.015	61.01% ± 0.015	61.01% ± 0.015
	D1-D4	69.37% ± 0.010	70.81% ± 0.009	**74.03%** ± **0.009**	71.30% ± 0.009	71.31% ± 0.009
	D1-D5	73.91% ± 0.010	73.91% ± 0.010	81.65% ± 0.009	**81.67%** ± **0.009**	73.91% ± 0.010

Table 7
Predictive accuracy using SGD (SVM) Classifier with 95% confidential interval showed the highest performance score for each drift detection method and without drift detection methods (Only classifier).

Dataset Name	Dataset parts	SGD (SVM) Classifier	DDM	EDDM	ADWIN	Page Hinkley
Yelp NYC	D1-D1	60.74% ± 0.012	61.87% ± 0.012	**62.35**% ± 0.012	60.74% ± 0.012	60.74% ± 0.012
	D1-D2	67.31% ± 0.005	67.93% ± 0.005	69.02% ± 0.005	**76.38**% ± 0.005	67.31% ± 0.005
	D1-D3	71.84% ± 0.004	72.49% ± 0.004	73.92% ± 0.004	**84.85**% ± 0.003	71.84% ± 0.004
Yelp ZIP	D1-D1	66.48% ± 0.009	66.48% ± 0.009	66.70% ± 0.009	**71.49**% ± 0.009	66.48% ± 0.009
	D1-D2	73.61% ± 0.004	73.61% ± 0.004	74.21% ± 0.004	**85.75**% ± 0.004	73.61% ± 0.004
	D1-D3	78.57% ± 0.003	79.14% ± 0.003	79.27% ± 0.003	**91.35**% ± 0.003	80.45% ± 0.002
Yelp CHI	D1-D1	59.25% ± 0.019	**59.25**% ± 0.019	57.34% ± 0.019	59.25% ± 0.019	59.25% ± 0.019
	D1-D2	62.92% ± 0.009	62.92% ± 0.009	63.62% ± 0.009	**67.12**% ± 0.009	62.92% ± 0.009
	D1-D3	64.13% ± 0.009	64.13% ± 0.009	65.84% ± 0.009	**68.17**% ± 0.009	65.10% ± 0.009
Yelp Consumer Electronic						
	D1-D1	57.17% ± 0.043	57.17% ± 0.043	57.17% ± 0.043	57.17% ± 0.043	57.17% ± 0.043
	D1-D2	56.33% ± 0.022	56.33% ± 0.022	54.68% ± 0.022	56.33% ± 0.022	56.33% ± 0.022
	D1-D3	58.08% ± 0.015	58.08% ± 0.015	**59.24**% ± 0.015	58.08% ± 0.015	58.08% ± 0.015
	D1-D4	67.09% ± 0.010	67.09% ± 0.010	**69.03%** ± **0.010**	68.86% ± 0.010	67.10% ± 0.010
	D1-D5	72.31% ± 0.010	72.31% ± 0.010	76.67% ± 0.009	**76.72%** ± **0.009**	72.31% ± 0.010

Nemenyi test with a confidence interval of $\alpha = 0.05$, which was 1.53. Figs. 11–13 display results from the Yelp NYC, Yelp ZIP and Yelp CHI real-world datasets. These results indicate that the variance of the base learners is dissimilar; this rejects the null hypothesis. For all datasets, ADWIN performed significantly better (the lowest average ranking) than DDM and Page Hinkley, using SVM as a base learner. The ADWIN and EDDM methods were hugely better in performance using SVM as a base learner in term of predictive accuracy. EDDM, DDM and Page Hinkley have similar statistical performance.

Similarly, ADWIN is significantly better (the lowest average ranking) DDM in all datasets, using LR as a base learner. The ADWIN and EDDM methods were hugely better in performance than LR as a base learner in term of predictive accuracy. EDDM, Page Hinkley and DDM have the same statistical performance.

Lastly, using PNN as a base learner, ADWIN was significantly better than the DDM best method for all the datasets with the lowest average ranking. EDDM, Page Hinkley and DDM have the same statistical performance.

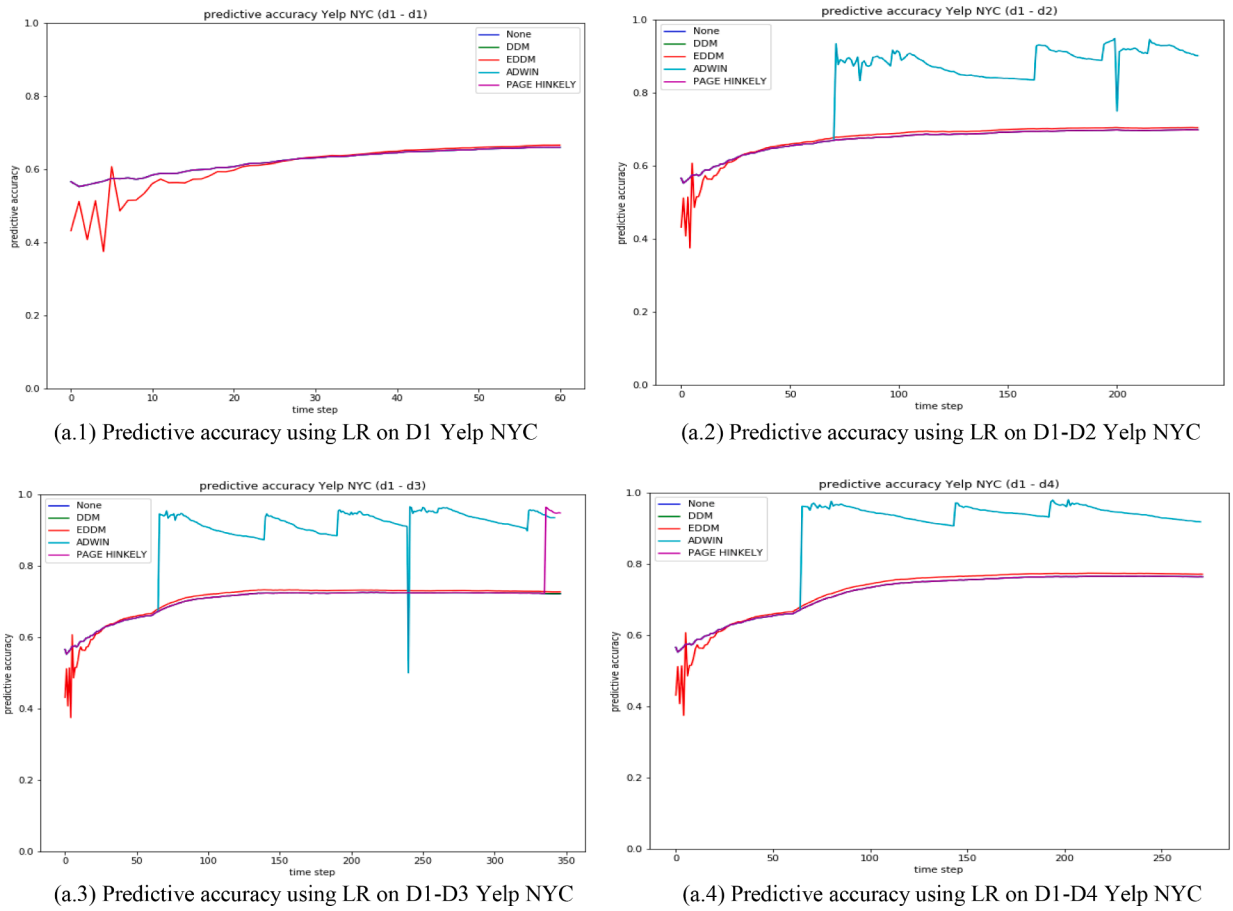
5.2.2. Number of Drifts

As Yelp datasets are considered to be real-world datasets, it is unknown if they have drift or not and there is no ground truth for drifts, if they exist. We do not know whether drifts occur; it is challenging obtaining the location of drift in these datasets due to velocity and unbinds size of real-world data streaming. We cannot use the false positive and negative for drift detector. As a consequence, we only utilise predictive accuracy, number of drifts and evaluation time as evaluation metrics.

The number of drifts for Yelp real-world datasets, as shown in Table 8, can show the ability of each concept drift detection algorithm in detecting concept drift. The results showed that using SVM as a base learner achieved the best results with EDDM methods on Yelp NYC, where LR and PNN had the worst results with Page Hinkley and DDM drift detection methods.

On Yelp ZIP, using LR a base learner with EDDM achieved the best results, while PNN as the base learner with Page Hinkley and DDM had the worst results. While for Yelp CHI, SVM had the best results with EDDM while PNN and LR and SVM had the worst results with Page Hinkley and DDM methods. For Yelp CHI, SVM had the best results with EDDM while PNN and LR and SVM had the worst results with Page Hinkley and DDM methods. Lastly, for Yelp consumer electronic dataset, using SVM, a base learner with EDDM achieved the best results, while PNN as the base learner with Page Hinkley and DDM had the worst results.

Despite this observation that the concept drift detection is better than with no detection method, the results given in the table below show the ability of drift detection methods to detect the number of alarms which can lead to higher classification performance. However, it does not necessarily imply that with no drift method are outperformed by drift detection methods.



(a.1) Predictive accuracy using LR on D1 Yelp NYC

(a.2) Predictive accuracy using LR on D1-D2 Yelp NYC

(a.3) Predictive accuracy using LR on D1-D3 Yelp NYC

(a.4) Predictive accuracy using LR on D1-D4 Yelp NYC

Fig. 7. Predictive accuracies in the real-world Yelp NYC dataset partitions using LR, SVM and PNN base learners where the x-axis is present the time steps and y-axis present the predictive accuracy.

Pesaranghader and Viktor (2016) and Pesaranghader et al. (2018b) found that the highest number of drifts did not lead to the highest accuracy. This is precisely what occurs during our experiments, where the EDDM detected the highest number of drifts and achieved the second highest accuracy after ADWIN algorithm.

Huang et al. (2015) and Pesaranghader and Viktor (2016) argued that there are two reasons the drift detection methods lead to the same accuracies, while some of them detect less drift than others. First, the drift detectors caused less false positives when the drift detection algorithm had not detected any drifts or detect less number than other methods. Second, having fewer drifts detected leads to lower accuracy and suggests significant false positives. Therefore, depending on the findings in this literature, it is more likely that EDDM caused false positives, the number of points incorrectly considered as drifts over the total number of points which are not drifts, whilst Page Hinkley caused false negatives, the number of drifts incorrectly left unidentified over the total number of drifts in a stream. Lastly, it is worth mentioning that EDDM frequently detects drift in the early stages if the distance between two classification errors is small (Pesaranghader et al., 2018b), and this has precisely occurred in our work.

5.2.3. Evaluation time

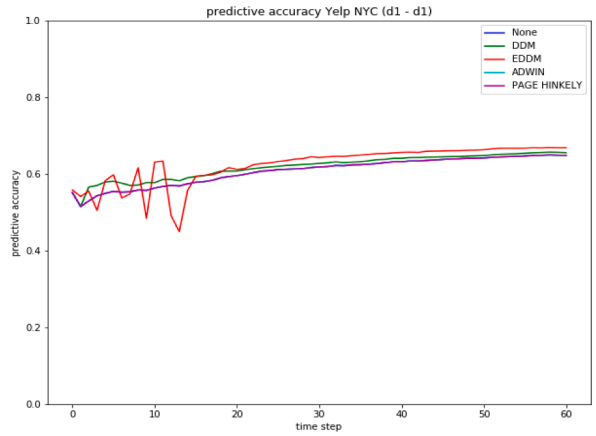
Tests were performed on a MacBook Pro computer, running a 2.7 GHz Intel Core i7. 16 GB of main memory and running the mac operating system. The evaluation times (run-time per second) are presented in Table 9 for four real-world Yelp datasets. As shown in the table 9, ADWIN was the fastest method on all the datasets as it compares all sub-windows of its sliding window for drift detection. These are represented as dynamic and fast-reactive in the presence of drift. While DDM was the worst dataset due to the creation of the base learner was trained since the warning level. We can conclude that the best methods depend on both the dataset and base learner.

6. Conclusions

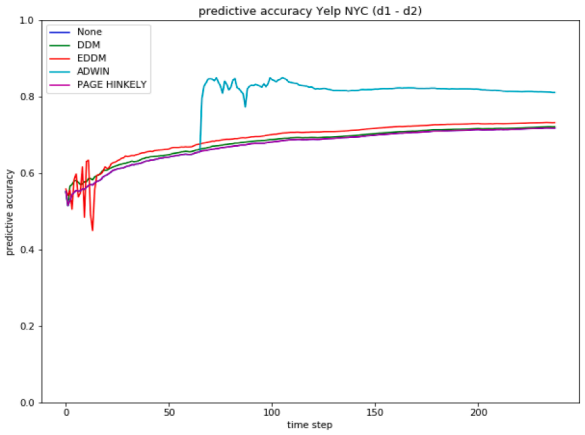
We analysed one significant question centred around fake reviews detection that required a diverse evaluation technique to arrive at a solution. To provide the answers, we conducted a comprehensive analysis to assess the correlation between the concept drift and classification performance using four real-world datasets. All of these experiments were implemented using python language. To the best of our knowledge, this is an innovative paper which studies the correlation between concept drift score and classification performance in fake reviews detection. Moreover, we divided the experiments into three parts:

- Will the performance of the classification techniques affect sorting the reviews in chronological order when changed over-time?

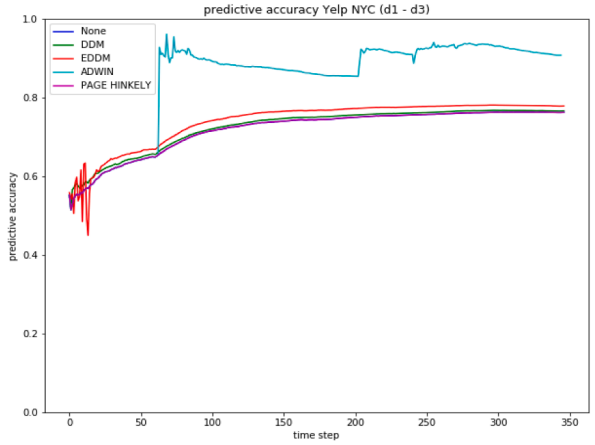
We used three machine learning methods to study the characteristics of the changes in the reviews over time. Results for the classification performance provided enough evidence indicating that the performance for all methods decreased over time due to the time and nature of the reviews. Therefore, we recommend updating the prediction model frequently. Furthermore, these experiments were conducted using different classifiers where the LR is significantly better than SVM and PNN in Yelp ZIP, Yelp NYC and Yelp CHI datasets while SVM and PNN



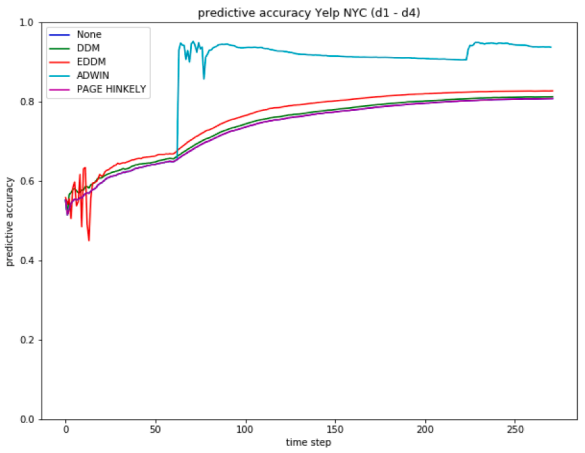
(b.1) Predictive accuracy using SVM on D1 Yelp NYC



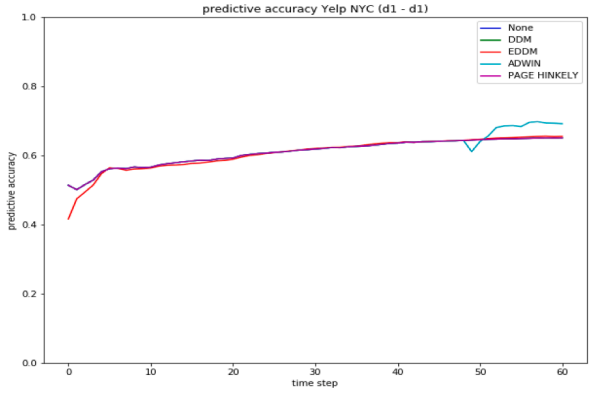
(b.2) Predictive accuracy using SVM on D1-D2 Yelp NYC



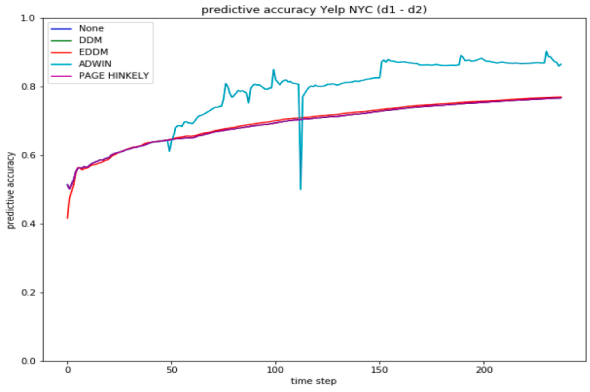
(b.3) Predictive accuracy using SVM on D1-D3 Yelp NYC



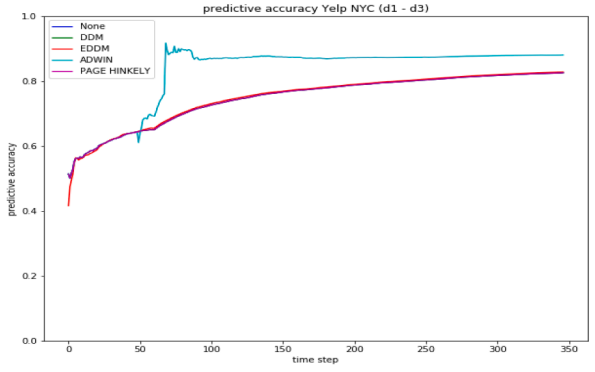
(b.4) Predictive accuracy using SVM on D1-D4 Yelp NYC



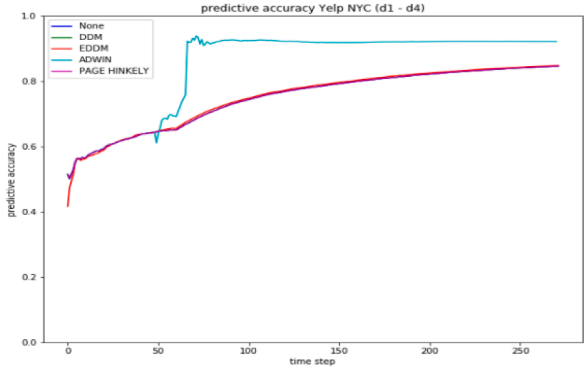
(c.1) Predictive accuracy using PNN on D1 Yelp NYC



(c.2) Predictive accuracy using PNN on D1-D2 Yelp NYC

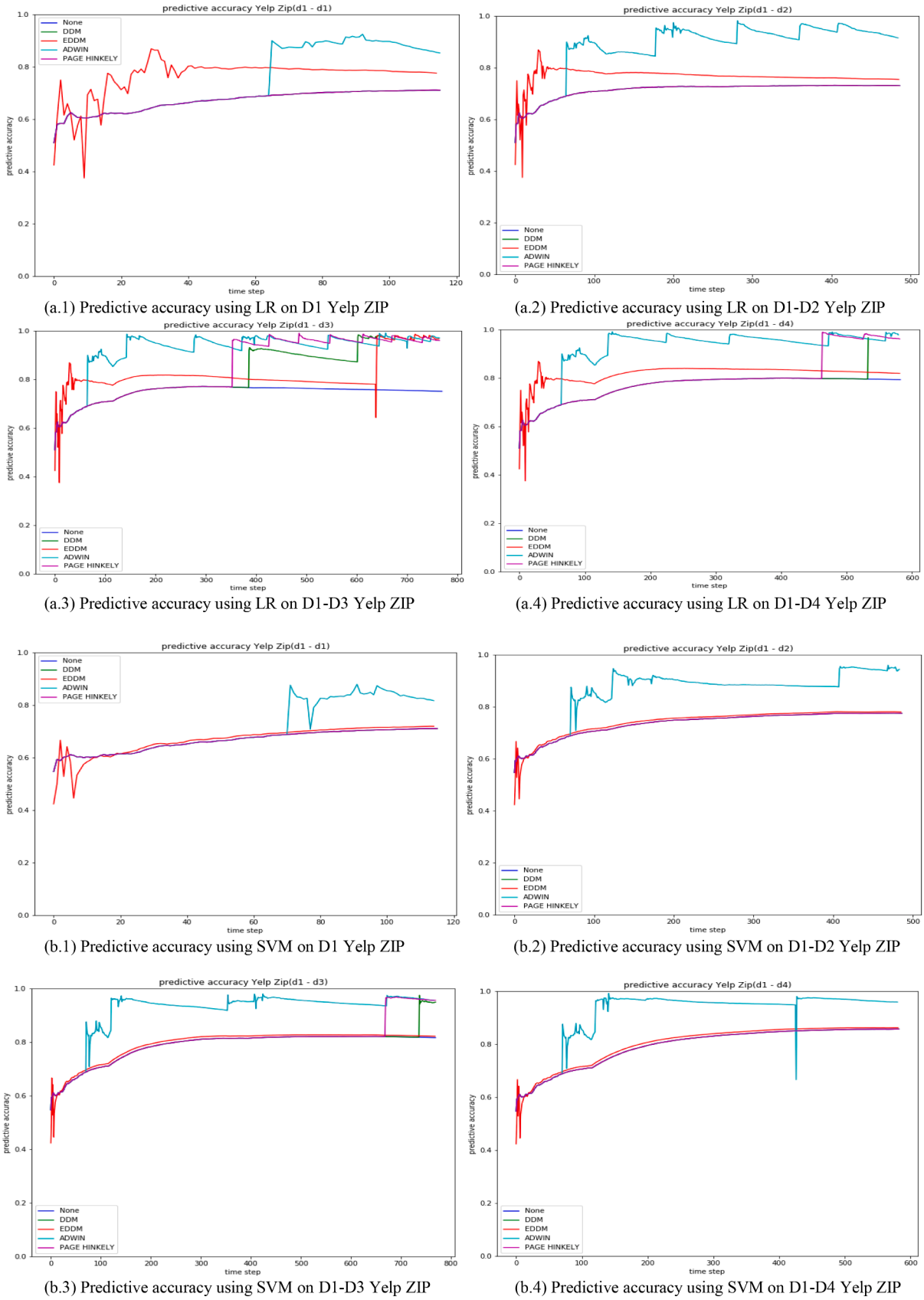


(c.3) Predictive accuracy using PNN on D1-D3 Yelp NYC



(c.4) Predictive accuracy using PNN on D1-D4 Yelp NYC

Fig. 7. (*continued*).



(a.1) Predictive accuracy using LR on D1 Yelp ZIP

(a.2) Predictive accuracy using LR on D1-D2 Yelp ZIP

(a.3) Predictive accuracy using LR on D1-D3 Yelp ZIP

(a.4) Predictive accuracy using LR on D1-D4 Yelp ZIP

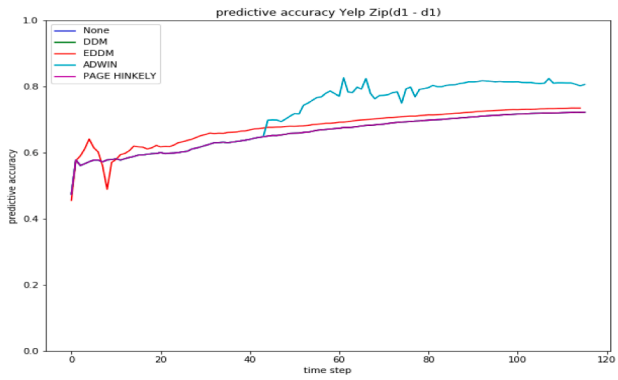
(b.1) Predictive accuracy using SVM on D1 Yelp ZIP

(b.2) Predictive accuracy using SVM on D1-D2 Yelp ZIP

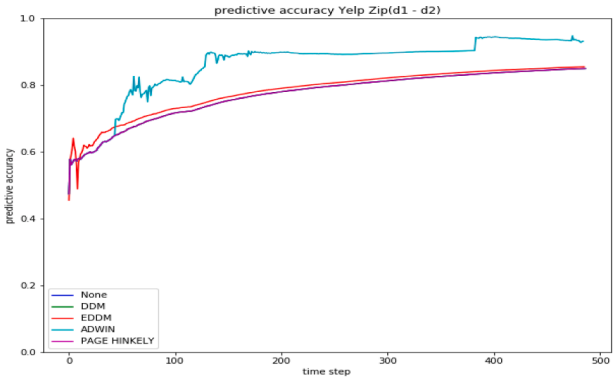
(b.3) Predictive accuracy using SVM on D1-D3 Yelp ZIP

(b.4) Predictive accuracy using SVM on D1-D4 Yelp ZIP

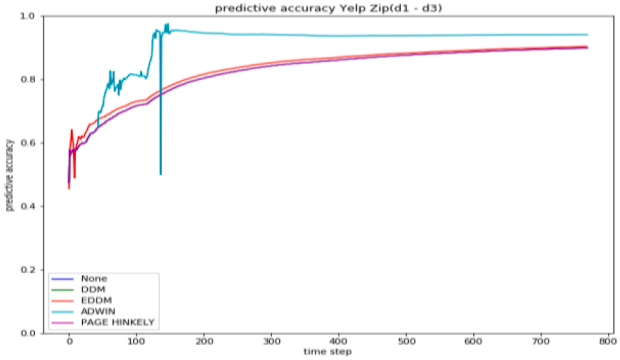
Fig. 8. Predictive accuracies in the real-world Yelp ZIP dataset partitions using LR, SVM and PNN base learners where the x-axis is present the time steps and y-axis present the predictive accuracy.



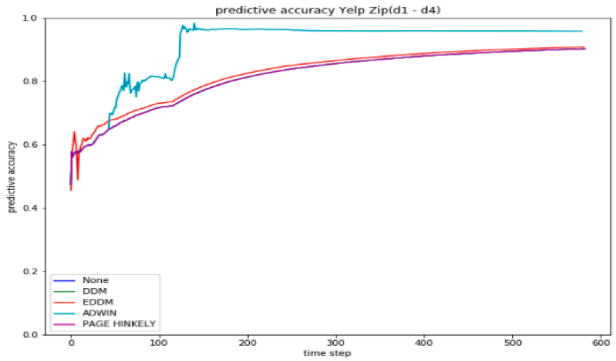
(c.1) Predictive accuracy using PNN on D1 Yelp ZIP



(c.2) Predictive accuracy using PNN on D1-D2 Yelp ZIP

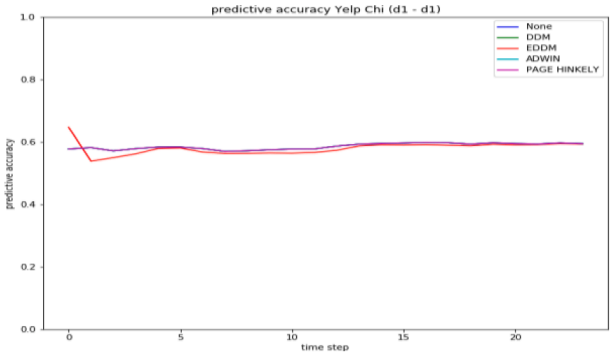


(c.3) Predictive accuracy using PNN on D1-D3 Yelp ZIP

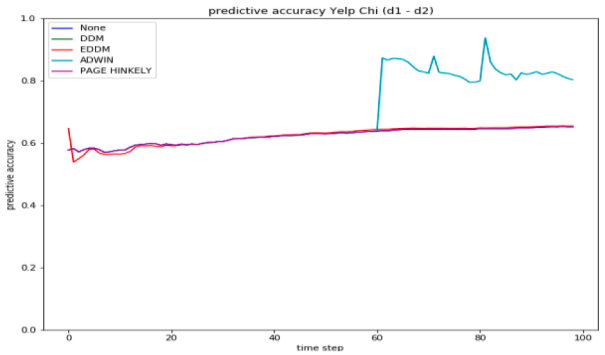


(c.4) Predictive accuracy using PNN on D1-D4 Yelp ZIP

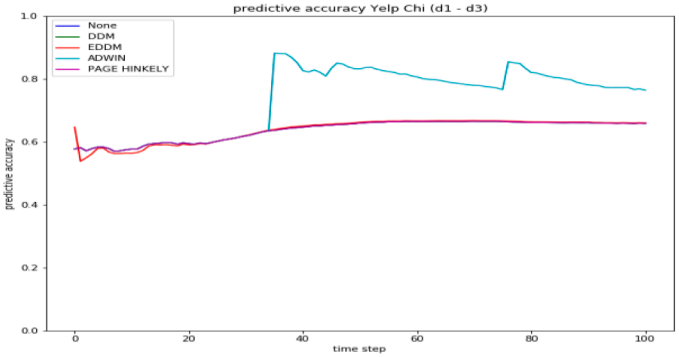
Fig. 8. (*continued*).



(a.1) Predictive accuracy using LR on D1 Yelp CHI

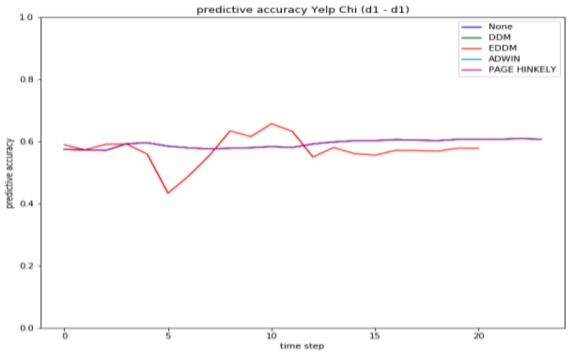


(a.2) Predictive accuracy using LR on D1-D2 Yelp CHI

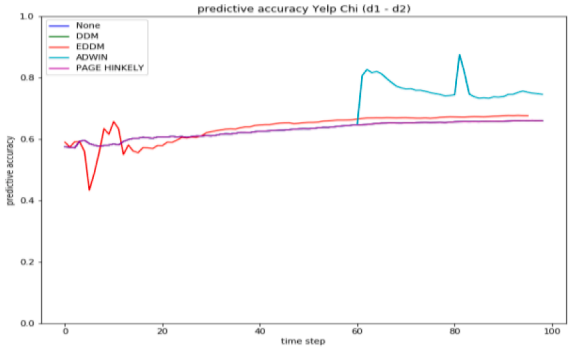


(a.3) Predictive accuracy using LR on D1-D3 Yelp CHI

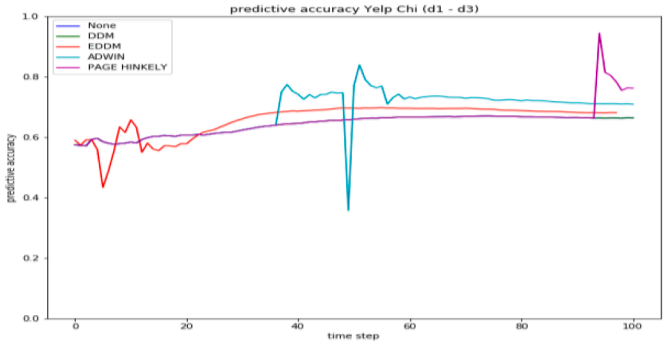
Fig. 9. Predictive accuracies in the real-world Yelp CHI dataset partitions using LR, SVM and PNN base learners where the x-axis is present the time steps and y-axis present the predictive accuracy.



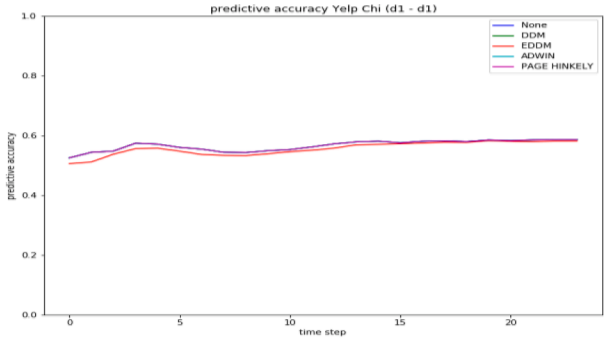
(b.1) Predictive accuracy using SVM on D1 Yelp CHI



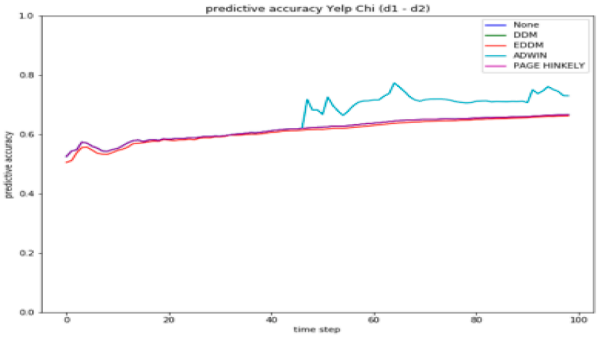
(b.2) Predictive accuracy using SVM on D1-D2 Yelp CHI



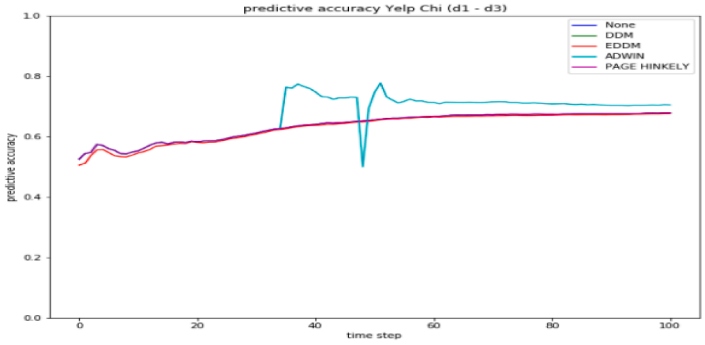
(b.3) Predictive accuracy using SVM on D1-D3 Yelp CHI



(c.1) Predictive accuracy using PNN on D1 Yelp CHI



(c.2) Predictive accuracy using PNN on D1-D2 Yelp CHI



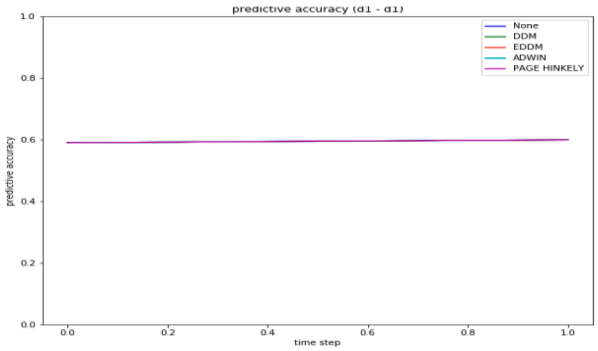
(c.3) Predictive accuracy using PNN on D1-D3 Yelp CHI

Fig. 9. (*continued*).

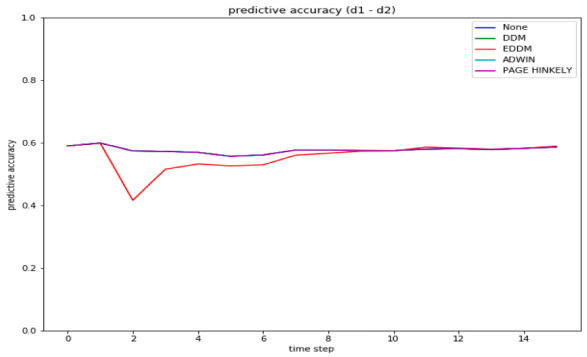
are significantly better than LR in Yelp consumer electronic datasets.

- Is there any significant concept drift problem in the fake review datasets?

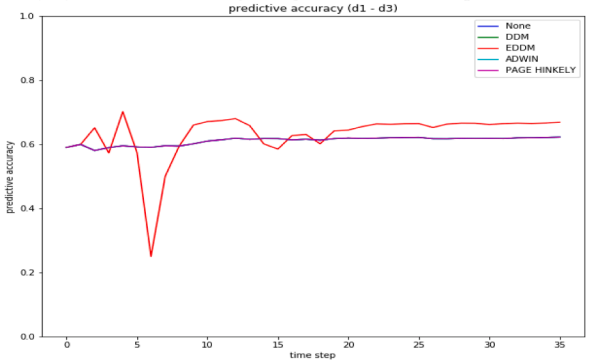
  In terms of concept drift detection, four benchmark drift detection methods (DDM, EDDM, Page Hinkley and ADWIN) with three base learners (SVM, LR and PNN) were used to analyse the concept drift in the fake review datasets in a fully supervised setting. The experimental results showed that there is a significant concept drift problem in fake review detection due to the spammers' behaviour which indicates that there is a temporal dependency between reviews of a stream.



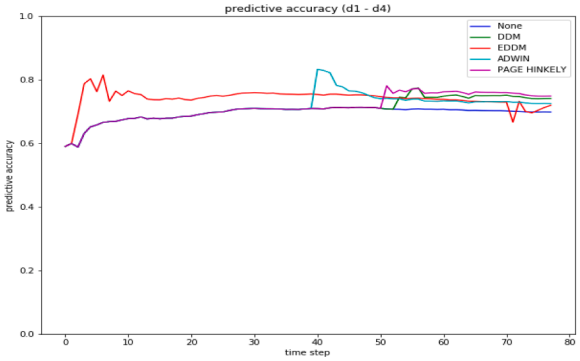
(a.1) Predictive accuracy using LR on D1-D1 Yelp consumer



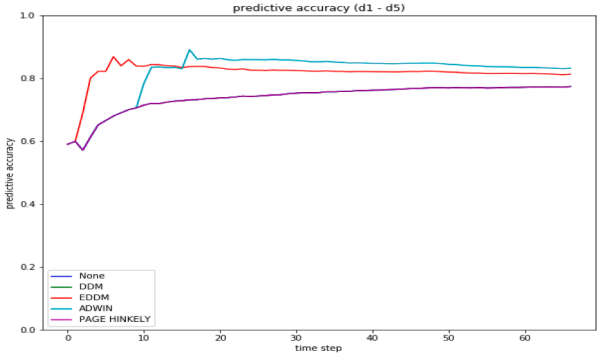
(a.2) Predictive accuracy using LR on D1-D2 Yelp consumer



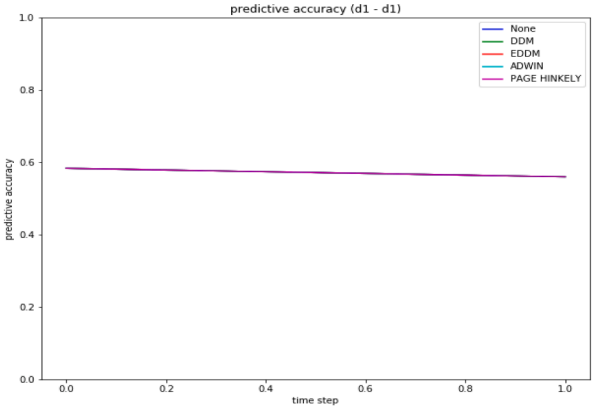
(a.4) Predictive accuracy using LR on D1-D3 Yelp consumer



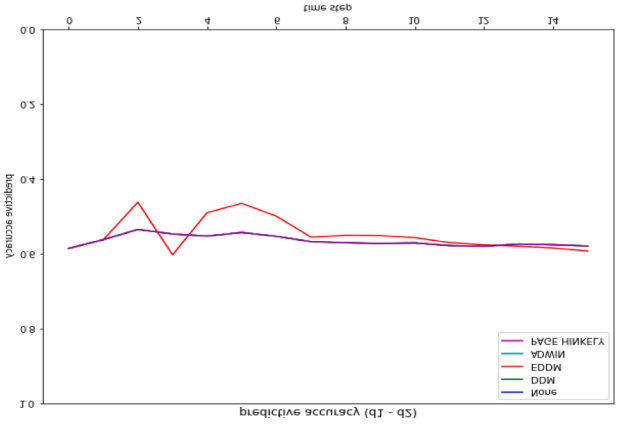
(a.4) Predictive accuracy using LR on D1-D4 Yelp consumer



(a.5) Predictive accuracy using LR on D1-D5 Yelp consumer

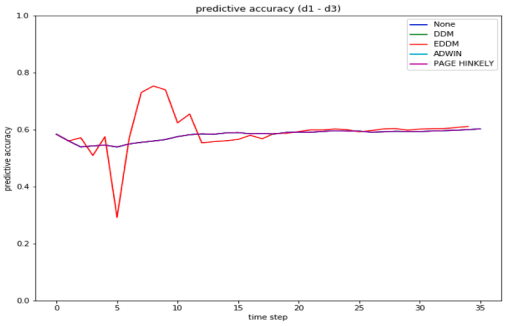


(b.1) Predictive accuracy using SVM on D1-D1 Yelp consumer

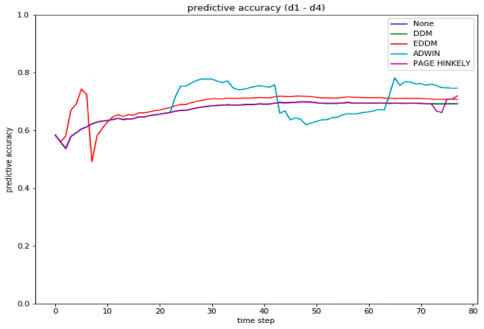


(b.2) Predictive accuracy using SVM on D1-D2 Yelp consumer

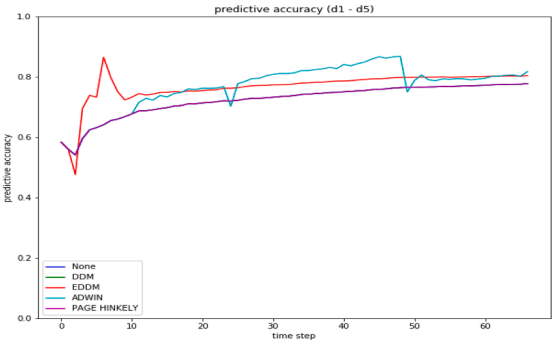
Fig. 10. Predictive accuracies in the real-world Yelp Consumer Electronic dataset partitions using LR, SVM and PNN base learners where the x-axis is present the time steps and y-axis present the predictive accuracy.



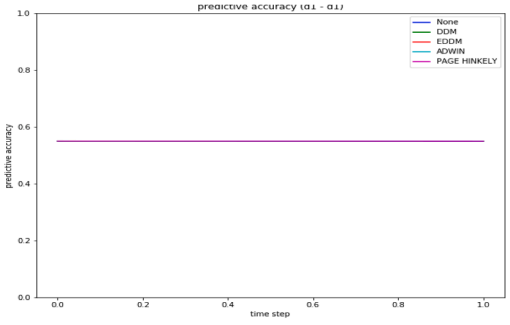
(b.3) Predictive accuracy using SVM on D1-D3 Yelp consumer



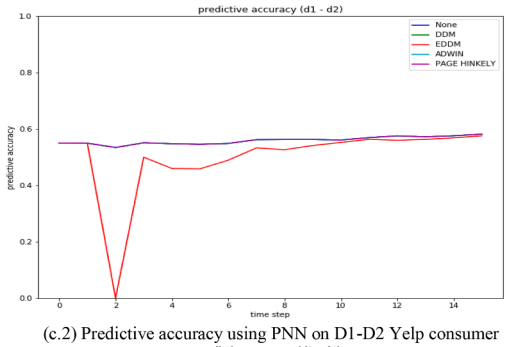
(b.4) Predictive accuracy using SVM on D1-D4 Yelp consumer



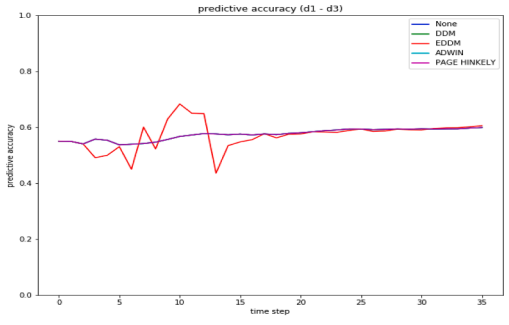
(b.5) Predictive accuracy using SVM on D1-D1 Yelp consumer



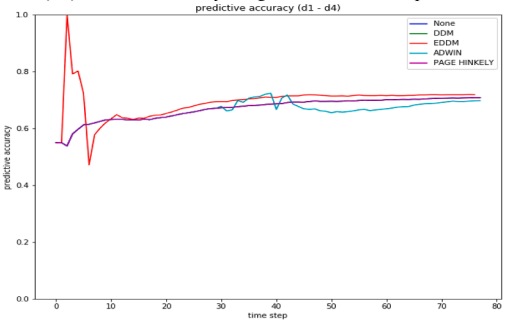
(c.1) Predictive accuracy using PNN on D1-D1 Yelp consumer



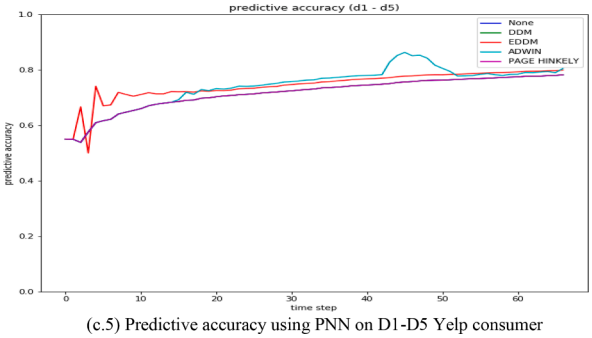
(c.2) Predictive accuracy using PNN on D1-D2 Yelp consumer



(c.3) Predictive accuracy using PNN on D1-D3 Yelp consumer



(c.4) Predictive accuracy using PNN on D1-D4 Yelp consumer



(c.5) Predictive accuracy using PNN on D1-D5 Yelp consumer

Fig. 10. (*continued*).

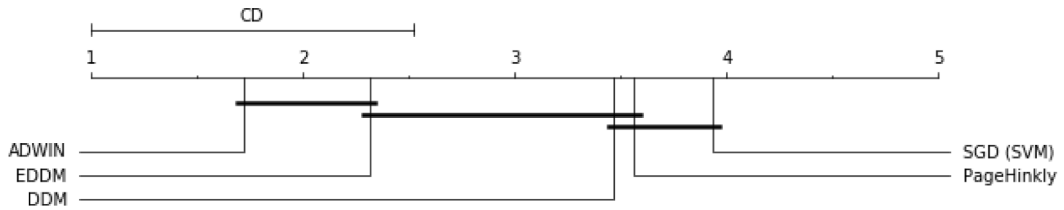


Fig. 11. Comparison of all classifiers against each other using Nemenyi test with SVM as a base learner which showed the statistical difference between various concept drift detection methods at (p-value = 0.05). where the classifiers are not connected.

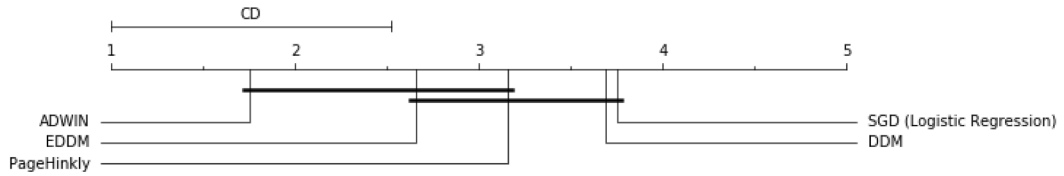


Fig. 12. Comparison of all classifiers against each other using Nemenyi test with LR as a base learner which showed the statistical difference between various concept drift detection methods at (p-value = 0.05). where the classifiers are not connected.

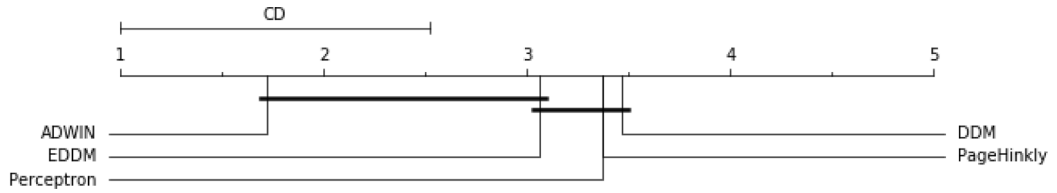


Fig. 13. Comparison of all classifiers against each other using Nemenyi test with PNN as a base learner which showed the statistical difference between various concept drift detection methods at (p-value = 0.05). where the classifiers are not connected.

Table 8
Number of drifts detected in four real world Yelp datasets partitions.

Dataset Name	Dataset parts	DDM			EDDM			ADWIN			Page Hinkley		
Yelp NYC		SVM	LR	PNN	SVM	LR	PNN	SVM	LR	PNN	SVM	LR	PNN
	D1-D1	1	0	0	14	6	2	0	0	1	0	0	0
	D1-D2	1	0	0	14	6	2	6	14	9	0	0	0
	D1-D3	1	0	0	14	6	2	10	16	5	0	1	0
	D1-D4	1	0	0	14	6	2	7	10	8	0	0	0
Yelp ZIP													
	D1-D1	0	0	0	10	27	7	5	5	8	0	0	0
	D1-D2	0	0	0	10	27	7	12	17	16	0	0	0
	D1-D3	1	2	0	10	35	7	14	22	15	1	7	0
	D1-D4	0	3	0	10	27	7	12	14	14	0	2	0
Yelp CHI													
	D1-D1	0	0	0	8	2	1	0	0	0	0	0	0
	D1-D2	0	0	0	8	2	1	2	5	5	0	0	0
	D1-D3	0	0	0	8	2	1	5	5	3	1	0	1
Yelp Consumer Electronic													
	D1-D1	0	0	0	2	1	0	0	0	0	0	0	0
	D1-D2	0	0	0	4	1	3	0	0	0	0	0	0
	D1-D3	0	0	0	14	12	12	0	0	0	0	0	0
	D1-D4	0	1	0	6	5	6	4	1	2	1	1	0
	D1-D5	0	0	0	4	3	3	4	2	3	0	0	0

Regarding predictive accuracy, the results indicate that:

- Using PNN as a base learner, ADWIN had the best average predictive accuracies with lowest average ranking in the Yelp ZIP, Yelp NYC and Yelp CHI datasets. EDDM had the best average predictive accuracies with second-lowest average ranking in the Yelp consumer electronic dataset. DDM was the worst in both metrics.
- Using SVM as a base learner, ADWIN and EDDM had the best predictive accuracies in the all datasets. ADWIN was the lowest average ranking, while Page Hinkley was the worst in both metrics.
- Using LR as a base learner, ADWIN had the best predictive accuracies with lowest average ranking in the Yelp ZIP, Yelp NYC and Yelp CHI datasets. ADWIN and EDDM had the best predictive accuracies with the second-lowest average ranking in the Yelp consumer electronic dataset. DDM was the worst in both metrics.

Table 9
Evaluation time (Run time per second) on four real world Yelp datasets which represents the evaluation time for each concept drift detection methods.

Dataset Name	Dataset parts	DDM			EDDM			ADWIN			Page Hinkley		
YelpNYC		SVM	LR	PNN	SVM	LR	PNN	SVM	LR	PNN	SVM	LR	PNN
	D1-D1	5.72	7.71	4.21	4.22	7.77	3.94	4.21	8.39	4.53	3.88	8.23	4.05
	D1-D2	34.58	55.18	26.38	28.69	54.5	25.05	18.66	60.21	16.6	25.08	56.58	33.01
	D1-D3	57.76	118.57	52.94	64.37	116.72	51.08	24.41	127.09	36.76	45.86	119.51	55.16
	D1-D4	38.39	72.85	44.12	35.24	77.65	32.04	25.76	80.24	25.04	43.13	74.81	30.88
Yelp ZIP													
	D1-D1	8.84	8.53	7.97	8.53	6.57	8.17	9.44	6.97	6.99	9.39	7.913	8.55
	D1-D2	90.73	76.8	77.14	89.27	72.58	82.1	46.79	31.91	40.16	92.2	77.27	80.6
	D1-D3	202.36	85.87	192.17	210.59	196.45	185.51	78.78	51.44	143.02	177.54	67.14	190.2
	D1-D4	131.41	95.5	116.37	122.94	145.73	111.81	57.93	40.52	81.36	124.77	76.94	114.64
Yelp CHI													
	D1-D1	1.35	1.57	1.69	3.53	1.71	1.73	3.33	3.01	2.72	1.55	2.09	1.53
	D1-D2	12.56	7.89	8.13	9.76	8.6	8.02	7.07	7.42	6.76	11.18	8.89	7.93
	D1-D3	19.8	11.09	8.37	16.22	17.29	8.86	7.01	9.54	7.14	7.76	8.91	8.73
Yelp Consumer Electronic													
	D1-D1	0.78	0.17	0.15	0.78	0.21	0.18	0.99	0.18	0.15	0.88	0.16	0.16
	D1-D2	1.12	1.32	1.01	1.14	0.99	0.98	1.10	0.95	0.96	1.00	0.87	0.91
	D1-D3	2.28	2.19	2.05	2.04	2.12	2.04	2.27	2.26	2.19	2.04	2.04	2.05
	D1-D4	6.43	5.74	4.98	5.67	5.66	4.83	4.62	4.88	4.65	5.27	4.42	5.07
	D1-D5	4.53	4.69	4.15	4.11	3.89	4.01	4.16	4.15	3.89	4.70	4.15	4.10

- Analysing most of the tested datasets, ADWIN, and EDDM were the best methods due to both metrics, predictive accuracy and average ranking.

Regarding evaluation time, using PNN as a base learner, ADWIN was the fastest method in 11 out of 16 of the real-world Yelp datasets partitions, which was faster than a base learner in some real-world datasets, while DDM was the worst in most of the real-world Yelp datasets for each arriving instance.

Regarding the number of drifts, the results showed that using SVM as a base learner achieved the best results with EDDM methods on Yelp NYC where LR and PNN had the worst results with Page Hinkley and DDM. On Yelp ZIP dataset, EDDM with LR as the base learner achieved the best results, while using PNN as a base learner with Page Hinkley and DDM had the worst results. For Yelp CHI dataset, the SVM had the best results with EDDM, while PNN and LR and SVM had the worst results with Page Hinkley and DDM methods. Lastly, for Yelp consumer electronic dataset, EDDM with SVM achieved the best results while DDM and Page Hinckley with PNN had the worst results.

- Is there any correlation between the concept drift and the classification performance of reviews?

The experimental results in this paper indicate that there is a strong negative correlation between the concept drift problem and classification performance since the concept drift negatively affects the prediction model performance.

To conclude, it is essential to mention that the experiments reported in this paper represent a first step towards building a more effective model for fake review detection and determine which drift detection methods are best at solving concept drift in the real-world data stream.

Our future work will be proposed a new efficient method to handle the concept drift problem in fake reviews detection.

CRediT authorship contribution statement

Rami Mohawesh: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Son Tran:** Software, Writing - review & editing, Validation, Resources, Supervision, Project administration. **Robert Ollington:** Writing - review & editing, Validation, Resources, Supervision, Project administration. **Shuxiang Xu:** Writing - review & editing, Validation, Resources, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aggarwal, C. C. (2005). On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering, 17*, 587–600.

Al Najada, H., & Zhu, X. (2014). iSRD: Spam review detection with imbalanced data distributions. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)* (pp. 553–560). IEEE.

Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006). Early drift detection method. In *Fourth international workshop on knowledge discovery from data streams* (pp. 77–86).

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management, 56*, 1234–1244.

Barros, R. S., Cabral, D. R., Gonçalves, P. M., Jr, & Santos, S. G. (2017). RDDM: Reactive drift detection method. *Expert Systems with Applications, 90*, 344–355.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *International conference on discovery science* (pp. 1–15). Springer.

Bifet, A., & Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining* (pp. 443–448). SIAM.

Bifet, A., Gavaldà, R., Holmes, G., & Pfahringer, B. (2018). *Machine learning for data streams: With practical examples in MOA*. MIT Press.

Bouchachia, A., & Vanaret, C. (2013). GT2FC: An online growing interval type-2 self-learning fuzzy classifier. *IEEE Transactions on Fuzzy Systems, 22*, 999–1018.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*, 2493–2537.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data, 2*, 23.

Das, Bijoyan, & Chakraborty, Sarit (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. *arxiv*. arXiv:1806.06407.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General), 147*, 278–290.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Drzadzewski, G., & Tompa, F. W. (2016). Partial materialization for online analytical processing over multi-tagged document collections. *Knowledge and Information Systems, 47*, 697–732.

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies, 8*, 1–119.

Frías-Blanco, I., del Campo-Ávila, J., Ramos-Jimenez, G., Morales-Bueno, R., Ortiz-Díaz, A., & Caballero-Mota, Y. (2014). Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Transactions on Knowledge and Data Engineering, 27*, 810–823.

Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In *Brazilian symposium on artificial intelligence* (pp. 286–295). Springer.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR), 46*, 1–37.

Harris, C. G. (2012). Detecting deceptive opinion spam using human computation. *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., & Flores, J. J. G. (2017). Cross-domain deception detection using support vector networks. *Soft Computing, 21*, 585–595.

Ho-Dac, N. N., Carson, S. J., & Moore, W. L. (2013). The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? *Journal of Marketing, 77*, 37–53.

Huang, D. T. J., Koh, Y. S., Dobbie, G., & Bifet, A. (2015). Drift detection using stream volatility. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 417–432). Springer.

Jindal, N., & Liu, B. (2007). Analyzing and detecting review spam. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (pp. 547–552). IEEE.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230). ACM.

Jing, Y. (2014). Research of deceptive opinion spam recognition based on deep learning. In East China Normal Univ., Shanghai, China, Tech. Rep.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137–142). Springer.

Karumanchi, A., Fu, L., & Deng, J. (2018). Prediction of Review Sentiment and Detection of Fake Reviews in Social Media. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE) (pp. 181-186): The Steering Committee of The World Congress in Computer Science, Computer….

Khurshid, F., Zhu, Y., Xu, Z., Ahmad, M., & Ahmad, M. (2018). Enactment of ensemble learning for review spam detection on selected features. *International Journal of Computational Intelligence Systems, 12*, 387–394.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion, 37*, 132–156.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196).

Li, H., Chen, Z., Mukherjee, A., Liu, B., & Shao, J. (2015). Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. *Ninth international AAAI conference on web and social Media*.

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1566–1576).

Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. *Neurocomputing, 254*, 33–41.

Li, L., Ren, W., Qin, B., & Liu, T. (2015). Learning document representation for deceptive opinion spam detection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 393–404). Springer.

Lin, Y., Zhu, T., Wang, X., Zhang, J., & Zhou, A. (2014). Towards online review spam detection. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 341–342). ACM.

Liu, A., Song, Y., Zhang, G., & Lu, J. (2017). Regional concept drift detection and density synchronized drift adaptation. *IJCAI International Joint Conference on Artificial Intelligence*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*, 1–167.

Liu, W., Jing, W., & Li, Y. (2019). Incorporating feature representation into BiLSTM for deceptive review detection. *Computing*, 1–15.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering, 31*, 2346–2363.

Lu, N., Zhang, G., & Lu, J. (2014). Concept drift detection via competence models. *Artificial Intelligence, 209*, 11–28.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62*, 3412–3427.

Méndez, J. R., Iglesias, E. L., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2005). Tokenising, stemming and stopword removal on anti-spam filtering domain. In *Conference of the Spanish Association for Artificial Intelligence* (pp. 449–458). Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mitchell, T. M. (1997). *Machine learning*. McGraw-hill New York.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013a). Fake review detection: Classification and analysis of real and pseudo reviews. UIC-CS-03-2013. Technical Report.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013b). What yelp fake review filter might be doing? In Seventh international AAAI conference on weblogs and social media.

Nguyen, H.-L., Woon, Y.-K., & Ng, W.-K. (2015). A survey on data stream clustering and classification. *Knowledge and information systems, 45*, 535–569.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 309–319). Association for Computational Linguistics.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100–115.

Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. *Computational Linguistics, 35*, 311–312.

Pears, R., Sakthithasan, S., & Koh, Y. S. (2014). Detecting concept change in dynamic data streams. *Machine Learning, 97*, 259–293.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*, 559–572.

Pesaranghader, A., Viktor, H., & Paquet, E. (2018a). Reservoir of diverse adaptive learners and stacking fast hoeffding drift detection methods for evolving data streams. *Machine Learning, 107*, 1711–1743.

Pesaranghader, A., & Viktor, H. L. (2016). Fast hoeffding drift detection method for evolving data streams. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 96–111). Springer.

Pesaranghader, A., Viktor, H. L., & Paquet, E. (2018b). McDiarmid drift detection methods for evolving data streams. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–9). IEEE.

Prasad, B. R., & Agarwal, S. (2017). Critical parameter analysis of Vertical Hoeffding Tree for optimized performance using SAMOA. *International Journal of Machine Learning and Cybernetics, 8*, 1389–1402.

Rajaraman, A., Leskovec, J., & Ullmann, J. D. (2014). Mining data streams. In *Mining of Massive Datasets* (2nd ed.,, pp. 165–173). Cambridge University Press.

Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 985–994). ACM.

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences, 385*, 213–224.

Ren, Y., & Ji, D. (2019). Learning to detect deceptive opinion spam: A survey. *IEEE Access, 7*, 42934–42945.

Ren, Y., Wang, R., & Ji, D. (2016). A topic-enhanced word embedding for Twitter sentiment classification. *Information Sciences, 369*, 188–198.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386.

Ross, G. J., Adams, N. M., Tasoulis, D. K., & Hand, D. J. (2012). Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters, 33*, 191–198.

Sakthithasan, S., Pears, R., & Koh, Y. S. (2013). One pass concept change detection for data streams. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 461–472). Springer.

Schlimmer, J. C., & Granger, R. H. (1986). Incremental learning from noisy data. *Machine Learning, 1*, 317–354.

Scott, J. (1988). Social network analysis. *Sociology, 22*, 109–127.

Šilić, A., & Bašić, B. D. (2012). Exploring classification concept drift on a large news text corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 428–437). Springer.

Silva, R. M., Alberto, T. C., Almeida, T. A., & Yamakami, A. (2017). Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Expert Systems with Applications, 83*, 314–325.

Taloba, A. I., Eisa, D., & Ismail, S. S. (2018). A Comparative Study on using Principle Component Analysis with Different Text Classifiers. arXiv preprint arXiv: 1807.03283.

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work. *Computer Science Department, Trinity College Dublin, 106*, 58.

Wang, H., Fan, W., Yu, P. S., & Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 226–235).

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23*, 69–101.

Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., & Roli, F. (2015). Support vector machines under adversarial label contamination. *Neurocomputing, 160*, 53–62.

Zhang, P., Zhu, X., & Shi, Y. (2008). Categorizing and mining concept drifting data streams. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 812–820).

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning* (p. 116).

Zhang, Y., Chu, G., Li, P., Hu, X., & Wu, X. (2017). Three-layer concept drifting detection in text data streams. *Neurocomputing, 260*, 393–403.

Zhao, S., Xu, Z., Liu, L., & Guo, M. (2017). Towards accurate deceptive opinion spam detection based on word order-preserving CNN. arXiv preprint arXiv:1711.09181.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*, 133–148.