TABLE DES MATIÈRES

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

EC	Commerce électronique (E-commerce)
RSs	Systèmes de réputation (Reputation Systems)
TRS	Système de confiance et de réputation (Trust and Reputation System)
SA	Analyse de sentiments (Sentiment Analysis)
DP	Processus de détection (Detection Process)
KNN	K plus proche voisin (K-Nearest Neighbors)
SVM	Machine à vecteurs de support (Support Vector Machine)
NB	Naïve bayésien (Naïve Bayes)
DT	Arbre de décision (Decision Tree)
LR	Régression logistique (Logistic Regression)
SAM	Modèle d’analyse de sentiments (Sentiment Analysis Model)
OM	Extraction d’opinion (Opinion Mining)
STWV	String To Word Vector
FS	Sélection des caractéristiques (Feature Selection)
TPR	Vrais avis positifs (True Positive Reviews)
FPR	Faux avis positifs (False Positive Reviews)
TNR	Vrais avis négatifs (True Negative Reviews)
FNR	Faux avis négatifs (False Negative Reviews)
TNR	Vrais avis neutres (True Neutral Reviews)
FNR	Faux avis neutres (False Neutral Reviews)
NLP	Traitement du langage naturel (Natural Language Processing)
PFP	Pourcentage de feedback positif (Positive Feedback Percentage)
CMP	Pourcentage de complicité et de manipulation (Collusion and manipulation Percentage)
SAS	Système d’analyse statistique (Statistical Analysis System)

INTRODUCTION

0.1 Contexte et motivation

La réputation et la confiance sont des valeurs primordiales, et elles jouent un rôle essentiel en permettant à de multiples parties d'établir des relations mutuellement bénéfiques. Par définition, la réputation est l'opinion du public envers une personne, un groupe d'individus ou une organisation Hoffman *et al.* (2007).

L'objectif d'un système de confiance et de réputation (Trust and Reputation System(TRS)) est de s'assurer que les valeurs de confiance et de réputation reflètent de manière appropriée les actions prises par les entités du système et ne peuvent pas être manipulées ou accessibles par des entités non autorisées Fraga *et al.* (2012).

Les systèmes de réputation sont présents partout, mais ce sont souvent des éléments sous-alimentés et mal conçus des plateformes du web social. Cependant, ils jouent un rôle crucial dans l'établissement de la confiance, la promotion de la qualité, l'amélioration de la collaboration et l'instauration d'un sentiment d'allégeance. Les systèmes de réputation sont la pièce du puzzle qui peut faire la différence entre l'échec et le succès. Lorsque les systèmes de réputation décident de valoriser certains aspects du comportement, les individus sont amenés à essayer de jouer le système à leur avantage. Il est donc important de choisir des mesures fiables et difficiles à manipuler.

Dans cette étude, nous identifions les défis qui affaiblissent les systèmes de confiance et de réputation. Nous considérons chaque étape du fonctionnement de ces systèmes, à savoir la génération, la distribution et l'agrégation des retours d'information ("feedback"). Chacun de ces éléments doit être protégé contre diverses menaces adverses. À titre d'exemple, la fiabilité concernant l'exactitude ("accuracy") de la réputation est une exigence importante pour le composant d'agrégation Tavakolifard (2012).

Les systèmes de réputation constituent une méthode précieuse pour mesurer la crédibilité des vendeurs ou la qualité des produits dans l'environnement du commerce électronique (e-commerce (EC)). Récemment, les plates-formes du commerce électronique (CE), par exemple les places de marché électroniques, offrent un environnement dynamique qui rassemble des millions d'acteurs pour faire du commerce de biens et de services. Les acheteurs et les vendeurs bénéficient ainsi d'opportunités jamais offertes auparavant, impliquant une variété presque infinie de produits. Quel que soit le type de produit recherché (livres anciens, nouvelles technologies ou instruments hautement spécialisés), l'acheteur trouvera la plupart du temps sur le Web un partenaire de transaction approprié. Toutefois, cet "univers d'étrangers" pose également de nombreux défis Dellarocas (2006). Contrairement aux transactions traditionnelles en face à face, les acheteurs ne sont pas en mesure de connaître la qualité réelle des produits ni la crédibilité d'un vendeur. Comme le paiement anticipé est une pratique courante dans de nombreux cas, les acheteurs sont souvent confrontés à des risques élevés. Pour surmonter ce problème, de nombreux systèmes de commerce électronique encouragent les clients à fournir un retour d'information sur une transaction en démontrant leur bonne volonté. Les systèmes de réputation recueillent toutes les évidences, regroupent les données d'entrée et fournissent une ou plusieurs valeurs de réputation en tant que sortie ("output"). De cette manière, les systèmes de réputation peuvent aider les acheteurs à décider à qui faire confiance et quels produits ou services ils choisissent.

Selon une étude récente de Sänger & Pernul (2014), les vendeurs ayant la meilleure réputation obtiennent des prix plus élevés et ont un nombre de ventes accru. Toutefois, la promotion d'une participation digne de confiance incite également les acteurs malveillants à pousser injustement leur réputation pour en tirer plus de profit.

0.2 Problématique

Les plates-formes du CE offrent un environnement dynamique qui rassemble des millions d'acteurs pour le commerce des services et des produits. Les acheteurs et les vendeurs bénéficient d’opportunités jamais vues auparavant, notamment une variété presque infinie de produits. Quel que soit l’objet de la transaction (livres anciens, nouvelles technologies ou instruments hautement spécialisés), l’acheteur trouvera la plupart du temps sur le Web un partenaire commercial approprié. Toutefois, cet "univers d’étrangers" pose également de nombreux défis aux systèmes de réputation en incitant à la bonne volonté et à la qualité des services, et en sanctionnant les mauvais comportements et les services de mauvaise qualité. En conséquence, cette étude se concentrera sur trois problèmes principaux, comme indiqué ci-dessous.

L’une des difficultés principales de la SA est de savoir comment extraire les sentiments de l’opinion, et comment détecter les fausses critiques positives et les fausses critiques négatives dans les avis (opinion reviews). De plus, les avis obtenus des utilisateurs peuvent être classés en avis positifs ou négatifs, qui peuvent être utilisés par un consommateur dans la sélection d’un produit. Selon une étude récente réalisée par Diekmann *et al.* (2014), les vendeurs ayant la meilleure réputation ont un nombre de ventes accru. Toutefois, la promotion d’une participation digne de confiance incite également les acteurs malveillants de pousser injustement leur réputation pour obtenir plus d’avantages. Les évaluations ou évaluations (ratings) malhonnêtes sont déjà devenues un sérieux problème dans la pratique. Un autre problème majeur est le manque de crédibilité des évaluations en retour (feedback reviews), par lequel les utilisateurs pourraient créer des évaluations fantômes en retour pour soutenir leur réputation. Ainsi, nous aurons le sentiment que ces avis et évaluations sont injustes. Les principaux défis auxquels l’AS est confrontée aujourd’hui sont de savoir comment détecter les avis négatifs injustes, les avis neutres injustes et les avis positifs injustes provenant des avis d’opinion.

Les évaluations injustes sont données individuellement ou collectivement Swamynathan *et al.* (2010) où les évaluations collectives injustes sont qualifiées de complicité Sun & Liu (2012) ; Swamynathan *et al.* (2010) et sont beaucoup plus compliquées et beaucoup plus difficiles à détecter que les évaluations uniques injustes Sun & Liu (2012). Le problème de la "toute excellente réputation" (all excellent reputation) est courant dans le domaine du commerce électronique. Un autre problème est que les vendeurs peuvent rédiger des avis injustes pour approuver ou rejeter tout produit ciblé, car une meilleure réputation entraîne des profits plus élevés. Les évaluations malhonnêtes sont déjà devenues un problème dangereux dans la pratique.

Pour cette raison, dans le présent travail, nous nous concentrons sur la compréhension et l'identification des scores de évaluations injustes, des problèmes de toute bonne réputation et de la détection des complicités et des manipulations.

0.3 Objectifs de la recherche

Dans ce projet de recherche, notre objectif principal est de développer un nouveau mécanisme et de construire un nouveau modèle pour surmonter les défis du système de réputation, y compris les faux avis en retour ("feedback reviews"), les avis en retour injustes, la complicité et la manipulation, et d'évaluer l'efficacité du mécanisme proposé pour traiter la complicité et la manipulation effectuées par les deux parties : le client et le vendeur. Cela permettra d'établir la confiance entre le client et le vendeur.

Pour résumer, nous cherchons à atteindre les sous-objectifs suivants :

1. Investiguer et définir la technique de détection des faux avis en utilisant les techniques d'apprentissage supervisé ;
2. Détecter les avis injustes des consommateurs sur un produit, et améliorer l'exactitude en utilisant des algorithmes d'analyse de sentiments et des techniques d'apprentissage supervisé ;
3. Offrir une solution nouvelle et complète pour concevoir un nouveau modèle afin d'obtenir le système de réputation le plus précis possible, qui réponde aux problèmes existants, tels que la collusion et la manipulation et le problème de "bonne réputation" qui sont

actuellement rencontrés par les systèmes de réputation. Alors que les modèles actuels de réputation appliqués reposent principalement sur les évaluations globales des articles, ils n'impliquent pas les avis des clients dans leur évaluation. Inversement, peu de modèles de réputation se concentrent uniquement sur les avis globaux des produits sans tenir compte des évaluations fournies par les clients. Cette recherche vise à calculer une évalutation des retours d'information sur la base des avis en retour des clients. Par la suite, afin d'obtenir des scores de réputation précis, nous proposons une méthode de calcul simple qui calcule les évaluations et les avis en retour pour obtenir des évaluations et des avis en retour réels, après avoir détecté les évaluations et les avis en retour injustes, par opposition aux sites web Amazon ou eBay qui calculent les scores de réputation à partir de fausses évaluations et de faux avis en retour.

0.4 Méthodologie de recherche

Notre méthodologie a été organisée comme indiqué dans la figure 0.1 autour des étapes suivantes :

La première étape : Collecte de données

À cette étape, nous avons basé notre expérience sur l'analyse de la valeur du sentiment de l'ensemble de données standard en utilisant des algorithmes d'apprentissage automatique. Nous avons utilisé les ensembles de données originaux des critiques de films et de produits pour tester nos méthodes de classification des critiques, et ces ensembles de données ont été utilisés à l'origine dans Pang & Lee (2004), Xu *et al.* (2016), Zhang *et al.* (2010). Plus de détails sont décrits dans le chapitre 2. Section 2.3 et chapitre 3. Section 3.3 et chapitre 4. Section 4.3.1.

La deuxième étape : Analyse

Pour interpréter les données, l'outil d'analyse statistique peut être implémenté de différentes manières, afin de trouver des relations, des différences ou des descriptions de données. En outre, les méthodes d'analyse dépendront des objectifs de l'étude. Plusieurs analyses peuvent être effectuées au cours de la phase initiale d'analyse des données, comme indiqué ci-dessous :

1. Nettoyage ("Cleaning") et prétraitement des données

Il s'agit du premier processus d'analyse des données où la mise en correspondance des enregistrements (" records matching "), la déduplication, la suppression des ponctuations, la suppression des mots vides (" stop words ") et la segmentation des colonnes sont effectuées pour nettoyer les données brutes. La phase de prétraitement comprend des opérations préliminaires qui aident à transformer les données avant la tâche de SA effective. Afin de démontrer, à travers notre schéma proposé, l'effet du prétraitement sur les modèles de classification. le prétraitement des données jouant un rôle très important dans le processus d'exploration des données et les techniques d'apprentissage machine, nous avons divisé le prétraitement des données comme suit :

A. StringToWordVecto

Préparer nos données pour l'apprentissage implique de les transformer en utilisant le filtre StringToWordVector, qui est le principal outil d'analyse de texte dans WEKA. Ce filtre permet de configurer les différentes étapes de l'extraction des termes.

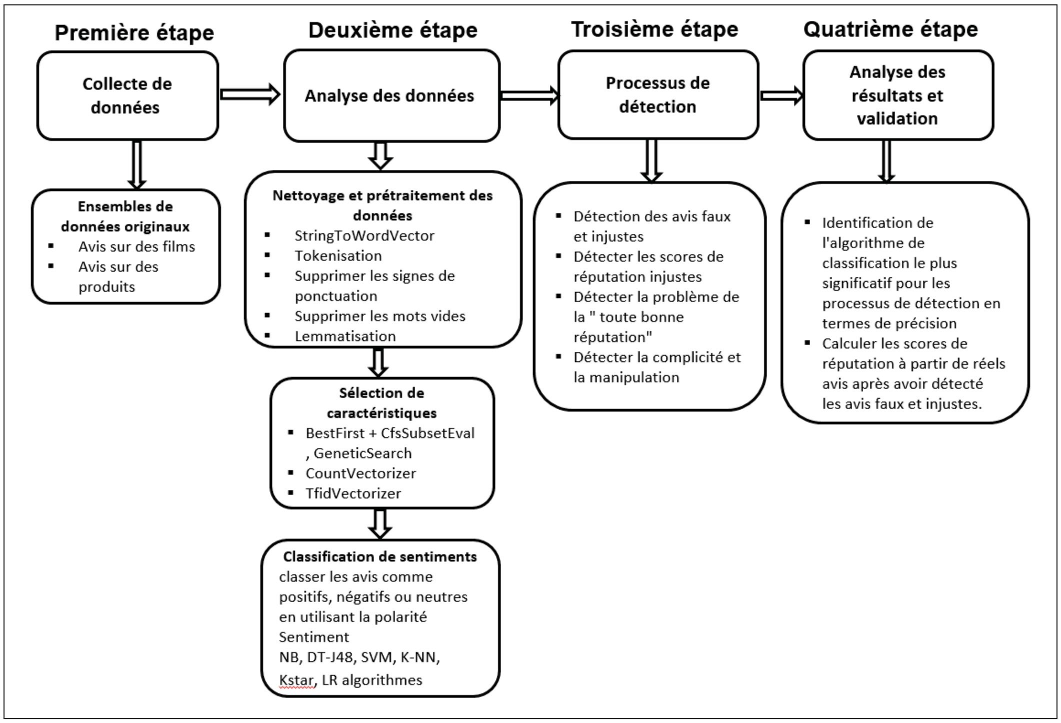

Première étape
Deuxième étape
Troisième étape
Quatrième étape
Collecte de données
Analyse des données
Processus de détection
Analyse des résultats et validation
Ensembles de données originaux
Avis sur des films
Avis sur des produits
Nettoyage et prétraitement des données
StringToWordVector
Tokenisation
Supprimer les signes de ponctuation
Supprimer les mots vides
Lemmatisation
Détection des avis faux et injustes
Détecter les scores de réputation injustes
Détecter la problème de la " toute bonne réputation"
Détecter la complicité et la manipulation
Identification de l'algorithme de classification le plus significatif pour les processus de détection en termes de précision
Calculer les scores de réputation à partir de réels avis après avoir détecté les avis faux et injustes.
Sélection de caractéristiques
BestFirst + CfsSubsetEval , GeneticSearch
CountVectorizer
TfidVectorizer
Classification de sentiments
classer les avis comme positifs, négatifs ou neutres en utilisant la polarité Sentiment
NB, DT-J48, SVM, K-NN, Kstar, LR algorithmes

Figure 0.1 Processus de la méthodologie de recherche

B. Tokenization

Dans ce processus, une fois les données extraites des ensembles de données, nous transformons les phrases en mots, afin qu’elles soient faciles à comprendre et à compter.

C. Lemmatisation

La lemmatisation désigne les techniques de normalisation des mots dans le domaine du traitement du langage naturel (TLN) qui sont utilisées pour préparer les textes, les mots et les documents en vue d’un traitement plus avancé. La principale fonction de la lemmatisation est le processus de conversion des mots en leurs mots racines. Dans notre étude, nous avons utilisé la lemmatisation, sans steaming, parce que tout au long de notre mise en œuvre, nous avons comparé entre les deux, en appliquant sur certains mots, comme "feeding" et "flying" : la conversion de "feeding" en utilisant steaming est "feed", et la conversion de "feeding" en utilisant la lemmatisation est aussi "feed". Cependant, la conversion de "flynig" en utilisant le steaming est "fli", et la conversion de "flynig" en utilisant la lemmatisation est "fly". Certains mots spéciaux utilisant le steaming n'ont pas de sens et cela aura un effet sur la classification des sentiments, et pour cette raison, nous utilisons la lemmatisation et non le steaming.

2. Sélection de caractéristiques

Afin de sélectionner les caractéristiques prosodiques les plus importantes et d'optimiser la performance de la classification, un évaluateur de sous-ensemble ("subet evaluator") a été utilisé. Les évaluateurs de sous-ensembles prennent un sous-ensemble de caractéristiques et renvoient un nombre, qui mesure une qualité du sous-ensemble et guide la recherche ultérieure. Pour la sélection de la méthode, l'outil d'exploration de données WEKA a été utilisé dans ce travail qui comprend CfsSubsetEval, GeneticSearch + BestFirst, et ces caractéristiques ont présenté la meilleure performance dans l'ensemble de données. Nous avons également implémenté CountVectorizer et TfidfVectorizer comme une sélection de caractéristiques en utilisant Scikit-Learn :

CountVectorizer : Le CountVectorizer offre une méthode pratique permettant de marquer (tokenizing) une compilation de texte et de construire la terminologie des mots connus.

TfidVectorizer : La caractéristique TfidfVectorizer permet de marquer les documents, d'apprendre le vocabulaire et la pondération inverse de la fréquence des documents, et de coder les nouveaux documents.

Dans notre étude, l'algorithme de régression logistique (Logistic Regression) avec la sélection de la caractéristique CountVectorizer a obtenu de meilleures performances qu'avec la caractéristique TfidfVectorizer.

3. Sentiment Classification

Dans cette phase, nous utilisons des algorithmes de classification des sentiments, qui ont été appliqués dans de nombreux domaines, tels que le commerce, la médecine, les médias, la biologie, etc. Une méthode de classification des sentiments est déterminée par un entraînement sur un ensemble de données connu, et en classant les avis en retour comme positifs ou négatifs. Il existe de nombreuses techniques différentes dans les méthodes de classification, comme NB, DT-J48, SVM,K-NN, et les réseaux neuronaux. Dans cette étude, nous avons utilisé six classificateurs supervisés populaires : NB, DT-J48, SVM, K-NN, KStar, les algorithmes LR. En fait, la régression logistique est un algorithme robuste pour la classification à deux classes. Dans notre étude, nous avons utilisé un algorithme de régression logistique avec des méthodes de sélection à deux caractéristiques ("two-feature selection"). Un algorithme de classification vraiment rapide et bien connu est la Régression logistique (LR), également connue sous le nom de fonction logistique, et est utilisée pour attribuer des observations à un ensemble discret de classes. Dans notre travail, nous utilisons cet algorithme avec CountVectorizer pour la sélection des caractéristiques, et nous avons trouvé qu'il s'agissait de la méthode la plus appropriée et la plus précise.

La troisième étape : Processus de détection

Cette étape consiste à prédire les résultats des modèles lors du test des ensembles de données, puis à générer une matrice de confusion ("confusion matrix"). La matrice de confusion affiche les méthodes dans lesquelles le modèle de classification est confondu lors de la réalisation des prédictions Hinton *et al.* (2015). Les avis sont classés sur la base de la matrice de confusion générée en positif, négatif ou neutre. La matrice de confusion montre le nombre de prédictions réelles et fausses obtenues avec des données connues, et pour chaque algorithme utilisé dans cette étude, il y a une évaluation de performance différente. La matrice de confusion représente également une partie particulièrement importante de notre recherche, puisqu'elle nous permet de classer l'ensemble des données Amazon des avis en avis justes ou injustes.

La quatrième étape : Analyse des résultats et validation

À ce stade, nous avons comparé les différentes exactitudes fournies par les ensembles de données avec divers algorithmes de classification et nous avons identifié l'algorithme de classification le plus significatif pour détecter les faux avis positifs et négatifs, les avis positifs injustes et les avis négatifs injustes, les scores de réputation injustes, la problème de la " toute bonne réputation ", la complicité et la manipulation. En outre, nous avons calculé les scores de réputation à partir de véritables avis en retour après avoir détecté les faux avis et les avis injustes.

0.5 Contributions de la thèse

Afin d'analyser les ensembles de données des avis sur les films et les ensembles de données des avis sur Amazon, nous avons construit trois modèles en utilisant des algorithmes de classification des sentiments. Ces modèles nous aideront à utiliser plusieurs scénarios pour valider le système proposé, qui permettra d'améliorer les systèmes de réputation. Les modèles ci-dessous présentent les principales composantes de notre approche.

Le premier modèle : Détection de faux avis sur les avis sur des films par l'analyse des sentiments (Fake Reviews Detection on Movie Reviews through Sentiment Analysis)

Pour atteindre notre objectif, nous avons analysé un ensemble de données de revues sur des films en utilisant l'outil Weka pour la classification des textes. Dans l'approche proposée, comme le montre la figure 2.1, nous avons suivi certaines étapes qui sont impliquées dans la SA en utilisant les approches décrites dans le chapitre **??**. Dans cette approche, nous avons étudié l'exactitude de tous les algorithmes de classification des sentiments, et la manière de déterminer quel algorithme est le plus précis. En outre, nous avons pu détecter de faux avis positifs et de faux avis négatifs grâce à des processus de détection.

Le deuxième modèle : Détection des avis injustes sur les avis Amazon grâce à l'analyse de sentiments (Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis)

Notre approche a été organisée en quelques étapes, comme le montre la figure 3.1, qui impliquent les approches de classification des sentiments supervisés à l'aide de l'outil Weka pour la classification des textes, comme décrit dans la section **??** du chapitre. Dans cette approche, nous avons étudié l'exactitude ("accuracy"), la précision ("precision") et le rappel des algorithmes de classification des sentiments. En outre, nous avons pu détecter des avis négatifs injustes, des avis neutres injustes et des avis positifs injustes en utilisant les processus de détection de cette méthode.

Le troisième modèle : Construction d'un modèle d'analyse du sentiment et calcul des scores de réputation (Building Sentiment Analysis Model and Compute Reputation Scores)

Notre approche a été organisée en plusieurs étapes, comme le montre la figure 1-31a, qui impliquent les approches de classification des sentiments supervisés à l'aide de l'outil Scikit-Learn in Python pour la classification des textes, comme décrit dans le chapitre 4. Dans cette approche, nous avons introduit une revue des systèmes existants d'analyse de réputations et de sentiments, et le développement pertinent de ces approches. Nous avons souligné les problèmes importants auxquels les acheteurs pourraient être confrontés en ce qui concerne la réputation des vendeurs, y compris les évaluations et les avis injustes ("unfair"). Ensuite, nous avons illustré certaines solutions potentielles capables de calculer les scores de réputation sans inclure les avis positifs injustes et les avis négatifs injustes dans un environnement de commerce électronique.

En raison de l'originalité de la présente thèse, plusieurs publications liées au travail de recherche ont été produites. La partie suivante énumère les documents susmentionnés dans l'ordre chronologique :

- Elmurngi, E., & Gherbi, A. (2017). An empirical study on detecting fake reviews using machine learning techniques. In 2017 seventh international conference on innovative computing technology (intech) (pp. 107–114).
- Elmurngi, E., & Gherbi, A. (2017). Detecting fake reviews through sentiment analysis using machine learning techniques. IARIA/data analytics, 65–72.
- Elmurngi, E., & Gherbi, A. (2018). Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques. International Journal on Advances in Systems and Measurements, 11(1 & 2), 196–207.
- Elmurngi, E. I., & Gherbi, A. (2018). Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques. JCS, 14 (5), 714–726.
- Elmurngi, E. I., & Gherbi, A. (2020). Building Sentiment Analysis Model and Compute Reputation Scores in E-Commerce Environment Using Machine Learning Techniques. International Journal of Organizational and Collective Intelligence (IJOCI), 10(1), 32-62.

0.6 Organisation de la thèse

L'organisation de cette thèse est divisée en quatre chapitres, comme indiqué dans la figure 0.2. Tout d'abord, une revue des travaux connexes, suivie de trois approches proposées et se terminant par la conclusion et les travaux futurs. Cette thèse est basée sur un manuscrit, chaque chapitre présentant une contribution différente. Les contributions sont énumérées ci-dessous :

- Au chapitre 2 (Fake Reviews Detection on Movie Reviews through Sentiment Analysis using Supervised Learning Techniques), détection de faux avis sur les avis de films par l'analyse de sentiments à l'aide de techniques d'apprentissage supervisées. Ce travail a été publié dans l'International Journal on Advances in Systems and Measurements (IARIA) ;
- Dans le chapitre 3 (Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques), Détection des avis injustes sur les avis Amazon par l'analyse de sentiments à l'aide des techniques d'apprentissage supervisé. Ce travail a été publié dans le Journal of Computer Science (JCS) ;
- Dans le chapitre 4 (Building Sentiment Analysis Model and Compute Reputation Scores in E-commerce Environment using Machine Learning Techniques), Construction d'un modèle d'analyse de sentiments et calcul de scores de réputation dans un environnement de commerce électronique à l'aide de techniques d'apprentissage automatique. Ce travail a été publié dans l'International Journal of Organizational and Collective Intelligence (IGI Global).

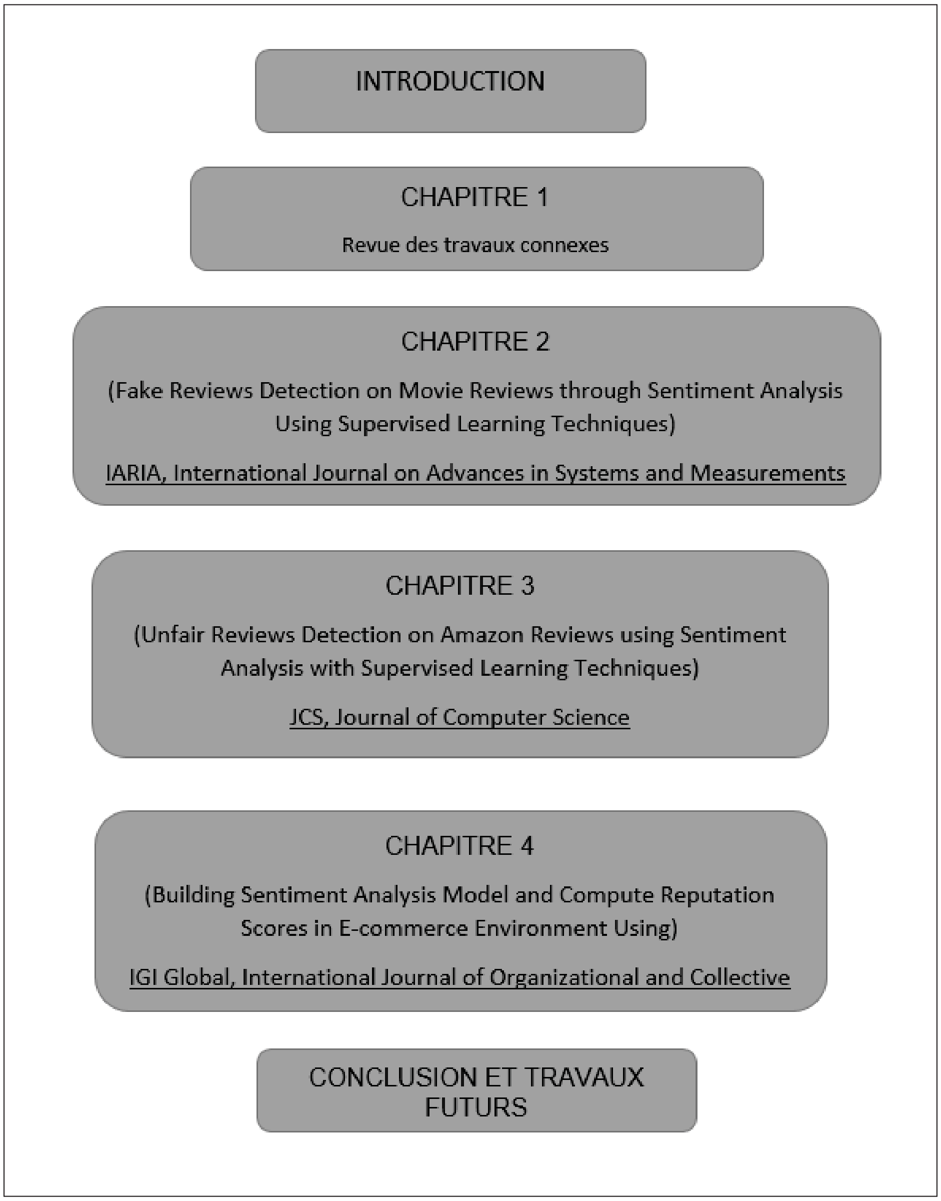

INTRODUCTION
CHAPITRE 1
Revue des travaux connexes
CHAPITRE 2
(Fake Reviews Detection on Movie Reviews through Sentiment Analysis Using Supervised Learning Techniques)
IARIA, International Journal on Advances in Systems and Measurements
CHAPITRE 3
(Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques)
JCS, Journal of Computer Science
CHAPITRE 4
(Building Sentiment Analysis Model and Compute Reputation Scores in E-commerce Environment Using)
IGI Global, International Journal of Organizational and Collective
CONCLUSION ET TRAVAUX FUTURS

Figure 0.2 Structure de la thèse. La ligne en gras et soulignée indique que le contenu a été publié dans une revue à comité de lecture ("peer review")

CHAPITRE 1

CONCEPTS DE BASE ET TRAVAUX CONNEXES

1.1 Introduction

L'objectif de ce chapitre est de situer la présente étude dans le contexte d'autres études sur l'évaluation de la vulnérabilité des systèmes de réputation. Cette étude utilise une méthode statistique d'évaluation de la vulnérabilité ; les travaux connexes mettent l'accent sur les études qui ont appliqué des méthodes statistiques à ce point. Enfin, ce chapitre compare ces approches en soulignant leurs forces et leurs faiblesses.

1.2 Définition du système de réputation

La réputation, en général, est une information utilisée pour porter un jugement de valeur sur une chose ou un bien dans son contexte pendant une période limitée.

La première définition des systèmes de réputation : Les systèmes de réputation (Reputation Systems (RSs)) sont l'un des mécanismes établis pour aider les consommateurs à prendre une décision concernant des achats en ligne Gutowska & Sloane (2009).

La deuxième définition des systèmes de réputation : Les systèmes de réputation (RS) sont un système qui collecte, distribue et rassemble les informations en retour (feedback) sur le comportement des participants.

1.3 Définition de l'analyse de sentiments

L'analyse du sentiment (Sentiment Analysis (SA)), connue également sous le nom de "extraction d'opinion" (Opinion Mining (OM)), est le domaine d'étude qui analyse les opinions, les évaluations et les sentiments envers des entités telles que les services, les individus, les enjeux, les sujets, et leurs attributs Liu (2012).

1.4 Système de réputation pour les applications du commerce électronique (e-commmerce)

La réputation des places de marché (marketplace) du commerce électronique joue désormais un rôle essentiel dans la décision de lancer une transaction et la tarification des produits ou services sur les places de marché en ligne comme eBay.com. Les informations en retour ("Feedback") d'eBay calculent la réputation d'un utilisateur comme la somme de ses évaluations à vie. Les profils de réputation sont conçus pour prédire les performances futures et aider les utilisateurs. Les vendeurs ayant une excellente réputation peuvent exiger des prix plus élevés pour leurs produits tandis que les détenteurs de mauvaise réputation attirent moins d'acheteurs Resnick & Zeckhauser (2002). De nombreuses études telles que Jøsang *et al.* (2006); Liu *et al.* (2011); Ehsaei (2012) proposent des architectures de systèmes de réputation de confiance (Trust Reputation Systems(TRS)) ainsi que différentes méthodes pour calculer le score de la réputation liée à un produit. D'autre part, peu de travaux de recherche sur les TRS ont pris en compte l'analyse sémantique des informations en retour ("feedback") et surtout le degré de confiance de l'utilisateur dans le calcul des scores de confiance des produits.

1.5 Revues textuelles pour fournir un avis détaillé sur le produit

La plupart des modèles de réputation disponibles dépendent des données numériques disponibles dans différents domaines ; un exemple est celui de l'évaluation (rating) dans le commerce électronique. En outre, la plupart des modèles de réputation se concentrent uniquement sur les évaluations globales des produits, sans tenir compte des avis fournis par les clients Xu *et al.* (2016). D'autre part, la plupart des sites Web permettent aux consommateurs d'ajouter des avis textuels pour donner un opinion détaillée sur le produit Tian *et al.* (2014a), Tian *et al.* (2014b). Ces avis sont mis à la disposition des consommateurs, qui dépendent de plus en plus des avis plutôt que des évaluations. Grâce aux modèles de réputation qui pourraient utiliser les méthodes de SA pour extraire les opinions des utilisateurs et utiliser ces données dans le système de réputation. Ces informations peuvent inclure les opinions des consommateurs sur différentes caractéristiques Abdel-Hafez & Xu (2013) Abdel-Hafez *et al.* (2012). Il existe un grand nombre d'études sur l'exploration de textes ("text mining") pour analyser les informations en retour ou les avis des clients. L'étude menée par Hu & Liu (2004) et Gupta *et al.* (2009) intègre l'utilisation du traitement du langage naturel pour extraire des paires de noms et d'adjectifs en phrases par le biais du marquage des parties du discours (Parts-of-speech (POS) tagging) et de l'exploration des règles d'association sur les avis des consommateurs de produits afin de trouver des caractéristiques fréquentes et peu fréquentes pour vérifier les caractéristiques du produit. Une autre étude de Chinsha & Joseph (2014) a réalisé une exploration de texte et de règles linguistiques pour analyser les avis et détecter l'orientation des opinions.

1.6 Problèmes liés à l'analyse du sentiment

De nos jours, l'analyse de sentiments est un domaine de recherche très populaire. De nombreux travaux sont réalisés, mais il n'existe pas encore de méthode suffisamment bonne pour classer les sentiments. Pour de nombreux auteurs, la moyenne des résultats est légèrement supérieure à 85%, mais cela ne suffit pas si nous avons besoin de résultats plus précis.

L'objectif principal de l'analyse des sentiments est d'analyser les avis et de tester les scores des sentiments. Cette analyse est divisée en trois niveaux Thomas (2013) : niveau document Yessenalina *et al.* (2010), niveau phrase Farra *et al.* (2010), niveau mot/termeEngonopoulos *et al.* (2011) ou niveau aspect Zhou & Song (2015). Les processus séquentiels sont l'évaluation de l'analyse des sentiments et la détection de la polarité des sentiments.

Plusieurs enjeux doivent être pris en compte lors de la conduite du SA Vinodhini & Chandrasekaran (2012). Deux enjeux majeurs sont abordés. Premièrement, le point de vue (ou l'opinion) observé comme négatif dans une situation peut être considéré comme positif dans une autre situation. Deuxièmement, les gens n'expriment pas toujours leurs opinions de la même manière. La plupart des techniques de traitement de texte courantes utilisent le fait que des modifications mineures entre les deux fragments de texte ne sont pas susceptibles de changer le sens réel Vinodhini & Chandrasekaran (2012).

L'analyse de sentiments des données des médias sociaux a également été appliquée pour évaluer les produits, comme expliqué par les auteurs de Oelke *et al.* (2009). Chaque auteur propose ses propres méthodes pour évaluer les opinions. Malheureusement, la plupart des outils ou algorithmes d'analyse de sentiments sont encore au stade de la recherche. Jusqu'à présent, il n'existe aucun algorithme qui puisse fournir des résultats 100 % précis pour l'analyse de sentiments. Il y a encore plusieurs débats entre différents chercheurs qui tentent de prouver que leur solution est plus parfaite que les autres.

Cette thèse se concentre sur les enjeux les plus importants dans la phase d'évaluation de sentiments, qui ont un impact significatif sur le score de sentiments et la détection de la polarité.

1.7 Détection des faux avis ("Fake Reviews")

Le problème principal pour identifier empiriquement les faux avis est que nous ne pouvons pas observer directement si un avis est faux. La situation est encore compliquée par l'absence de norme unique pour déterminer ce qui rend un avis "faux".

Le filtrage et l'identification des faux avis ont une signification substantielle Jindal & Liu (2008). Dans Moraes *et al.* (2013) les auteurs ont proposé une technique pour catégoriser un avis textuel sur un seul sujet. Le niveau de document classé par sentiments est appliqué pour indiquer qu'un sentiment est négatif ou positif. Les techniques d'apprentissage supervisé comprennent deux phases, la sélection et l'extraction de la catégorisation des avis en utilisant des modèles d'apprentissage tels que le MVC. Extraire la meilleure et la plus précise approche, et simultanément catégoriser le texte des commentaires écrits des clients en opinions négatives ou positives. Il s'agit d'un domaine de recherche majeur qui a attiré l'attention. Bien qu'il soit encore dans une phase d'introduction, il y a eu beaucoup de travail lié à plusieurs langues Liu *et al.* (2005) Fujii & Ishikawa (2006) Ku *et al.* (2006). Notre travail a utilisé plusieurs algorithmes d'apprentissage supervisé tels que SVM, NB, KNN-IBK et DT-J48 pour la classification de sentiments des textes afin de détecter les faux avis.

Un problème récemment apparu avec les avis en ligne est que certains avis en ligne sont faux. Bien que la plupart des plateformes en ligne disposent de leurs propres algorithmes de détection des faux avis Diesner *et al.* (2017), ces algorithmes ont parfois une portée limitée et ne filtrent que

16 % des faux avis publiés Luca & Zervas (2016). Il est donc clairement nécessaire d'améliorer les algorithmes existants et d'élaborer de nouvelles approches. Certaines études ont tenté de le faire (Diesner *et al.* (2017); Zhang *et al.* (2016)). À cette fin, plusieurs méthodologies ont été utilisées, dont certaines seront examinées dans la suite de cette thèse.

1.8 Filtrage des avis injustes (unfair reviews)

L'identification et le filtrage des examens injustes ont une importance fondamentale.

(Jindal & Liu (2008); Moraes *et al.* (2013)) a proposé une méthode pour catégoriser l'avis textuel d'un sujet donné. L'analyse de sentiments au niveau du document s'applique pour indiquer un sentiment positif, neutre ou négatif. Les algorithmes d'apprentissage supervisé comprennent deux étapes, l'extraction et surtout la sélection des avis à l'aide de modèles d'apprentissage supervisé, comme l'algorithme NB. Cependant, nous avons besoin de l'analyse de sentiments (AS) pour chaque classe des avis contenant la caractéristique du produit, afin de classer les avis du consommateur comme des avis négatifs, des avis neutres ou des avis positifs. Nous devons également détecter les avis positifs injustes, les avis neutres injustes et les avis négatifs injustes en utilisant plusieurs algorithmes de classification par apprentissage supervisé.

1.9 Importance de la régression logistique (LR) sur les techniques de classification de sentiments

Les chercheurs Gamal *et al.* (2019) et Lin *et al.* (2015) ont présenté une étude empirique sur la classification de sentiments et la régression logistique qui est construite pour combiner différentes méthodes d'apprentissage machine et obtenir une performance exceptionnelle en matière de précision et de rappel.

Notre travail en 2018 comme le montre le chapitre 3 a utilisé des techniques de classification de sentiments sur un ensemble de données d'avis de consommateurs. Les expérimentations ont été réalisées en utilisant des algorithmes de classification : NaïveBayes (NB), Arbre de décision (DT-J48), Régression logistique (LR) et Machine à vecteurs de support (SVM) pour la classification de sentiments en utilisant trois ensembles de données d'avis. Les résultats des expérimentations montrent que l'algorithme de régression logistique (LR) est plus performant et plus exact que les trois autres classificateurs, non seulement pour la classification de textes, mais aussi pour la détection des avis injustes. Notre travail en 2019, comme le montre le chapitre 4, a utilisé l'algorithme de régression logistique avec deux sélections de caractéristiques différentes, afin d'analyser deux ensembles de données différents provenant d'Amazon. Nous avons pu détecter les avis positifs injustes et les avis négatifs injustes, la problématique de " toute excellente réputation ", ainsi que la complicité et la manipulation par le biais de nos processus.

1.10 Étude comparative de différents algorithmes de classification

Nous avons réalisé des études comparatives sur les algorithmes de classification afin de prouver la meilleure méthode pour détecter les faux avis en utilisant différents ensembles de données tels que les ensembles de données des News Group, les documents textuels, les ensembles de données des avis sur des films prouvent (Chu *et al.* (2016); Singh *et al.* (2013)) que NB et DKV (Distributed Keyword Vector) sont précis sans mots vides (stopwords) tandis que Hassan *et al.* (2011) trouve que NB est précis avec l'utilisation de mots vides. En utilisant les mêmes ensembles de données, Kalaivani & Shunmuganathan (2013) constate que le SVM est précis avec les mots vides, tandis que Pang & Lee (2004) constate que le SVM est précis uniquement sans mots vides. Cependant, les résultats de notre étude empirique de 2017, comme le montre le chapitre 2, prouvent que le SVM est robuste et précis à la fois avec et sans mots-clés, et aussi pour la détection de faux avis. En outre, nos travaux de 2018, comme le montre le chapitre 3, fournissent une comparaison de quatre algorithmes d'apprentissage automatique supervisé : Naïve Bayes (NB), Arbre de décision (DT-J48), Régression logistique (LR) et Machine à vecteurs de support (SVM) pour la classification de sentiments en utilisant trois ensembles de données des avis, ils utilisent également ces méthodes pour détecter les avis positifs injustes et les avis négatifs injustes. Ils ont constaté que l'algorithme de régression logistique (LR) est plus précis que les algorithmes Naïve Bayes (NB), Machine à vecteur de support (SVM) et Arbre de décision (DTJ48), tant pour la classification des textes que pour la détection des avis injustes.

Nos travaux en 2020, comme le montre le chapitre 4, nous avons proposé un modèle d'analyse de sentiments (SAM) basé sur un algorithme de régression logistique avec deux sélections de caractéristiques différentes, afin d'analyser deux ensembles de données différents provenant d'Amazon. Nous avons pu détecter les avis positifs injustes et les avis négatifs injustes, le problème de la " toute excellente réputation ", ainsi que la complicité et la manipulation par le biais de nos processus. En outre, notre méthode expérimentale a étudié l'exactitude, la précision et le rappel de l'algorithme de régression logistique avec deux sélections de caractéristiques, et la manière de déterminer quelle sélection de caractéristiques est la plus précise ("accurate") et prend le moins de temps.

CHAPITRE 2

ARTICLE 1 : FAKE REVIEWS DETECTION ON MOVIE REVIEWS THROUGH SENTIMENT ANALYSIS USING SUPERVISED LEARNING TECHNIQUES

Elshrif Elmurngi [a], Abdelouahed Gherbi [b]

[a, b] Département de Génie logiciel et des technologies de l'information, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article publié dans « International Journal on Advances in Systems and Measurements », juin 2018
ISSN : 1942-261x

2.1 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.

2.2 Introduction

Sentiment analysis (SA) is one of the significant domains of machine learning techniques Elmurngi & Gherbi (2017a). 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 Liu (2012). "The sentiment is usually formulated as a two-class classification problem, positive and negative" Liu (2012). 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 Medhat *et al.* (2014), 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 Pang *et al.* (2002) and Pang & Lee (2004). 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 Malbon (2013). 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 Xia *et al.* (2011). SVM is a type of learning algorithm that represents supervised machine learning approaches Barbu (2012), and it is an excellent successful prediction approach. The SVM is also a robust classification approach Esposito (2014). A recent research presented in Medhat *et al.* (2014) 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 Kalaivani & Shunmuganathan (2013)-Singh *et al.* (2013). 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 Hall *et al.* (2009), 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 2.3 presents the related works. Section 2.4 shows the methodology. Section 2.5 explains the experiment results, and finally, Section 2.6 presents the conclusion and future works.

2.3 Related Work

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.

2.3.1 Sentiment analysis issues

There are several issues to consider when conducting SA Vinodhini & Chandrasekaran (2012). 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 Vinodhini & Chandrasekaran (2012).

2.3.2 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 Xu *et al.* (2016). On the other hand, most websites allow consumers to add textual reviews to provide a detailed opinion about the product Tian *et al.* (2014a) Tian *et al.* (2014b). 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 Abdel-Hafez & Xu (2013) and Abdel-Hafez *et al.* (2012).

2.3.3 Detecting Fake Reviews Using Machine Learning

Filter and identification of fake reviews have substantial significance Jindal & Liu (2008). Moraes *et al.* (2013) 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 Liu *et al.* (2005)-Ku *et al.* (2006). Our work used several supervised learning algorithms such as SVM, NB, KNNIBK, K* and DT-J48 for Sentiment Classification of text to detect fake reviews.

2.3.4 A Comparative Study of different Classification algorithms

Table **??** 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 Chu *et al.* (2016) and Singh *et al.* (2013). While Hassan *et al.* (2011) finds that NB is accurate and a better choice, but it is not oriented for detecting fake reviews. Using the same datasets, Kalaivani & Shunmuganathan (2013) finds that SVM is accurate with stopwords method, but it does not focus on detecting fake reviews, while Pang & Lee (2004) 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 Elmurngi & Gherbi (2017a). 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.

Tableau 2.1 A Comparative Study of Different Classification Algorithms

Reference	Data Source	Size of dataset	Using Supervised Learning	Language	Classifiers	Detecting Fake Review	Measures	Using stopwords	The best method
Kalaivani & Shunmuganathan (2013)	Movie Reviews dataset	2000 Movie Reviews	Yes	English	NB,SVM, kNN	NO	Accuracy, Precision and recall	NO	SVM
Pang & Lee (2004)	Movie Reviews dataset	2000 Movie Reviews	Yes	English	NB, SVM	NO	Accuracy ,t-test	NO	SVM
Hassan *et al.* (2011)	News Group dataset	20 categories with 1000 documents	Yes	English	NB, SVM	NO	Micro-average and macro-average F measure	NO	NB
Chu *et al.* (2016)	Movie Reviews dataset	4000 movie reviews	Yes		NB, SVM, K-NN LLR, Delta TFIDF, LDASVM, TFIDF, DKV	NO	precision, recall, Fscore as metric, and Accuracy	NO	DKV
Singh *et al.* (2013)	Movie Reviews dataset	1400, 2000 Movie Reviews	Yes	English	NB, SVM	NO	Accuracy, Fmeasure and Entropy	NO	NB
Elmurngi & Gherbi (2017a)	Movie Reviews dataset	1400, 2000 Movie Reviews	Yes	English	NB, SVM, IBK, K*,DT-J48	NO	Precision, and Accuracy	NO	SVM
This work	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**

2.4 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 2.1, 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 Singh *et al.* (2013), which is frequently conceded as the standard gold dataset for the researchers working in the field of the Sentiment Analysis.

Tableau 2.2 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-)

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 2.2.

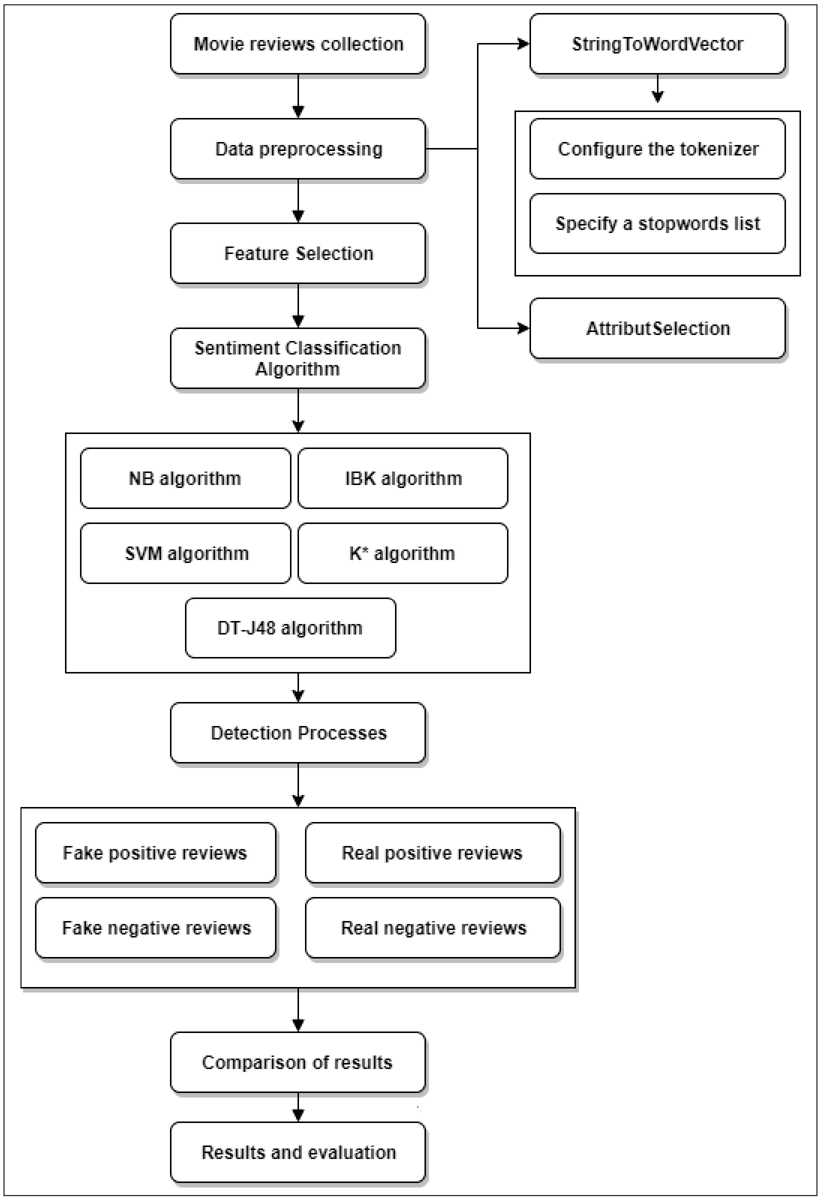


Figure 2.1 Steps and Techniques used in Sentiment Analysis

Step 2 : Data preprocessing

The preprocessing phase includes two preliminary operations, shown in Figure 2.1, 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 singlewords, 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 Cortes & Vapnik (1995). 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 nonparametric 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 (Song *et al.* (2007); Bhattacharya *et al.* (2012); Latourrette (2000); Zhang (2010)).

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,2.1-2.8 shows numerical parameters that could be applied following measures to evaluate the Detection Process (DP) performance. In Table 2.3, 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.

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

Tableau 2.3 The Confusion Matrix

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

$$Fake\ Positive\ Reviews\ Rate = \frac{FP}{TN + FP} \tag{2.1}$$

$$Fake\ Negative\ Reviews\ Rate = \frac{FN}{TP + FN} \tag{2.2}$$

$$Real\ Positive\ Reviews\ Rate = \frac{TP}{TP + FN} \tag{2.3}$$

$$Real\ Negative\ Reviews\ Rate = \frac{TN}{TN + FP} \tag{2.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$Recall = \frac{TP}{TP\ \ FN} \tag{2.7}$$

$$F - measure = \frac{2 * Precision * Recall}{Recall + Precision} \tag{2.8}$$

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.

2.5 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.

2.5.1 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 2.4, the confusion matrix displays results for movie reviews dataset v1.0.

Tableau 2.4 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. 2.5 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%.

The graph in Figure 2.2 displays a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

Tableau 2.5 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

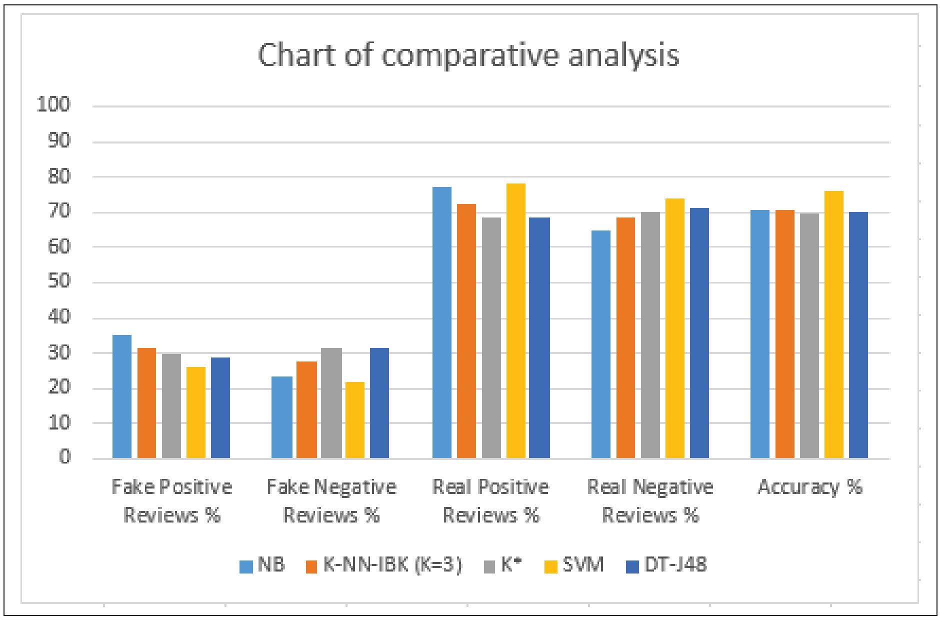


Figure 2.2 Comparative analysis of all methods

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

The graph in Figure 2.3 displays accuracy rate of NB, SVM, (K-NN, k=3), DT-J48 algorithms. We obtained a higher accuracy of SVM algorithm than other algorithms.

Tableau 2.6 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

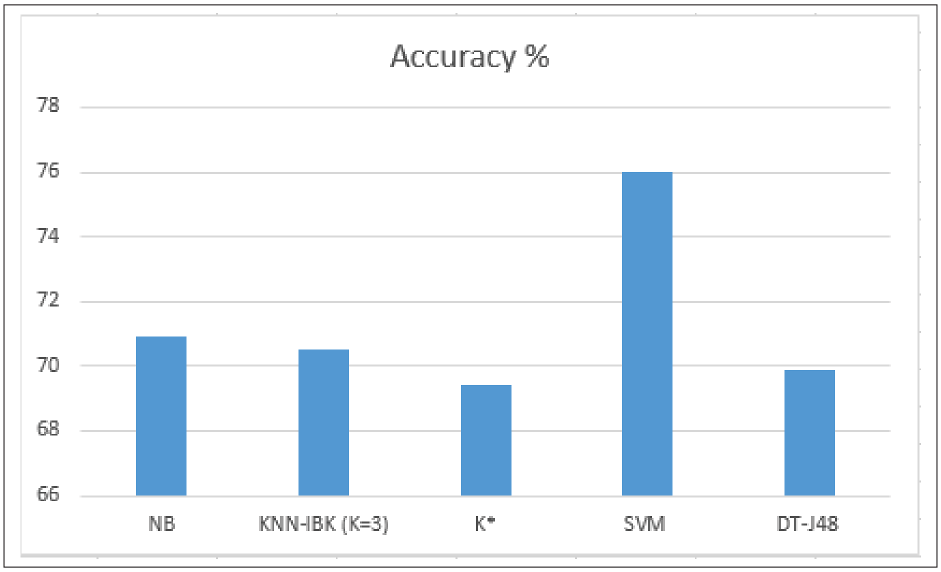


Figure 2.3 Accuracy of different algorithms

Table 2.7 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.

Tableau 2.7 Time Taken to Build the Model

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

Tableau 2.8 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

Table 2.8 and Figure 2.4 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.

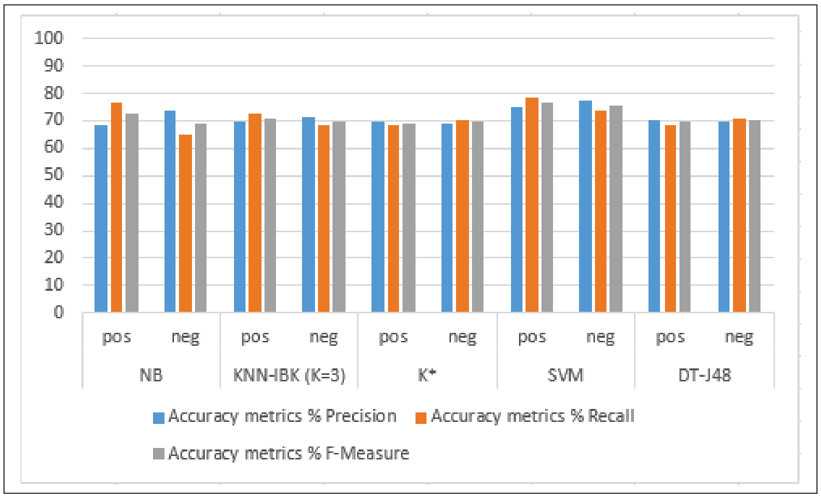


Figure 2.4 Comparison of metrics obtained from various multi-label classifiers

2.5.2 Experimental results 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 2.9 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.

Tableau 2.9 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.

Tableau 2.10 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 2.5 shows a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

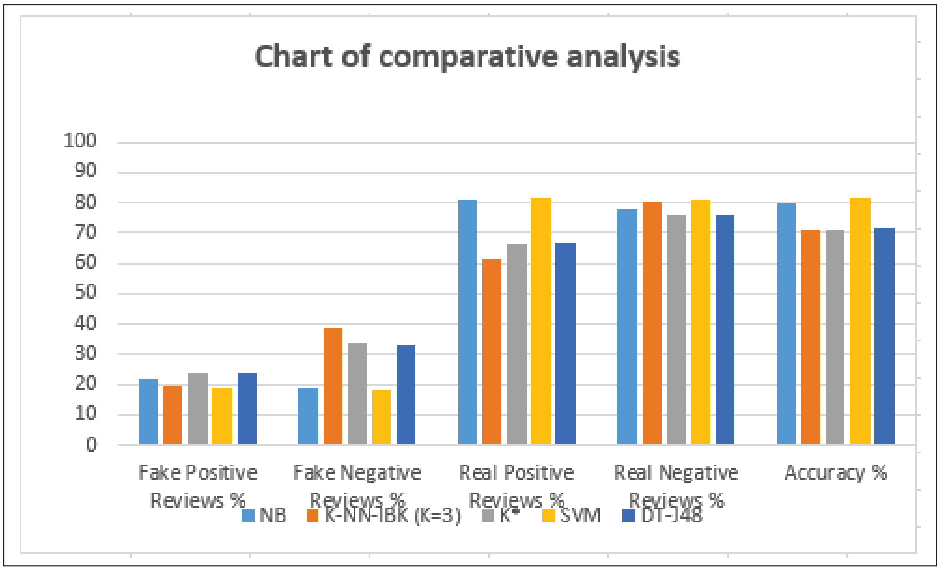


Figure 2.5 Comparative analysis of all methods

The comparison in Table 2.11 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, K*, and DT-J48 algorithms.

The graph in Figure 2.6 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.

Tableau 2.11 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

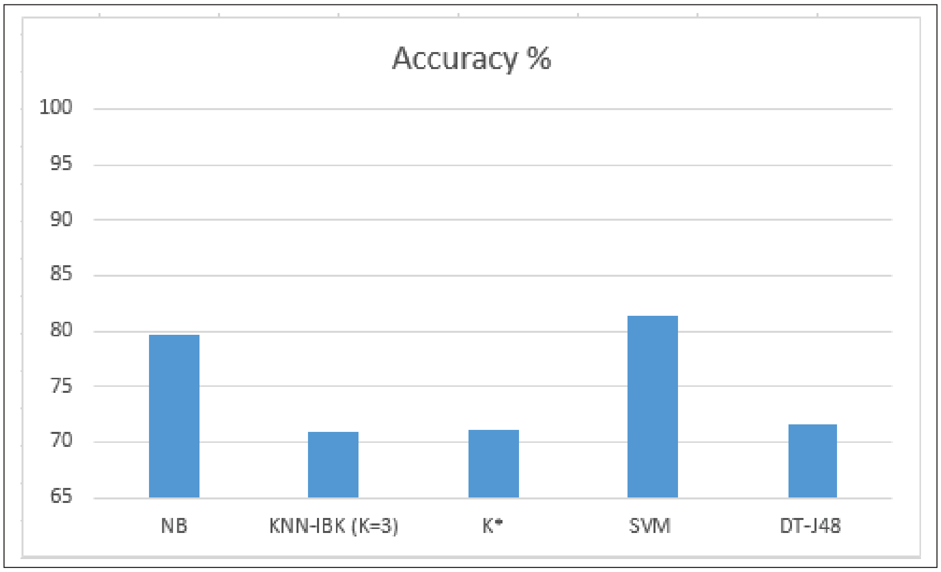


Figure 2.6 Graph showing the accuracy of different algorithms

Table 2.12 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.

Tableau 2.12 Time Taken to Build the Model

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

Table 2.13 and Figure 2.7 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.

Tableau 2.13 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

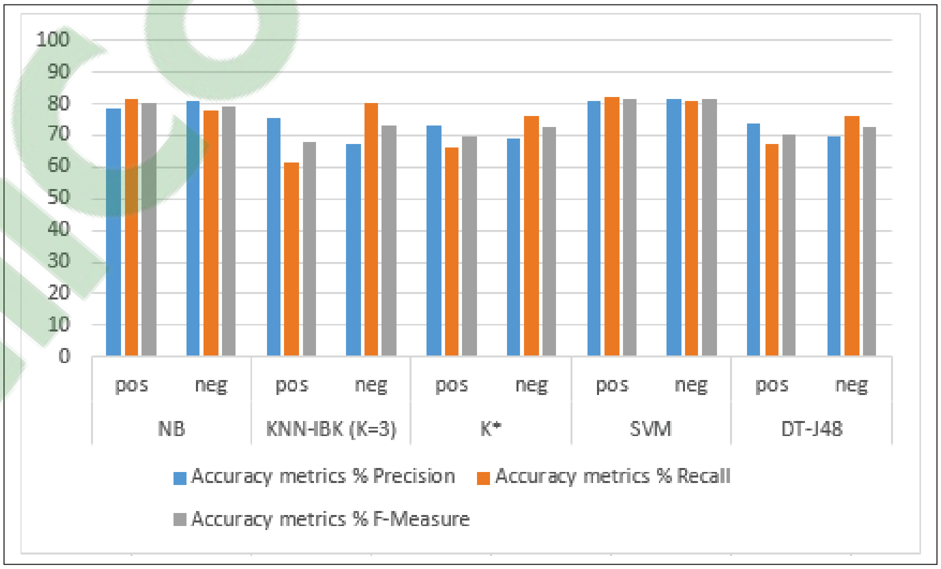


Figure 2.7 Comparison of metrics obtained from various multi-label classifiers

2.5.3 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 2.14, the confusion matrix displays results for movie reviews dataset v3.0.

Tableau 2.14 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 2.15 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%.

The graph in Figure 2.8 displays a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy for comparative analysis of all different algorithms.

Tableau 2.15 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

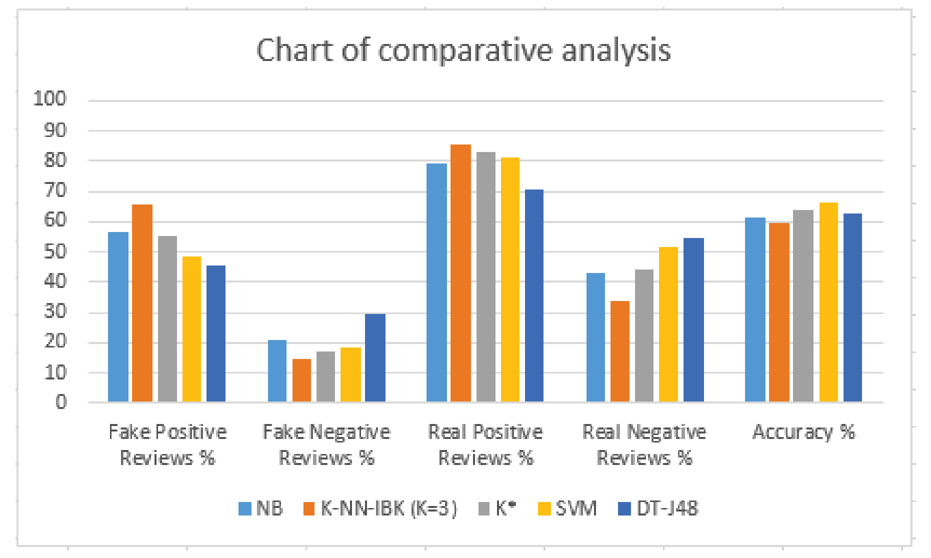


Figure 2.8 Comparative analysis of all methods

The comparison in Table 2.16 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

Tableau 2.16 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 2.9 displays accuracy rate of NB, SVM, (K-NN, k=3), DT-J48 algorithms. We obtained a higher accuracy of SVM algorithm than other algorithms.

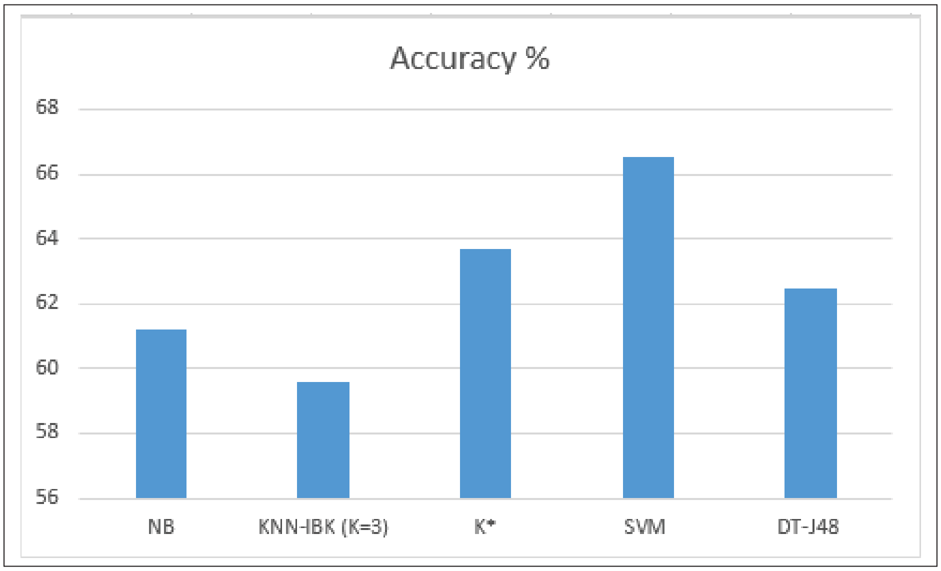


Figure 2.9 Graph showing the accuracy of different algorithms

Tableau 2.17 Time Taken to Build the 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 2.17 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 2.18 and Figure 2.10 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.

Tableau 2.18 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

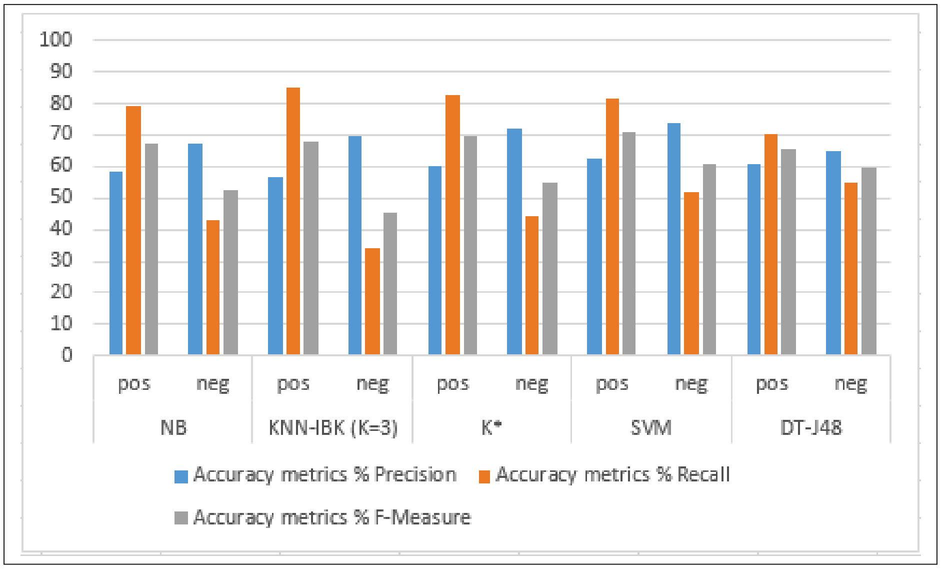


Figure 2.10 Comparison of metrics obtained from various multi-label classifiers

2.5.4 Discussion

Table 2.19 and Figure 2.11 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.

Tableau 2.19 The Best Result of 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**

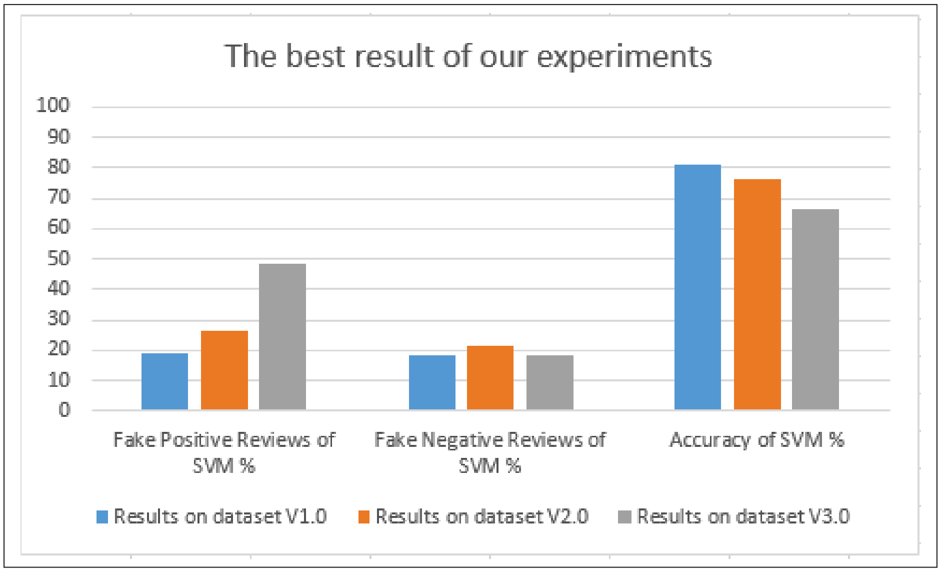


Figure 2.11 Summary of our experiments

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.

2.6 Conclusions 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.

2.7 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.

CHAPITRE 3

ARTICLE 2 : UNFAIR REVIEWS DETECTION ON AMAZON REVIEWS USING SENTIMENT ANALYSIS WITH SUPERVISED LEARNING TECHNIQUES

Elshrif Elmurngi [a], Abdelouahed Gherbi [b]

[a, b] Département de Génie logiciel et des technologies de l'information, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article publié dans « Journal of Computer Science », Mai 2018
DOI : 10.3844/jcssp.2018.714.726

3.1 Abstract

Reputation and trust are significantly important and play a pivotal role in enabling multiple parties to establish relationships that achieve mutual benefit especially in an E-Commerce (EC) environment. There are several factors negatively affecting the sight of customers and sellers in terms of reputation. For instance, lack of credibility in providing feedback reviews, by which users might create phantom feedback reviews to support their reputation. Thus, we will feel that these reviews and ratings are unfair. In this study, we have used Sentiment Analysis (SA) which is now the subject generating the most interest in the field of text analysis. One of the major challenges confronting SA today is how to detect unfair negative reviews, unfair neutral reviews and unfair positive reviews from opinion reviews. Sentiment classification techniques are used against a dataset of consumer reviews. Precisely, we provide comparison of four supervised machine learning algorithms : Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) for sentiment classification using three datasets of reviews, including Clothing, Shoes and Jewelry reviews, Baby reviews as well as Pet Supplies reviews. In order to evaluate the performance of sentiment classification, this work has implemented accuracy, precision and recall as a performance measure. Our experiments' results show that the Logistic Regression (LR) algorithm is the best classifier with the highest accuracy as compared to the other three classifiers, not merely in text classification, but in unfair reviews detection as well.

Keywords : Reputation Systems, Sentiment Analysis (SA), E-commerce (EC), Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR), Support Vector Machine (SVM)

3.2 Introduction

Nowadays, a large number of user reviews are made on almost everything that is present on the websites of the e-commerce environment, such as Amazon and eBay etc. Reviews may contain user reviews on products, destined to help other users in their buying decision making. Huge numbers of reviews exist, which makes it difficult for a consumer to read them all and make a decision. Furthermore, if the consumer reads some of the product reviews, it is difficult for them to distinguish between fair and unfair reviews. Likewise, user reviews are an important source of information for consumers. However, depending on their credibility, they can increase or decrease the reputation of products or websites.

Sentiment Analysis (SA) aims at determining the opinion of reviewers. With the growing popularity of websites such as Amazon.com where people can state their opinion on different products and rate them, e-commerce is replete with reviews and ratings. Thus, it is easy to find reviews on specific products. In this context, Reputation Systems for E-Commerce are considered as a collective measure to establish trustworthiness towards reviews or ratings coming from members of a community.

Reputation systems present a prominent technique to quantify the trustworthiness of vendors or the quality of products in E-Commerce (EC) environment. Recently e-commerce platforms, such as electronic marketplaces, have become a hot environment that allows millions of actors to trade goods and services by bringing them together. Purchasers and vendors are thereby offered incomparable opportunities to endless varieties of products. Regardless of whether Purchasers are looking for brand new technologies, highly specialized instruments or any other desired products, they will find a suitable transaction partner on the Web in most of the times. However, this “universe of strangers” also poses many issues Dellarocas (2005). In contrast with traditional person-to-person transactions in e-commerce, purchasers do neither get a complete feel of the products' actual quality nor do they get to know of the trustworthiness of a vendor. To tackle these issues, many e-commerce systems promote customers to provide feedback on a transaction describing their online shopping experience. Reputation systems process this information by collecting the feedback, aggregating the input data and providing one or more reputation values as output. In this way, reputation systems can assist purchasers in deciding which products or services to choose and whom to trust.

According to a recent study carried out by Diekmann *et al.* (2014),vendors with the best reputation have an increased number of sales. However, promoting trustworthy participation also bears an incentive for malicious actors to push their reputation unfairly to gain more benefit. Dishonest reviews or ratings have already become a serious problem in practice.Thus, in this research, our primary goal is detecting unfair reviews on Amazon reviews through Sentiment Analysis using supervised learning techniques in an E-Commerce environment. Our research is fundamentally focused at the document level of Sentiment Analysis, precisely on datasets of Amazon reviews. Sentiment Analysis methods will have a fundamental positive effect on reputation systems, especially inunfair reviews detection processesin an e-commerce environment and other domains. Feedback reviews in e-commerce is an important source of information for customers to reduce product uncertainty when making purchasing decisions. However, with increasing volume of feedback reviews, customers sometimes make product buying decisions based on unfair or fake feedback reviews.

One recent research provided in Medhat *et al.* (2014) introduces a survey on different SA algorithms, however, it only concentrates on using algorithms in diverse languages, with no focus on unfair reviews detection (Kalaivani & Shunmuganathan (2013); Singh *et al.* (2013)). Detecting unfair rating and unfair reviews have been studied in several works, including (Dellarocas (2000); Wu *et al.* (2010)). The methods that are used include : Clustering ratings into unfairly lowratings and unfairly high ratings and using third-party ratings on the producers of ratings, where ratings from less reputable producers are then assumed as unfair.

This research presents four supervised machine learning algorithms that include Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) in order to classify an opinion document that is put in comparison with three distinct Amazon reviews datasets. This research also spots unfair positive reviews, unfair neutral reviews and unfair negative reviews with the use of this method. The main goals of our study is to classify the document polarity of Amazon reviews datasets as fair or unfair reviews, with the use of Sentiment Analysis algorithms and supervised learning techniques.

The conducted experiments through sentiment classification algorithms have shown the performance measures of precision, recall and accuracy. In three cases (Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset and Pet Supplies dataset), we have applied NB, DT-J48, LR and SVM classifiers. These classifiers provide a useful perspective for understanding and evaluating many learning algorithms.

We can summarize the main contributions of this study as follows :

- This study use the Weka tool, an open source software for implementing machine learning algorithms Hall *et al.* (2009), to apply sentiment classification with the NB, DT-J48, LR and SVM algorithm which classifies the Amazon reviews datasets into unfair and fair reviews ;
- The sentiment classification algorithms are applied with stopwords removal, using three different Amazon reviews datasets. We observed that it is more effective to use the stopwords removal method than not using stopwords and that is also more efficient to detect unfair reviews ;
- This work implement several analysis on various Amazon reviews datasets to getthe supervised learning algorithmswith regard to precision, recall and exactitude.

The remainder of this paper is organized as per the following : Section 2 shows the related works. Section 3 presents the applied methodology. Section 4 displays the results of the experiment and lastly, Section 5 presents our conclusion and future studies.

3.3 Related Work

The majority of reputation models have been focused only on the overall products' ratings without taking into consideration their views provided by consumers Xu *et al.* (2016). Conversely, some of the reputation models have been focused solely on the overall products' reviews without taking into consideration the ratings provided by consumers. Furthermore, most E-commerce websites let their customers add textual reviews in order to give their opinion about the product in details (Tian *et al.* (2014b); Abdel-Hafez & Xu (2013)). Consumers can read these reviews and users are more and more dependent on reviews rather than on ratings. Through the Reputation, sentiment analysis methods could be used by models to extract the opinions of users and use the corresponding data in the reputation system, data that can include opinions about various features (Abdel-Hafez & Xu (2013); Gaber *et al.* (2012)).

Detection processes of sentiment classification based on a machine learning technique can clearly be expressed as a supervised learning technique with three classes : negative, neutral and positive. The testing and training data used in the existing research is commonly from reviews Gaber *et al.* (2012).

There is fundamental importance in the identification and filtering of unfair reviews (Jindal & Liu (2008); Moraes *et al.* (2013)) proposed a method to categorize the textual review of a given topic. The document level sentiment analysis is applied for stating a positive, neutral or negative sentiment. Supervised learning algorithms consist of two stages, extraction and especially reviews' selection using supervised learning models, such as NB algorithm. However, we need the Sentiment Analysis (SA) for each class of the reviews feedback containing the product feature, in order to classify the customer feedback reviews as negative reviews, neutral reviews or positive reviews. We need also to detect unfair positive reviews, unfair neutral reviews and unfair negative reviews by using several supervised learning classification algorithms.

A major research field has emerged around the subject of how to extract the best and most accurate method and simultaneously categorize the customers' written reviews into negative or positive opinions. Such research is still in introductory preliminary phase, but much work has been done in relation to several languages (Liu *et al.* (2005); Ku *et al.* (2006)).

A survey on various applications and SA algorithms was introduced in a recent research presented in Medhat *et al.* (2014), however, it only concentrates on using algorithms in various languages and does not concentrate on the detections of unfair reviews (Kalaivani & Shunmuganathan (2013); Singh *et al.* (2013)).

Supervised learning is a type of machine learning that requires learning from a set of training data. However, a dataset of the product is usually represented as a corpus of documents that possesses text processing challenges to be overcome before a classification model Shankar & Lin (2011). Cases of text processing techniques are stopword removal and tokenization. The common classification techniques for document analysis include Support Vector Machine (Elmurngi & Gherbi (2017b); Wen & Li (2007)), Logistic Regression Cheng & Hüllermeier (2009), Decision Tree Rajput & Arora (2013).

In this study, we present four supervised machine learning algorithms to classify the sentiment that is compared using three different Amazon reviews datasets. We also use these methods to detect unfair positive reviews and unfair negative reviews. Our study's main goal is to classify Amazon reviews datasets into fair reviews or unfair reviews with the use of Sentiment Analysis algorithms and supervised learning techniques.

The results of the conducted experiments have shown their accuracy and performance via four sentiment classification algorithms in order to detect unfair reviews. We have performed our experiments using three different datasets : The Clothing, Shoes and Jewelry reviews dataset and Baby reviews dataset. We have found that the Logistic Regression (LR) algorithm is more accurate as compared to the Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT-J48) algorithms, as much in text classification as in unfair reviews detection.

3.4 Methodology

Our methodology was organized in the next six steps, as shown in Fig. 3.1, steps that involve the supervised sentiment classification approaches using Weka tool for text classification as described below.

Tableau 3.1 Number of reviews and ratings of dataset

Dataset	Reviews	Ratings
Clothing, Shoes and Jewelry	278,677	278,677 (1to5 scores)
Baby	160,792	160,792 (1to5 scores)
Pet Supplies	157,836	157,836 (1to5 scores)

Tableau 3.2 Datasets before and after cleaning

	Before cleaning			After cleaning		
Dataset	View of a dataset	Class rating	Number of reviews	View of a dataset	Class rating	Number of reviews
Clothing, shoes and jewelry	ReviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star, 2 star 3 star 4 star, 5 star	26655 30425 221597	ReviewText, overall (rating of the product)	Negative Neutral Positive	23019 30423 221578
Baby	ReviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star, 2 star 3 star 4 star, 5 star	17012 17255 126525	ReviewText, overall (rating of the product)	Negative Neutral Positive	17001 17252 126479
Pet supplies	ReviewerID, asin (ID of the product), reviewerName, helpful (rating of the review), reviewText, overall (rating of the product), summary (summary of the review), unixReviewTime, reviewTime	1 star, 2 star 3 star 4 star, 5 star	17655 15933 124248	ReviewText, overall (rating of the product)	Negative Neutral Positive	12314 8106 118203

Step One : Amazon Reviews Collection

We have based our experiment on analyzing the standard dataset's sentiment value using machine learning algorithms. We have used the Amazon reviews' original dataset to test our reviews classification methods. Amazon.com has many different kinds of products, but here we would focus on three datasets : Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset

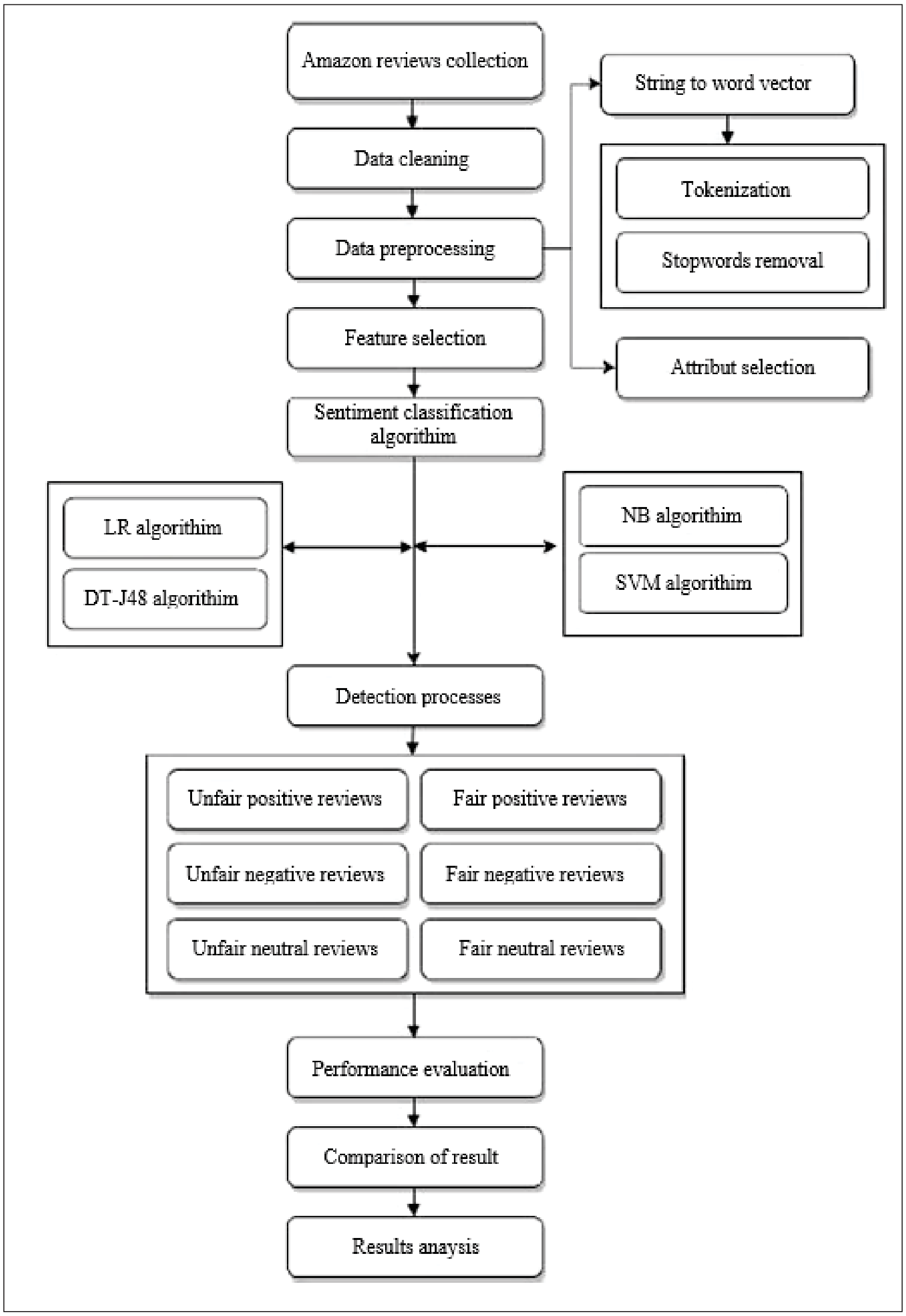


Figure 3.1 Steps used in the supervised learning approach

and Pet Supplies dataset. The datasets are available and have been collected by (McAuley and Leskovec, 2013). Table 3.1 describes a summary of the three collected datasets.

Step Two : Data Cleaning

The dataset used in our experiment is obtained from Amazon product data and was divided into five scales rating : 1 star, 2 stars, 3 stars, 4 stars and 5 stars. The original dataset is not easy to model and usually not so clean. We have deleted some blank rows that cause confusion in the analysis process. The datasets before and after cleaning are listed in Table 3.2 and are separated to apply the sentiment classification classifiers after cleaning datasets.

Step Three : Data Preprocessing

Data preprocessing is a significant step in the text mining process and plays an important part in a number of supervised learning techniques. We have broken down data preprocessing as per the following :

1) StringToWordVector (STWV)

StringToWordVector filter is the main text analysis tool in Weka and it makes the transformed datasets' attribute value either Positive, Negative or Neutral for all single-words, depending on the word appearing in the document or not. It's a filtration process which is used by the following two sub-processes : Stopwords Removal and Tokenization.

2) Stopwords Removal and Tokenization

Stopwords are common words that must be filtered out, before training the classifier. Some of those words are common words (e.g., "the," "a," "I," "of," "you," "and," "it") but do not add any significant information to our labeling scheme and do not add value to a sentence's meaning, but instead they bring confusion to our classifier.

3) Attribute Selection

Attribute selection in machine learning, also known as feature selection, is the process of selecting a subset of relevant features for use in model construction. Attributes selection can significantly increase the classification accuracy and make it better.

Step Four : Feature Selection

Feature Selection (FS) methods in sentiment analysis have got a significant role in increasing classification accuracy and identifying relevant attributes (Koncz and Paralic, 2011). Our research has implemented one feature selection method (BestFirst + CfsSubsetEval, GeneticSearch) largely used for the SA classification task with Stopwords Removal. Our analysis of Amazon reviews datasets with feature selection method found the use of Logistic Regression (LR) algorithm gave more accuracy in the classification task.

Step Five : Sentiment Classification Algorithms

For this step, the Sentiment classification algorithm was used to classify documents as positive, negative, or neutral. In our study, we used four popular supervised classifiers such as NB, DT-J48, LR and SVM classifiers.

Naïve Bayes(NB)

In machine learning Techniques, The NB algorithm is based on the Bayes rule of conditional probability with independence assumptions between the features.

Decision Tree (DT-J48)

The DT is a predictive machine-learning technique that decides the target value of a new sample based on several attribute values of the available data. DT-J48 is the implementation of Ross Quinlan's Iterative Dichotomiser 3 algorithm, used to generate a decision tree from a dataset.

Logistic Regression (LR)

The LR is a classification algorithm, also called the logistic function, used to assign observations to a discrete set of classes. logistic regression is actually a robust technique for two-class and multiclass classification. It is a simple, fast and popular classification technique. In our study, we used this algorithm and found it to be the best and most accurate method.

Support Vector Machine (SVM)

The SVM is supervised learning techniques with related learning algorithms that analyze dataset used for classification. In recent years, the SVM has been among the most widely used and most popular classifiers with supervised learning techniques.

Step Six : Detection Processes

This step consists in predicting the models output on testing the datasets and then generating a confusion matrix that classifies the reviews into positive, negative or neutral ones. The following attributes are involved in the results :

- True Positive Reviews (TPR) : Fair Positive Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Positive ;
- False Positive Reviews (FPR) : Unfair Positive Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Positive ;
- True Negative Reviews (TNR) : Fair Negative Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Negative ;
- False Negative Reviews (FNR) : Unfair Negative Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Negative ;
- True Neutral Reviews (TNR) : Fair Neutral Reviews found in the testing data and defined as the number of sentences that are correctly predicted by the classification model as Neutral ;
- False Neutral Reviews (FNR) : Unfair Neutral Reviews found in the testing data and defined as the number of sentences that are incorrectly predicted by the classification model as Neutral.

In Table 3.3, the confusion matrix shows the number of fair and unfair predictions made by the model compared with the actual classifications, equations 1 to 9 displays numerical parameters that could be applied following measures to evaluate the performance of detection process. For each algorithm used in this study, there is a different confusion matrix and evaluation of performance.

The confusion matrix represents a particularly significant part of our research since it lets us classify the Amazon datasets reviews into unfair or fair reviews. The confusion matrix is applied to each of the two algorithms mentioned in Step 4.

Step Seven : Comparison of Results

Here, we compared the different accuracy and precision provided by the Amazon reviews datasets using different classification algorithms and identified which algorithm was the most significant in the detection of Unfair positive and negative and Neutral Reviews.

3.5 Experimentsand Result Analysis

In this section, we present our experimental results from four different supervised machine learning algorithms to classify sentiment of three datasets, which are Clothing, Shoes and Jewelry reviews dataset, Baby reviews dataset and Pet Supplies dataset. Moreover and at the same time, we have used the same approaches to detect unfair reviews using Weka 3.8 tool, which is the latest stable version.

3.5.1 Confusion Matrix

Using the confusion matrix is one of the approaches used to evaluate the performance of a classifier. For a given set of a classifier and a document, there are six possible outcomes : True negative, false negative, true neutral and false neutral, true positive and false positive. If the document is labelled negative and is classified as negative, then it is counted as fair negative, else, if it is classified as positive then it is counted unfair positive. Likewise, if a document is labelled positive and is classified as positive, then it is counted as fair positive, else, if it is classified as negative, then it is calculated as unfair negative. Similarly, if a document is labelled neutral and is classified as neutral, then it is calculated as fair neutral, else, if it is classified as negative or positive, then it is calculated as unfair negative or positive.

The confusion matrix displays the number of fair and unfair predictions acquired from the classification model in comparison with the actual results. The confusion matrix is obtained by implementing NB, DT-J48, LR, SVM algorithms.

Table 3.4, 3.5 and 3.6 display confusion matrix for the Clothing, Shoes and Jewelry reviews dataset and the Baby reviews dataset, respectively.

Tableau 3.3 The confusion matrix

		Predicted class A Fair	Predicted class B Unfair	Predicted class C Unfair
Actual class A	Fair	True Negative Reviews (TNR)	False Neutral Reviews (FNeR)	False Positive Reviews (FPR)
Actual class B	Unfair	False Negative Reviews (FNR)	True Neutral Reviews (TNeR)	False Positive Reviews (FPR)
Actual class C	Unfair	False Negative Reviews (FNR	False Neutral Reviews (FNeR)	True Positive Reviews (TPR)
Unfair Negative Reviews Rate = FNR/(TNR + FNeR + FPR) (3.1)				
Unfair Neutral Reviews Rate = FNeR/(FNR + TNeR + FPR) (3.2)				
Unfair Positive Reviews Rate = FPR/(FNR + FNeR + TPR) (3.3)				
Fair Negative Reviews Rate = TNR/(TNR + FNeR + FPR) (3.4)				
Fair Neutral Reviews Rate = TNeR/(TNeR + FPR + FNR) (3.5)				
Fair Positive Reviews Rate = TPR/(TPR + FNeR+FNR) (3.6)				
Accuracy = TPR + TNR + TNeR/(TNR + FNRclassB + FNRclassC + FNeR + TNeR + FNeR + FPRclaasA + FPRclassB + TPR) (3.7)				
Precision = TNR/(TNR + FNR class B + FNRclass C) (3.8)				
Recall = TNR/(TNR + TNeR + FPR) (3.9)				

3.5.2 Evaluation Parameters

For us to establish the performance evaluation of the four Classification algorithms, we use an experiment on three different product reviews in terms of Unfair Negative Reviews predictive value, Unfair Neutral Reviews predictive value, Unfair Positive Reviews predictive value, Fair Negative Reviews predictive value, Fair Neutral Reviews predictive value, Fair Positive Reviews predictive value. Table 3.7, 3.8 and 3.9 display the evaluation parameters' results for four different classifiers and provide a summary of the experiment's recordings.

Tableau 3.4 Confusion matrix on clothing, shoes and jewelry

Algorithms			Predicted class A Fair	Predicted class B Unfair	Predicted class C Unfair
NB	Actual class A	Negative	6304	2118	14597
	Actual class B	Neutral	3551	3794	23078
	Actual class C	Positive	2118	5387	211621
DT-J48	Actual class A	Negative	5183	1310	16526
	Actual class B	Neutral	2713	2248	25462
	Actual class C	Positive	2979	2008	216591
LR	Actual class A	Negative	5006	1129	16884
	Actual class B	Neutral	2354	2151	25918
	Actual class C	Positive	2470	1806	217302
SVM	Actual class A	Negative	2835	86	20098
	Actual class B	Neutral	1386	84	28953
	Actual class C	Positive	1879	101	219598

Tableau 3.5 Confusion matrix on baby reviews dataset

Algorithms			Predicted class A Fair	Predicted class B Unfair	Predicted class C Unfair
NB	Actual class A	Negative	353	3253	1267
	Actual class B	Neutral	172	14234	7030
	Actual class C	Positive	46	6707	15925
DT-J48	Actual class A	Negative	322	3479	1072
	Actual class B	Neutral	237	14800	6399
	Actual class C	Positive	94	7545	15039
LR	Actual class A	Negative	380	3427	1066
	Actual class B	Neutral	199	15131	6106
	Actual class C	Positive	50	7610	15018
SVM	Actual class A	Negative	303	3610	960
	Actual class B	Neutral	188	15633	5615
	Actual class C	Positive	122	8179	14377

The graph in Fig. 3.2, 3.3 and 3.4 show a rate of Unfair Negative Reviews predictive value, Unfair Neutral Reviews predictive value, Unfair Positive Reviews predictive value, Fair Negative Reviews predictive value, Fair Neutral Reviews predictive value and Fair Positive Reviews predictive value from the comparative analysis of four different algorithms.

Tableau 3.6 Confusion matrix on pet supplies dataset

Algorithms			Predicted class A Fair	Predicted class B Unfair	Predicted class C Unfair
NB	Actual class A	Negative	5436	919	5959
	Actual class B	Neutral	2341	1059	4706
	Actual class C	Positive	2956	1141	19857
DT-J48	Actual class A	Negative	5554	523	6237
	Actual class B	Neutral	2275	534	5297
	Actual class C	Positive	2829	541	20584
LR	Actual class A	Negative	5220	438	6656
	Actual class B	Neutral	2094	513	5499
	Actual class C	Positive	2317	426	21211
SVM	Actual class A	Negative	4150	252	7912
	Actual class B	Neutral	1537	308	6261
	Actual class C	Positive	1683	196	22075

Classifier Evaluation Metrics : Accuracy and Precision and Recall for Various Datasets

Table 3.10 displays the results of evaluation parameters for four different Classification algorithms, including : NB, DT-J48, LR, SVM algorithms and provides a summary of this experiment's results.

Tableau 3.7 Evaluation parameters on clothing, shoes and jewelry dataset

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	3.2	3.1	70.5	27.4	12.5	95.5
DT-J48	2.3	1.4	78.6	22.5	7.4	97.7
LR	1.9	1.2	80.1	21.7	7.1	98.1
SVM	1.3	0.1	91.8	12.3	0.3	99.1

Tableau 3.8 Evaluation parameters on baby reviews dataset

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	3.1	2.1	74.8	27.2	7.6	96.3
DT-J48	2.5	0.6	81.3	24.1	2.8	98.0
LR	2.2	0.8	80.6	24.1	3.9	98.0
SVM	2.5	0.1	86.0	20.6	0.4	98.2

Tableau 3.9 Evaluation parameters on pet supplies dataset

Algorithms	Unfair negative reviews %	Unfair neutral reviews %	Unfair positive reviews %	Fair negative reviews %	Fair neutral reviews %	Fair positive reviews %
NB	16.5	5.7	52.2	44.1	13.1	82.9
DT-J48	15.9	2.9	56.5	45.1	6.6	85.9
LR	13.8	2.4	59.5	42.4	6.3	88.5
SVM	10	1.2	69.4	33.7	3.8	92.2

Tableau 3.10 Comparison of accuracy, precision, recall and time taken to the build model (in seconds) of classifiers on baby reviews dataset

Evaluation metrics %				
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	43.7	27.4	17.71
	neu	33.6	12.5	
	pos	84.9	95.5	
DT-J48	neg	47.7	22.5	261.55
	neu	40.4	7.4	
	pos	83.8	97.7	
LR	neg	50.9	217.0	83.81
	neu	42.3	7.1	
	pos	83.5	98.1	
SVM	neg	46.5	123.0	11561.03
	neu	31.0	0.3	
	pos	81.7	99.1	

The Comparison of accuracy of different classifiers on Clothing, Shoes and Jewelry reviews dataset in Table 3.11 indicates that the LR algorithm outperformed NB, DT-J48, SVM algorithms. The graph in Fig. 3.5 displays Accuracy of evaluation parameters for NB, DT-J48, Logistic Regression, SVM algorithms, as applied on the Musical Instruments reviews dataset. The Logistic Regression algorithms classification accuracy outperformed other algorithms.

The Comparison of accuracy of different classifiers on Baby reviews dataset and Clothing, Shoes and Jewelry reviews dataset and pet supplies reviews dataset in Table 3.12, 3.13 and 3.14 indicate that the LR algorithm outperformed NB, DT-J48, SVM algorithms.

Tableau 3.11 Classification Accuracy of different algorithms

Algorithms	Accuracy %
NB	80.61
DT-J48	81.45
LR	**81.61**
SVM	80.90

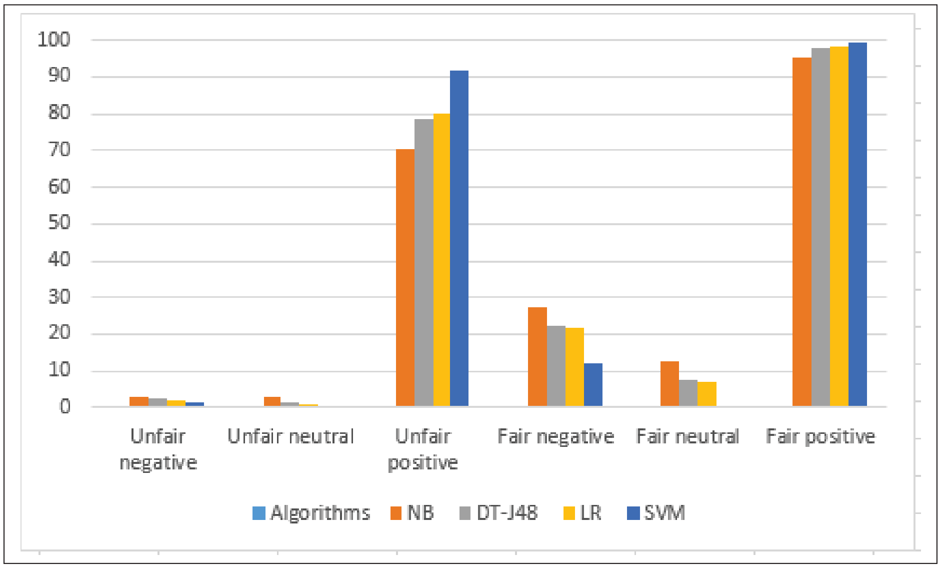


Figure 3.2 Graph showing the evaluation parameters on clothing, shoes and jewelry dataset

The graph shown in Fig. 3.6 displays Accuracy of evaluation parameters for NB, DT-J48, LR, SVM algorithms, as applied on the Baby reviews dataset. The Logistic Regression algorithm's classification accuracy outperformed other algorithms.

The Comparison of accuracy of different classifiers on Baby reviews dataset in Table 3.15 indicates that the LR algorithm outperformed NB, DT-J48, SVM algorithms.

The graph shown in Fig. 3.7 displays Accuracy of evaluation parameters for NB, DT-J48, LR, SVM algorithms, as applied on the Baby reviews dataset. The Logistic Regression algorithm's classification accuracy outperformed other algorithms.

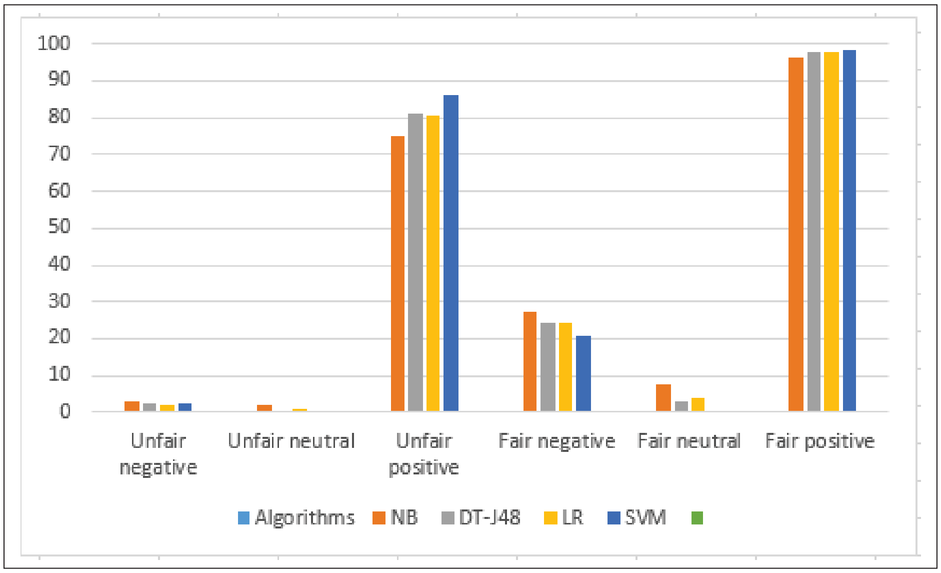


Figure 3.3 Graph showing the evaluation parameters on baby reviews dataset

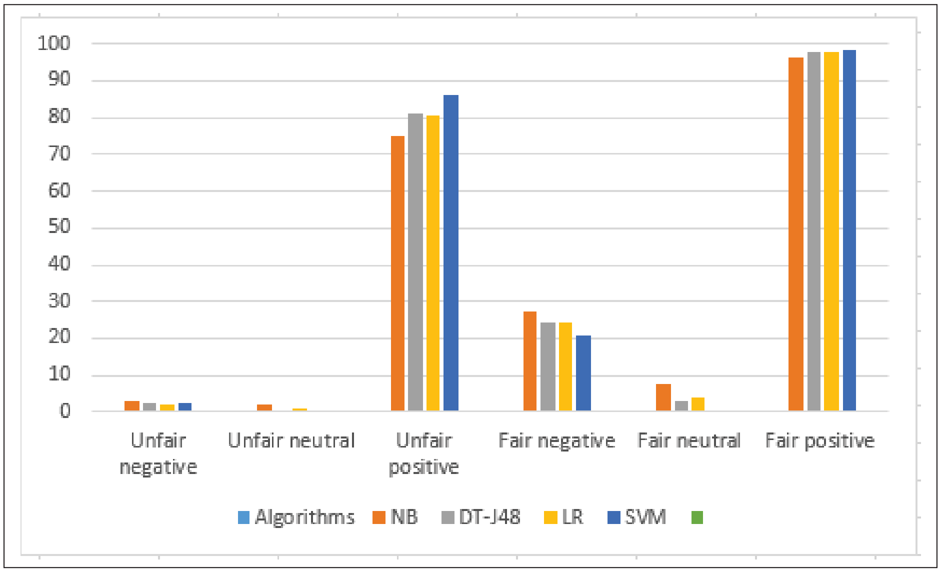


Figure 3.4 Graph showing the evaluation parameters on pet supplies dataset

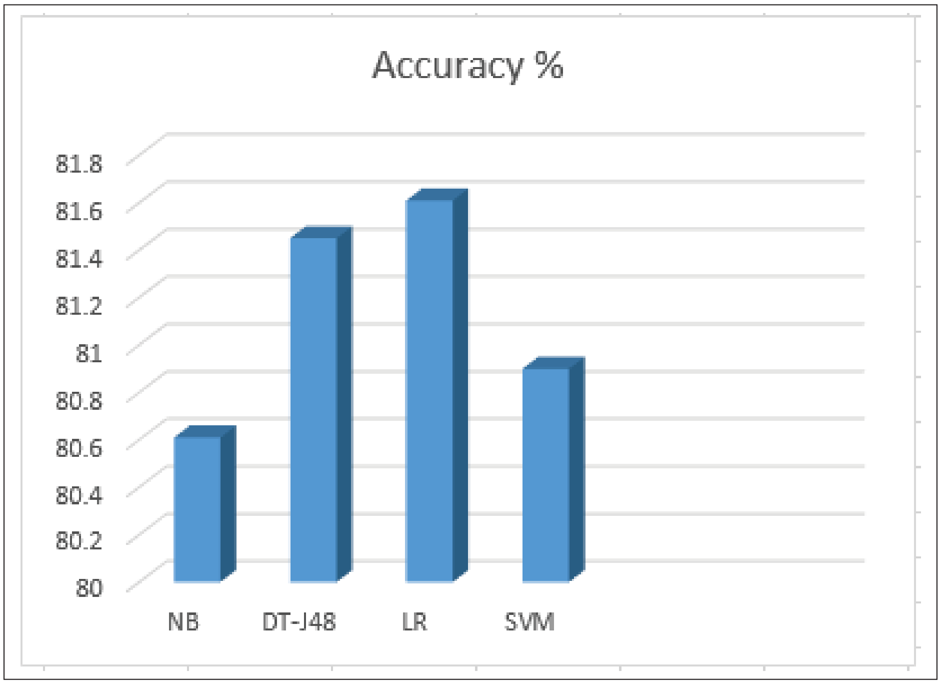


Figure 3.5 Comparison of accuracy of different classifiers on clothing, shoes and jewelry reviews dataset

Tableau 3.12 Classification Accuracy of different algorithms

Evaluation metrics %				
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	51.0	27.2	10.45
	neu	30.8	7.6	
	pos	82.6	96.3	
DT-J48	neg	53.6	24.1	97.05
	neu	36.6	2.8	
	pos	81.7	98.0	
LR	neg	56.1	24.1	67.78
	neu	36.2	3.9	
	pos	81.8	98.0	
SVM	neg	49.6	20.6	11561.03
	neu	34.5	0.4	
	pos	80.8	98.1	

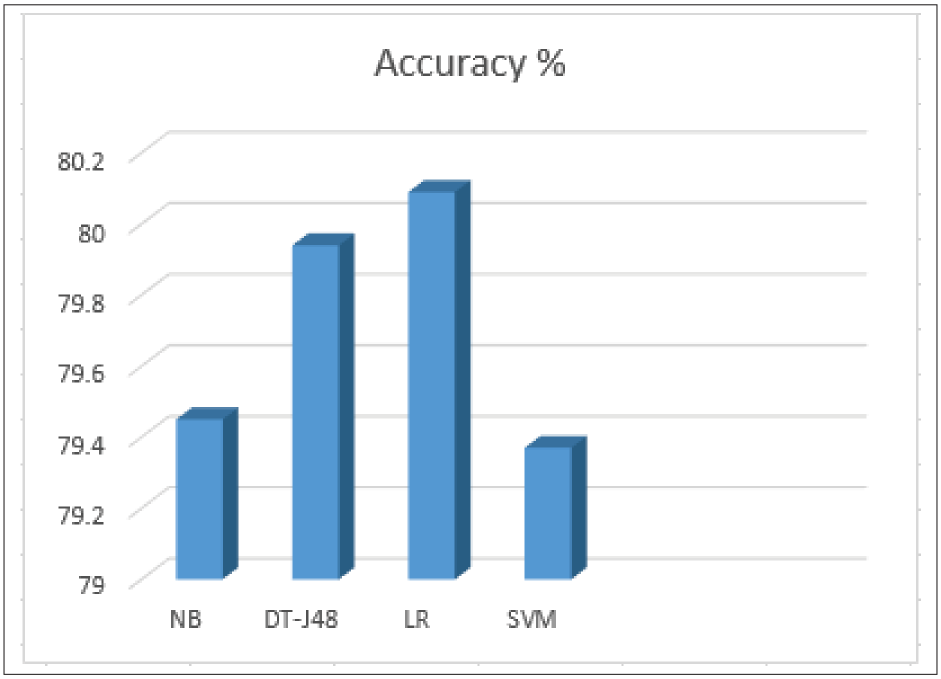


Figure 3.6 Comparison of accuracy of different classifiers on baby reviews dataset

Tableau 3.13 Classification Accuracy of different algorithms

Algorithms	Accuracy %
NB	79.45
DT-J48	79.94
LR	**80.09**
SVM	79.37

3.6 Discussion

Table 3.16 and Fig. 3.8 show the summary of experimental results. The experiments include four supervised machine learning algorithms, NB, DT-J48, LR, SVM algorithms to the Amazon product reviews datasets. This study could observe that well-trained supervised machine learning techniques were able to perform very useful classifications on reviews sentiment polarities (Negative, Neutral, Positive). In matters of accuracy, LR turned out to be the best algorithm

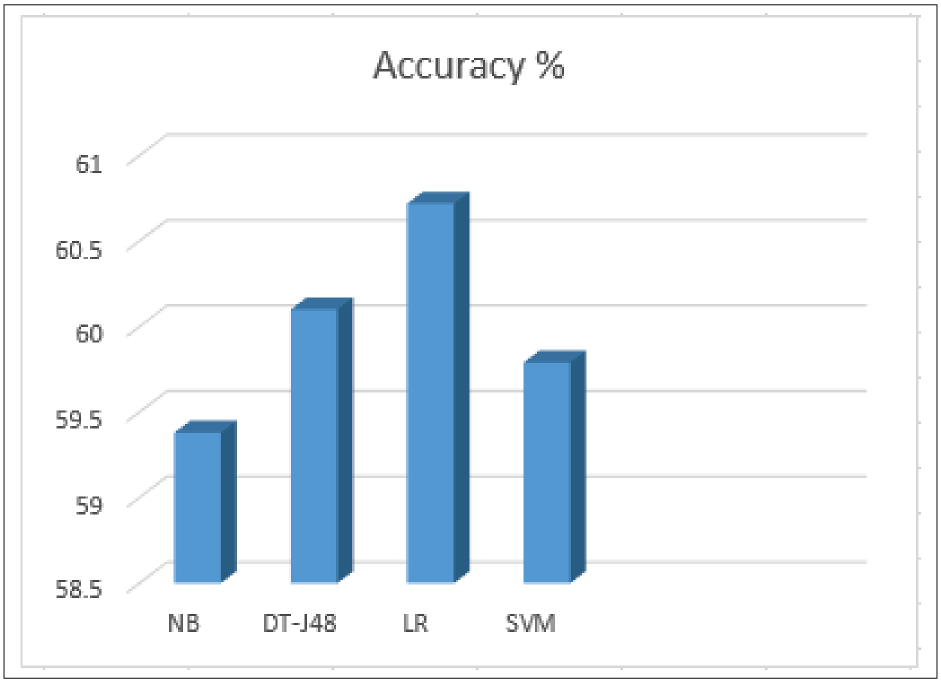


Figure 3.7 Comparison of accuracy of different classifiers on pet supplies dataset

Tableau 3.14 Comparison of accuracy, precision, recall and time taken to the build model (in seconds) of classifiers on baby reviews dataset

Evaluation metrics %				
Algorithms	Class	Precision	Recall	Time taken to the build model (seconds)
NB	neg	50.6	44.1	1.94
	neu	34.0	13.1	
	pos	65.1	82.9	
DT-J48	neg	52.1	45.1	18.89
	neu	33.4	6.6	
	pos	64.1	85.9	
LR	neg	54.2	42.4	12.34
	neu	37.3	6.3	
	pos	63.6	88.5	
SVM	neg	56.3	33.7	16085.65
	neu	40.7	3.8	
	pos	60.9	92.2	

for all tests, as it correctly classified 81.61% on Clothing, Shoes and Jewelry reviews dataset and 80.09% on Baby reviews dataset and 60.72% on Pet Supplies reviews dataset. Also, in our

Tableau 3.15 Classification Accuracy of different algorithms

Algorithms	Accuracy %
NB	59.38
DT-J48	60.10
LR	**60.72**
SVM	59.79

experimental results, we observed that the detection rate of unfair positive reviews is greater than the detection rate of unfair negative reviews and unfair neutral reviews.

Tableau 3.16 Performance evaluation rate and accuracy for unfair reviews detection

Experiments	Unfair negative reviews of LR%	Unfair neutral reviews of LR%	Unfair positive reviews of LR%	Accuracy of LR %
Results on clothing, shoes and jewelry reviews dataset	1.9	1.2	80.1	81.61
Results on Baby reviews dataset	2.2	0.8	80.6	80.09
Results on Pet Supplies	13.8	2.4	59.5	60.72

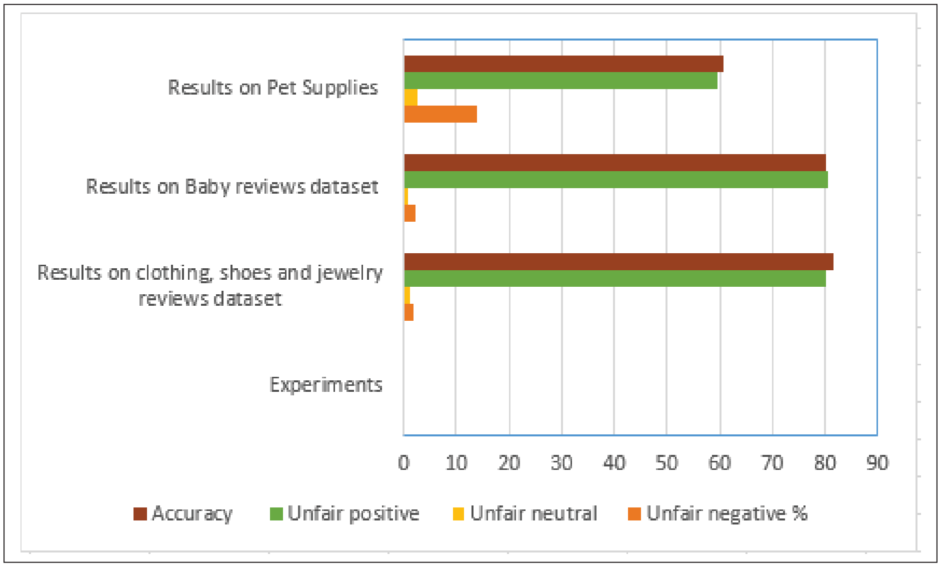


Figure 3.8 Summary of experimental results

In conclusion, from this analysis and through detecting of unfair positive reviews that the e-commerce domain is facing a problem of “all good reputation”, making it difficult for purchasers to select credible sellers.

3.7 Conclusions and Future Work

In this research, we proposed NB, DT-J48, LR and SVM algorithms to analyze Amazon reviews datasets. We also presented sentiment classification methods and we carried out our experiments using three different datasets of Amazon reviews with stopwords removal.

Our experimental approaches studied the accuracy, precision and recall of sentiment classification algorithms. Moreover, we were able to detect unfair negative reviews, unfair neutral reviews and unfair positive reviews using the detection processes of this method.

The main contributions of this study are summarized as follows :

- Firstly, this study compares different sentiment classification algorithms in Weka tool, which are used to classify Amazon reviews datasets into fair and unfair reviews ;
- Secondly, this study implements one feature selection method used for the SA classification task and tests with Stopwords Removal to find the best-supervised learning algorithm in terms of accuracy.

For future work, we wish to extend this work to use more recent snapshot Amazon reviews datasets as well as different feature selection methods. Additionally, we may use sentiment classification methods to detect unfair reviews and unfair ratings using different tools, such as Statistical Analysis System (SAS) or software machine learning library (scikit-learn) and then we would evaluate our work performance using these tools.

3.8 Acknowledgment

This study was supported and funded by the Libyan Ministry of Education and Canadian Bureau for International Education (CBIE).

CHAPITRE 4

ARTICLE 3 : BUILDING SENTIMENT ANALYSIS MODEL AND COMPUTE REPUTATION SCORES IN E-COMMERCE ENVIRONMENT USING MACHINE LEARNING TECHNIQUES

Elshrif Elmurngi [a], Abdelouahed Gherbi [b]

[a, b] Département de Génie logiciel et des technologies de l'information, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article publié par IGI Global « International Journal of Organizational and Collective Intelligence», janvier 2020
DOI : 10.4018/IJOCI.2020010103

4.1 Abstract

Online reputation systems are a novel and active part of e-commerce environments such as eBay, Amazon, etc. These corporations use reputation reporting systems for trust evaluation by measuring the overall feedback ratings given by buyers, which enables them to compute the reputation score of their products. Such evaluation and computation processes are closely related to sentiment analysis and opinion mining. These techniques incorporate new features into traditional tasks, like polarity detection for positive or negative reviews. The "all excellent reputation" problem is common in the e-commerce domain. Another problem is that sellers can write unfair reviews to endorse or reject any targeted product since a higher reputation leads to higher profits. Therefore, the purpose of the present work is to use a statistical technique for excluding unfair ratings and to illustrate its effectiveness through simulations. Also, the authors have calculated reputation scores from users' feedback based on a sentiment analysis model (SAM). Experimental results demonstrate the effectiveness of the approach.

Keywords : Reputation Systems, Sentiment Analysis Model (SAM), E-commerce (EC), Logistic Regression (LR).

4.2 Introduction

E-commerce has become one of the major way of shopping for products ranging from simple electronics to valuable items. In online shopping, the customers often depend on other customers' feedback posted through a rating system, before deciding on buying a product Mukherjee *et al.* (2012). Online feedback-based rating systems, also known as online reputation systems, are systems in which users provide ratings to items they bought. Based on the feedback of the product, the consumer decides whether to buy the product or not. This motivates the seller to promote or demote a product of their interest depending on their competitors' product, by posting rating scores which are unfair (Harmon (2004); Brown & Morgan (2006)). For example, fraudulent sellers may try to increase their income by submitting positive feedback which increases their product rating Harmon (2004). In addition, occasional sellers on eBay boost their reputations unfairly by selling or buying feedbacks Brown & Morgan (2006). Unfair rating scores are given singularly or collectively Swamynathan *et al.* (2010) where collective unfair ratings are referred to as collusion (Sun & Liu (2012); Swamynathan *et al.* (2010)) and are more complicated and much difficult to detect than the single unfair ratings Sun & Liu (2012). For that reason, in the present work, we are focusing on understanding and identifying unfair rating scores, all good reputation problems, and collusion and manipulation detection.

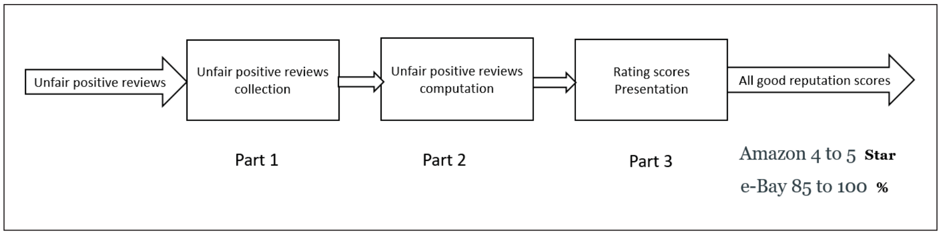

Unfair positive reviews
Unfair positive reviews collection
Unfair positive reviews computation
Rating scores Presentation
All good reputation scores
Part 1
Part 2
Part 3
Amazon 4 to 5 Star
e-Bay 85 to 100 %

Figure 4.1 Generic process of "all good reputation" problem

An online rating system needs truthful feedback in order to work properly. An important part of a rating system is creating honest and representative feedback. It should not only have qualitative, quantitative facts opinion-based process, but also should detect situations where some users may try to mislead the system by providing unfair positive and negative reviews, which is likely to lead to collusion and manipulation. Figure 4.1 Generic process of a reputation system and consists of three main components : feedback collection, computation and rating scores presentation (Sun & Liu (2012); Noorian & Ulieru (2010)). Feedback Collection is responsible for collecting feedback from community and feeding it to the rating system. A feedback is the opinion of an evaluator on the quality of an item or a person. Generally, a feedback can be expressed either as a number, such as a textual review, or a rating score Adler *et al.* (2011). A numeric feedback can be a negative or positive number chosen from either a discrete list of options, for example, an integer choice between 1 and 5 representing quality of an item. Feedback computation is responsible for computation of all feedbacks received from evaluators to calculate correct rating scores for people and items. Investigating the feedback computation part after using sentiment classification algorithmis the focus of this study.

According to Reyes-Menendez *et al.* (2019), it is improbable for all consumers to be satisfied with a product, very positive ratings might be interpreted by users as incredible information. Consequently, it is recommended that companies ensure that the feedback made to them on platforms such as TripAdvisor, contains only actual comments,as well as several less positive comments ; this can generate greater customer credibility and reliability of an offered product.

Figure 4.2 shows how unfair positive and negative reviews affects negatively on reputation scores computation. If the collection part collects unfair positive and negative reviews, the computation part will compute unfair positive and negative reviews and then the rating scores presentation part will present unfair rating scores as output.

Both the research community and the e-commerce industries has accepted unfair reviews to be a crucial challenge to the e-commerce industry Feng *et al.* (2012) and Breure (2013) ; Sussin & Thompson (2012). Any (positive or negative) review that is a unfair review and not an actual consumer's honest opinion will affect reputation scores negatively.

The main objective of this study is to offer a novel and comprehensive solution for designing a new model to obtain the most accurate reputation system, which addresses the existing issues, such as collusion and manipulation and the ”all good reputation” issue that is being currently

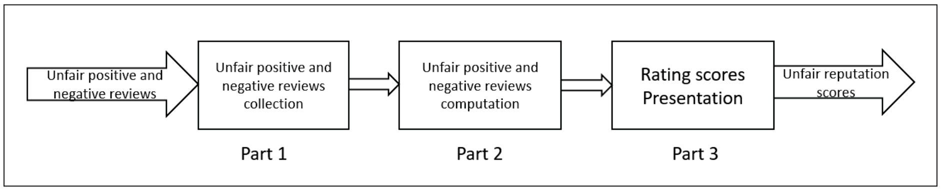

Unfair positive and negative reviews
Unfair positive and negative reviews collection
Unfair positive and negative reviews computation
Rating scores Presentation
Unfair reputation scores
Part 1
Part 2
Part 3

Figure 4.2 Impact unfair positive and negative reviews on the reputation system

encountered by reputation systems. While applied reputation models currently rely mainly on the overall ratings of items, they do not involve customer reviews in their assessment. Conversely, few of the reputation models focus only on the overall reviews of products without considering the ratings provided by the customers. This research aims to compute feedback rating based on feedback reviews. Subsequently, in order to get accurate reputation scores, we propose a simple calculation method that calculates feedback ratings and feedback reviews to obtain real feedback ratings and real feedback reviews, after detecting unfair feedback ratings and unfair feedback reviews, as opposed to , Amazon or eBay websites that calculate reputation scores from unfair feedback ratings and unfair feedback reviews. As mentioned above and based on the limitations of the existing methods employed, our main contributions to enhance Reputation systems are summarized as follows :

1. This study uses the scikit-learn machine in Python tool, an open source software for implementing machine learning algorithms Brunner & Kim (2016), to apply sentiment classification with the Logistic regression algorithm which classifies the Amazon reviews datasets into unfair and fair reviews and unfair and fair ratings.

2. The sentiment classification algorithm is applied with CountVectorizer Selection and TfidfVectorizer Selection, using two different Amazon reviews datasets. We observed that it is more effective to use the CountVectorizer Selection method than using TfidfVectorizer Selection and that it is more efficient to detect unfair reviews, ”all good reputation” issues, collusion, and manipulation.

3. We propose a statistical method to detect unfair reputation scores. Subsequently, we have designed and implemented a logistic regression algorithm to calculate new ratings from real feedback ratings and real feedback reviews in order to obtain fair reputation scores.

4. To evaluate the effectiveness of the proposed mechanism :

- Reducing the collusion and manipulation done by, both sides, customer and seller ;
- Establishing the confidence between customer and seller ;
- Provide the developer with the ability to improve current reputation systems and take into consideration all of the issues focused on in this study.

The remainder of this paper is organized as per the following : Section 4.3 shows Background and Related Work. Section 4.4 presents the applied methodology. Section 4.5 displays the results of the experiment and lastly, Section 4.6 presents our conclusion and future studies.

4.3 Background and Related Work

In this study, we demonstrate some background regarding Reputation System issues in E-commerce environment and discuss some related work to set the present study in the context of other studies.

4.3.1 Background

4.3.1.1 Definition of Reputation system

The Reputation, in general, is information used to make a value judgment about one thing within the context for a limited period Farmer & Glass (2010). As a first definition, a Reputation system can be considered as one of the established mechanisms to help customers in making decision in online shopping Gutowska & Sloane (2009).The second definition of a Reputation system is a process that collects, distributes and aggregates feedback about participants's behavior.

4.3.1.2 Benefits and Limitations of Reputation system

We have identified the benefits and limitations of reputation systems in general as the following :

Benefits

- Several websites currently provide a rating system for products, which allows customers to rate their online shopping experiences. The online ratings are aggregated and collected by Reputation systems to calculate all reputation for products, users, or services Resnick *et al.* (2000a) ;
- The insight gained from customer ratings about a product or service being good or bad can help improve customer satisfaction ;
- Provide the developer with the ability to improve current reputation systems and take into consideration all of the issues focused on in this study ;
- Creating opportunities to listen and involve customers to promote a particular brand ;
- Valuable insight can be gained about competitors by obtaining their customer's perception about their products and services ;
- The buyer can be aided by reputation systems to select the best seller for their transaction and avoiding getting cheated by the seller ;
- Marketing expenses can be reduced by knowing how the customers can be reached ;
- The services which require less time and money can be employed, thus reducing internal costs.

Limitations

- Manipulation of a reputation system cannot be detected if there are no robust technical mechanisms Jøsang (2012) ;
- It is difficult to stop collusion and potential attack due to the behavior of malicious identities Saini *et al.* (2014) ;

- On online social networks, the mitigating and detecting the manipulated Reputation effect is an important drawback of Reputation systems (RSs) Aggarwal (2016) ;
- The “all good reputation” problem is common in the e-commerce field, making it hard for buyers to choose credible sellers Jha *et al.* (2017) ;
- Elmurngi & Gherbi (2017b), Elmurngi & Gherbi (2018) and Barbado *et al.* (2019) presented sentiment classification techniques, and they have detected unfair reviews. However, they have not computed reputation scores, after detecting unfair negative reviews and unfair positive reviews. In our study, we have detected unfair positive reviews and unfair negative reviews and computed reputation scores and, after having computed reputation scores, this article have detected other issues such as “all good reputation” problems, collusion and manipulation issues.

4.3.1.3 Generic Architecture of a reputation system

According to Resnick *et al.* (2000a) a rating expresses an opinion as a result of a transaction through the feedback of the customer. Reputation systems, through monitoring, collect, combine and distribute this feedback. Figure 4.3 shows the main components of a reputation system and its actors : The collector gathers ratings from agents called raters. The goal of a rating is called ratee. This information is aggregated and processed by the processor. The algorithm used by the processor to calculate an aggregated representation of an agent’s reputation is the metric of the reputation system. The distributor makes the outcome available to other requesting agents.

4.3.1.4 Sentiment Analysis

Sentiment analysis (SA), also called opinion mining, is an approach to natural language Processing (NLP) that extracts subjective information behind a body of text Gamal *et al.* (2019). Text mining techniques consist of huge repository of unorganized data. To extract latent public opinion and sentiment through analysing this data is a challenging task. The aim of Sentiment Analysis is to study the reviews and evaluate the scores of sentiments. This analysis can be

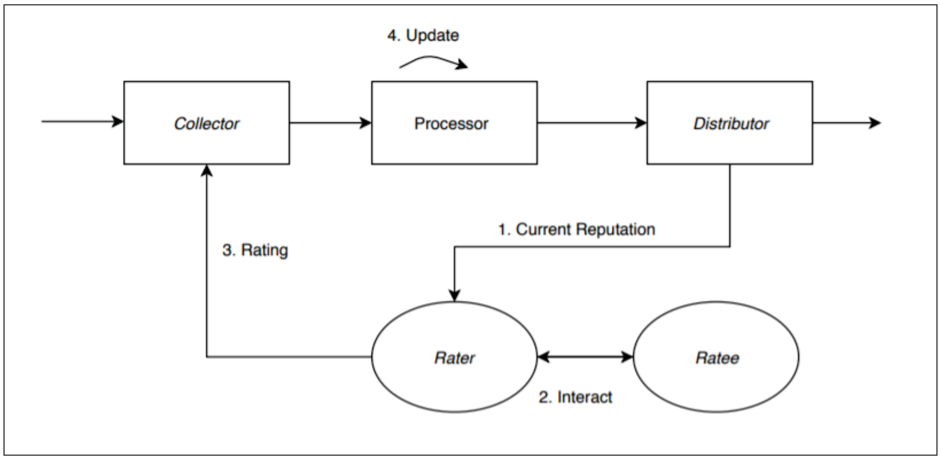


Figure 4.3 Architecture of a reputation system

divided into several levels : document level Moraes *et al.* (2013), sentence level Shoukry & Rafea (2012), sentence level Engonopoulos *et al.* (2011), word/term level or aspect level Zhou & Song (2015). The main aim of the analysis is to predict the sentiment inclination (i.e. positive, negative or neutral) by studying opinion words or sentiments and expressions in sentences and documents.

4.3.2 Related Work

The aim of this section is to set the present study in the context of other studies of reputation system vulnerabilities evaluation. This section employs a statistical method to vulnerability assessment; the related work emphasizes those studies that have applied statistical methods to this issue. Finally, this section compares these approaches outlining their weaknesses.

4.3.2.1 The Fundamental Problems and Available Solutions on Reputation Systems

There are three stages of operation on reputation systems, namely : first stage is feedback generation, second stage is feedback distribution, and third stage is feedback aggregation. Each stage of these components needs protecting against a variety of adversarial threats.

- A. Feedback Generation Stage

  Representative feedback is one of the most significant tasks in a reputation system. Actually, users will occasionally try to trick the system and we have identified some issues as follows :

  Unfair feedback reviews : This issue results from incorrectly presenting some users feedback reviews, which leads to create some errors in the system.

  Unfair Review Detection : Presently, Amazon website uses some machine learning algorithms to select relevant features and decide the final rating of a product. However, it does not apply any algorithm to detect whether a review is unfair or not. Few websites, like Yelp.com and Fakespot.com, can be used to detect unfair reviews online, but there is no particular algorithm to filter reviews Mane *et al.* (2017).

- B. Feedback Distribution Stage

  In Tavakolifard & Almeroth (2012), the important issues regarding the Feedback Distribution stage include “Reputation lag problem”, “Lack of portability between systems,” “Inability to filter or search,” “Categorization”. Once the reputation feedback is collected and processed accurately without any malicious effect, the next problem is how to get this feedback to the ones in need of it to make their decisions. Some of the issues : After the collection and processing of reputation information, two of the important issues mentioned by Resnick *et al.* (2000b) for feedback distribution are ”Lack of portability between systems” and “Categorization.” A lack of portability implies that there is not a widespread sharing of feedback between systems. When an e-commerce environment does allow importing for feedback, there is a bias towards only the importation of positive feedback. Categorization implies reputation could promote systems by providing better granularity. For instance, a user might have a good reputation in one area and a bad reputation in another area.

- C. Feedback Aggregation Stage

  There are some challenges in displaying and aggregating feedback so that it is truly useful in impacting future decisions about whom to trust. Some of the problems in this stage of the process include :

  Inaccurate equations : Some of E-commerce websites, such as eBay, use a simple reputation schema that could be misleading for users. For instance, an equal reputation score would be assigned to two users on eBay, one of them with ten negative ratings and 100 positive ratings ; while the other with no negative ratings and 90 positive ratings. This is likely to lead to raise the vulnerability issue caused by “increased trust by increased volume” Tavakolifard & Almeroth (2012).

The publishing of false rumors : This issue occurs when the reputation of the feedback providers is not considered. According to Hoffman *et al.* (2009), one approach to this issue is to employ statistical methods to build Bayesian framework and robust formulations as an example that can be reasoned about in a precise method.

4.3.2.2 Reputation System for E-Commerce Applications

In online marketplaces like eBay, reputations of E-Commerce Marketplaces now act as an essential role in the decision to start a transaction and in the pricing of goods or services. A user’s reputation as the sum of the lifetime ratings are computed by eBay’s Feedback. These lifetime ratings then create reputation profiles which are tailored to forecast future performance and to aid users. Resnick & Zeckhauser (2002) The sellers that have excellent reputations can request for higher prices of their products, whereas poor reputation holders can only interest a fewer buyers. Resnick & Zeckhauser (2002) shows that a seller’s reputation, given by review scores in online marketplace, does not influence the listing price or possibility of consumer purchases. They proved that changing a review score or reputation score has an effect on the buyer’s decision, but the task of the visual cues is still substantial for tourism industry.

4.3.2.3 Textual Reviews to Provide Detailed Opinion about the Product

Current reputation models mostly rely on numerical data from different fields, such as ratings in e-commerce. These reputation models only consider the overall ratings of the products without taking into account reviews given by customers Xu *et al.* (2016). On the other hand, many online websites admit consumers to give textual reviews of their opinion About the product (Tian *et al.* (2014a,b)).Thus, some of the reputation models only use the textual reviews of products, not taking into account ratings which were provided by customers. Consumers can read these reviews, and now an increasing number of users rely on these reviews more than the ratings. Reputation models could use sentiment analysis methods to obtain users opinion which can be used by reputation systems having included consumers opinion on different features (Abdel-Hafez & Xu (2013); Abdel-Hafez *et al.* (2012)).

4.3.2.4 Sentiment Analysis Based on User Behavior

Many researchers are working on opinion mining and sentiment analysis on textual information gathered from social network platforms. These opinions and sentiments are being used to improve business productivity Iqbal *et al.* (2015).

To understand the behaviors of people are difficult and complex. In order to understand people's behaviors, it takes many resources to collect and analyze a large amount of information such as posts, comments, clicking likes, and sharing of thoughts. However, the difficulty is to get real business and customer data, since it is challenging to get confidential data Chang (2018). A Social Network Analysis Platform is designed to understand the strength of the relationship between social networks and sentiment analysis. Karyotis *et al.* (2018) aim to show the usefulness of implementing the proposed fuzzy emotion representation model and to demonstrate its effectiveness and applicability in big data settings. However, the authors approach does not use sentiment analysis and opinion mining for detecting unfair feedback ratings and unfair feedback reviews from a Social Network. This study proposes a new method to build a Sentiment Analysis Model and compute Reputation scores in an e-commerce environment using machine learning techniques.

4.3.2.5 Importance of Logistic Regression (LR) on Sentiment Classification Techniques

The researchers Elmurngi & Gherbi (2017b) used Sentiment classification techniques against a dataset of consumer reviews. The experiments were carried out using classification algorithms : Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) for sentiment classification using three datasets of reviews. The experiments' results show that the Logistic Regression (LR) algorithm achieves better performance and is the best classifier with the highest accuracy as compared to the other three classifiers, not merely in text classification, but in unfair reviews detection as well. The researchers Gamal *et al.* (2019) and Lin *et al.* (2015) presented an empirical study on Sentiment Classification and Logistic Regression which is constructed to combine different machine learning methods and get an outstanding performance in precision and recall.

4.3.2.6 The Impact of Feature Selection on Classification Accuracy

Techniques for feature selection are very beneficial for text classification in general and specifically in sentiment analysis. Fattah (2017), Guyon & Elisseeff (2003). These techniques rank features by given less weightage to non-informative features so that these features can be removed while valuable features are given more weightage to be kept for better classification accuracy and efficiency based on a specific measure. In this work, we study two feature selection techniques with the Logistic Regression, including CountVectorizer selection and TfidfVectorizer selection.

4.4 Methodology and the proposed approach

Our methodology was organized in the next six steps, as shown in Figure 4.4, steps that involve the supervised sentiment classification approaches using the scikit-learn in python tool for text classification, as described below.

4.4.1 Amazon Reviews Collection

Datasets of Amazon are used by many researchers such as Catherine & Cohen (2017), Elmurngi & Gherbi (2018), Ling *et al.* (2014), Tan *et al.* (2016). The datasets are available and have been collected and released by McAuley & Leskovec (2013). We have based our experiment on analyzing the standard datasets of Amazon reviews to sentiment value using Logistic regression algorithm and classification methods. Datasets of Amazon have many different kinds of products, however here we focus on two datasets : The Baby reviews dataset, and The Sports and Outdoors dataset, with raw data size of 30.5 MB and 65.1 MB, respectively. According to Chen *et al.* (2015) the category list provides all reviews of top customers. The kinds containing most reviews are “Sports and Outdoors” (296,337 reviews), “Baby” (160,792 reviews), while consumers’ reviews in “Digital Music” (64,706 reviews), “Musical Instruments” (10,261 reviews), and “Amazon Instant Video” (37,126 reviews) have the least reviews. Each product review of dataset is provided with the following labels : ReviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime.

4.4.2 Data Cleaning and Preprocessing

The Python programming language with Scikit-learn is a free software machine learning library that was used for the cleaning and preprocessing process. This language was chosen because of the ability to deal with several languages, and the available libraries and packages for English text. Data cleaning and preprocessing was done in some steps in order to get the dataset cleaned and ready for learning by the classification algorithm.

Figure 3 shows all the steps and their order for the cleaning process. The datasets used in our experiment are obtained from Amazon product and was divided into five scales rating : 1 star, 2 stars, 3 stars, 4 stars and 5 stars. The original datasets are not cleaned and not easy to model for classification. We have separated the datasets before cleaning and applied the sentiment classification classifiers after cleaning and after using Logistic regression algorithm.

Pre-processing of data is an important step in the text mining process and plays a significant part in a number of supervised learning techniques, Figure 4.5 and Table 4.1 show all the steps for the Pre-processing of data process using SAM. Pre-processing of datasets are majorly done in three steps, as per the following :

Step One : Tokenization

In this process, after the data is retrieved from the datasets, we tokenize the sentences into words, so that it is easily to understand and count.

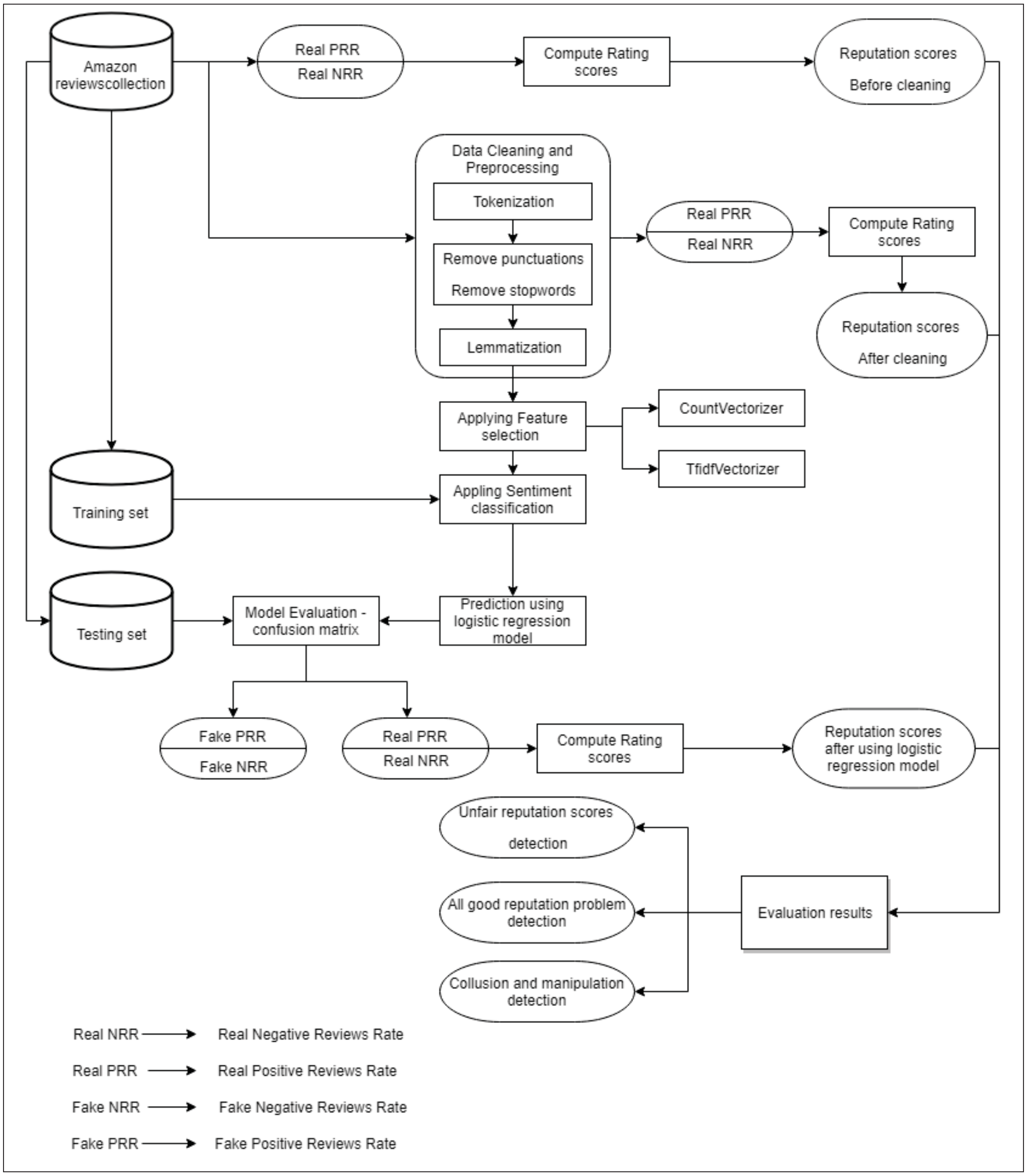

Amazon reviewscollection
Real PRR
Real NRR
Compute Rating scores
Reputation scores Before cleaning
Data Cleaning and Preprocessing
Tokenization
Remove punctuations
Remove stopwords
Lemmatization
Real PRR
Real NRR
Compute Rating scores
Reputation scores After cleaning
Applying Feature selection
CountVectorizer
TfidfVectorizer
Training set
Appling Sentiment classification
Testing set
Model Evaluation - confusion matrix
Prediction using logistic regression model
Fake PRR
Fake NRR
Real PRR
Real NRR
Compute Rating scores
Reputation scores after using logistic regression model
Unfair reputation scores detection
All good reputation problem detection
Collusion and manipulation detection
Evaluation results
Real NRR → Real Negative Reviews Rate
Real PRR → Real Positive Reviews Rate
Fake NRR → Fake Negative Reviews Rate
Fake PRR → Fake Positive Reviews Rate

Figure 4.4 Research methodology

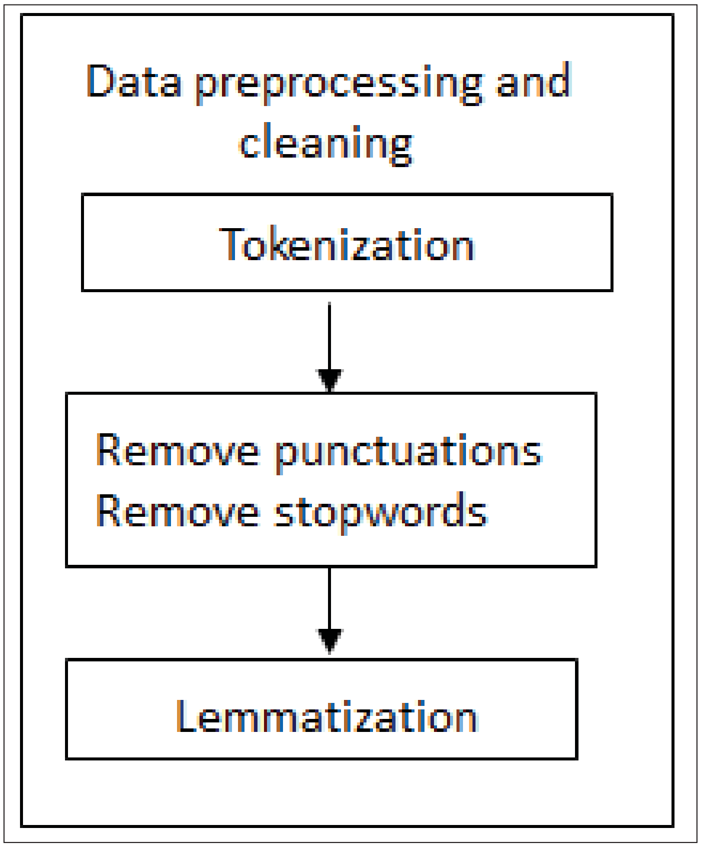


Figure 4.5 Data Preprocessing Steps using SAM

Step Two : Punctuation and Stopwords Removal

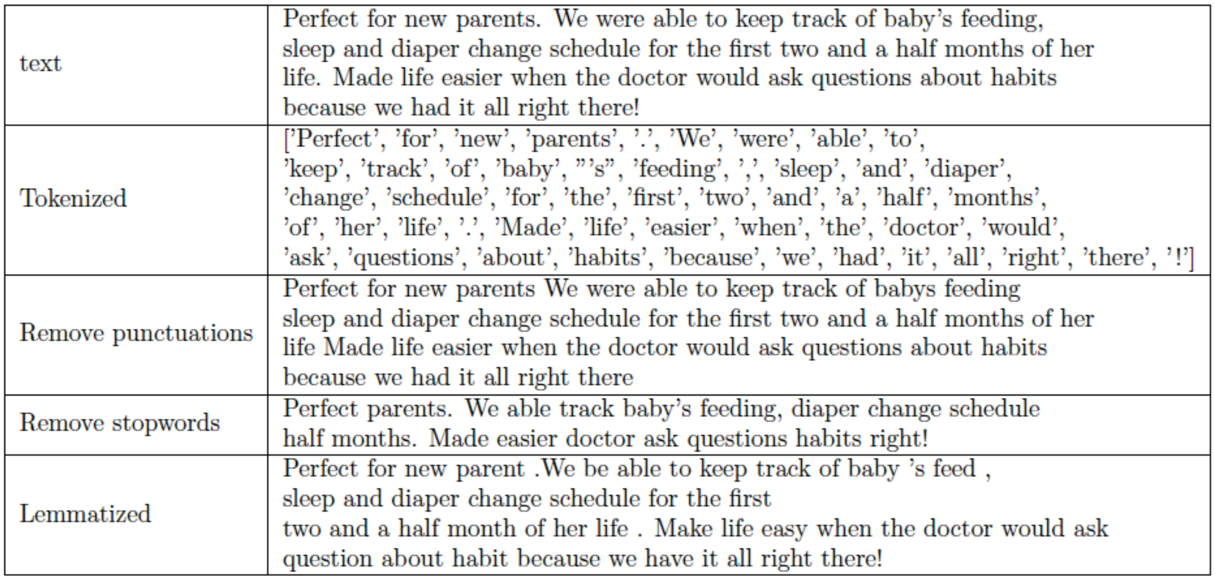
Punctuation Removal

Punctuation is a string containing numbers, whitespace, and letters, including periods, semicolons, and commas. Basically, we believe that removing punctuation from a string is the best way in Python or any other tool in machine learning techniques.

Stopwords Removal

Stopwords are the English words, which does not add much meaning to a sentence, and must be filtered out. For example, the words such as “the,” “he,” “a,” “of,” “you,” “and,” “have,” etc.

Tableau 4.1 Examples showing data preprocessing steps of text summarization



text	Perfect for new parents. We were able to keep track of baby's feeding, sleep and diaper change schedule for the first two and a half months of her life. Made life easier when the doctor would ask questions about habits because we had it all right there!
Tokenized	['Perfect', 'for', 'new', 'parents', '.', 'We', 'were', 'able', 'to', 'keep', 'track', 'of', 'baby', "'s", 'feeding', ',', 'sleep', 'and', 'diaper', 'change', 'schedule', 'for', 'the', 'first', 'two', 'and', 'a', 'half', 'months', 'of', 'her', 'life', '.', 'Made', 'life', 'easier', 'when', 'the', 'doctor', 'would', 'ask', 'questions', 'about', 'habits', 'because', 'we', 'had', 'it', 'all', 'right', 'there', '!']
Remove punctuations	Perfect for new parents We were able to keep track of babys feeding sleep and diaper change schedule for the first two and a half months of her life Made life easier when the doctor would ask questions about habits because we had it all right there
Remove stopwords	Perfect parents. We able track baby's feeding, diaper change schedule half months. Made easier doctor ask questions habits right!
Lemmatized	Perfect for new parent .We be able to keep track of baby 's feed , sleep and diaper change schedule for the first two and a half month of her life . Make life easy when the doctor would ask question about habit because we have it all right there!

Step Three : Lemmatization :

Lemmatization is called word normalization techniques in the field of Natural Language Processing (NLP) that are used to prepare text, words, and documents for further processing. The main function of lemmatization is the process of converting the words into their root words. In our study, we have used lemmatization, not steaming, because throughout our implementation we have compared between both, applying some words, such as feeding and flying : the conversion of feeding using steaming is feed, and the conversion of feeding using lemmatization is also feed. However, the conversion of flying using steaming is fli, and the conversion of flying using lemmatization is fly. Some special words using steaming are not given meaning and it will have an effect on sentiment classification and that is the reason why we are using lemmatization and not steaming.

4.4.3 Feature Selection

In data pre-processing, in order to gain efficient data reduction, feature selection (FS) methods in sentiment analysis can be employed. This helps to find precise data models and in finding the important attributes Koncz & Paralic (2011). In the recent paper, the list of features applied to classify sentiments include N-gram features, this feature has been the baseline in most related research (Agarwal & Mittal (2016); Dashtipour *et al.* (2016)). In this study, bigrams are applied as feature sets , and these features are consisting of every two consecutive words and capable of incorporating some contextual information.

Our research has implemented two-feature selection methods as show bellow :

CountVectorizer :

The CountVectorizer gives an easy method for tokenizing a compilation of text and for building terminology of known words.

TfidVectorizer :

The TfidfVectorizer will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents.

In our study, the Logistic Regression algorithm with CountVectorizer feature selection achieved better performance than TfidfVectorizer.

4.4.4 Sentiment Classification Method

To apply sentiment classification, several supervised methods have been applied to these systems using supervised methods based on manually labelled samples Pang *et al.* (2008). The sentiment in sentiment classification method is determined by training on a known dataset, and classifying feedback reviews as positive or negative. Actually, logistic regression is a robust algorithm fortwo-class classification. In our study, we used Logistic Regression algorithm with two-feature selection methods. A really fast and well-known classification algorithm is Logistic Regression (LR), also known as the logistic function, and is used to assign observations to a discrete set of classes. In our work, we employ this algorithm with CountVectorizer as feature selection and found that it is the most suitable and precise method.

4.4.5 Detection Processes

The original datasets, which we used in our study, are not labelled as positive and negative reviews. However, we prepared the datasets and we labeled them as positive and negative reviews based on ratings scores provided by users, as shown in Tables **??** and Table **??**. In our study, we built a sentiment classifier to identify whether the review has a positive or negative sentiment. The Logistic Classifier model used the CountVectorizer Selection and TfidfVectorizer Selection from the training data to develop a model predicting True positive reviews, False positive reviews, True negative reviews, and True negative reviews. We noted the numbers of positive and negative reviews from the original datasets, as shown in Tables. **??** and Table **??**, which are not the same numbers of positive and negative reviews found when we used the sentiment classification, as shown in Tables 7a and Table 7b.

In order to evaluate the performance of our classification model and test our results, there is a very common method called confusion matrix, which is shown in Table 4.2. The confusion matrix displays the methods in which the classification model is confused when making predictions Hinton *et al.* (2015). Reviews are classified based on the generated confusion matrix to positive and negative. In our classification of reviews, unfair is made up of the set of reviews considered to be False, which in this case involves both False positive reviews and False Negative reviews. On the other hand, Real is defined to combine the set of reviews to be considered as True, which in this case involves both True positive reviews and True Negative reviews. The fair reviews and unfair reviews are determined according to equations 1 to 4 as shown in Table 4.3.

- True Positive Reviews (TPR) : when the actual class of the positives reviews point was 1 (True) and the predicted is also 1 (True);
- True Negative Reviews (TNR) : when the actual class of the negatives reviews point was 0 (False) and the predicted is also 0 (False);
- False Positive Reviews (FPR) : when the actual class of the positives reviews point was 0 (False) and the predicted is 1 (True);

- False Negative Reviews (FNR) : when the actual class of the negatives reviews point was 1 (True) and the predicted is 0 (False).

Tableau 4.2 The confusion matrix

	Predicted actual reviews Fair	Predicted actual reviews Unfair
Actual reviews Fair	True Negative Reviews (TNR)	False Positive Reviews (FPR)
Actual reviews Unfair	False Negative Reviews (FNR)	True Positive Reviews (TPR)

Evaluation measures from the confusion matrix, (1)-(7) as shown in Table 3 display numerical parameters that apply below the mentioned measures to assess the Detection Process performance. In Table 4.2 the confusion matrix shows the Predicted actual reviews and Predicted unactual reviews forecasting found through known data, and for each algorithm used in this study are different confusion matrix and performance evaluation.

For each feature selection with Logistic Regression algorithm used in our study different Performance evaluation and confusion matrix.

Tableau 4.3 Evaluation measures from the confusion matrix

Measure	Formula
Unfair PRR	FP / (TN + FP)
Unfair NRR	FN / (TP + FN)
Fair PRR	TP / (TP + FN)
Fair NRR	TN / (TN + FP)
ACC	TP + TN/ (TP + TN + FN + FP)
PREC	TP / (TP + FP)
REC	TP/(TP+FN)
Unfair PRR : Unfair Positive Reviews Rate ; Unfair NRR : Unfair Negative Reviews Rate ; Real PRR : Real Positive Reviews Rate ; Fair NRR : Fair Negative Reviews Rate ; ACC : Accuracy ; PREC : Precision ; REC : Recall ; TPR : True Positive Reviews ; TNR : True Negative Reviews ; FPR : False Positives Review ; FNR : False Negatives Review	

4.4.6 Calculation Processes

Feedback scores, stars and percentages on eBay or amazon are the original and best-known marketplace reputation system. The system has become more developed in recent years. In Figure 4.6a and Figure 4.6b show an example of eBay overall rating and an example of Amazon overall rating, respectively. Feedback is generally reciprocal ; users almost always give positive feedback. The techniques for calculating rater's credibility in most of the existing models are not sufficient either. The authors in Malik & Bouguettaya (2009) propose an extremely complicated method to calculate rater's credibility. In our study, we have used a simple calculation method in order to compute reputation scores from real feedback reviews, after detecting unfair reviews.

4.4.7 Reputation Scores Calculation

On eBay

On eBay, the reputation score is represented by the Positive Feedback Percentage (PFP), which is computed through the transaction which ended within the last year based on the total number of positive and negative Feedback ratings using this formula :

$$PFP = \frac{Positive}{Positive\ negative} \tag{4.1}$$

We have transferred positive Feedback percentage to positive Feedback star in order to calculate the reputation scores calculation on amazon using this formula :

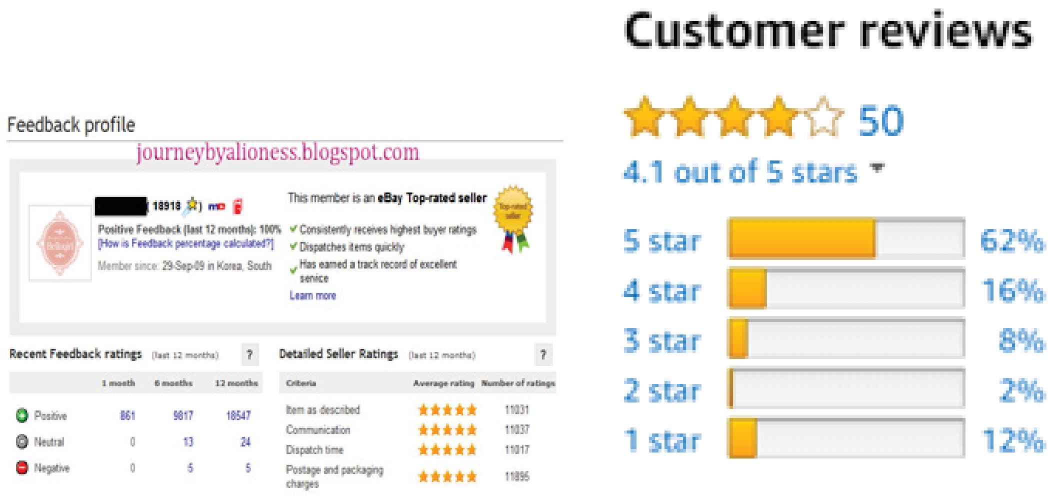
$$Reputation\ scores = (PFP * 5) \setminus 100 \tag{4.2}$$

Example :

Positive : 850   Negative : 10

PFP = 850 \ (850 10) = 98.8

$Reputation\ Scores$ = (98.8 ∗ 5) \ 100 = 4.9 $Scores$



Feedback profile

journeybyalioness.blogspot.com

[redacted] 18918 ☆)

Positive Feedback (last 12 months): 100%
[How is Feedback percentage calculated?]
Member since: 29-Sep-09 in Korea, South

This member is an **eBay Top-rated seller**
✓ Consistently receives highest buyer ratings
✓ Dispatches items quickly
✓ Has earned a track record of excellent service
Learn more

Recent Feedback ratings (last 12 months) ?

	1 month	6 months	12 months
● Positive	861	9817	18547
● Neutral	0	13	24
● Negative	0	5	5

Detailed Seller Ratings (last 12 months) ?

Criteria	Average rating	Number of ratings
Item as described	★★★★★	11031
Communication	★★★★★	11037
Dispatch time	★★★★★	11017
Postage and packaging charges	★★★★★	11896

Customer reviews

★★★★☆ 50

4.1 out of 5 stars

5 star	62%
4 star	16%
3 star	8%
2 star	2%
1 star	12%

a) Example of overall rating from Amazon.com   b) Example of overall rating from ebay.com

Figure 4.6 Examples of overall ratings

On Amazon

Amazon calculates a product’s star ratings using a machine learned model instead of a raw data average. However, we did not find the formula that Amazon uses to calculate the Reputation scores. In our study, we have used the same formula which eBay used to calculate the positive Feedback percentage and then we have transferred the positive Feedback percentage to the positive Feedback star in order to create the reputation scores calculation on Amazon, as shown in equation 9.

Example 1 :

$reviews1star \quad reviews2star = Negative\ reviews$

$reviews4star \quad reviews5star = Positive\ reviews$

Negative reviews = 70

Positive reviews = 285

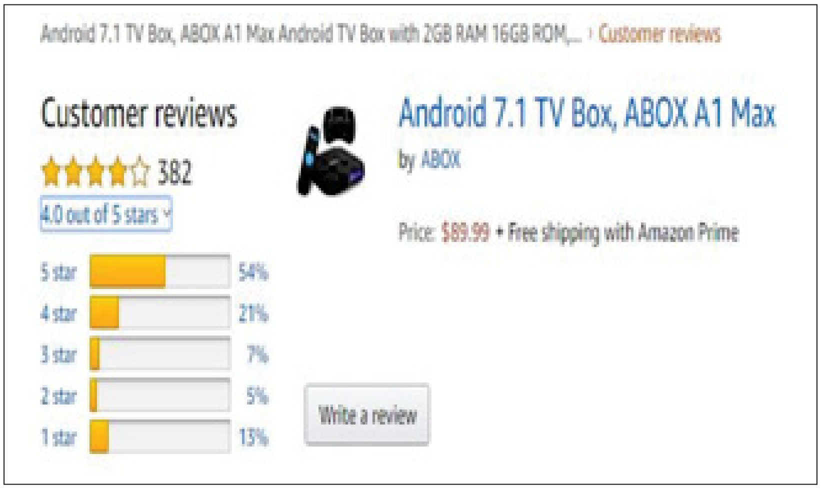
$PFP = Positives \setminus (Positives \ \ negatives)$

$PFP = 285 \setminus (285 \ \ 70) = 80.28$

$Reputation\ scores = (PFP * 5) \setminus 100$

$Reputation\ scores = (80.28 * 5) \setminus 100 = 4.0\ Scores$

Figure 4.7 displays a webpage screenshot of customer reviews, showing 1-10 of 206 reviews (5 star), Showing 1-10 of 79 reviews (4 star), Showing 1-10 of 27 reviews (3 star), Showing 1-10 of 18 reviews (2 star), and Showing 1-10 of 52 reviews (1 star).



Android 7.1 TV Box, ABOX A1 Max Android TV Box with 2GB RAM 16GB ROM,... › Customer reviews

Customer reviews

★★★★☆ 382

4.0 out of 5 stars

5 star	54%
4 star	21%
3 star	7%
2 star	5%
1 star	13%

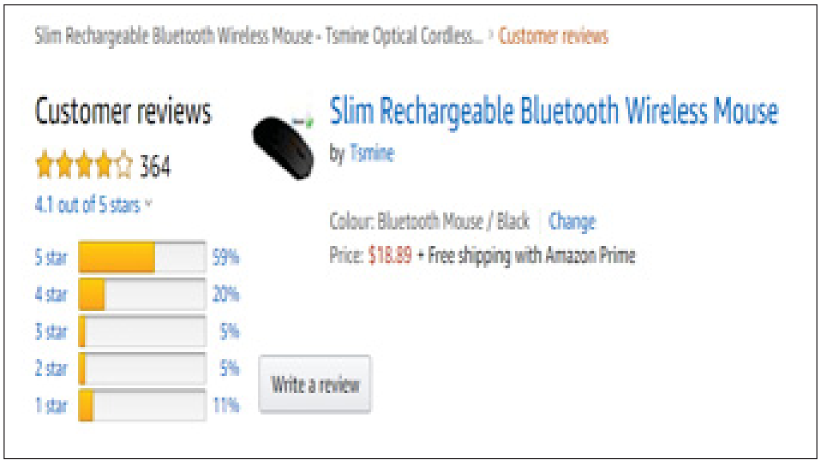
Android 7.1 TV Box, ABOX A1 Max

by ABOX

Price: $89.99 + Free shipping with Amazon Prime

Figure 4.7 A review example of Android 7.1 TV Box, ABOX A1 Max from Amazon.com

Example 2 :



Slim Rechargeable Bluetooth Wireless Mouse - Tsmine Optical Cordless... › Customer reviews

Customer reviews

★★★★☆ 364

4.1 out of 5 stars

5 star	59%
4 star	20%
3 star	5%
2 star	5%
1 star	11%

Slim Rechargeable Bluetooth Wireless Mouse

by Tsmine

Colour: Bluetooth Mouse / Black | Change

Price: $18.89 + Free shipping with Amazon Prime

Figure 4.8 A review example of Slim Rechargeable Bluetooth Wireless Mouse from Amazon.com

Figure 4.8 displays a webpage screenshot of customer reviews, showing 1-10 of 216 reviews (5 star), showing 1-10 of 72 reviews (4 star), showing 1-10 of 18 reviews (3 star), showing 1-10 of 17 reviews (2 star), and showing 1-10 of 41 reviews (1 star).

4.4.8 Evaluation Results

In this step, we have compared the different Reputation scores before cleaning, after cleaning, and after using Logistic Regression algorithm with diverse datasets from amazon.com, which are Baby reviews dataset, and Sports and Outdoors dataset. Accuracy and time required for execution by the Logistic Regression technique is observed with two different feature selections. The expected result is to obtain the detection of unfair reputation scores detection, all good reputation issue, and collusion and manipulation.

4.5 Experiments and Result Analysis

In this section, we present experimental results based on a logistic regression technique borrowed by machine learning with two different feature selections to classifying sentiment on two different real datasets, which are Baby reviews dataset, and Sports and Outdoors dataset. In addition, we have used the same method at the same time to detect unfair reviews, and “all excellent reputation” problem and collusion and manipulation using the scikit-learn, which is the free software machine-learning library of the Python programming language.

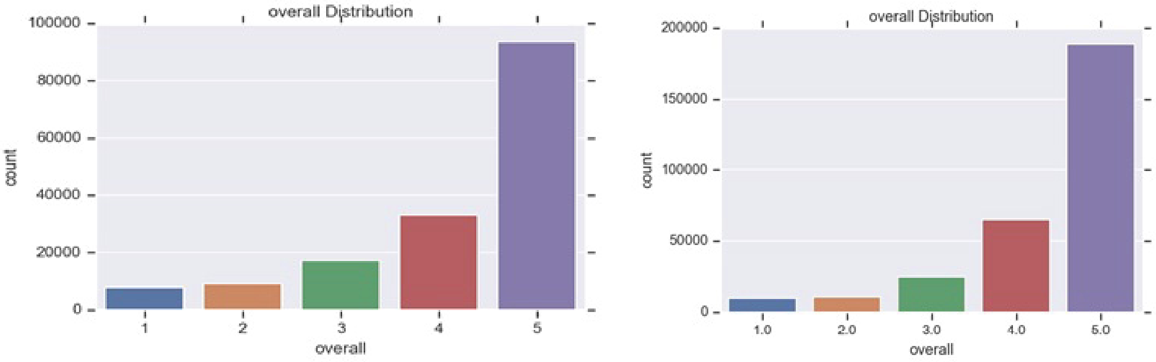
The datasets that were used in our experiments come in json file and we have converted them to csv files before preparing the datasets for learning.

4.5.1 Basic Statistics of All Reviews Datasets and Overall Distribution of Ratings

Amazon’s product reviews and ratings are a very important business. Customers on Amazon often make purchasing decisions based on those reviews, and a single bad review can cause a potential purchaser to reconsider. The primary difference between the two distributions is that there is a significantly higher proportion of Amazon customers giving only 5-star reviews. On dataset of Baby reviews, we first analyzed all reviews dataset and overall distribution of ratings. Table 4.4 and Figure 4.9a show that 58% of the reviews have an overall rating of 5 Star, 20% of the reviews have an overall rating of 4 Star, 11% of the reviews have an overall rating of 3 Star, 6% of the reviews have an overall rating of 2 Star, and 5% of the reviews have an overall rating of 1 Star. On the dataset of Sports and Outdoors reviews, we first analyzed all reviews dataset and overall distribution of ratings. 4.5 and Figure 4.9b show that 64% of the reviews have an overall rating of 5 Star, 22% of the reviews have an overall rating of 4 Star, 8% of the reviews have an overall rating of 3 Star, 3% of the reviews have an overall rating of 2 Star, and 3% of the reviews have an overall rating of 1 Star.

The distribution of ratings among the reviews show that most of the reviewers have given 5-star and 4-star ratings with relatively very few giving 1-star and 2-star ratings.

We prepared datasets on Baby reviews and Sports and Outdoors reviews and we Labeled them as Positive and Negative. As shown in Table.4.6 , Label of ratings and Reviews as Positive and Negative on dataset of Baby reviews and Table 4.7, Label of ratings and Reviews as Positive and Negative on dataset of Sports and Outdoors reviews.



a) Distribution of ratings on on Baby reviews dataset

b) Distribution of ratings on Sports and Outdoors reviews dataset

Figure 4.9 Distribution of ratings on datasets

Tableau 4.4 Number of reviews and ratings on Baby reviews dataset

Actual (Star)	No. of Reviews
5.0 Star	93526
4.0 Star	32999
3.0 Star	17255
2.0 Star	9193
1.0 Star	7819

Tableau 4.5 Number of reviews and ratings on Sports and Outdoors reviews datase

Actual (Star)	No. of Reviews
5.0 Star	188208
4.0 Star	64809
3.0 Star	24071
2.0 Star	10204
1.0 Star	9045

Tableau 4.6 Label of ratings and Reviews as Positive and Negative on Baby reviews dataset

Actual (Star)	No. of Reviews	Label of ratings
5.0 Star	126525	Positive
4.0 Star		
3.0 Star	———	———
2.0 Star	17012	Negative
1.0 Star		

Tableau 4.7 Label of ratings and Reviews as Positive and Negative on Sports and Outdoors reviews dataset

Actual (Star)	No. of Reviews	Label of ratings
5.0 Star	253017	Positive
4.0 Star		
3.0 Star	———	———
2.0 Star	19249	Negative
1.0 Star		

4.5.2 Reputation Scores before Cleaning

With the polarity results, Table 4.8 and Table 4.9 provide the basic of Sentiment and Number of reviews, which gives a high-level insight into what percentage and visualization of Sentiment and number of reviews are positive or negative, shown in Figure 10a and Figure 10b.

We have calculated Reputation scores before cleaning as the following :

1- On dataset of Baby reviews

$Reputation\ scores = (PFP * 5scores) \backslash 100$

$PFP = Positives \backslash (Positives\ \ negatives)$

$PFP = 126525 \backslash (126525\ \ 17012) = 88.14$

$Reputation\ scores = (88.14 * 5) = 4.4\ Scores$

2- On dataset of Sports and Outdoors reviews

$Reputation\ scores = (PFP * 5scores) \setminus 100$

$PFP = Positives \setminus (Positives\ \ negatives)$

$PFP = 253017 \setminus (253017\ \ 19249) = 92.93$

$Reputationscores = (92.93 * 5) = 4.6\ Scores$

Tableau 4.8 Sentiment and Number of reviews on Baby reviews dataset

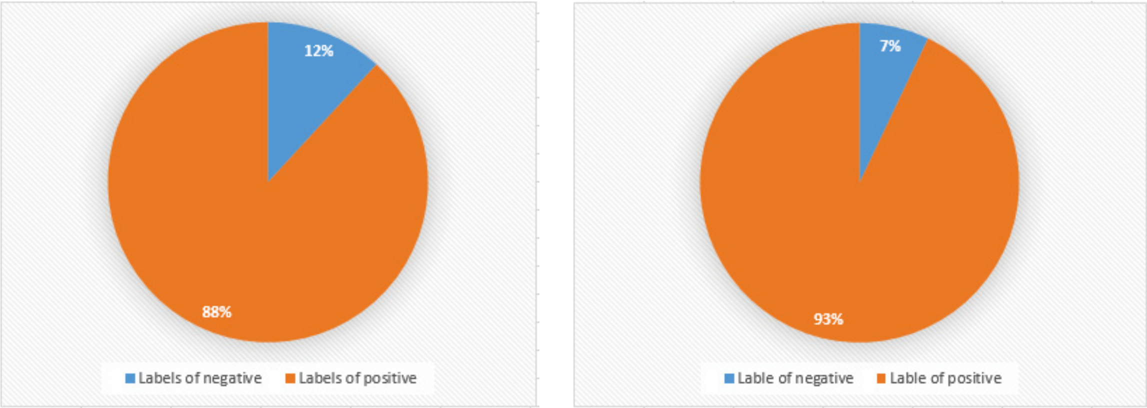
Sentiment	No. of Reviews
NEGATIVE	17012
POSITIVE	126525

Tableau 4.9 Sentiment and Number of reviews on Sports and Outdoors reviews dataset

Sentiment	No. of Reviews
NEGATIVE	17012
POSITIVE	126525

4.5.3 Reputation Scores after Cleaning

After we cleaned and preprocessed the datasets of products reviews and used sentiment analysis, we obtained Real Positive Reviews and Real Negative Review as show in Table 4.10 and Figure 4.11a on dataset of Baby reviews and Table 4.11 and Figure 4.11b on dataset of Sports and Outdoors reviews.



a) percentage of labels and number of reviews on Baby reviews dataset

b) percentage of labels and number of reviews on Sports and Outdoors reviews dataset

Figure 4.10 percentage of Labels and Number of reviews

We have calculated Reputation scores after cleaning as the following :

1- On dataset of Baby reviews

$Reputation\ scores = (PFP * 5scores) \setminus 100$

$PFP = Positives \setminus (Positives\ \ negatives)$

$PFP = 125348 \setminus (125348\ \ 34079) = 78.62$

$Reputationscores = (78.62 * 5) = 3.9\ Scores$

2- On dataset of Sports and Outdoors reviews

$Reputationscores = (PFP * 5scores) \setminus 100$

$PFP = Positives \setminus (Positives\ \ negatives)$

$PFP = 253017 \setminus (253017\ \ 43320) = 85.38$

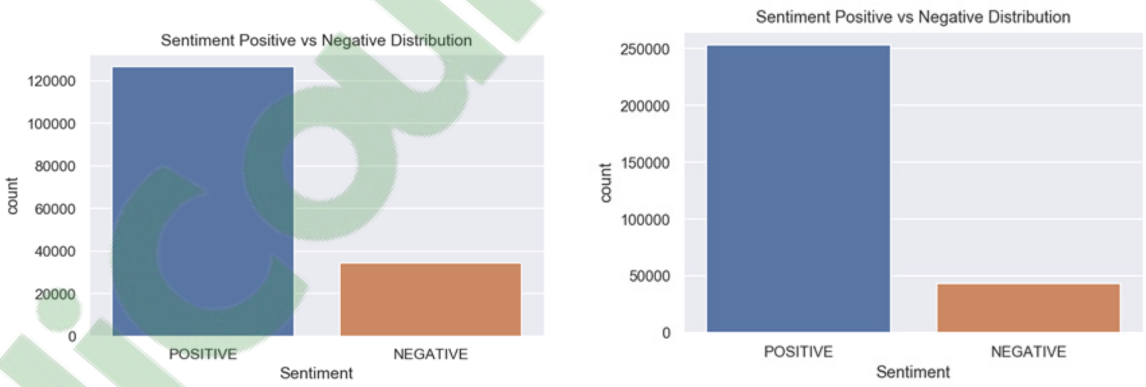
$Reputationscores = (85.38 * 5) = 4.2\ Scores$

Tableau 4.10 Sentiment and Number of reviews on Baby reviews dataset

Sentiment	No. of Reviews
NEGATIVE	34079
POSITIVE	125348

Tableau 4.11 Sentiment and Number of reviews on Sports and Outdoors reviews dataset

Sentiment	No. of Reviews
NEGATIVE	43320
POSITIVE	253017



a) Histogram of average sentiment on dataset of Baby reviews

b) Histogram of average sentiment on dataset of Sports and Outdoors reviews

Figure 4.11 Distribution of ratings on datasets

4.5.4 Training and Testing Sets

To build our model a training set is carried out in the dataset and for validation of our model a test set is built. We list the statistics of the data set in Table 4.12 and Table 4.13. The dataset of Baby reviews was randomly divided into training set and test set. Training data contains 120594 reviews and the testing set contains 40198 reviews. Dataset of Sports and Outdoors reviews was randomly divided into training set and test set. Training data contains 222252 reviews and the testing set contains 74085 reviews.

Tableau 4.12 Number of training and testing on Baby reviews dataset

Training	Testing
120594	40198

Tableau 4.13 Number of training and testing on Sports and Outdoors reviews datase

Training	Testing
120594	40198

4.5.5 Reputation Scores after Cleaning and after Using Logistic Regression Algorithm

Building a sentiment classifier to identify whether the review has positive or negative sentiment, the Logistic Classifier model will use the CountVectorizer Selection and TfidfVectorizer Selection from the training data to develop a model to predict True positive reviews, False positive reviews, True negative reviews, and True negative reviews. In this study, through Confusion matrix as shown in Figure 4.12a , 4.12b , 4.12c , and 4.12d we use True positive reviews as positive sentiment and True negative reviews as negative sentiment as show in Table 4.14 and Table 4.15 and Table 4.16 and Table 4.17 Then we calculate Reputation scores with CountVectorizer and TfidfVectorizer Selection on Baby reviews dataset, and Sports and Outdoors dataset, as explained in next steps.

1- CountVectorizer on Baby reviews dataset

$Reputation\ scores = PFP * 5 scores 100$

$PFP = Positives Positives\ \ negatives$

Tableau 4.14 Sentiment and Number of reviews

Sentiment	No. of Reviews
NEGATIVE	4935
POSITIVE	30128

$PFP = 3012830128\ \ 4935 = 60.18$

$Reputation\ scores = 60.18 * 5100 = 3.0\ Scores$

2- TfidfVectorizer on Baby reviews dataset

Tableau 4.15 Sentiment and Number of reviews

Sentiment	No. of Reviews
NEGATIVE	3314
POSITIVE	30907

$Reputation\ scores = PFP * 5scores100$

$PFP = PositivesPositives\ \ negatives$

$PFP = 3090730907\ \ 3314 = 90.31$

$Reputation\ scores = 60.18 * 5100 = 4.5\ Scores$

3- CountVectorizer on Sports and Outdoors reviews dataset

Tableau 4.16 Sentiment and Number of reviews

Sentiment	No. of Reviews
NEGATIVE	8121
POSITIVE	65964

$Reputation\ scores = PFP * 5scores100$

$PFP = Positives Positives\ \ negatives$

$PFP = 6596465964\ \ 8121 = 89.03$

$Reputation\ scores = 89.03 * 5100 = 4.4\ Scores$

4- TfidfVectorizer on Sports and Outdoors reviews dataset

Tableau 4.17 Sentiment and Number of reviews

Sentiment	No. of Reviews
NEGATIVE	8963
POSITIVE	65122

Reputation scores = (PFP* 5 scores) / 100

PFP = Positives / (Positives + negatives)

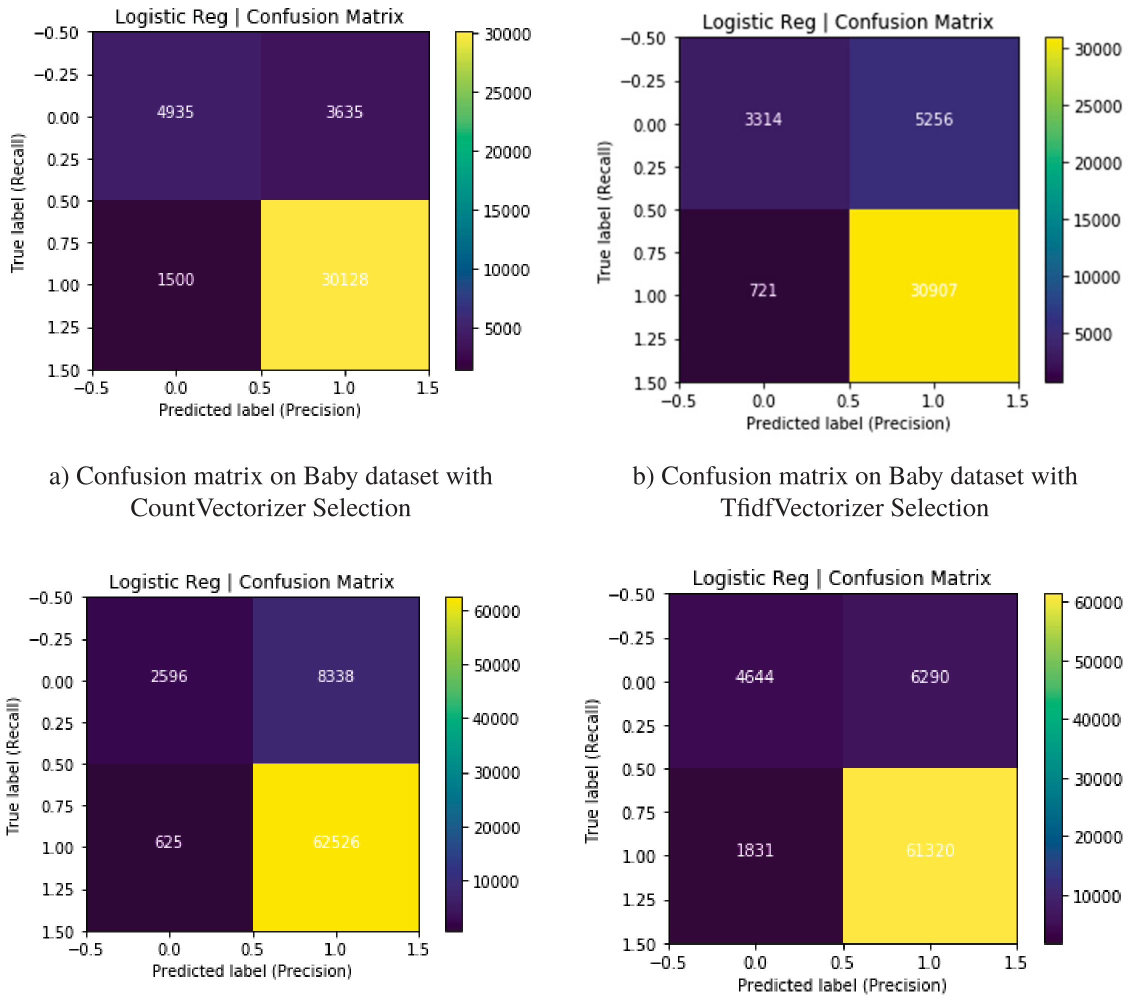
PFP = 65122/(65122+8963) = 87.90

Reputation scores = (87.90 * 5) / 100 = 4.3 Scores

4.5.6 Unfair Reviews Detection Methods

The main goal of opinion unfair reviews Detection is to identify each unfair review. So there are mainly two methods to detect unfair reviews : unfair Positive Reviews value and unfair Negative Reviews value. One method using logistic regression algorithm with CountVectorizer Selection, and the second method using TfidfVectorizer Selection.

Table 4.18 displays the results of evaluation parameters for two different features selection and provides a summary of recordings of unfair reviews rate and fair reviews rate obtained from the experiment.



a) Confusion matrix on Baby dataset with CountVectorizer Selection

b) Confusion matrix on Baby dataset with TfidfVectorizer Selection

c) Confusion matrix on Sports and Outdoors dataset with TfidfVectorizer Selection

d) Confusion matrix on Sports and Outdoors dataset with CountVectorizer Selection

Figure 4.12 Confusion matrix on datasets

The graph in Figures 4.13 and Figure 4.18 show the percentage of evaluation parameters with two features selection, CountVectorizer Selection and TfidfVectorizer Selection and we identfied a rate of unfair Negative Reviews,unfair Positive Reviews, fair Negative Reviews, fair Positive Reviews for comparative analysis of logistic regression algorithm.

We have used CountVectorizer Selection for detection processes because, throughout our implementation, we have compared between CountVectorizer Selection and TfidfVectorizer

Tableau 4.18 Evaluation parameters for two different methods

Logistic regression algorithm on Baby reviews dataset				
Features	Unfair Positive Reviews Rate	Unfair negative Reviews Rate	Fair Positive Reviews Rate	Fair negative Reviews Rate
CountVectorizer	42.41	4.74	95.25	57.58
TfidfVectorizer	61.33	2.27	97.72	38.66

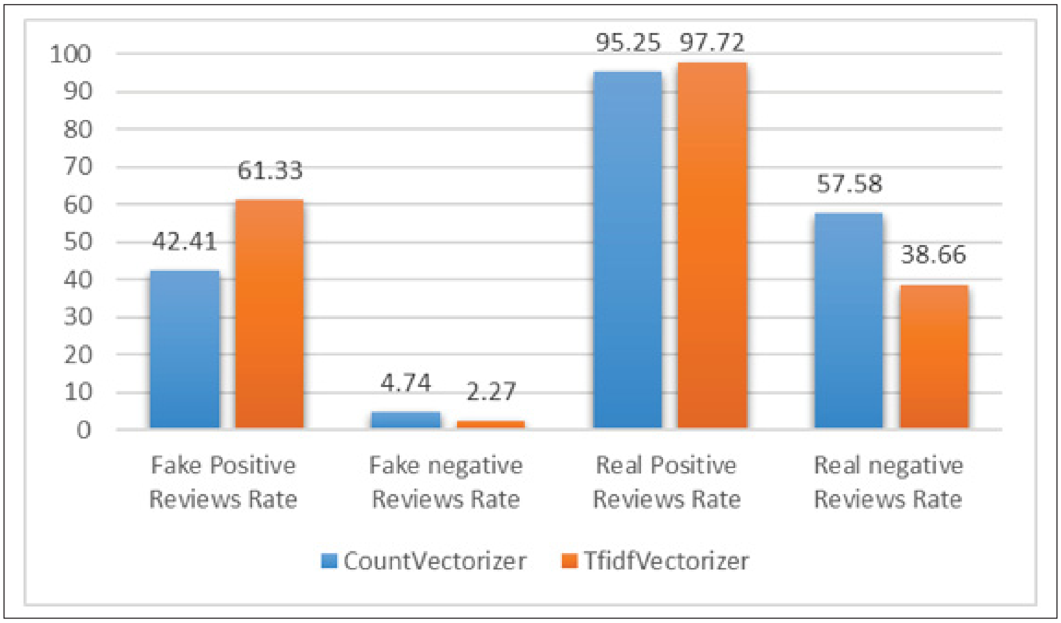


Figure 4.13 Percentage of evaluation parameters with two features selection

Tableau 4.19 A Comparison of the accuracy and Time taken to the build model on Baby reviews dataset

Logistic regression algorithm (on Baby reviews dataset)		
Features	Accuracy %	Time taken to build model (seconds)
CountVectorizer	87.22	44.2
TfidfVectorizer	85.13	45.6

Tableau 4.20 A Comparison of the accuracy and Time taken to the build model on Sports and Outdoors reviews dataset

Logistic regression algorithm (on Sports and Outdoors reviews dataset)		
Features	Accuracy %	Time taken to build model (seconds)
CountVectorizer	89.03	87
TfidfVectorizer	87.90	96

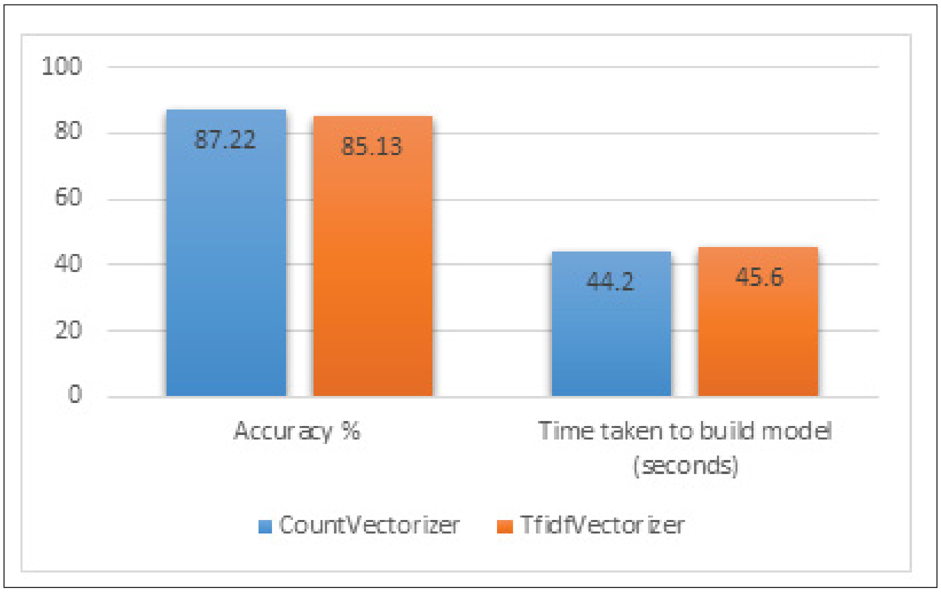


Figure 4.14 Accuracy and Time taken to the build model (seconds) Baby reviews dataset

Selection with Logistic Regression algorithm and we found more accuracy and less time with CountVectorizer Selection, and applied it for Baby and Sports and Outdoors datasets, as shown

Tableau 4.21 Comparison result of Precision, Recall on Baby reviews dataset

Features	Precision	Recall
CountVectorizer	89.23	95.25
TfidfVectorizer	85.13	85.46

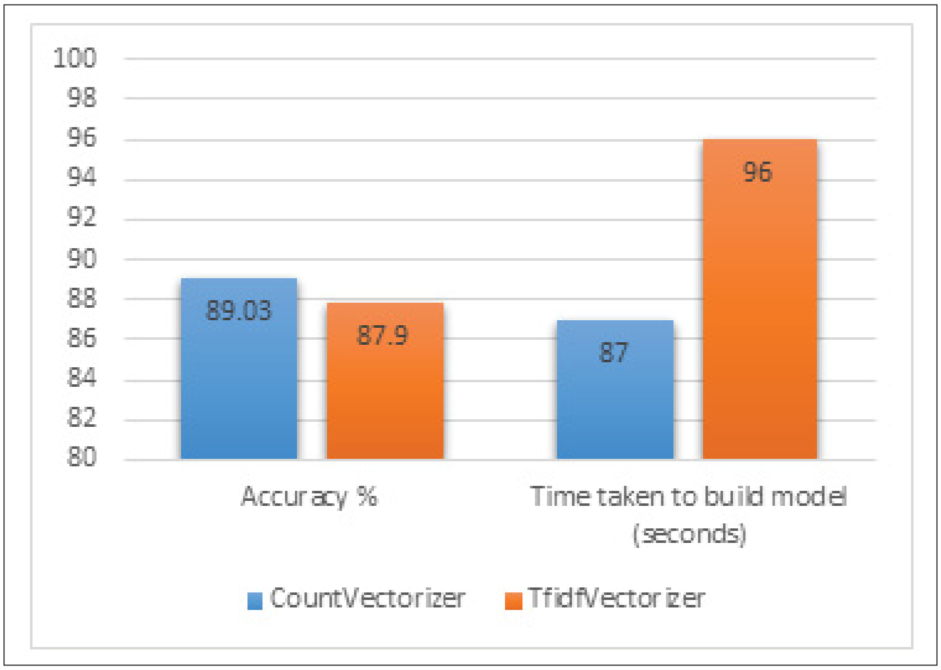


Figure 4.15 Accuracy and Time taken to the build model (seconds) on Sports and Outdoors reviews dataset

Tableau 4.22 Comparison result of Precision, Recall on Sports and Outdoors reviews dataset

Features	Precision	Recall
CountVectorizer	89.23	97.10
TfidfVectorizer	85.46	99.01

in Table 4.19 and Table 4.20 and Figures 4.14 and 4.15. In the Table 4.21, Table 4.22, Figure 4.16 and Figure 4.17 present the percentage of comparison result of Precision, Recall on Baby

Tableau 4.23 Evaluation parameters for two different methods

Logistic regression algorithm on Sports and Outdoors reviews dataset				
Features	Unfair Positive Reviews Rate	Unfair negative Reviews Rate	Fair Positive Reviews Rate	Fair negative Reviews Rate
CountVectorizer	57.52	42.47	2.89	97.10
TfidfVectorizer	76.25	23.74	0.98	99.01

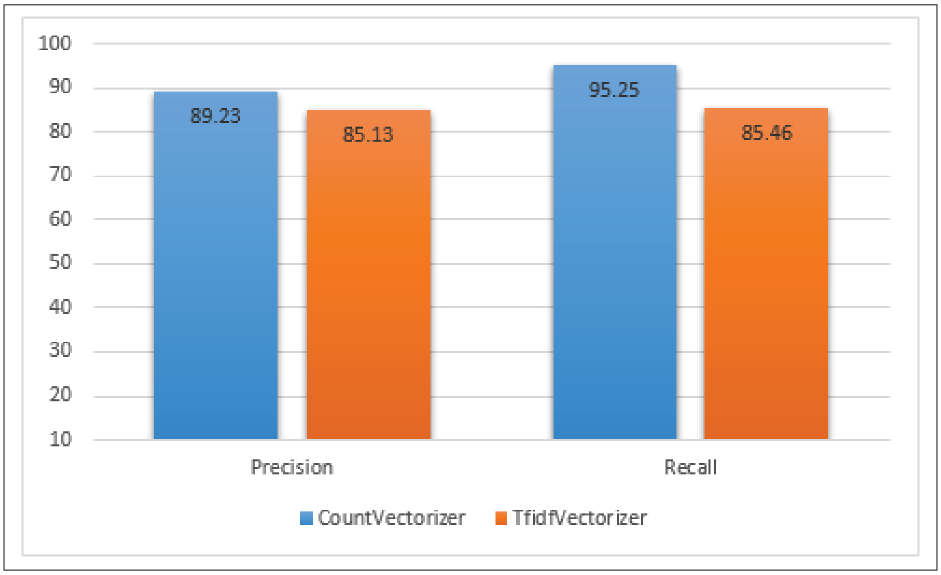


Figure 4.16 Percentage of comparison result of Precision, Recall on Baby reviews dataset

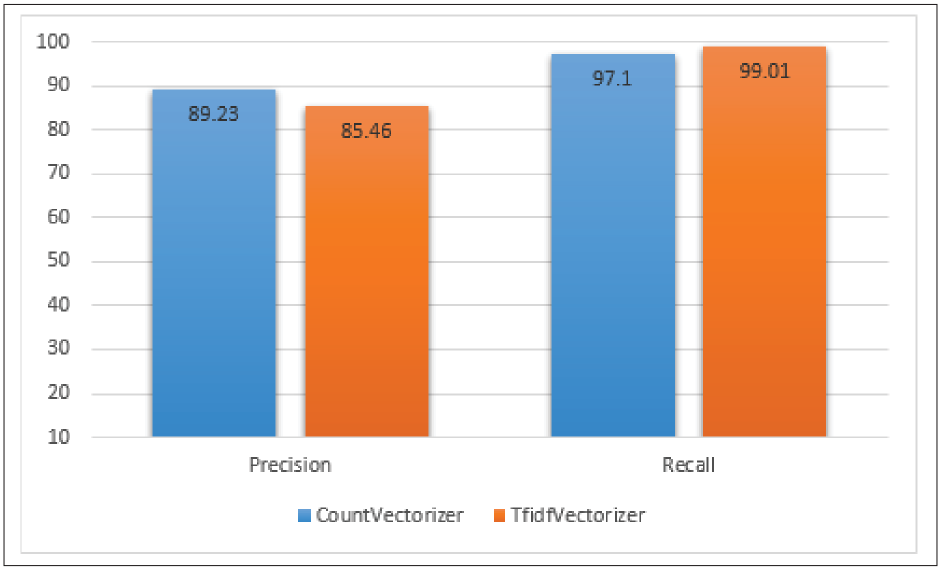


Figure 4.17 Percentage of comparison result of Precision, Recall on Sports and Outdoors reviews dataset

and Sports and Outdoors datasets, and all of these metrics are calculated for each Features selection of CountVectorize and TfidfVectorizer.

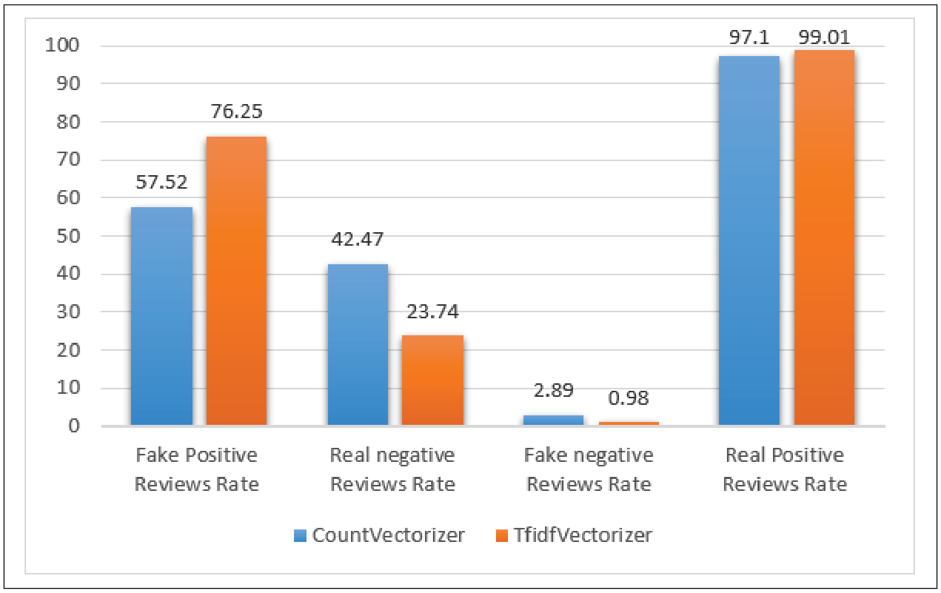


Figure 4.18 Percentage of evaluation parameters with two features selection

4.5.7 Unfair Reputation Scores Detection with Countvectorizer Selection

In common reputation systems, most evaluation and protection process used are quite non-transparent, only giving the summed up reputation value which does not expose much details on how it is calculated. In a study carried out and based on a user-centric method, more than half of the respondents complained of this lack of transparency. The user experience can be improved in reputation system by enhanced transparency. Table 4.25 shows Fair reputation scores and Unfair reputation scores using CountVectorizer Selection with Baby reviews dataset, and Sports and Outdoors dataset. It is often the case that unfair reputation scores have a different statistical pattern than fair reputation scores.

Tableau 4.24 Positive feedback percentage before and after cleaning on two different datasets

Datasets	The positive Feedback Percentage before cleaning	The positive Feedback Percentage after cleaning and using LR algorithm
Baby	88.14	60.18
Sports and Outdoors	92.93	89.03

4.5.8 Collusion and Manipulation Detection with Countvectorizer Selection

Collusion and manipulation are illegal cooperation between seller and customer in order to cheat or deceive others. The seller may try to collude with the customer in order to increase his/her reputation likewise the seller may try to collude with the customer in order to decrease the reputation of another seller.

Our detection of collusion and manipulation depended on the best feature selection that was used with Logistic regression algorithm in this study. Through the Table 4.25, and after detecting Fair reputation scores and Unfair reputation scores through our methods for Reputation scores calculation, before cleaning and after using Logistic regression algorithm, we found on Baby and Sports and Outdoors datasets with CountVectorizer Selection, that there was collusion and manipulation between the seller and customer, and we have calculated Collusion and manipulation Percentage (CMP), as shown in equation 10.

$$CMP = \frac{Fair\ reputation\ scores - Unfair\ reputation\ scores}{Total\ scores} * 100 \tag{4.3}$$

CMP on baby dataset = $4.4 - 3.05 * 100$ = 2.8

CMP on Sports and Outdoors dataset = $4.6 - 4.45 * 100$ = 4

The percentage of Collusion and manipulation on baby dataset is 28, the percentage of Collusion and manipulation on Sports and Outdoors dataset is 4.

Tableau 4.25 Fair and unfair reputation scores detection on two different datasets

Datasets	Fair reputation scores	Unfair reputation scores
Baby	3.0	4.4
Sports and Outdoors	4.4	4.6

4.5.9 All Good Reputation Issue Detection

Through our result, we have compared between the positive Feedback percentage before cleaning and after cleaning and using LR algorithm as shows in Table 18, and we have found the positive feedback percentage after cleaning and using LR algorithm less than the positive Feedback percentage before cleaning. Moreover, we have compared between accuracy and Time taken to build the model with CountVectorizer Selection and TfidfVectorizer Selection as shown in Figure 4.25 and Figure 4.25, and we have found LR algorithm with CountVectorizer Selection less time and more accurate than LR algorithm with TfidfVectorizer Selection.

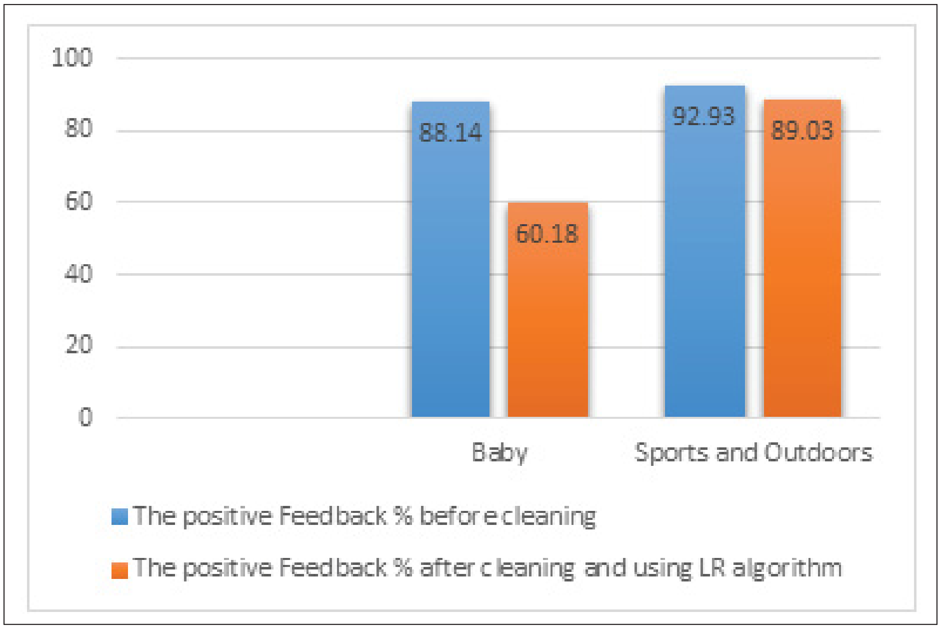


Figure 4.19 Comparison between the positive Feedback percentage before cleaning and after cleaning

4.6 Conclusions and Future Work

In this study, we introduced a review of the existing reputation and Sentiment Analysis systems, and the relevant development of these approaches. We pointed out the significant issues that the buyers might face in regard to the reputation of sellers, including unfair rating and unfair reviews. Then, we illustrated some potential solutions that are capable of computing reputation scores without including Unfair positive reviews and Unfair negative reviews in an E-Commerce environment. Natural Language Processing encourages the human-level understandings of the text reviews. In sentiment classification, significant text features are extracted to train binary classifiers to get positive or negative rating predictions Qiao (2019). In this work, we proposed a SAM based on a logistic regression algorithm with two different feature selections, in order to analyze two different datasets from Amazon. We were able to detect Unfair positive reviews and Unfair negative reviews, the "all excellent reputation" issue, and collusion and manipulation through our processes. Furthermore, our experimental method studied the accuracy, precision and recall of Logistic regression algorithm with two feature selections, and how to determine which feature selection is more accurate and takes less time.

We have used CountVectorizer Selection for our detection processes because, throughout our implementation, we have compared CountVectorizer Selection and TfidfVectorizer Selection with a Logistic Regression algorithm and we found that with CountVectorizer Selection our algorithm was more accurate and took less time. As for logistic regression algorithm classifier on Baby reviews dataset, TfidfVectorizer Selection (Acc = 85.13%) and CountVectorizer Selection (Acc = 87.22%) have the best-trained models. Likewise, for logistic regression algorithm classifier on Sports and Outdoors reviews dataset with TfidfVectorizer Selection (Acc = 87.90%) and CountVectorizer Selection (Acc = 89.03%) have the best-trained models.

We have calculated reputation scores from feedback reviews based on a SAM to obtain useful information from reviews. In addition, we chose an optimal feature selection in this study, to better detect Unfair positive reviews and Unfair negative reviews, the "all excellent reputation" issue, unfair reputation scores, and collusion and manipulation. In the future, we intend to refine our method. This study can be extended by improving some aspects and subtasks of our approach. In the following, we propose some suggestions and future extensions to our work :

- Add more and different feature selections to our approach, instead of CountVectorizer and TfidfVectorizer Selections. For example, Hofmann & Chisholm (2016) performed an initial basic analysis ; a more sophisticated approach using word n-grams is adopted to yield improvements in performance ;

- Try to combine our classifier with a stronger supervised learning method such as Support-Vector Machines (SVMs) to have better accuracy with sentiment analysis ;
- Apply sentiment classification algorithms in social commerce environments, such as Facebook or Twitter, to detect Unfair reviews, the ”all excellent reputation” issue, unfair reputation scores, and collusion and manipulation ;
- Use various tools such as R studio, Statistical Analysis System (SAS) to implement and evaluate the performance of our work.

4.7 Acknowledgment

The first author would like to thank the Libyan Ministry of Education and Canadian Bureau for International Education (CBIE).

CONCLUSION ET RECOMMANDATIONS

Cette thèse est composée de trois articles. Dans le premier article, nous avons proposé plusieurs méthodes pour analyser un ensemble de données des avis sur les films. Nous avons également présenté des algorithmes de classification de sentiments pour appliquer un apprentissage supervisé des avis sur les films situés dans deux ensembles de données différents. Nos approches expérimentales ont étudié l'exactitude, la précision, le rappel ("recall") et la mesure F ("F-Measure") de tous les algorithmes de classification de sentiments, et la manière de déterminer quel algorithme est le plus précis. En outre, nous avons pu détecter de faux avis positifs et de faux avis négatifs grâce à des processus de détection. Cinq algorithmes d'apprentissage supervisé pour la classification de sentiments de nos ensembles de données ont été comparés dans cet article : NB, K-NN, K*, SVM et DT-J48. En utilisant l'analyse de l'exactitude (accuracy) pour ces cinq techniques, nous avons trouvé que l'algorithme SVM est le plus exact pour classer correctement les avis dans les ensembles de données des avis sur les films. En outre, les processus de détection des faux avis positifs et des faux avis négatifs dépendent de la meilleure méthode.

Dans le deuxième article, nous avons proposé les algorithmes NB, DT-J48, LR et SVM pour analyser les ensembles de données des avis Amazon. Nous avons également présenté des méthodes de classification de sentiments et nous avons mené nos expérimentations en utilisant trois ensembles de données différents des avis Amazon avec suppression des mots vides. Nos approches expérimentales ont étudié l'exactitude, la précision et le rappel des algorithmes de classification de sentiments. De plus, nous avons pu détecter des avis négatifs injustes, des avis neutres injustes et des avis positifs injustes en utilisant les processus de détection de cette méthode.

Dans le troisième article, nous avons présenté une revue des systèmes existants d'analyse de la réputation et de sentiments, et le développement pertinent de ces approches. Nous avons souligné les problèmes importants auxquels les acheteurs pourraient être confrontés en ce qui concerne la réputation des vendeurs, notamment les évaluations et les avis inéquitables. Ensuite, nous avons illustré certaines solutions potentielles capables de calculer les scores de réputation sans inclure les avis positifs injustes et les avis négatifs injustes dans un environnement du commerce électronique. Le traitement du langage naturel encourage la compréhension des textes d'avis au niveau humain. Dans la classification de sentiments, les caractéristiques significatives du texte sont extraites pour former les classificateurs binaires à obtenir des prédictions d'évaluation positives ou négatives Qiao (2019). Dans ce travail, nous avons proposé un SAM basé sur un algorithme de régression logistique avec deux sélections de caractéristiques différentes, afin d'analyser deux ensembles de données différents provenant d'Amazon. Nous avons pu détecter des critiques positives injustes et des critiques négatives injustes, le problème de la "toute excellente réputation", ainsi que la complicité et la manipulation par le biais de nos processus. En outre, notre méthode expérimentale a étudié l'exactitude, la précision et le rappel de l'algorithme de régression logistique avec deux sélections de caractéristiques, et comment déterminer quelle sélection de caractéristiques est la plus importante et prend le moins de temps. Pour les travaux futurs, nous souhaitons étendre ces travaux pour utiliser des ensembles de données plus récents des avis d'Amazon ainsi que différentes sélections de fonctionnalités, au lieu des sélections CountVectorizer et TfidfVectorizer.

À titre d'exemple, certains chercheurs ont effectué une première analyse de base et ont adopté une approche plus sophistiquée, en utilisant le mot " n-grams ", afin d'améliorer les performances. En outre, nous pouvons utiliser des méthodes de classification de sentiments pour détecter les faux ou injustes avis et les fausses ou injustes évaluations à l'aide de différents outils, puis nous évaluerons la performance de notre travail à l'aide de ces outils. En outre, nous pouvons essayer d'appliquer des algorithmes de classification des sentiments dans des environnements de commerce social, tels que Facebook ou Twitter, pour détecter les avis faux ou injustes, le problème de "toute excellente réputation", les scores de réputation injustes, ainsi que la complicité et la manipulation. Nous pouvons également essayer de concevoir des outils qui peuvent être très efficaces pour détecter les avis faux ou injustes avec une exactitude raisonnable. Enfin, nous pouvons essayer de développer un processus similaire pour l'apprentissage non supervisé de données non étiquetées (unlabeled) afin de détecter les faux avis et de calculer ensuite les scores de réputation à partir des avis réels (vrais).

ANNEXE I

AN EMPIRICAL STUDY ON DETECTING FAKE REVIEWS USING MACHINE LEARNING TECHNIQUES

Elshrif Elmurngi [a], Abdelouahed Gherbi [b]

[a, b] Département de Génie logiciel et des technologies de l'information, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Communication publié dans «International Conference on Innovative Computing Technology», November 2017

DOI : 10.1109/INTECH.2017.8102442

1. Abstract

Reputation systems in E-commerce (EC) play a substantial role that allows various parties to achieve mutual benefits by establishing relationships. The reputation systems aim at helping consumers in deciding whether to negotiate with a given party. Many factors negatively influence the sight of the customers and the vendors in terms of the reputation system. For instance, lack of honesty or effort in providing the feedback reviews, by which users might create phantom feedback from fake reviews to support their reputation. Moreover, the opinions obtained from users can be classified into positive or negative which can be used by a consumer to select a product. In this paper, we study online movie reviews using Sentiment Analysis (SA) methods in order to detect fake reviews. Text classification and SA methods are applied on a real conducted dataset of movie reviews. Specifically, we compare four supervised machine learning algorithms : Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN-IBK), and Decision Tree (DT-J48) for sentiment classification of reviews in two different situations without stopwords and with stopwords methods are employed. The measured results show that for both methods the SVM algorithm outperforms other algorithms, and it reaches the highest accuracy not only in text classification but also to detect fake reviews.

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

2. Introduction

Sentiment Analysis (SA), also known as Opinion Mining (OM), 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 Liu (2012).

In this study, we sometimes consider the time is more valuable than money, therefore instead of spending times in reading and figuring out the positivity or negativity of review we can use automated techniques for sentiment analysis.

The basic of Sentiment Analysis is classifying the polarity (positive and negative) of a given text at the levels of document, sentence, and aspect whether the expressed opinion in three levels is positive or negative.

The aim of sentiment analysis is to find opinions from reviews and then classify these opinions based upon polarity.

According to Medhat *et al.* (2014), in Sentiment Analysis there are three major classification levels : the first level is document level, the second level is sentence level, and the third level is aspect level. The document level Sentiment Analysis aims to classify an opinion document as a negative or positive opinion. It regards the whole record as a basic information unit. The sentence level using Sentiment Analysis aims to setup opinion stated in every sentence. The aspect level using Sentiment Analysis goals to categorize the sentiment on the specific aspects of entities.

The document is obtained from a dataset of movie reviews , and then a sentiment analysis technique is applied to classify the documents resultant as real positive and real negative reviews or fake positive and fake negative reviews. Real negative and fake positive reviews can lead to financial gains. This, unfortunately, gives strong incentives to write fake reviews that attempt to intentionally mislead readers by providing unfair reviews to several goods for the purpose of damaging their reputations. Detecting such fake reviews is a significant challenge. For example, fake consumer reviews in the e-commerce sector are not only affecting individual consumers but also corrupt purchaser's confidence in the online shopping Malbon (2013). Machine learning techniques and Sentiment Analysis methods will have a major positively effect on reputation systems, and especially to detection processes of fake reviews in an e-commerce and social commerce environments.

In machine learning-based techniques, algorithms such as SVM, NB, and DT-J48 are applied for the classification Xia *et al.* (2011). SVM is a type of learning algorithm that represents supervised machine learning approaches Barbu (2012), and it is an excellent successful prediction approach. The SVM also is a robust classifier approach Esposito (2014). One of the recent researchers has presented in Medhat *et al.* (2014) that introduce a survey on different applications and algorithms for Sentiment Analysis but it focused on algorithms used in various languages with stopwords and did not focus on without stopwords and the results are not accurate when without stopwords are considered. Also, the researchers did not focus on detecting fake reviews (Kalaivani & Shunmuganathan (2013); Pang & Lee (2004)). This paper presents four supervised machine learning approaches to classifying sentiment of our dataset which is compared with stopwords and without stopwords methods. We have also detected 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 review or fake review using Sentiment Analysis algorithms with supervised learning techniques.

The conducted experiments have shown the accuracy of results through sentiment classification algorithms, and we have found that SVM in both cases without stopwords and with stopwords is more accurate than other methods such as NB, KNN-IBK, and DT-J48.

The main contributions of this study are detailed in the Conclusion and Future Work section, but can be briefly summarized as follows : We compared different sentiment classification algorithms for labeling movie reviews as fake or real, and ranked the algorithms according to accuracy. We also found that the use of stopwords proved more efficient in the classification task.

The rest of this paper is organized as follows. Section 3 presents the related works. Section 4 shows the methodology, section 5 explains the experiment results, and finally, section 6 presents the conclusion and future work.

3. Related Works

This work belongs to the set of studies on reputation systems evaluation vulnerability. This study employs statistical methods to evaluate the performance of detection mechanism for fake reviews and evaluate the accuracy of this detection ; here we emphasize our literature review on studies that applied statistical methods to this issue.

3.1 Sentiment Analysis Issues

There are several issues accounted in conducting of Sentiment analysis Vinodhini & Chandrasekaran (2012). In the first major issue, the viewpoint (or opinion) observed as negative in a situation possibly be considered positive in another situation. In the second major issue, the people don't always have same express views in a similar approach. Most common text processing employed the fact the minor changes between the two text fragments don't change the actual sense, accurately Vinodhini & Chandrasekaran (2012).

3.2 Textual Reviews to Provide Detailed Opinion About the Product

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 focused only on the overall ratings of products without considering reviews which provided by customers Xu *et al.* (2016). On the other hand, most websites allow consumers to add textual reviews to provide a detailed opinion about the product Tian *et al.* (2014a), Tian *et al.* (2014b). These reviews are available for customers' to read, and customers' now depend increasingly on reviews rather than on ratings. Through the Reputation models that could use SA methods to extract users' opinions and use this data in the reputation system. This information may include consumers' opinions about different features (Abdel-Hafez & Xu (2013); Abdel-Hafez *et al.* (2012)).

3.3 Detecting Fake Reviews Using Machine Learning

Filter and identification of fake reviews have substantial significance Jindal & Liu (2008). In Moraes *et al.* (2013) authors 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 techniques comprise of two phases, selection, and extraction of reviews categorization utilizing learning models such as SVM.

Extract the best and accurate approach, and simultaneously categorize the customers' written reviews text into negative or positive opinions. It 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 (Liu *et al.* (2005); Pang *et al.* (2002); Fujii & Ishikawa (2006); Ku *et al.* (2006)). Our work used several supervised learning algorithms such as SVM, NB, KNN-IBK, and DT-J48 for Sentiment Classification of text to detect fake reviews.

Tableau-A I-1 Provides a Comparative Study by authors using different Classification algorithms

Reference	Data Source	Size of dataset	Using Supervised learning	Using Unsupervised learning	Language	Classification algorithms	without stopwords	With stopwords	The best method
Liu *et al.* (2005)	News Group dataset	20 categories with 1000	Yes	No	English	NB,SVM	No	Yes	NB
Pang *et al.* (2002)	Movie Reviews dataset	2000 Movie Reviews	Yes	No	English	NB,SVM,IBK DT	No	Yes	SVM
Fujii & Ishikawa (2006)	Movie Reviews dataset	4000 Movie Reviews	Yes	No	Chinese	NB,SVM,KNN LLR,Delta TFIDF, LDA-SVM, TFIDF,DKV	Yes	No	NB and DKV
Ku *et al.* (2006)	Movie Reviews dataset	1400 Movie Reviews, 2000 Movie Reviews	Yes	No	English	NB,SVM	Yes	No	NB
Hassan *et al.* (2011)	Movie Reviews dataset	2000 Movie Reviews	Yes	No	English	NB,SVM	Yes	No	SVM
This work	Movie Reviews dataset	2000 Movie Reviews	Yes	No	English	NB,SVM, KNN-IBK,DT-J48	Yes	Yes	SVM Robust and very accurate

3.4 Comparative Study of Different Classification Algorithms

Table-A I-1 shows comparative studies on classification algorithms to prove the best method for detecting fake reviews using different dataset such as News Group dataset, Text documents, Movie Reviews dataset proves Chu *et al.* (2016), Singh *et al.* (2013) that NB and DKV (Distributed Keyword Vector) are accurate without stopwords while Hassan *et al.* (2011) finds that NB is accurate for stopwords. Using same data sets Kalaivani & Shunmuganathan (2013) finds that SVM is accurate for with stopwords while Pang & Lee (2004) finds that SVM is only accurate without stopwords. However, in our empirical study results prove that SVM is robust and accurate for both with and without stopwords, and also for detecting fake reviews.

4. Methodology

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

Step 1 : Movie Reviews Collection

To provide an exhaustive study of machine learning algorithms, the experiment based on analyzing the sentiment value of the standard dataset. We use the original data set of the movie review to test our methods of reviews classification. This dataset is available and has been used in Pang & Lee (2004). The dataset of movie reviews is available obtained from the Internet Movie Database (IMDb), and this dataset consists of 2000 reviews, and are uniform in 1000 positive and 1000 negative.

Step 2 : Data Pre-Processing

The pre-processing phase includes preliminary operations which help in transforming the data before the actual SA task. To demonstrate the effect of pre-processing on the classification models data preprocessing plays a very significant in many supervised learning, through our

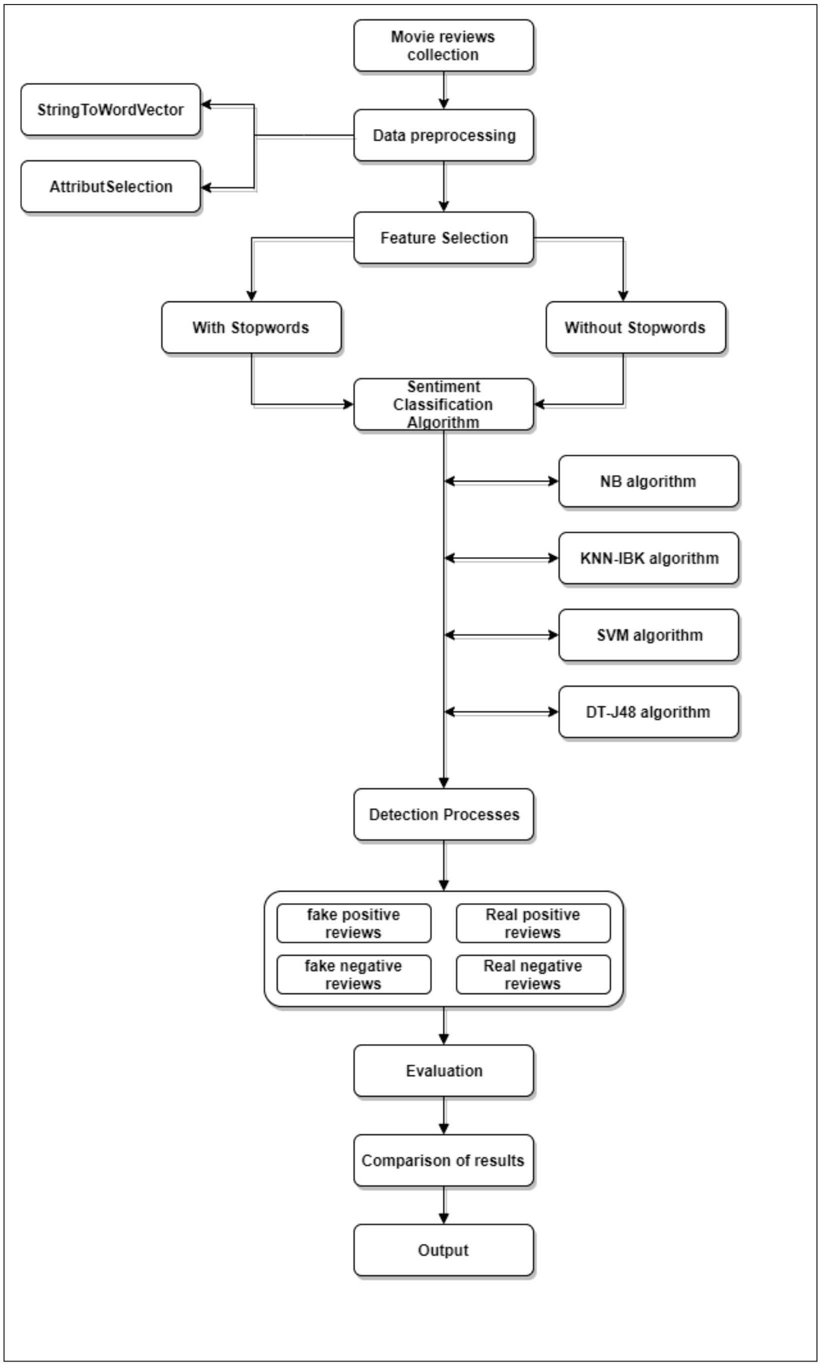


Figure-A I-1 Flowchart of Proposed Work

proposed scheme we divided data preprocessing as the following :

A. StringToWordVector

To preparing our data for learning, which involves transforming it by using the StringToWordVector filter, and which is the main tool for text analysis in WEKA. This filter allows configuring the different steps of the term extraction. Indeed, we should be able to see something such as the following :

- Configure the tokenizer We need to do Feature extraction using machine learning technique that is converting the normal text to a set of features to make the provided document classifiable ;
- Specify a stopwords list Stop words list are the words we want to filter out before training the classifier. Several of the most commonly used stop words in English, they could be “a,” “the”, “of,” “I,” “you,” “it,” “and.” These are usually high-frequency words that aren’t giving any additional information to our labeling, but rather they actually confuse our classifier. In this study, we used a 630 English stopwords list. Stop word removal can help us in reducing the memory requirement while classifying the reviews.

B. Attribute Selection

Removing the poorly describing attributes can be valuable to get improved classification accuracy. Because not all attributes are relevant to the classification work, and irrelevant attributes can even decrease the performance of some algorithms. We should perform attribute selection before training the classifier. Attribute selection with supervised learning differs from unsupervised learning, where in the latter case, data have no goal attribute.

Step 3 : Feature Selection

In this study, we implemented four feature selection methods widely used for the classification task of Sentiment Analysis with Stopwords and without Stopwords methods. The results differ from one method to another.

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 four popular supervised classifiers : NB, DT-J48, SVM, K-NN, 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 data set. Also, the NB classifier has fast decisions making process.

2) Support Vector Machine (SVM)

SVM in machine learning is supervised learning models with the related learning algorithm, which examines data and identify patterns, used for regression and classification analysis. 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 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, similarity measure for voting, and, k parameter (Song *et al.* (2007); Bhattacharya *et al.* (2012); Latourrette (2000); Zhang (2010)).

A set of vectors and class labels which are related with 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=l, k=3, k=5, k=7. These numbers decide how many neighbors influence the classification.

4) Decision Tree (DT-J48)

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

Step 5 : Detection Processes

After training, the next step is to predict the output of the model on the testing dataset, and a confusion matrix generated which classifies the review as positive or negative. We are defining as Fake the set of reviews that are found to be False (False Positive or False Negative) and defining as Real the set of reviews that are found to be True (True positive and True Negative). The Fake and Real reviews are determined according to equations a trough d. 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) is events which are Real and is effectively labeled as Real, true positive(TP) is 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. According to the confusion matrix, A I-1 to A I-6 shows numerical parameters that apply following measures to evaluate the Detection Process (DP) performance. In Table-A I-2 the confusion matrix shows the counts of real and fake predictions obtained with known data, and for each algorithm used in this study is different performance evaluation and confusion matrix.

Tableau-A I-2 The confusion matrix

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

$$Fake\ Positive\ Reviews\ Rate = \frac{FP}{TN + FP} \tag{A I-1}$$

$$Fake\ Negative\ Reviews\ Rate = \frac{FN}{TP + FN} \tag{A I-2}$$

$$Real\ Positive\ Reviews\ Rate = \frac{TP}{TP + FN} \tag{A I-3}$$

$$Real\ Negative\ Reviews\ Rate = \frac{TN}{TN + FP} \tag{A I-4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{A I-5}$$

$$Precision = \frac{TP}{TP + FP} \tag{A I-6}$$

For each algorithm different Performance evaluation and confusion matrix.

Step 6 : Comparison of Results

In this section, we present experimental results from four different supervised machine learning approaches to classifying sentiment of our dataset which is compared with stopwords and without stopwords methods. Also, we have used the same techniques at the same time to detect fake reviews.

5. Experimental Results

In this section, we present experimental results from four different supervised machine learning approaches to classifying sentiment of our dataset which is compared with stopwords and without stopwords methods. Also, we have used the same techniques at the same time to detect fake reviews.

A. Without Stopwords

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, DT-J48 algorithms. Table-A I-3 displays confusion matrix for respectively Movie review 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.

2) Evaluation Parameters and Accuracy for all Methods

Four 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-A I-4 shows the results of evaluation parameters for all methods and provides a

Tableau-A I-3 Confusion matrix for all methods

Classification algorithms	SA	Real	Fake
NB	Real	806	194
	Fake	177	823
KNN-IBK (K=1)	Real	766	234
	Fake	426	574
KNN-IBK (K=3)	Real	800	200
	Fake	382	618
KNN-IBK (K=5)	Real	817	183
	Fake	370	630
KNN-IBK (K=7)	Real	824	176
	Fake	366	634
SVM	Real	812	188
	Fake	177	823
(DT-J48)	Real	743	257
	Fake	286	714

summary of recordings obtained from the experiment. Where, SVM surpasses for best accuracy among the other classification algorithms with 81.75%. The tabulated observations list the readings as well as accuracies obtained for a specific supervised learning algorithm on a dataset of a movie review.

Tableau-A I-4 Evaluation parameters and accuracy for all methods

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Precision %	Accuracy %
NB	19.4	17.7	82.3	80.6	80.9	81.45
KNN-IBK (K=1)	23.4	42.6	57.4	76.6	71	67
KNN-IBK (K=3)	20	38.2	61.8	80	75.6	70.9
KNN-IBK (K=5)	18.3	37	63	81.7	77.5	72.35
KNN-IBK (K=7)	17.6	36.6	63.4	82.4	78.3	72.9
SVM	18.8	17.7	82.3	81.2	81.4	**81.75**
(DT-J48)	25.7	28.6	71.4	74.3	73.5	72.85

The graph in Figure-A I-2 shows a rate of Fake Positive Reviews, Fake negative Reviews, Real Positive Reviews, Real negative Reviews, Accuracy, and Precision for comparative analysis of all different algorithms.

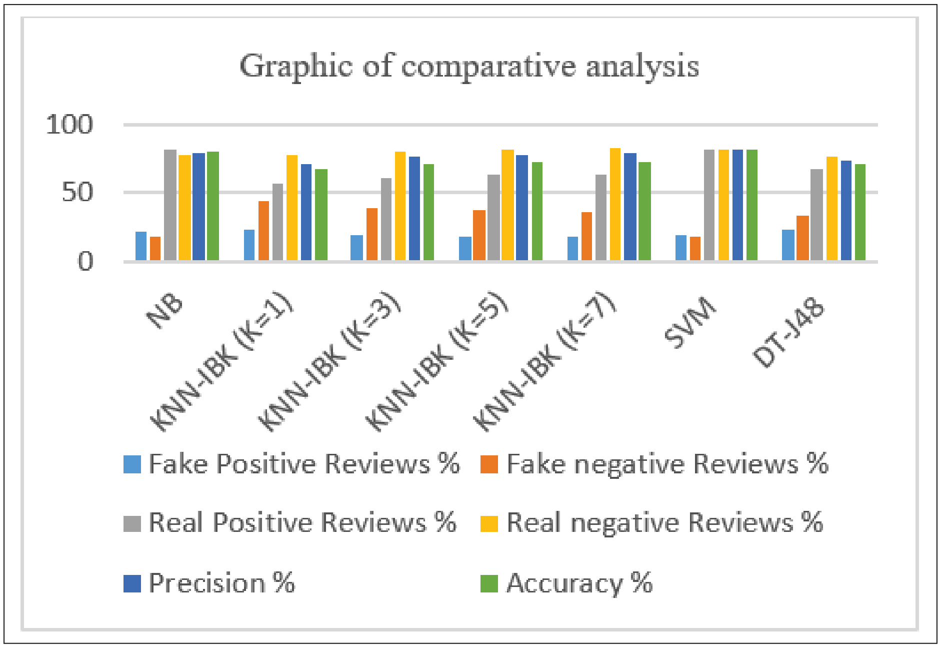


Figure-A I-2 Graphic of comparative analysis of all methods

The comparison in Table-A I-5 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

Tableau-A I-5 Comparison of Accuracy of classifiers

Classification algorithms	Accuracy %
NB	81.45
KNN-IBK (K=1)	67
KNN-IBK (K=3)	70.9
KNN-IBK (K=5)	72.35
KNN-IBK (K=7)	72.9
SVM	**81.75**
DT-J48	72.85

The graph in Figure-A I-3 shows accuracy rate of NB, SVM, (K-NN, k=l, k=3, k=5, k=7), DT-J48 algorithms. We obtained a high accuracy of SVM algorithm than other algorithms.

Table-A I-6 shows the time taken to build prediction model by each algorithm. As evident from the table, K-NN takes the shortest amount of time of 0 seconds to create a model and SVM takes the longest amount of time of 1.58 seconds to build a model.

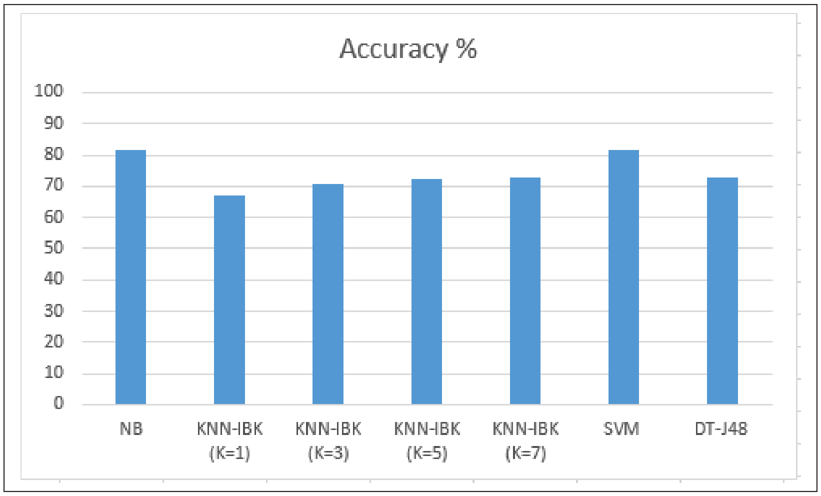


Figure-A I-3 The accuracy of different algorithms

Tableau-A I-6 Time taken to build model and accuracy for all classification algorithms

Classification algorithms	Time taken to build model (Seconds)
NB	0.05
KNN-IBK (K=1)	0
KNN-IBK (K=3)	0.01
KNN-IBK (K=5)	0
KNN-IBK (K=7)	0
SVM	**1.58**
DT-J48	0.93

B. With Stopwords

1. Confusion Matrix for All Methods

The previous section compared different algorithms without the usage of stopwords. In this section, the algorithms were made to do a sentimental analysis on data with stopwords. From the results (refer Table-A I-7) the confusion matrix displays for respectively Movie review dataset.

2. Evaluation Parameters and Accuracy for All Methods

Four main performance evaluation measures have been introduced for Classification algorithms.

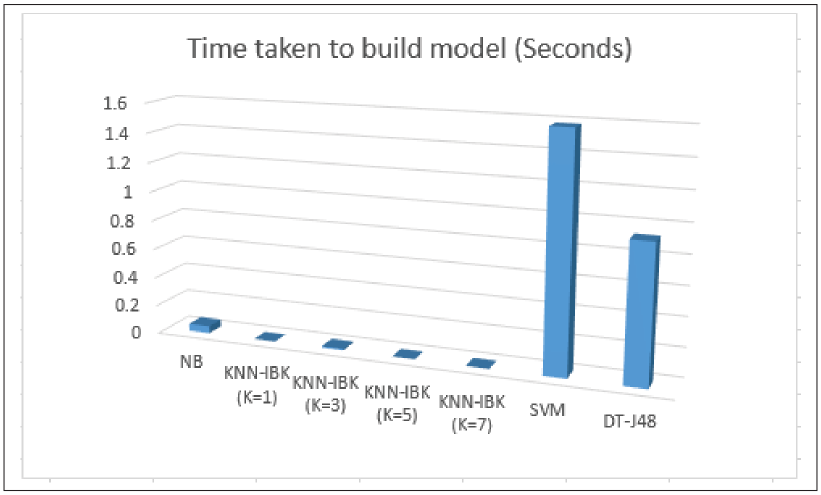


Figure-A I-4 Time taken to build model (Seconds) : without stopwords

Tableau-A I-7 Confusion matrix for all methods

Classification algorithms	SA	Real	Fake
NB	Real	781	219
	Fake	187	813
KNN-IBK (K=1)	Real	771	229
	Fake	435	565
KNN-IBK (K=3)	Real	804	196
	Fake	387	613
KNN-IBK (K=5)	Real	816	184
	Fake	372	628
KNN-IBK (K=7)	Real	824	176
	Fake	366	634
SVM	Real	809	191
	Fake	182	818
(DT-J48)	Real	762	238
	Fake	330	670

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-A I-8 displays the results of evaluation parameters for all methods and provides a summary of recordings obtained from the experiment. As a results, SVM surpasses for best accuracy among the other classification algorithms with 81.35%.

Tableau-A I-8 An evaluation of all methods using different parameters : with stopwords

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Precision %	Accuracy %
NB	21.9	18.7	81.3	78.1	78.8	79.7
KNN-IBK (K=1)	22.9	43.5	56.5	77.1	71.1	66.8
KNN-IBK (K=3)	19.6	38.7	61.3	80.4	75.8	70.85
KNN-IBK (K=5)	18.4	37.2	62.8	81.6	77.3	72.2
KNN-IBK (K=71)	17.6	36.6	63.4	82.4	78.3	72.9
SVM	19.1	18.2	81.8	80.9	81.1	**81.35**
(DT-J48)	23.8	33	67	76.2	73.8	71.6

The graph in Figure-A I-5 displays a rate of Fake Positive Reviews, Fake negative Reviews, Real Positive Reviews, Real negative Reviews, Accuracy, and Precision for comparative analysis of all different algorithms.

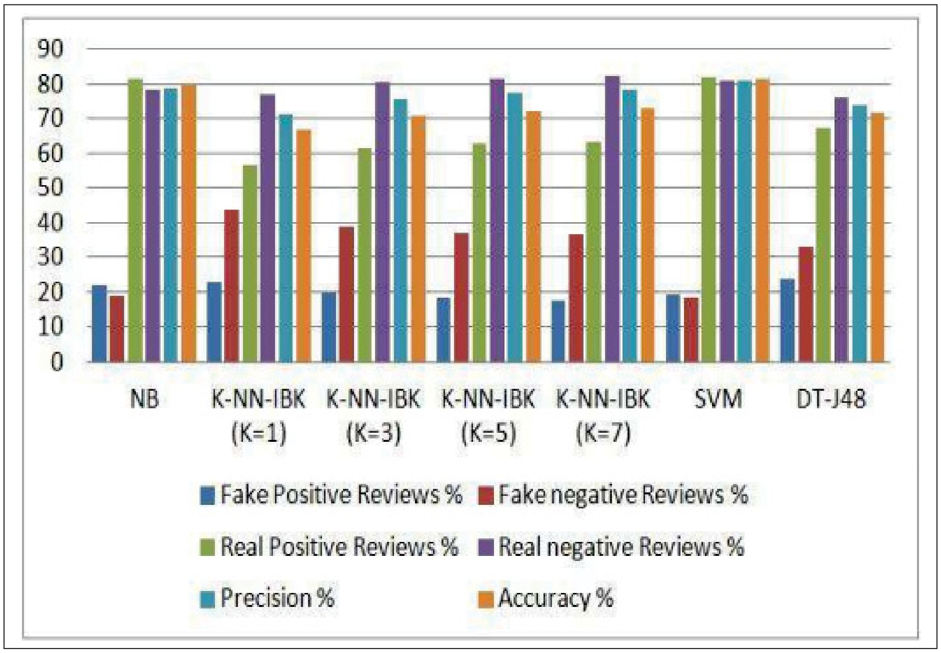


Figure-A I-5 Comparative analysis of all methods

The comparison in Table-A I-9 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

Tableau-A I-9 Comparison of Accuracy of classifiers

Classification algorithms	Accuracy %
NB	79.7
KNN-IBK (K=1)	66.8
KNN-IBK (K=3)	70.85
KNN-IBK (K=5)	72.2
KNN-IBK (K=7)	72.9
SVM	**81.35**
DT-J48	71.6

The graph in Figure-A I-10 displays accuracy rate of NB, SVM, (K-NN, k=1, k=3, k=5, k=7), DT-J48 algorithms. We obtained a high accuracy of SVM algorithm than other algorithms.

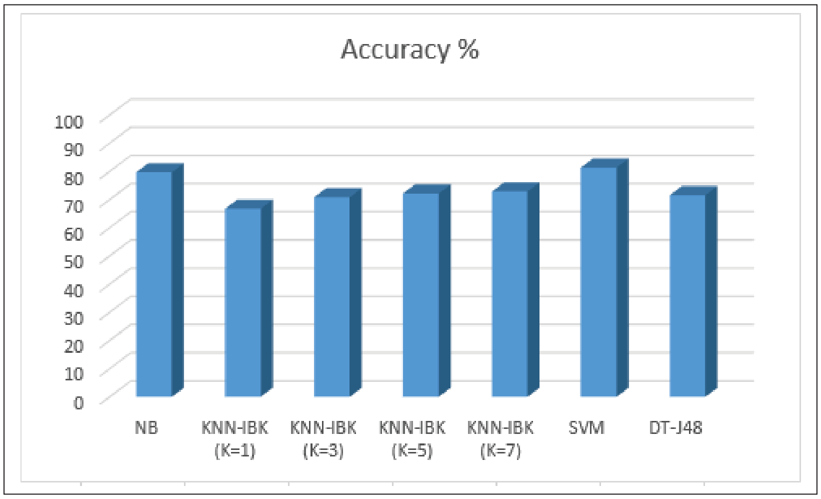


Figure-A I-6 The accuracy of different algorithms

Table-A I-10 displays the time taken to build prediction model by each algorithm. As evident from the table, K-NN takes the shortest amount of time of 0 seconds to create a model and SVM takes the longest amount of time of 14.84 seconds to build a model.

C. The Summary of Our Experiments

Table-A I-11 and Figure-A I-7 present the summary of the experiments, where SVM is the best algorithm by accuracy for all tests with stopwords and without stopwords. It can be inferred

Tableau-A I-10 Time taken to build model : with stopwords

Classification algorithms	Time taken to build model (Seconds)
NB	0.11
KNN-IBK (K=1)	0
KNN-IBK (K=3)	0.01
KNN-IBK (K=5)	0
KNN-IBK (K=7)	0
SVM	14.84
DT-J48	0.34

Tableau-A I-11 The best result of our experiments by accuracy

Features and Parameters	Fake Positive Reviews of SVM %	Fake Negative Reviews of SVM %	Precision of SVM %	Accuracy of SVM %
without stopwords	**18.8**	**17.7**	81.4	**81.75**
with stopwords	**19.1**	**18.2**	81.1	**81.35**

that SVM does not agree with other algorithms. SVM tends to be more accurate than other methods in comparison. The presented study emphasizes that the accuracy of SVM tends to be higher when using the without stopwords feature. However, the detection process of Fake Positive Reviews and Fake Negative Reviews offers less promising results when compared to using the with stopwords feature.

6. Conclusion and Future Work

In this paper, we proposed several methods to analyze a dataset of movie reviews and presented sentiment classification algorithms and supervised learning used in our work with stopwords and without stopwords methods. Our experimental approaches studied the accuracy of all sentiment classification algorithms, and how to determine which algorithm is more accurate. Furthermore,

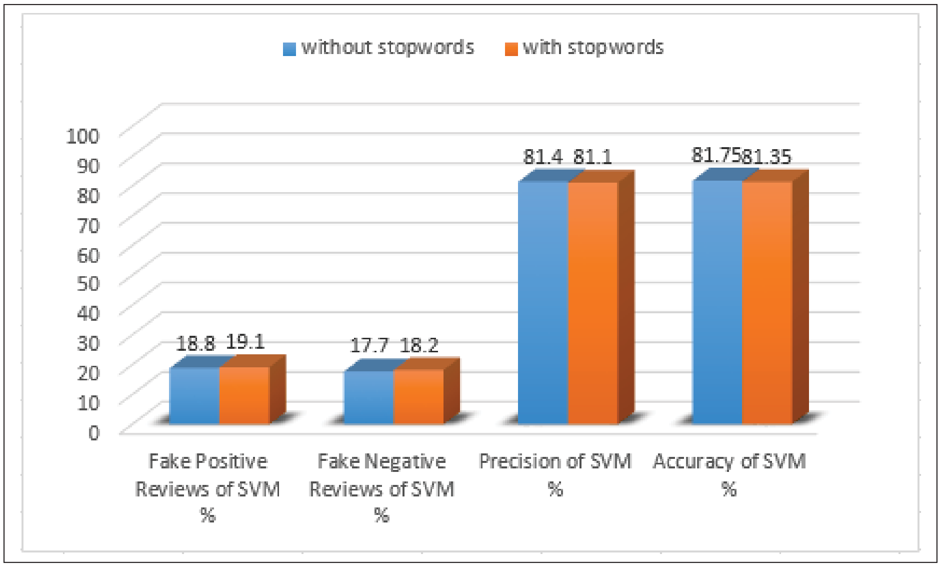


Figure-A I-7 The summary of our experiments

we were able to detect fake positive review, and fake negative review through detection processes are shown in our results.

Four supervised learning algorithms to classifying sentiment of our dataset have been compared in this paper with stopwords and without stopwords. The first algorithm is NB, the second algorithm is SVM, and the third algorithm is K-NN, and the fourth algorithm is DT-J48. Through all of these algorithms also we have detected fake positive reviews and fake negative reviews. In this paper, our experiments have shown the accuracy of results through sentiment classification algorithms, and we have found that SVM algorithm in both cases stopwords and without stopwords are more accurate than other methods. Also, detection processes for fake positive reviews and fake negative reviews depend on the best and more accurate method that used in this study.

The main contributions of this study are summarized as follows :

- This study compares different sentiment classification algorithms in Weka tool, which are used to classify movie reviews dataset into fake and real reviews ;

- This study applies the sentiment classification algorithms using without-stopwords and with-stopwords methods. We rea lizedthat without-stopwords method is more efficient not only in text categorization but also to detect fake reviews ;
- This study performs several analysis and tests to find the best-supervised learning algorithm in terms of accuracy.

Finally, in our future work, we would like to extend this work to use other datasets such as Amazon dataset or eBay dataset or different dataset of a movie review and use different feature selection methods. Furthermore, we may apply sentiment classification algorithms to detect fake reviews for other aspects in the same area such as collusion and manipulation issues.

ANNEXE II

DETECTING FAKE REVIEWS THROUGH SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

Elshrif Elmurngi [a], Abdelouahed Gherbi [b]

[a, b] Département de Génie logiciel et des technologies de l’information, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Communication publié par IARIA, « The Sixth International Conference on Data Analytics », November 2017
ISSN : 2308-4464

1. Abstract

Recently, 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 two different datasets, including movie review dataset V2. 0 and movie reviews dataset V1. 0. 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.

2. Introduction

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 Liu (2012). "The sentiment is usually formulated as a two-class classification problem, positive and negative" Liu (2012). 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 Medhat *et al.* (2014), 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 Pang *et al.* (2002) and Pang & Lee (2004). 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 [4]. 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 Xia *et al.* (2011). SVM is a type of learning algorithm that represents supervised machine learning approaches Barbu (2012), and it is an excellent successful prediction approach. The SVM is also a robust classification approach Esposito (2014). A recent research presented in Medhat *et al.* (2014) 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 Kalaivani & Shunmuganathan (2013); Pang & Lee (2004); Hassan *et al.* (2011); Chu *et al.* (2016); Singh *et al.* (2013). 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 of results through sentiment classification algorithms. In both cases (movie reviews dataset V2.0 and movie reviews dataset V1.0), 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 Hall *et al.* (2009), 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 two different datasets with stopwords. We realized that using the stopwords 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.

The rest of this paper is organized as follows. Section 3 presents the related works. Section 4 shows the methodology. Section 5 explains the experiment results, and finally, Section 6 presents the conclusion and future works.

3. 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 Vinodhini & Chandrasekaran (2012). 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 Vinodhini & Chandrasekaran (2012).

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 Xu *et al.* (2016). On the other hand, most websites allow consumers to add textual reviews to provide a detailed opinion about the product Tian *et al.* (2014a,b). 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 Abdel-Hafez & Xu (2013) and Abdel-Hafez *et al.* (2012).

C. Detecting Fake Reviews Using Machine Learning

Filter and identification of fake reviews have substantial significance Jindal & Liu (2008); Moraes *et al.* (2013) 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 Liu *et al.* (2005); Fujii & Ishikawa (2006); Ku *et al.* (2006). Our work used several supervised learning algorithms such as SVM, NB, KNNIBK, K* and DT-J48 for Sentiment Classification of text to detect fake reviews.

D. A Comparative Study of different Classification algorithms

Table-A II-11 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 Chu *et al.* (2016) and Singh *et al.* (2013). While Hassan *et al.* (2011) finds that NB is accurate and a better choice, but it is not oriented for detecting fake reviews. Using the same datasets, [8] finds that SVM is accurate with stopwords method, but it does not focus on detecting fake reviews, while Pang & Lee (2004) finds that SVM is only accurate without using stopwords method, and also without detecting fake reviews. However, in our empirical study, results in both cases with movie reviews dataset V2.0 and with movie reviews dataset V1.0 prove that SVM is robust and accurate for detecting fake reviews.

Tableau-A II-1 A comparative study of different classification algorithms

Reference	Data Source	Using Supervised learning	Using Unsupervised learning	Classification algorithms	without stopwords	With stopwords	The best method
Abdel-Hafez *et al.* (2012)	News Group dataset	Yes	No	NB, SVM	No	Yes	NB
Jindal & Liu (2008)	Text documents	Yes	No	NB,SVM,IBK,DT	No	Yes	SVM
Moraes *et al.* (2013)	Movie Reviews dataset	Yes	No	NB, SVM,K-NN	Yes	No	NB
Liu *et al.* (2005)	Movie Reviews dataset	Yes	No	NB, SVM	Yes	No	NB
Fujii & Ishikawa (2006)	Movie Reviews dataset	Yes	No	NB, SVM	Yes	No	SVM
Our study	Movie Reviews dataset	Yes	No	NB,SVM,IBK,J48	Yes	Yes	SVM Robust and very accurate

4. 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 II-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 Singh *et al.* (2013) , 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 V2.0 which consists of 2000 movie reviews out of which 1000 reviews are positive, and 1000 reviews are negative. The second dataset is known as movie reviews dataset V1.0, which consists of total 1400 movie reviews, 700 of which are positive and 700 of which are negative. A summary of the two datasets collected is described in Table-A II-12.

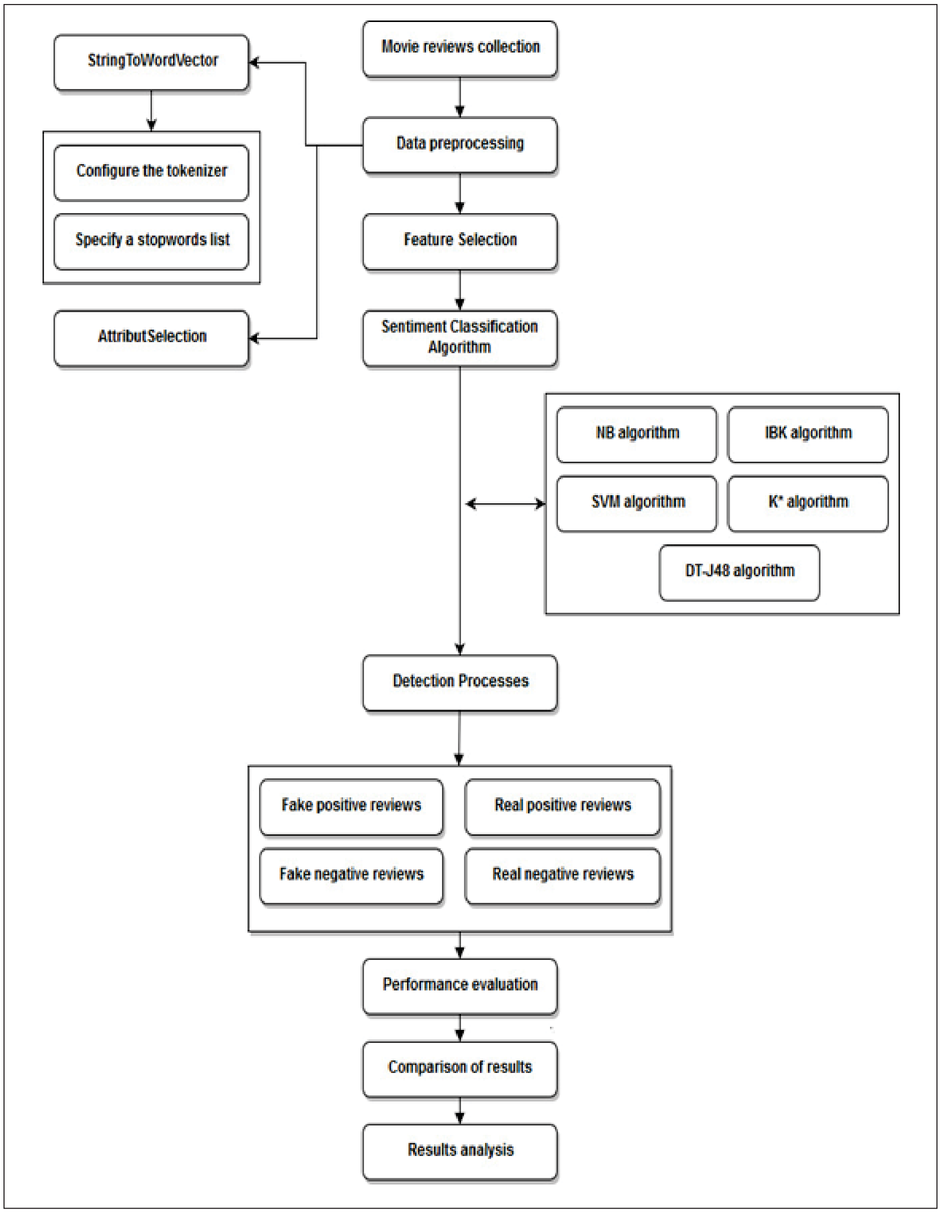


Figure-A II-1 Steps and Techniques used in Sentiment Analysis

Step 2 : Data Pre-Processing

The preprocessing phase includes two preliminary operations, shown in Figure II-1, that 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 :

Tableau-A II-2 Description of dataset

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

1) StringToWordVector

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 singlewords, 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 :

- Configure the tokenizer

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

- Specify a stopwords list

  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 dataset V2.0. 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 five feature selection methods 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)

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 [24]. 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 nonparametric 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 (Song *et al.* (2007); Bhattacharya *et al.* (2012); Latourrette (2000); Zhang (2010)).

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, A II-1-A II-6 shows numerical parameters that could be applied following measures to evaluate the Detection Process (DP) performance. In Table-A II-3, 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.

Tableau-A II-3 The confusion matrix

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

$$Fake\ Positive\ Reviews\ Rate = \frac{FP}{TN + FP} \qquad \text{(A II-1)}$$

$$Fake\ Negative\ Reviews\ Rate = \frac{FN}{TP + FN} \qquad \text{(A II-2)}$$

$$Real\ Positive\ Reviews\ Rate = \frac{TP}{TP + FN} \tag{A II-3}$$

$$Real\ Negative\ Reviews\ Rate = \frac{TN}{TN + FP} \tag{A II-4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{A II-5}$$

$$Precision = \frac{TP}{TP + FP} \tag{A II-6}$$

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.

5. Experiments and result analysis

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

A. 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-A II-4 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.

Tableau-A II-4 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	**R**eal	**8**09	**1**91
	Fake	**1**82	**8**18
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-A II-5 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.

Tableau-A II-5 Evaluation parameters and accuracy for all methods

Classification algorithms	Fake Positive Reviews %	Fake Negative Reviews %	Real Positive Reviews %	Real Negative Reviews %	Precision %	Accuracy %
NB	21.9	18.7	81.3	78.1	78.8	79.7
K-NN-IBK (K=3)	19.6	38.7	61.3	80.4	75.8	70.85
K*	24	33.7	66.3	76	73.4	71.15
SVM	19.1	18.2	81.8	80.9	81.1	**81.35**
DT-J48	23.8	33	67	76.2	73.8	71.6

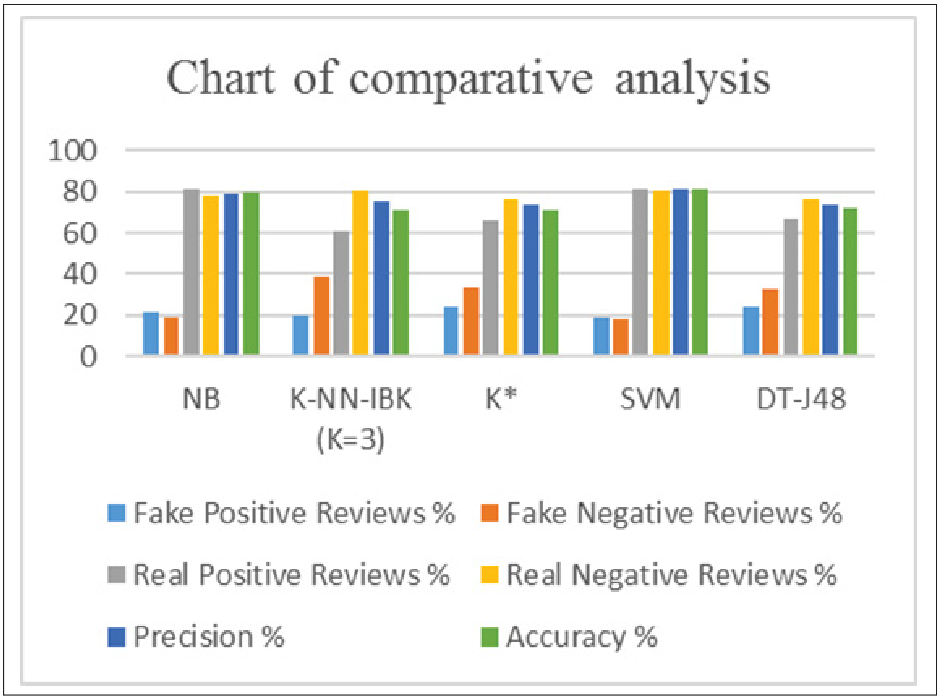


Figure-A II-2 Comparative analysis of all methods

The graph in Figure II-2 shows a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy, and Precision for comparative analysis of all different algorithms.

The comparison in Table-A II-6 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, K*, and DT-J48 algorithms.

Tableau-A II-6 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 II-3 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.

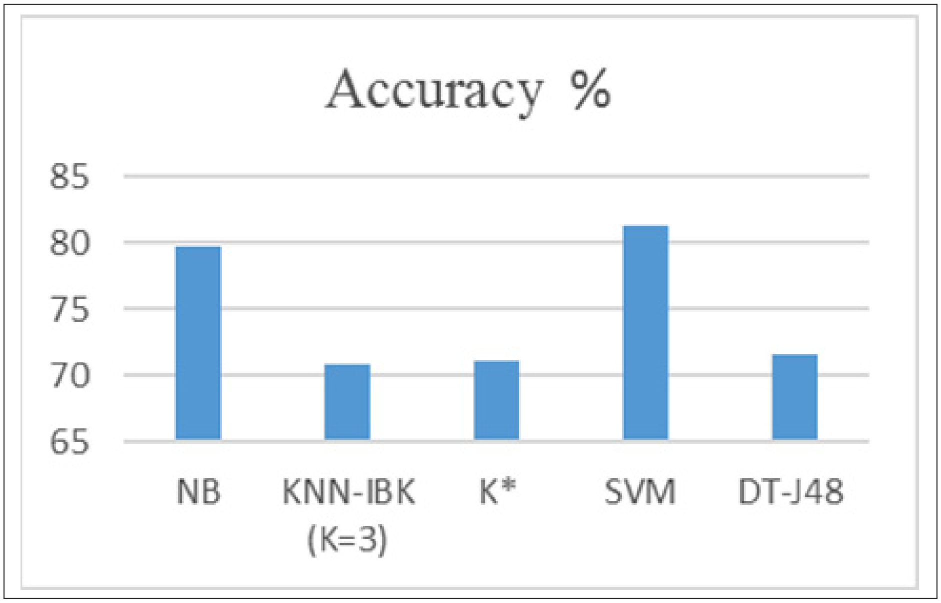


Figure-A II-3 Graph showing the accuracy of different algorithms

Table-A II-7 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 seconds to create a model and SVM takes the longest amount of time of 14840 seconds to build a model.

Tableau-A II-7 Time taken to build the model

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

B. 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-A II-8, the confusion matrix displays results for movie reviews dataset v1.0.

Tableau-A II-8 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-A II-9 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%.

Tableau-A II-9 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 II-4 displays a rate of Fake Positive Reviews, Fake Negative Reviews, Real Positive Reviews, Real Negative Reviews, Accuracy, and Precision for comparative analysis of all different algorithms.

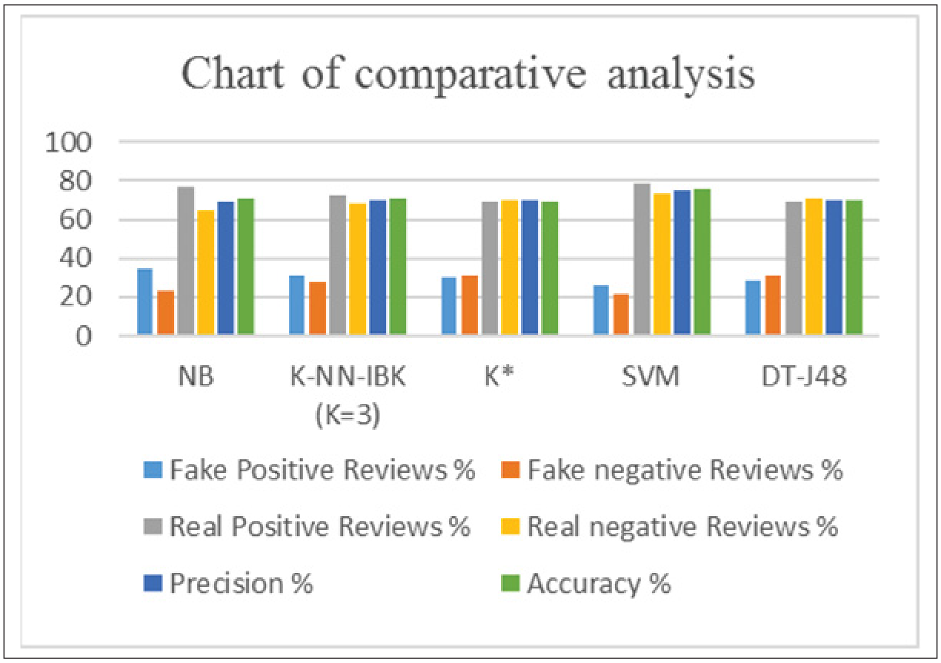


Figure-A II-4 Comparative analysis of all methods

The comparison in Table-A II-10 indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, and DT-J48 algorithms.

Tableau-A II-10 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 II-5 displays accuracy rate of NB, SVM, (K-NN, k=3), DT-J48 algorithms. We obtained a higher accuracy of SVM algorithm than other algorithms.

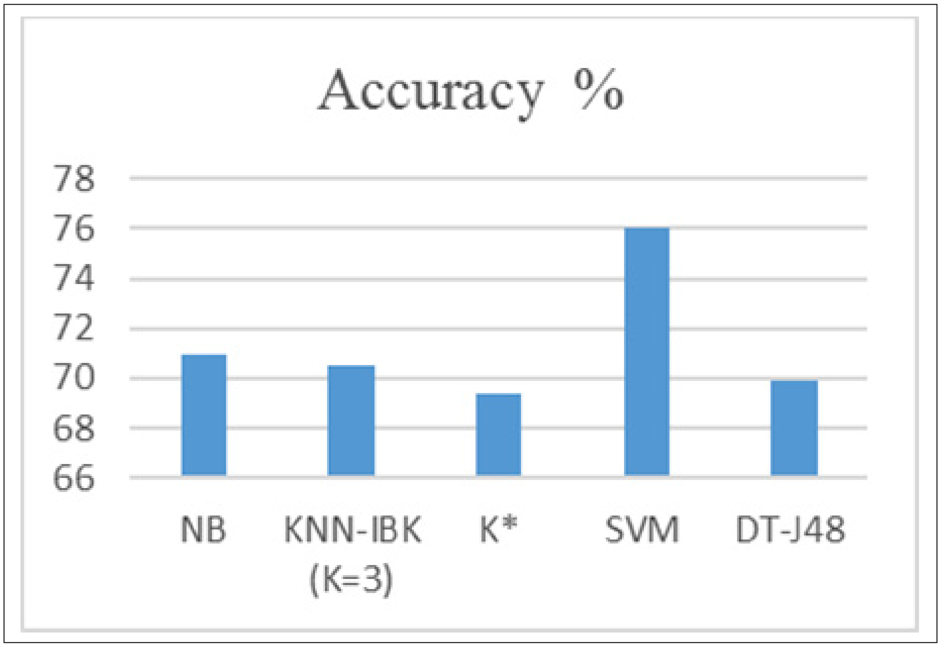


Figure-A II-5 Accuracy of different algorithms

Table-A II-11 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 seconds to create a model and SVM takes the longest amount of time of 4.24 seconds to build a model.

Tableau-A II-11 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

C. Discussion

Table-A II-12 and Figure II-6 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 V2.0 and 76% of the reviews in dataset V1.0. SVM tends to be more accurate than other methods.

Tableau-A II-12 The best result of our experiments

Experiments	Fake Positive Reviews of SVM %	Fake Negative Reviews of SVM %	Precision of SVM %	Accuracy of SVM %
Results on dataset V2.0	**19.1**	**18.2**	81.1	**81.35**
Results on dataset V1.0	26.3	**21.7**	74.9	**76**

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-A II-12.

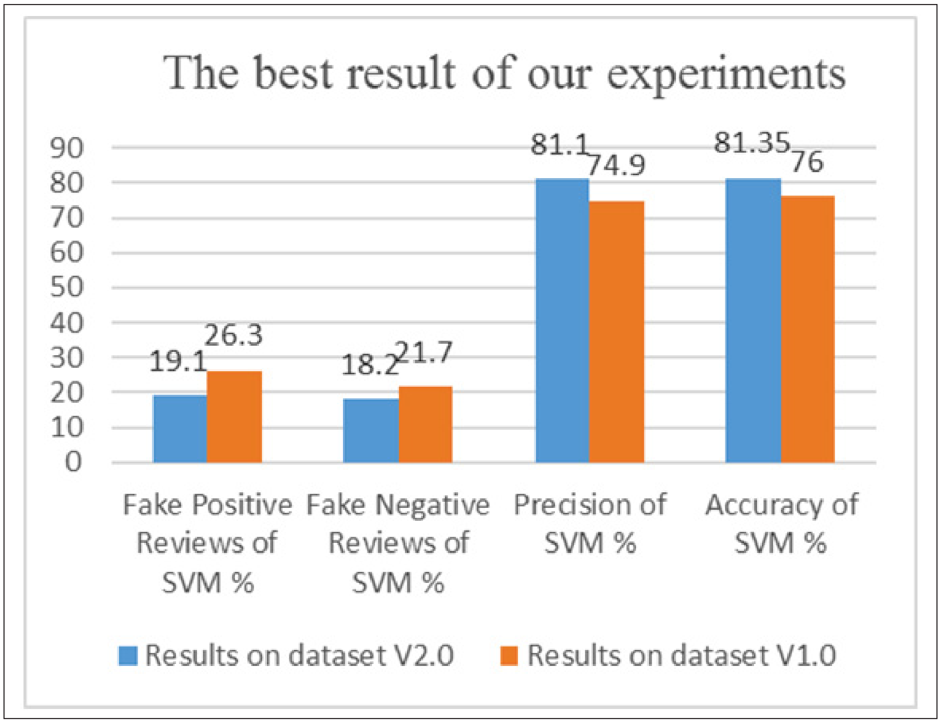


Figure-A II-6 Accuracy of different algorithms

6. Conclusion and futur work

In this paper, 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 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., V2.0 and V1.0. 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 to detect fake reviews using various tools such as Python and R or R studio, Statistical Analysis System (SAS), and Stata ; then we will evaluate the performance of our work with some of these tools.

7. 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.

BIBLIOGRAPHIE

Abdel-Hafez, A. & Xu, Y. (2013). A survey of user modelling in social media websites. *Computer and Information Science*, 6(4), 59–71.

Abdel-Hafez, A., Xu, Y. & Tjondronegoro, D. (2012). Product Reputation Model : An Opinion Mining Based Approach. *SDAD@ ECML/PKDD*, pp. 16–27.

Adler, B. T., de Alfaro, L., Kulshreshtha, A. & Pye, I. (2011). Reputation systems for open collaboration. *Communications of the ACM*, 54(8), 81.

Agarwal, B. & Mittal, N. (2016). Prominent feature extraction for review analysis : an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(3), 485–498.

Aggarwal, A. (2016). Detecting and mitigating the effect of manipulated reputation on online social networks. *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 293–297.

Barbado, R., Araque, O. & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4), 1234–1244.

Barbu, T. (2012). SVM-based human cell detection technique using histograms of oriented gradients. *cell*, 4(11).

Bhattacharya, G., Ghosh, K. & Chowdhury, A. S. (2012). An affinity-based new local distance function and similarity measure for kNN algorithm. *Pattern Recognition Letters*, 33(3), 356–363.

Breure, E. (2013). “Hotel Reviews- Can we trust them ?

Brown, J. & Morgan, J. (2006). Reputation in online auctions : The market for trust. *California Management Review*, 49(1), 61–81.

Brunner, R. J. & Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, 80, 1947–1956.

Catherine, R. & Cohen, W. (2017). Transnets : Learning to transform for recommendation. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pp. 288–296.

Chang, V. (2018). A proposed social network analysis platform for big data analytics. *Technological Forecasting and Social Change*, 130, 57–68.

Chen, Y., Chai, Y., Liu, Y. & Xu, Y. (2015). Analysis of review helpfulness based on consumer perspective. *Tsinghua Science and Technology*, 20(3), 293–305.

Cheng, W. & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3), 211–225.

Chinsha, T. & Joseph, S. (2014). Aspect based opinion mining from restaurant reviews. *International Journal of Computer Applications*, 975, 8887.

Chu, C.-H., Wang, C.-A., Chang, Y.-C., Wu, Y.-W., Hsieh, Y.-L. & Hsu, W.-L. (2016). Sentiment analysis on Chinese movie review with distributed keyword vector representation. *2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 84–89.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A. & Zhou, Q. (2016). Multilingual sentiment analysis : state of the art and independent comparison of techniques. *Cognitive computation*, 8(4), 757–771.

Dellarocas, C. (2000). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. *Proceedings of the 2nd ACM conference on Electronic commerce*, pp. 150–157.

Dellarocas, C. (2005). Reputation mechanism design in online trading environments with pure moral hazard. *Information systems research*, 16(2), 209–230.

Dellarocas, C. (2006). Reputation mechanisms. *Handbook on Economics and Information Systems*, pp. 2006.

Diekmann, A., Jann, B., Przepiorka, W. & Wehrli, S. (2014). Reputation formation and the evolution of cooperation in anonymous online markets. *American sociological review*, 79(1), 65–85.

Diesner, J., Ferrari, E. & Xu, G. (2017). *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.

Ehsaei, F. G. (2012). Acceptance of feedbacks in reputation systems : the role of online social interactions. *Information Management and Business Review*, 4(7), 391–401.

Elmurngi, E. & Gherbi, A. (2017a). Detecting fake reviews through sentiment analysis using machine learning techniques. *IARIA/data analytics*, 65–72.

Elmurngi, E. & Gherbi, A. (2017b). An empirical study on detecting fake reviews using machine learning techniques. *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, pp. 107–114.

Elmurngi, E. I. & Gherbi, A. (2018). Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques. *JCS*, 14(5), 714–726.

Engonopoulos, N., Lazaridou, A., Paliouras, G. & Chandrinos, K. (2011). ELS : a word-level method for entity-level sentiment analysis. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pp. 12.

Esposito, G. (2014). *LP-type methods for Optimal Transductive Support Vector Machines*. Gennaro Esposito, PhD.

Farmer, R. & Glass, B. (2010). *Building web reputation systems*. " O'Reilly Media, Inc.".

Farra, N., Challita, E., Assi, R. A. & Hajj, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. *2010 IEEE international conference on data mining workshops*, pp. 1114–1119.

Fattah, M. A. (2017). A Novel Statistical Feature Selection Approach for Text Categorization. *Journal of Information Processing Systems*, 13(5).

Feng, S., Xing, L., Gogar, A. & Choi, Y. (2012). Distributional footprints of deceptive product reviews. *Sixth International AAAI Conference on Weblogs and Social Media*.

Fraga, D., Bankovic, Z. & Moya, J. M. (2012). A taxonomy of trust and reputation system attacks. *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 41–50.

Fujii, A. & Ishikawa, T. (2006). A system for summarizing and visualizing arguments in subjective documents : Toward supporting decision making. *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 15–22.

Gaber, M., Cocea, M., Weibelzahl, S., Menasalvas, E. & Labbé, C. (2012). Proceedings of the First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012).

Gamal, D., Alfonse, M., M El-Horbaty, E.-S. & M Salem, A.-B. (2019). Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains. *Machine Learning and Knowledge Extraction*, 1(1), 224–234.

Gupta, A., Tenneti, T. & Gupta, A. (2009). Sentiment based Summarization of Restaurant Reviews. *Final Year Project*.

Gutowska, A. & Sloane, A. (2009). Modelling the B2C Marketplace : Evaluation of a Reputation Metric for e-commerce. *International Conference on Web Information Systems and Technologies*, pp. 212–226.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software : an update. *ACM SIGKDD explorations newsletter*, 11(1), 10–18.

Harmon, A. (2004). Amazon glitch unmasks war of reviewers, the New York Times. February.

Hassan, S., Rafi, M. & Shaikh, M. S. (2011). Comparing svm and naive bayes classifiers for text categorization with wikitology as knowledge enrichment. *2011 IEEE 14th International Multitopic Conference*, pp. 31–34.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv :1503.02531*.

Hoffman, K., Zage, D. & Nita-Rotaru, C. (2007). A Survey of attacks on Reputation Systems.

Hoffman, K., Zage, D. & Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, 42(1), 1.

Hofmann, M. & Chisholm, A. (2016). *Text mining and visualization : case studies using open-source tools*. CRC Press.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177.

Iqbal, S., Zulqurnain, A., Wani, Y. & Hussain, K. (2015). The survey of sentiment and opinion mining for behavior analysis of social media. *arXiv preprint arXiv :1610.06085*.

Jha, V., Ramu, S., Shenoy, P. D. & Venugopal, K. (2017). Reputation Systems : Evaluating Reputation Among All Good Sellers. *Data-Enabled Discovery and Applications*, 1(1), 8.

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

Jøsang, A. (2012). Robustness of trust and reputation systems : Does it matter ? *IFIP International Conference on Trust Management*, pp. 253–262.

Jøsang, A., Hayward, R. & Pope, S. (2006). Trust network analysis with subjective logic. *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*, pp. 85–94.

Kalaivani, P. & Shunmuganathan, K. (2013). Sentiment classification of movie reviews by supervised machine learning approaches. *Indian Journal of Computer Science and Engineering*, 4(4), 285–292.

Karyotis, C., Doctor, F., Iqbal, R., James, A. & Chang, V. (2018). A fuzzy computational model of emotion for cloud based sentiment analysis. *Information Sciences*, 433, 448–463.

Koncz, P. & Paralic, J. (2011). An approach to feature selection for sentiment analysis. *2011 15th IEEE International Conference on Intelligent Engineering Systems*, pp. 357–362.

Ku, L.-W., Liang, Y.-T. & Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. *Proceedings of AAAI*, pp. 100–107.

Latourrette, M. (2000). Toward an explanatory similarity measure for nearest-neighbor classification. *European Conference on Machine Learning*, pp. 238–245.

Lin, Y., Lei, H., Wu, J. & Li, X. (2015). An empirical study on sentiment classification of chinese review using word embedding. *arXiv preprint arXiv :1511.01665*.

Ling, G., Lyu, M. R. & King, I. (2014). Ratings meet reviews, a combined approach to recommend. *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 105–112.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1–167.

Liu, B., Hu, M. & Cheng, J. (2005). Opinion observer : analyzing and comparing opinions on the web. *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351.

Liu, Y., Zhang, J. & Zhu, Q. (2011). Design of a reputation system based on dynamic coalition formation. *International conference on social informatics*, pp. 135–144.

Luca, M. & Zervas, G. (2016). Fake it till you make it : Reputation, competition, and Yelp review fraud. *Management Science*, 62(12), 3412–3427.

Malbon, J. (2013). Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36(2), 139–157.

Malik, Z. & Bouguettaya, A. (2009). Rater credibility assessment in web services interactions. *World Wide Web*, 12(1), 3–25.

Mane, S. B., Assar, K., Sawant, P. & Shinde, M. (2017). Product rating using opinion mining. *Int. J. Comput. Eng. Res. Trends*, 4(5), 161–168.

McAuley, J. & Leskovec, J. (2013). Hidden factors and hidden topics : understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172.

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications : A survey. *Ain Shams engineering journal*, 5(4), 1093–1113.

Moraes, R., Valiati, J. F. & Neto, W. P. G. (2013). Document-level sentiment classification : An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633.

Mukherjee, A., Liu, B. & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*, pp. 191–200.

Noorian, Z. & Ulieru, M. (2010). The state of the art in trust and reputation systems : a framework for comparison. *Journal of theoretical and applied electronic commerce research*, 5(2), 97–117.

Oelke, D., Hao, M., Rohrdantz, C., Keim, D. A., Dayal, U., Haug, L.-E. & Janetzko, H. (2009). Visual opinion analysis of customer feedback data. *2009 IEEE Symposium on Visual Analytics Science and Technology*, pp. 187–194.

Pang, B. & Lee, L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp. 271.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86.

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135.

Qiao, R. (2019). *Yelp Review Rating Prediction : Sentiment Analysis and the Neighborhood-Based Recommender*. (Ph.D. thesis, UCLA).

Rajput, S. & Arora, A. (2013). Designing spam model-classification analysis using decision trees. *International Journal of Computer Applications*, 75(10), 6–12.

Resnick, P. & Zeckhauser, R. (2002). Trust among strangers in Internet transactions : Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce* (pp. 127–157). Emerald Group Publishing Limited.

Resnick, P., Kuwabara, K., Zeckhauser, R. & Friedman, E. (2000a). Reputation Systems. *Commun. ACM*, 43(12), 45–48. doi : 10.1145/355112.355122.

Resnick, P., Zeckhauser, R., Friedman, E. & Kuwabara, K. (2000b). Reputation systems. *Communications of the ACM*, 43(12), 45–45.

Reyes-Menendez, A., Saura, J. R. & Martinez-Navalon, J. G. (2019). The impact of e-WOM on Hotels Management Reputation : Exploring TripAdvisor Review Credibility with the ELM model. *IEEE Access*.

Saini, N. K., Sihag, V. K. & Yadav, R. C. (2014). A reactive approach for detection of collusion attacks in P2P trust and reputation systems. *2014 IEEE International Advance Computing Conference (IACC)*, pp. 312–317.

Sänger, J. & Pernul, G. (2014). Reusability for trust and reputation systems. *IFIP International Conference on Trust Management*, pp. 28–43.

Shankar, S. & Lin, I. (2011). Applying machine learning to product categorization. *Department of Computer Science, Stanford University*.

Shoukry, A. & Rafea, A. (2012). Sentence-level Arabic sentiment analysis. *2012 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 546–550.

Singh, V., Piryani, R., Uddin, A. & Waila, P. (2013). Sentiment analysis of Movie reviews and Blog posts. *2013 3rd IEEE International Advance Computing Conference (IACC)*, pp. 893–898.

Song, Y., Huang, J., Zhou, D., Zha, H. & Giles, C. L. (2007). Iknn : Informative k-nearest neighbor pattern classification. *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 248–264.

Sun, Y. & Liu, Y. (2012). Security of online reputation systems : The evolution of attacks and defenses. *IEEE Signal Processing Magazine*, 29(2), 87–97.

Sussin, J. & Thompson, E. (2012). The consequences of fake fans,'Likes' and reviews on social networks. *Gartner Research*, 2091515.

Swamynathan, G., Almeroth, K. C. & Zhao, B. Y. (2010). The design of a reliable reputation system. *Electronic Commerce Research*, 10(3-4), 239–270.

Tan, Y., Zhang, M., Liu, Y. & Ma, S. (2016). Rating-Boosted Latent Topics : Understanding Users and Items with Ratings and Reviews. *IJCAI*, 16, 2640–2646.

Tavakolifard, M. (2012). On some challenges for online trust and reputation systems.

Tavakolifard, M. & Almeroth, K. C. (2012). A taxonomy to express open challenges in trust and reputation systems. *Journal of Communications*, 7(7), 538–551.

Thomas, B. (2013). What Consumers Think about brands on social media, and what bunesses need to do about it. *Report, Keep Social Honest*.

Tian, N., Xu, Y., Li, Y., Abdel-Hafez, A. & Josang, A. (2014a). Generating product feature hierarchy from product reviews. *International Conference on Web Information Systems and Technologies*, pp. 264–278.

Tian, N., Xu, Y., Li, Y., Abdel-Hafez, A. & Jøsang, A. (2014b). Product Feature Taxonomy Learning based on User Reviews. *WEBIST (2)*, pp. 184–192.

Vinodhini, G. & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining : a survey. *International Journal*, 2(6), 282–292.

Wen, J. & Li, Z. (2007). Semantic smoothing the multinomial Naive Bayes for biomedical literature classification. *2007 IEEE International Conference on Granular Computing (GRC 2007)*, pp. 648–648.

Wu, G., Greene, D., Smyth, B. & Cunningham, P. (2010). Distortion as a validation criterion in the identification of suspicious reviews. *Proceedings of the First Workshop on Social Media Analytics*, pp. 10–13.

Xia, R., Zong, C. & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6), 1138–1152.

Xu, G., Cao, Y., Zhang, Y., Zhang, G., Li, X. & Feng, Z. (2016). TRM : computing reputation score by mining reviews. *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Yessenalina, A., Yue, Y. & Cardie, C. (2010). Multi-level structured models for document-level sentiment classification. *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 1046–1056.

Zhang, D., Zhou, L., Kehoe, J. L. & Kilic, I. Y. (2016). What online reviewer behaviors really matter ? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, 33(2), 456–481.

Zhang, S. (2010). KNN-CF Approach : Incorporating Certainty Factor to kNN Classification. *IEEE Intelligent Informatics Bulletin*, 11(1), 24–33.

Zhang, Y., Li, Q. & Lin, Z. (2010). A novel reputation computing model. *International Conference on Trusted Systems*, pp. 316–325.

Zhou, H. & Song, F. (2015). Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model. *The Twenty-Eighth International Flairs Conference*.