Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques

Elshrif Elmurngi, Abdelouahed Gherbi
Department of Software and IT Engineering
École de Technologie Supérieure
Montreal, Canada
Email: elshrif.elmurngi.1@ens.etsmtl.ca, abdelouahed.gherbi@etsmtl.ca

Abstract— **In recent years, Sentiment Analysis (SA) has become one of the most interesting topics in text analysis, due to its promising commercial benefits. One of the main issues facing SA is how to extract emotions inside the opinion, and how to detect fake positive reviews and fake negative reviews from opinion reviews. Moreover, the opinion reviews obtained from users can be classified into positive or negative reviews, which can be used by a consumer to select a product. This paper aims to classify movie reviews into groups of positive or negative polarity by using machine learning algorithms. In this study, we analyse online movie reviews using SA methods in order to detect fake reviews. SA and text classification methods are applied to a dataset of movie reviews. More specifically, we compare five supervised machine learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN-IBK), KStar (K*) and Decision Tree (DT-J48) for sentiment classification of reviews using three different datasets, including movie review dataset V1.0 and movie reviews dataset V2.0 and movie reviews dataset V3.0. To evaluate the performance of sentiment classification, this work has implemented accuracy, precision, recall and F-measure as a performance measure. The measured results of our experiments show that the SVM algorithm outperforms other algorithms, and that it reaches the highest accuracy not only in text classification, but also in detecting fake reviews.**

Keywords- Sentiment Analysis; Fake Reviews; Naïve Bayes; Support Vector Machine; k-Nearest Neighbor; KStar; Decision Tree -J48.

I. INTRODUCTION

Sentiment analysis (SA) is one of the significant domains of machine learning techniques [1]. Opinion Mining (OM), also known as Sentiment Analysis (SA), is the domain of study that analyzes people's opinions, evaluations, sentiments, attitudes, appraisals, and emotions towards entities such as services, individuals, issues, topics, and their attributes [2]. "The sentiment is usually formulated as a two-class classification problem, positive and negative" [2]. Sometimes, time is more precious than money, therefore, instead of spending time in reading and figuring out the positivity or negativity of a review, we can use automated techniques for Sentiment Analysis.

The basis of SA is determining the polarity of a given text at the document, sentence or aspect level, whether the expressed opinion in a document, a sentence or an entity aspect is positive or negative. More specifically, the goals of SA are to find opinions from reviews and then classify these opinions based upon polarity. According to [3], there are three major classifications in SA, namely: document level, sentence level, and aspect level. Hence, it is important to distinguish between the document level, sentence level, and the aspect level of an analysis process that will determine the different tasks of SA. The document level considers that a document is an opinion on its aspect, and it aims to classify an opinion document as a negative or positive opinion. The sentence level using SA aims to setup opinion stated in every sentence. The aspect level is based on the idea that an opinion consists of a sentiment (positive or negative), and its SA aims to categorize the sentiment based on specific aspects of entities.

The documents used in this work are obtained from a dataset of movie reviews that have been collected by [4] and [10]. Then, an SA technique is applied to classify the documents as real positive and real negative reviews or fake positive and fake negative reviews. Fake negative and fake positive reviews by fraudsters who try to play their competitors existing systems can lead to financial gains for them. This, unfortunately, gives strong incentives to write fake reviews that attempt to intentionally mislead readers by providing unfair reviews to several products for the purpose of damaging their reputation. Detecting such fake reviews is a significant challenge. For example, fake consumer reviews in an e-commerce sector are not only affecting individual consumers but also corrupt purchaser's confidence in online shopping [5]. Our work is mainly directed to SA at the document level, more specifically, on movie reviews dataset. Machine learning techniques and SA methods are expected to have a major positive effect, especially for the detection processes of fake reviews in movie reviews, e-commerce, social commerce environments, and other domains.

In machine learning-based techniques, algorithms such as SVM, NB, and DT-J48 are applied for the classification purposes [6]. SVM is a type of learning algorithm that represents supervised machine learning approaches [7], and it is an excellent successful prediction approach. The SVM is also a robust classification approach [8]. A recent research presented in [3] introduces a survey on different applications and algorithms for SA, but it is only focused on algorithms used in various languages, and the researchers did not focus on detecting fake reviews [9]-[13]. This paper presents five supervised machine learning approaches to classify the sentiment of our dataset, which is compared with two different datasets. We also detect fake positive reviews and

fake negative reviews by using these methods. The main goal of our study is to classify movie reviews as a real reviews or fake reviews using SA algorithms with supervised learning techniques.

The conducted experiments have shown the accuracy, precision, recall, and f-measure of results through sentiment classification algorithms. In three cases (movie reviews dataset V1.0 and movie reviews dataset V2.0 and movie reviews dataset V3.0), we have found that SVM is more accurate than other methods such as NB, KNN-IBK, KStar, and DT-J48.

The main contributions of this study are summarized as follows:

- Using the Weka tool [30], we compare different sentiment classification algorithms, which are used to classify the movie reviews dataset into fake and real reviews.
  - We apply the sentiment classification algorithms using three different datasets with stopwords removal. We realized that using the stopwords removal method is more efficient than without stopwords not only in text categorization, but also to detection of fake reviews.
  - We perform several analysis and tests to find the learning algorithm in terms of accuracy, precision, recall and F-Measure.

The rest of this paper is organized as follows. Section II presents the related works. Section III shows the methodology. Section IV explains the experiment results, and finally, Section V presents the conclusion and future works.

II. RELATED WORKS

Our study employs statistical methods to evaluate the performance of detection mechanism for fake reviews and evaluate the accuracy of this detection. Hence, we present our literature review on studies that applied statistical methods.

A. Sentiment analysis issues

There are several issues to consider when conducting SA [14]. In this section, two major issues are addressed. First, the viewpoint (or opinion) observed as negative in a situation might be considered positive in another situation. Second, people do not always express opinions in the same way. Most common text processing techniques employ the fact that minor changes between the two text fragments are unlikely to change the actual meaning [14].

B. Textual reviews

Most of the available reputation models depend on numeric data available in different fields; an example is ratings in e-commerce. Also, most of the reputation models focus only on the overall ratings of products without considering the reviews which are provided by customers [15]. On the other hand, most websites allow consumers to add textual reviews to provide a detailed opinion about the product [16] [17]. These reviews are available for customers to read. Also, customers are increasingly depending on reviews rather than on ratings. Reputation models can use SA methods to extract users' opinions and use this data in the Reputation system. This information may include consumers' opinions about different features [18] and [19].

C. Detecting Fake Reviews Using Machine Learning

Filter and identification of fake reviews have substantial significance [20]. Moraes et al. [21] proposed a technique for categorizing a single topic textual review. A sentiment classified document level is applied for stating a negative or positive sentiment. Supervised learning methods are composed of two phases, namely selection and extraction of reviews utilizing learning models such as SVM.

Extracting the best and most accurate approach and simultaneously categorizing the customers written reviews text into negative or positive opinions has attracted attention as a major research field. Although it is still in an introductory phase, there has been a lot of work related to several languages [22]-[24]. Our work used several supervised learning algorithms such as SVM, NB, KNN-IBK, K* and DT-J48 for Sentiment Classification of text to detect fake reviews.

D. A Comparative Study of different Classification algorithms

Table I shows comparative studies on classification algorithms to verify the best method for detecting fake reviews using different datasets such as News Group dataset, text documents, and movie reviews dataset. It alsoproves that NB and distributed keyword vectors (DKV) are accurate without detecting fake reviews [12] and [13]. While [11] finds that NB is accurate and a better choice, but it is not oriented for detecting fake reviews. Using the same datasets, [9] finds that SVM is accurate with stopwords method, but it does not focus on detecting fake reviews, while [10] finds that SVM is only accurate without using stopwords method, and also without detecting fake reviews. Sentiment Analysis is a very significant to detect fake reviews [1]. However, they used only supervisor learning techniques based on accuracy and precision. Fundamentally, classification accuracy and precision only are typically not enough information to obtain a good result. However, in our empirical study, results in three cases with movie reviews dataset V1.0 and movie reviews dataset V2.0 and movie reviews dataset V3.0 prove that SVM is robust and accurate for detecting fake reviews by evaluation of measuring the performance with accuracy, precision, F-measure and recall. However, in our empirical study, results in three cases with movie reviews dataset V1.0 and movie reviews dataset V2.0 and movie reviews dataset V3.0 prove that SVM is robust and accurate for detecting fake reviews.

TABLE I. A COMPARATIVE STUDY OF DIFFERENT CLASSIFICATION ALGORITHMS.

Reference	Year	Data Source	Size of dataset	Using Supervised Learning	Language	Classifiers	Detecting Fake Review	Measures	Using stopwords	The best method
[9]	2013	Movie Reviews dataset	2000 Movie Reviews	Yes	English	NB,SVM, kNN	NO	Accuracy, Precision and recall	NO	SVM
[10]	2004	Movie Reviews dataset	2000 Movie Reviews	Yes	English	NB, SVM	NO	Accuracy ,t-test	NO	SVM
[11]	2011	News Group dataset	20 categories with 1000 documents	Yes	English	NB, SVM	NO	Micro-average and macro-average F measure	Yes	NB
[12]	2016	Movie Reviews dataset	4000 movie reviews	Yes	Chinese	NB, SVM, K-NN LLR, Delta TFIDF, LDA-SVM, TFIDF, DKV	NO	precision, recall, F-score as metric, and Accuracy	NO	DKV
[13]	2013	Movie Reviews dataset	1400, 2000 Movie Reviews	Yes	English	NB, SVM	NO	Accuracy, F-measure and Entropy	NO	NB
[1]	2017	Movie Reviews dataset	1400, 2000 Movie Reviews	Yes	English	NB, SVM, IBK, K*,DT-J48	Yes	Precision, and Accuracy	Yes	SVM
This work	**2018**	Movie Reviews dataset	**1400,2000,10662 Movie Reviews**	Yes	English	NB, SVM, IBK, K*,DT-J48	Yes	**Precision, Accuracy, Recall, and F-Measure**	Yes	**SVM**

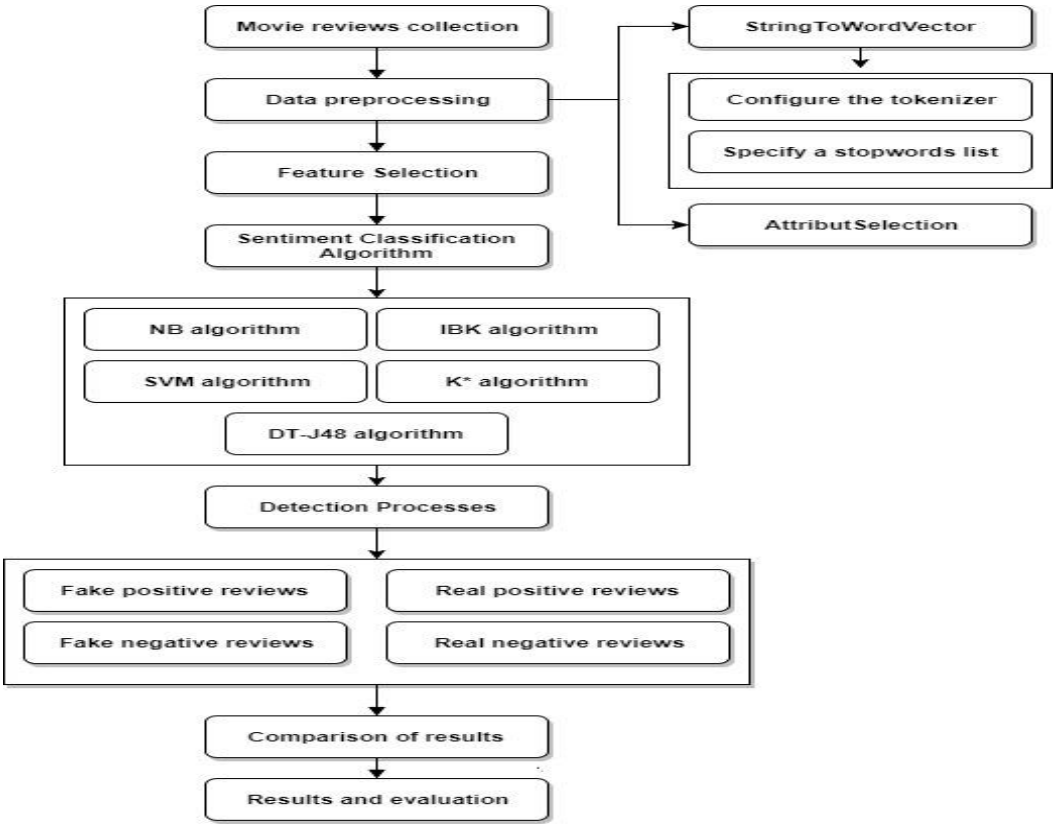

Figure 1. Steps and Techniques used in Sentiment Analysis

III. METHODOLOGY

To accomplish our goal, we analyze a dataset of movie reviews using the Weka tool for text classification. In the proposed methodology, as shown in Figure 1, we follow some steps that are involved in SA using the approaches described below.

Step 1: Movie reviews collection

To provide an exhaustive study of machine learning algorithms, the experiment is based on analyzing the sentiment value of the standard dataset. We have used the original dataset of the movie reviews to test our methods of reviews classification. The dataset is available and has been used in [13], which is frequently conceded as the standard gold dataset for the researchers working in the field of the Sentiment Analysis. The first dataset is known as movie reviews dataset V1.0 which consists of 1400 movie reviews out of which 700 reviews are positive, and 700 reviews are negative. The second dataset is known as movie reviews dataset V2.0, which consists of total 2000 movie reviews, 1000 of which are positive and 1000 of which are negative. The third dataset is known as movie reviews dataset V3.0, which consists of total 10662 movie reviews, 5331 of which are positive and 5331 of which are negative. A summary of the two datasets collected is described in Table II.

TABLE II. DESCRIPTION OF DATASET

Dataset	Content of the Dataset
Movie Reviews Dataset V1.0	1400 Movie Reviews (700+ & 700-)
Movie Reviews Dataset V2.0	2000 Movie Reviews (1000+ & 1000-)
Movie Reviews Dataset V3.0	10662 Movie Reviews (5331+ & 5331-)

Step 2: Data preprocessing

The preprocessing phase includes two preliminary operations, shown in Figure 1, which help in transforming the data before the actual SA task. Data preprocessing plays a significant role in many supervised learning algorithms. We divided data preprocessing as follows:

1) StringToWordVector

To prepare the dataset for learning involves transforming the data by using the StringToWordVector filter, which is the main tool for text analysis in Weka. The StringToWordVector filter makes the attribute value in the transformed datasets Positive or Negative for all single-words, depending on whether the word appears in the document or not. This filtration process is used for configuring the different steps of the term extraction. The filtration process comprises the following two sub-processes:

• Tokenization

This sub-process makes the provided document classifiable by converting the content into a set of features using machine learning.

• Stopwords Removal

The stopwords are the words we want to filter out, eliminate, before training the classifier. Some of those words are commonly used (e.g., "a," "the," "of," "I," "you," "it," "and") but do not give any substantial information to our labeling scheme, but instead they introduce confusion to our classifier. In this study, we used a 630 English stopwords list with movie reviews datasets. Stopwords removal helps to reduce the memory requirements while classifying the reviews.

2) Attribute Selection

Removing the poorly describing attributes can significantly increase the classification accuracy, in order to maintain a better classification accuracy, because not all attributes are relevant to the classification work, and the irrelevant attributes can decrease the performance of the used analysis algorithms, an attribute selection scheme was used for training the classifier.

Step 3: Feature Selection

Feature selection is an approach which is used to identify a subset of features which are mostly related to the target model, and the goal of feature selection is to increase the level of accuracy. In this study, we implemented one feature selection method (BestFirst + CfsSubsetEval, GeneticSearch) widely used for the classification task of SA with Stopwords methods. The results differ from one method to the other. For example, in our analysis of Movie Review datasets, we found that the use of SVM algorithm is proved to be more accurate in the classification task.

Step 4: Sentiment Classification algorithms

In this step, we will use sentiment classification algorithms, and they have been applied in many domains such as commerce, medicine, media, biology, etc. There are many different techniques in classification method like NB, DT-J48, SVM, K-NN, Neural Networks, and Genetic Algorithm. In this study, we will use five popular supervised classifiers: NB, DT-J48, SVM, K-NN, KStar algorithms.

1) Naïve Bayes(NB)

The NB classifier is a basic probabilistic classifier based on applying Bayes' theorem. The NB calculates a set of probabilities by combinations of values in a given dataset. Also, the NB classifier has fast decision-making process.

2) Support Vector Machine (SVM)

SVM in machine learning is a supervised learning model with the related learning algorithm, which examines data and identifies patterns, which is used for regression and classification analysis [25]. Recently, many classification algorithms have been proposed, but SVM is still one of the most widely and most popular used classifiers.

3) K-Nearest Neighbor (K-NN)

K-NN is a type of lazy learning algorithm and is a non-parametric approach for categorizing objects based on closest training. The K-NN algorithm is a very simple algorithm for all machine learning. The performance of the K-NN algorithm depends on several different key factors, such as a suitable distance measure, a similarity measure for voting, and, k parameter [26]- [29].

A set of vectors and class labels which are related to each vector constitute each of the training data. In the simplest way; it will be either positive or negative class. In this study, we are using a single number ''k'' with values of k=3. This number decides how many neighbors influence the classification.

4) KStar (K)*

K-star (K*) is an instance-based classifier. The class of a test instance is established in the class of those training instances similar to it, as decided by some similarity function. K* algorithm is usually slower to evaluate the result.

5) Decision Tree (DT-J48)

The DT-J48 approach is useful in the classification problem. In the testing option, we are using percentage split as the preferred method.

Step 5: Detection Processes

After training, the next step is to predict the output of the model on the testing dataset, and then a confusion matrix is generated, which classifies the reviews as positive or negative. The results involve the following attributes:

- True Positive: Real Positive Reviews in the testing data, which are correctly classified by the model as Positive (P).
- False Positive: Fake Positive Reviews in the testing data, which are incorrectly classified by the model as Positive (P).
- True Negative: Real Negative Reviews in the testing data, which are correctly classified by the model as Negative (N).
- False Negative: Fake Negative Reviews in the testing data, which are incorrectly classified by the model as Negative (N).

True negative (TN) are events which are real and are effectively labeled as real, True Positive (TP) are events which are fake and are effectively labeled as fake. Respectively, False Positives (FP) refer to Real events being classified as fakes; False Negatives (FN) are fake events incorrectly classified as Real events. The confusion matrix, (1)-(6) shows numerical parameters that could be applied following measures to evaluate the Detection Process (DP) performance. In Table III, the confusion matrix shows the counts of real and fake predictions obtained with known data, and for each algorithm used in this study there is a different performance evaluation and confusion matrix.

TABLE III. THE CONFUSION MATRIX

	Real	**Fake**
Real	True Negative Reviews (TN)	False Positive Reviews (FP)
Fake	False Negative Reviews (FN)	True Positive Reviews (TP)

Fake Positive Reviews Rate = FP/FP+TN (1)
Fake negative Reviews Rate = FN/TP+FN (2)
Real Positive Reviews Rate = TP/TP+FN (3)
Real negative Reviews Rate = TN/TN+FP (4)
Accuracy = TP+TN/TP+TN+FN+FP (5)
Precision = TP/TP+FP (6)
Recall = TP/(TP+FN) (7)
F-measure = 2×(Precision*Recall)/ (Recall + Precision) (8)

The confusion matrix is a very important part of our study because we can classify the reviews from datasets whether they are fake or real reviews. The confusion matrix is applied to each of the five algorithms discussed in Step 4.

Step 6: Comparison of results

In this step, we compared the different accuracy provided by the dataset of movie reviews with various classification algorithms and identified the most significant classification algorithm for detecting Fake positive and negative Reviews.

IV. EXPERIMENTS AND RESULT ANALYSIS

In this section, we present experimental results from five different supervised machine learning approaches to classifying sentiment of three datasets which is compared with movie reviews dataset V1.0 and movie reviews dataset V2.0 and movie reviews dataset V3.0. Also, we have used the same methods at the same time to detect fake reviews.

A. Experimental results on dataset v1.0

1. Confusion matrix for all methods

The previous section compared different algorithms with different datasets. In this section, the algorithms are applied to perform a sentiment analysis on another dataset. From the results presented in Table IV, the confusion matrix displays results for movie reviews dataset v1.0.

TABLE IV. CONFUSION MATRIX FOR ALL METHODS

Classification algorithms	SA	Real	Fake
NB	Real	455	245
	Fake	162	538
KNN-IBK (K=3)	Real	480	220
	Fake	193	507
K*	Real	491	209
	Fake	219	481
SVM	Real	**516**	**184**
	Fake	**152**	**548**
DT-J48	Real	498	202
	Fake	219	481

2. Evaluation parameters and accuracy for all methods

Five main performance evaluation measures have been introduced for Classification algorithms. These include Fake Positive Reviews predictive value, Fake Negative Reviews predictive value, Real Positive Reviews predictive value, Real Negative Reviews predictive value, accuracy and Precision. Table V displays the results of evaluation parameters for all methods and provides a summary of recordings obtained from the experiment. As a result, SVM surpasses for best accuracy among the other classification algorithms with 76%.

TABLE V. EVALUATION PARAMETERS AND ACCURACY FOR ALL METHODS

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Accuracy %
NB	35	23.1	76.9	65	70.9
K-NN-IBK (K=3)	31.4	27.6	72.4	68.6	70.5
K*	29.9	31.3	68.7	70.1	69.4
SVM	26.3	21.7	78.3	73.7	**76**
DT-J48	28.9	31.3	68.7	71.1	69.9

The graph in Figure 2 displays a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

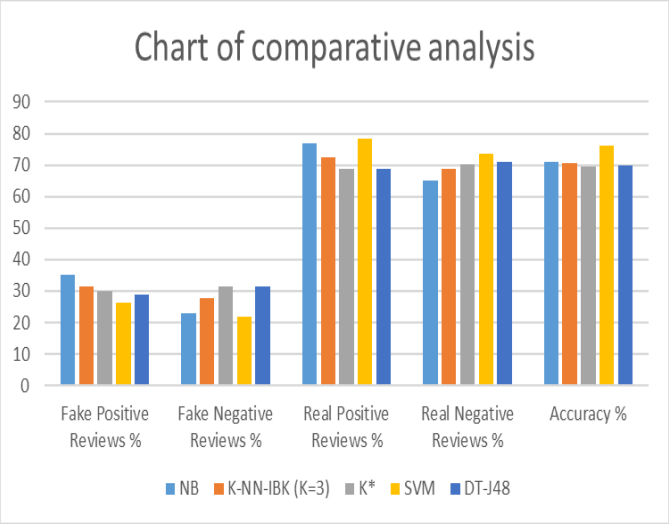

Figure 2. Comparative analysis of all methods

The comparison in Table VI indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

TABLE VI. COMPARISON OF ACCURACY OF CLASSIFIERS

Classification algorithms	Accuracy %
NB	70.9
KNN-IBK (K=3)	70.5
K*	69.4
SVM	**76**
DT-J48	69.9

The graph in Figure 3 displays accuracy rate of NB, SVM, (K-NN, k=3), DT-J48 algorithms. We obtained a higher accuracy of SVM algorithm than other algorithms.

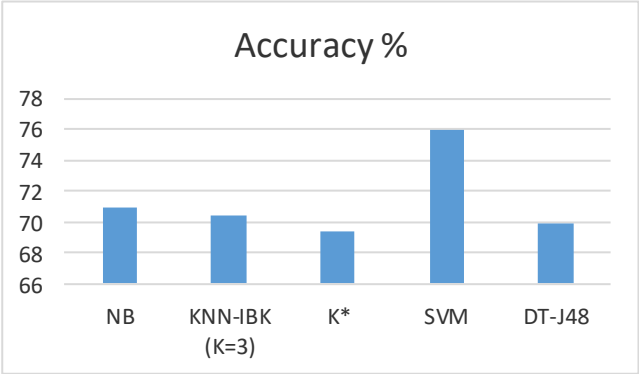

Figure 3. Accuracy of different algorithms

TABLE VII. TIME TAKEN TO BUILD MODEL

Classification algorithms	Time taken to build model (milliseconds)
NB	90
KNN-IBK (K=3)	0
K*	10
SVM	**4240**
DT-J48	330

Table VII displays the time taken by each algorithm to build prediction model. As it is evident from the table, K-NN takes the shortest amount of time of 0 milliseconds to create a model and SVM takes the longest amount of time of 4240 milliseconds to build a model.

TABLE VIII. COMPARISON RESULTS OF PRECISION, RECALL, AND F-MEASURE

classifier	class	Accuracy metrics %		
		Precision	Recall	F-Measure
NB	pos	68.7	76.9	72.6
	neg	73.7	65.0	69.1
KNN-IBK (K=3)	pos	69.7	72.4	71.1
	neg	71.3	68.6	69.9
K*	pos	69.7	68.7	69.2
	neg	69.2	70.1	69.6
SVM	pos	**74.9**	**78.3**	**76.5**
	neg	**77.2**	**73.7**	**75.4**
DT-J48	pos	70.4	68.7	69.6
	neg	69.5	71.1	70.3

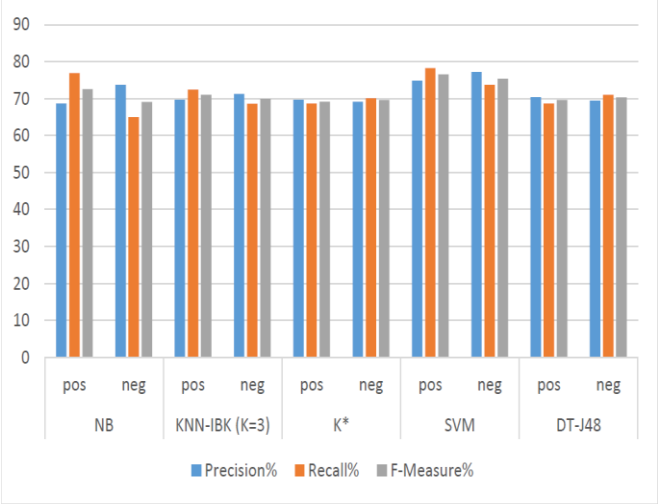

Figure 4. Comparison of metrics obtained from various multi-label classifiers

Table VIII and Figure 4 present the performance evaluation of precision, recall, and f-measure metrics, and all of these metrics are calculated for each class of positive and negative.

B. *Experimental result on dataset V2.0*

1) *Confusion matrix for all methods*

The number of real and fake predictions made by the classification model compared with the actual results in the test data is shown in the confusion matrix. The confusion matrix is obtained after implementing NB, SVM, K-NN, K*, DT-J48 algorithms. Table IX displays the results for confusion matrix for V2.0 dataset. The columns represent the number of predicted classifications made by the model. The rows display the number of real classifications in the test data.

TABLE IX. CONFUSION MATRIX FOR ALL METHODS

Classification algorithms	SA	Real	Fake
NB	Real	781	219
	Fake	187	813
KNN-IBK (K=3)	Real	804	196
	Fake	387	613
K*	Real	760	240
	Fake	337	663
SVM	Real	**809**	**191**
	Fake	**182**	**818**
DT-J48	Real	762	238
	Fake	330	670

2) *Evaluation parameters and accuracy for all methods*

Five main performance evaluation measures have been introduced for Classification algorithms. These include Fake Positive Reviews predictive value, Fake Negative Reviews predictive value, Real Positive Reviews predictive value, Real Negative Reviews predictive value, accuracy and Precision. Table X shows the results of evaluation parameters for all methods and provides a summary of recordings obtained from the experiment. SVM surpasses as the best accuracy among the other classification algorithms with 81.35%. The tabulated observations list the readings as well as accuracies obtained for a specific supervised learning algorithm on a dataset of a movie review.

TABLE X. EVALUATION PARAMETERS AND ACCURACY FOR ALL METHODS.

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Accuracy %
NB	21.9	18.7	81.3	78.1	79.7
K-NN-IBK (K=3)	19.6	38.7	61.3	80.4	70.85
K*	24	33.7	66.3	76	71.15
SVM	19.1	18.2	81.8	80.9	**81.35**
DT-J48	23.8	33	67	76.2	71.6

The graph in Figure 5 shows a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

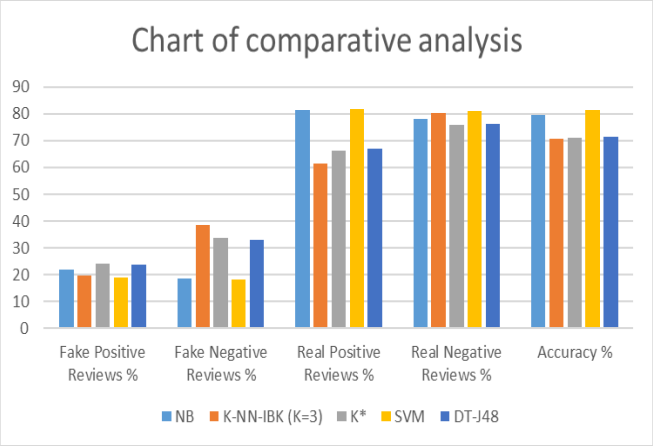

Figure 5. Comparative analysis of all methods

The comparison in Table XI indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, K*, and DT-J48 algorithms.

TABLE XI. COMPARISON OF ACCURACY OF CLASSIFIERS

Classification algorithms	Accuracy %
NB	79.7
KNN-IBK (K=3)	70.85
K*	71.15
SVM	**81.35**
DT-J48	71.6

The graph in Figure 6 shows accuracy rate of NB, SVM, (K-NN, k=3), and DT-J48 algorithms. We obtained a higher accuracy in SVM algorithm than in the other algorithms.

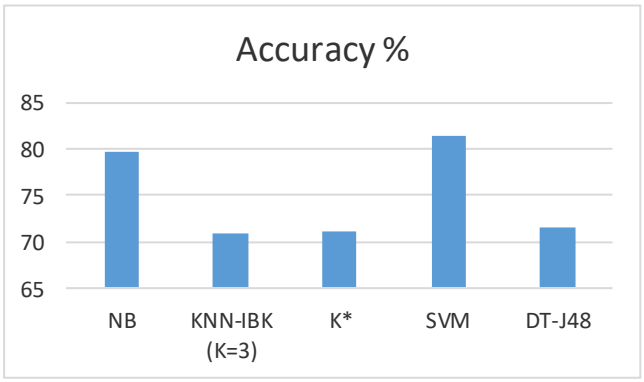

Figure 6. Graph showing the accuracy of different algorithms

Table XII shows the time taken by each algorithm to build prediction model. As it is evident from the table, K-star takes the shortest amount of time of 0 milliseconds to create a model and SVM takes the longest amount of time of **14840** milliseconds to build a model.

TABLE XII. TIME TAKEN TO BUILD MODEL

Classification algorithms	Time taken to build model (milliseconds)
NB	110
KNN-IBK (K=3)	10
K*	0
SVM	**14840**
DT-J48	340

TABLE XIII. COMPARISON RESULTS OF PRECISION, RECALL, AND F-MEASURE

classifier	class	Accuracy metrics %		
		Precision	Recall	F-Measure
NB	pos	78.8	81.3	80.0
	neg	80.7	78.1	79.4
KNN-IBK (K=3)	pos	75.8	61.3	67.8
	neg	67.5	80.4	73.4
K*	pos	73.4	66.3	69.7
	neg	69.3	76.0	72.5
SVM	pos	**81.1**	**81.8**	**81.4**
	neg	**81.6**	**80.9**	**81.3**
DT-J48	pos	73.8	67.0	70.2
	neg	69.8	76.2	72.8

Figure 7. Comparison of metrics obtained from various multi-label classifiers

Table XIII and Figure 7 present the performance evaluation of precision, recall, and f-measure metrics, and all of these metrics are calculated for each class of positive and negative.

C. Experimental results on dataset v3.0

1. Confusion matrix for all methods

The previous section compared different algorithms with different datasets. In this section, the algorithms are applied to perform a sentiment analysis on another dataset. From the results presented in Table XIV, the confusion matrix displays results for movie reviews dataset v3.0.

TABLE XIV. CONFUSION MATRIX FOR ALL METHODS

Classification algorithms	SA	Real	Fake
NB	Real	2303	3028
	Fake	1107	4224
KNN-IBK (K=3)	Real	1813	3518
	Fake	789	4542
K*	Real	2373	2958
	Fake	910	4421
SVM	Real	**2758**	**2573**
	Fake	**994**	**4337**
DT-J48	Real	2914	2417
	Fake	1571	3760

2. Evaluation parameters and accuracy for all methods

Five main performance evaluation measures have been introduced for Classification algorithms. These include Fake Positive Reviews predictive value, Fake Negative Reviews predictive value, Real Positive Reviews predictive value, Real Negative Reviews predictive value, accuracy and Precision. Table XV displays the results of evaluation parameters for all methods and provides a summary of recordings obtained from the experiment. As a result, SVM surpasses for best accuracy among the other classification algorithms with 66.5%.

TABLE XV. EVALUATION PARAMETERS AND ACCURACY FOR ALL METHODS

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Accuracy %
NB	56.8	20.8	79.2	43.2	61.2
K-NN-IBK (K=3)	66	14.8	85.2	34	59.6
K*	55.5	17.1	82.9	44.5	63.7
SVM	48.3	18.6	81.4	51.7	**66.5**
DT-J48	45.3	29.5	70.5	54.7	62.5

The graph in Figure 8 displays a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

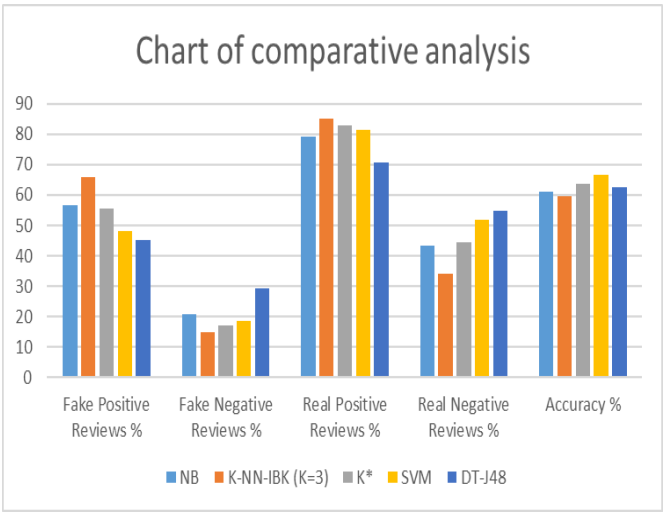

Figure 8. Comparative analysis of all methods

The comparison in Table XVI indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

TABLE XVI. COMPARISON OF ACCURACY OF CLASSIFIERS

Classification algorithms	Accuracy %
NB	61.2
KNN-IBK (K=3)	59.6
K*	63.7
SVM	**66.5**
DT-J48	62.5

The graph in Figure 9 displays accuracy rate of NB, SVM, (K-NN, k=3), DT-J48 algorithms. We obtained a higher accuracy of SVM algorithm than other algorithms.

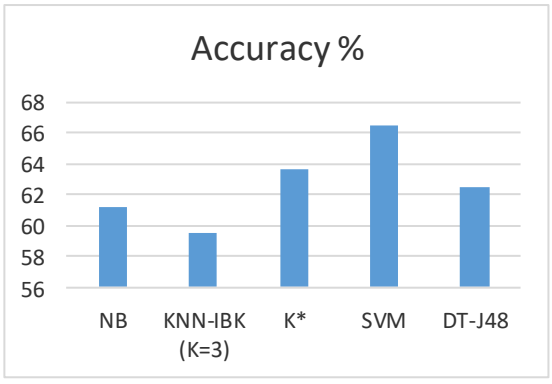

Figure 9. Accuracy of different algorithms

TABLE XVII. TIME TAKEN TO BUILD MODEL

Classification algorithms	Time taken to build model (milliseconds)
NB	680
KNN-IBK (K=3)	20
K*	10
SVM	**2,515,260**
DT-J48	11,480

Table XVII displays the time taken by each algorithm to build prediction model. As it is evident from the table, K* takes the shortest amount of time of 10 milliseconds to create a model and SVM takes the longest amount of time of 2,515,260 milliseconds to build a model.

TABLE XVIII. COMPARISON RESULTS OF PRECISION, RECALL, AND F-MEASURE

classifier	class	Accuracy metrics %		
		Precision	Recall	F-Measure
NB	pos	58.2	79.2	67.1
	neg	67.5	43.2	52.7
KNN-IBK (K=3)	pos	56.4	85.2	67.8
	neg	69.7	34	45.7
K*	pos	59.9	82.9	69.6
	neg	72.3	44.5	55.1
SVM	pos	**62.8**	**81.4**	**70.9**
	neg	**73.5**	**51.7**	**60.7**
DT-J48	pos	60.9	70.5	65.3
	neg	65	54.7	59.4

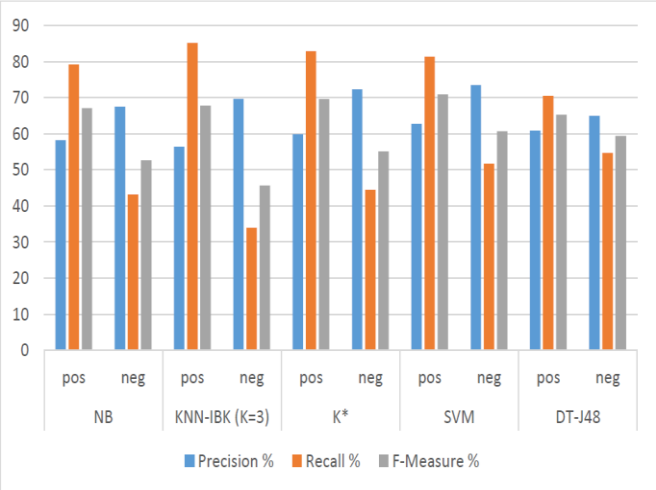

Figure 10. Comparison of metrics obtained from various multi-label classifiers

Table XVIII and Figure 10 present the performance evaluation of precision, recall, and f-measure metrics, and all of these metrics are calculated for each class of positive and negative.

D. *Discussion*

Table XIX and Figure 11 present the summary of the experiments. Five supervised machine learning algorithms: NB, SVM, K-NN, K*, DT-J48 have been applied to the online movie reviews. We observed that well-trained machine learning algorithms could perform very useful classifications on the sentiment polarities of reviews. In terms of accuracy, SVM is the best algorithm for all tests since it correctly classified 81.35% of the reviews in dataset V1.0 and 76% of the reviews in dataset V2.0 and 66.5% of the reviews in dataset V3.0. SVM tends to be more accurate than other methods.

TABLE XIX. THE BEST RESULT OF OUR EXPERIMENTS

Experiments	Fake Positive Reviews of SVM %	Fake Negative Reviews of SVM %	Accuracy of SVM %
Results on dataset V1.0	**19.1**	**18.2**	**81.35**
Results on dataset V2.0	**26.3**	**21.7**	**76**
Results on dataset V3.0	**48.3**	**18.6**	**66.5**

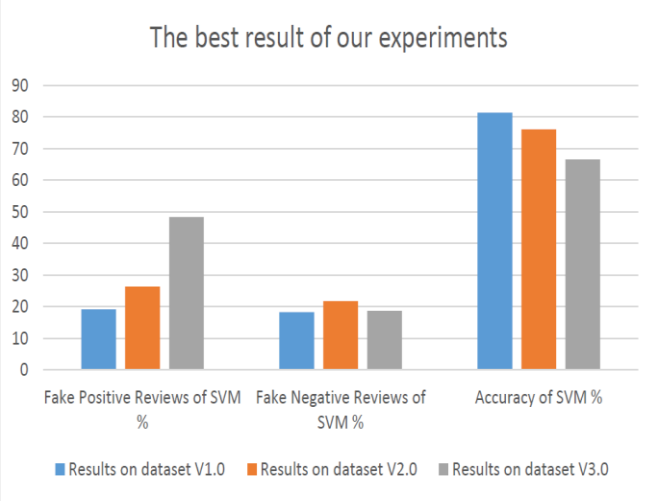

Figure 11. Summary of our experiments

The presented study emphasizes that the accuracy of SVM is higher for Movie Review dataset V2.0. However, the detection process of Fake Positive Reviews and Fake Negative Reviews offers less promising results for Movie Review dataset V2.0 in comparison to Movie Review dataset V1.0 as evident from Table XII.

V. CONCLUSION AND FUTURE WORK

In this research, we proposed several methods to analyze a dataset of movie reviews. We also presented sentiment classification algorithms to apply a supervised learning of the movie reviews located in two different datasets. Our experimental approaches studied the accuracy, precision, recall and F-Measure of all sentiment classification algorithms, and how to determine which algorithm is more accurate. Furthermore, we were able to detect fake positive reviews and fake negative reviews through detection processes.

Five supervised learning algorithms to classifying sentiment of our datasets have been compared in this paper: NB, K-NN, K*, SVM, and DT-J48. Using the accuracy analysis for these five techniques, we found that SVM algorithm is the most accurate for correctly classifying the reviews in movie reviews datasets, i.e., V1.0, V2.0 and V3.0. Also, detection processes for fake positive reviews and fake negative reviews depend on the best method that is used in this study.

For future work, we would like to extend this study to use other datasets such as Amazon dataset or eBay dataset and use different feature selection methods. Furthermore, we may apply sentiment classification algorithms with stopwords removal and stemming methods to detect fake reviews using various tools such as Python or R studio; then we will evaluate the performance of our work with some of these tools.

ACKNOWLEDGMENT

Mr. Elshrif Elmurngi would like to thank the Ministry of Education in Libya and Canadian Bureau for International Education (CBIE) for their support to his Ph.D. research work.

REFERENCES

[1] E. Elmurngi and A. Gherbi, “Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques,” IARIA/ DATA ANALYTICS 2017 – the Sixth International Conference on Data Analytics, ISBN: 978-1-61208-603-3, November, pp. 65–72, Barcelona, Spain, 2017.

[2] B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, 2012, pp. 1–167.

[3] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. 5, no. 4, 2014, pp. 1093–1113.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques,” in Proceedings of EMNLP, 2002, pp. 79–86. [Online]. Available: http://www.cs.cornell.edu/People/pabo/movie%2Dreview%2Ddata/

[5] J. Malbon, “Taking fake online consumer reviews seriously,” Journal of Consumer Policy, vol. 36, no. 2, 2013, pp. 139–157.

[6] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences, vol. 181, no. 6, 2011, pp. 1138–1152.

[7] T. Barbu, “Svm-based human cell detection technique using histograms of oriented gradients,” cell, vol. 4, 2012, p. 11.

[8] G. Esposito, LP-type methods for Optimal Transductive Support Vector Machines. Gennaro Esposito, PhD, 2014, vol. 3.

[9] P. Kalaivani and K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches," Indian Journal of Computer Science and Engineering, vol. 4, no. 4, pp. 285-292, 2013.

[10] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271. [Online]. Available from: http://www.cs.cornell.edu/People/pabo/movie%2Dreview%2Ddata/

[11] S. Hassan, M. Rafi, and M. S. Shaikh, “Comparing svm and naive bayesclassifiers for text categorization with wikitology as knowledge enrich-ment,” in Multitopic Conference (INMIC), 2011 IEEE 14th International.IEEE, 2011, pp. 31–34.

[12] C.-H. Chu, C.-A. Wang, Y.-C. Chang, Y.-W. Wu, Y.-L. Hsieh, and W.-L. Hsu, “Sentiment analysis on chinese movie review with distributed keyword vector representation,” in Technologies and Applications of Artificial Intelligence (TAAI), 2016 Conference on. IEEE, 2016, pp. 84–89.

[13] V. Singh, R. Piryani, A. Uddin, and P. Waila, “Sentiment analysis of movie reviews and blog posts,” in Advance Computing Conference (IACC), 2013 IEEE 3rd International. IEEE, 2013, pp. 893–898.

[14] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: a survey,” International Journal, vol. 2, no. 6, 2012, pp. 282–292.

[15] G. Xu, Y. Cao, Y. Zhang, G. Zhang, X. Li, and Z. Feng, “Trm: Computing reputation score by mining reviews.” in AAAI Workshop: Incentives and Trust in Electronic Communities, 2016.

[16] N. Tian, Y. Xu, Y. Li, A. Abdel-Hafez, and A. Josang, “Generating prod-uct feature hierarchy from product reviews,” in International Conference on Web Information Systems and Technologies. Springer, 2014, pp.264–278.

[17] N. Tian, Y. Xu, Y. Li, A. Abdel-Hafez, and A. Jøsang, “Product feature taxonomy learning based on user reviews.” in WEBIST (2), 2014, pp. 184–192.

[18] A. Abdel-Hafez and Y. Xu, “A survey of user modelling in social media websites,” Computer and Information Science, vol. 6, no. 4, 2013, p. 59.

[19] A. Abdel-Hafez, Y. Xu, and D. Tjondronegoro, “Product reputation model: an opinion mining based approach,” in SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data, 2012, p. 16.

[20] N. Jindal and B. Liu, "Opinion spam and analysis," in Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008, pp. 219–230.

[21] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," Expert Systems with Applications, vol. 40, no. 2, 2013, pp. 621–633.

[22] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and com-paring opinions on the web," in Proceedings of the 14th international conference on World Wide Web. ACM, 2005, pp. 342–351.

[23] A. Fujii and T. Ishikawa, "A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making," in Proceedings of the Workshop on Sentiment and Subjectivity in Text. Association for Computational Linguistics, 2006, pp. 15–22.

[24] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in Proceedings of AAAI, 2006, pp. 100–107.

[25] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, 1995.

[26] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "Iknn: Informativek-nearest neighbor pattern classification," in PKDD. Springer, 2007, pp.248–264.

[27] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "An affinity-based new local distance function and similarity measure for knn algorithm," Pattern Recognition Letters, vol. 33, no. 3, 2012, pp. 356–363.

[28] M. Latourrette, "Toward an explanatory similarity measure for nearest-neighbor classification," Machine Learning: ECML 2000, pp. 238–245.

[29] S. Zhang, "Knn-cf approach: Incorporating certainty factor to knn classification." IEEE Intelligent Informatics Bulletin, vol. 11, no. 1, 2010, pp. 24–33.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H.Witten, "The weka data mining software: an update," ACM SIGKDD Explorations newsletter, vol. 11, no. 1, 2009, pp. 10–18.