An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators

Malika Ben Khalifa, Zied Elouedi, Eric Lefevre

An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators

Malika BEN KHALIFA
Université de Tunis,
Institut Supérieur de gestion, LARODEC
Tunis, Tunisia
malikabenkhalifa2@gmail.com

Zied ELOUEDI
Université de Tunis,
Institut Supérieur de gestion, LARODEC
Tunis, Tunisia
zied.elouedi@gmx.fr

Eric LEFEVRE
Univ. Artois, EA 3926,
LGI2A,
62400 Béthune, France
eric.lefevre@univ-artois.fr

***Abstract*—The e-reputation is the key factor for the success of different companies and organizations. It is mainly influenced by the online reviews that have an important impact on the company's development. In fact, they affect the buying decision of the customer. Due to this attraction, the spammers post deceptive reviews to deliberately mislead the potential customers. Thus, the spammer detection becomes crucial to control the fake reviews, to protect the e-commerce from the fraudsters' activities and to ensure an equitable online competition. In this way, we propose a novel method based on the K-nearest neighbor algorithm within the belief function theory to handle the uncertainty involved by the suspicious behaviors' indicators. Our method relies on several spammers indicators used as features to perform the distinguishing between innocent and spammer reviewers. To evaluate our method performance and robustness, we test our approach on two large real-world labeled datasets extracted from yelp.com.**

***Index Terms*—Spammer detection, Online reviews, Fake reviews, Uncertainty, Classification, E-commerce.**

I. Introduction

Nowadays, internet gives the opportunity to people everywhere in the world to express and share their opinions and attitudes regarding products or services. These opinions called online reviews become one of the most important source of information thanks to their availability and visibility. They are increasingly used by both consumers and organizations. Positive reviews usually attract new customers and bring financial gain. However, negative ones damage the e-reputation of different business which lead to a loss. Reviewing has changed the face of marketing in this new area. Due to their important impact, companies invest money to overqualify their product to gain insights into readers preferences. For that, they rely on spammers to usually post deceptive reviews; positive ones to attract new customers and negative ones to damage the competitors' e-reputation. These fraudulent activities are extremely harmful for both companies and readers. Hence, detecting and analyzing the opinion spam becomes pivotal to save the e-commerce and to ensure trustworthiness and equitable competition between different products and services. Therefore, different researchers have given a considerable attention to this challenging problem. In fact, several researches [**?**], [**?**], [**?**], [**?**], [**?**] have been devoted to develop performing method capable of spotting fake reviews and stopping these misleading actions. These approaches can be classified into three global categories; spam review detection based on the reviews contents and linguistic features, group spammer detection based on the relational indicators and spammer detection.

Since spammers are the chief responsible of the appearance of deceptive reviews, spotting them is surly one of the most essential task in this field. Several approaches addressed this problem [**?**] and succeed to achieve significant results. The spammer detection techniques can be divided into two global categories; graph based method and behaviors indicators based methods.

One of the first studies that relies on the graph representation to detect fake reviews was proposed in [**?**]. This method attempted a spot of fake reviewers and reviews of online stores. This approach is based on a graph model composed by three types of nodes which are reviewers, reviews and stores. The spamming clues are composed through the interconnections and the relationships between nodes. The detection of these clues is based on the trustiness of reviewers, the honesty of reviews and the reliability of stores. Thanks to these three measures the method generates a ranking list of spam reviews and reviewers. This method was tested on real dataset extracted from resellerratings.com and labeled by human experts judged. However, the accuracy of this method is limited to 49%. Similar study was proposed in [**?**] based also on the review graph model. This method generates a suspicion score for each node in the review graph and updates these scores based on the graph connectivity using an iterative algorithm. This method was performing using a dataset labeled through human judgment. Moreover, the third graph related approach was introduced by [**?**] as an unsupervised framework. This method relies on a bipartite network composed by reviewers and products. The review can be positive or negative according to the rating. The method assumes that the spammers usually write positive reviews for a bad products and negative ones for good quality products. The authors use an iterative propagation algorithm as well as the correlations between nodes and assign a score to each vertex and update it using the loopy belief propagation (LBP). This method offers a list of scores to rank reviewers and products in order to get k clusters. Results were compared to two iterative classifiers, where they have shown performance.

The aspect of the behaviors indicators was introduced by [**?**]

to detect spammers. This method measures the spamming behaviors and accord a score to rank reviewers regarding the rating they give. It is essentially based on the assumption that fake reviewers target specific products and that their reviews rating deviates from the average rating associated to these products. Authors assume that this method achieved significant results. Another method proposed in [**?**] is based also on the rating behavior of the each reviewer. It focuses on the gap between the majority of the given rating and each reviewer's rating. This method uses the binomial regression to identify spammers. One of the most preferment studies was proposed by [**?**], which is essentially based on various spammers behavioral patterns. Since the spammers and the genuine reviewers display distinct behaviors, the proposed method models each reviewer's spamicity while observing his actions. It was formulated as an unsupervised clustering problem in a Bayesian framework. The proposed technique was tested on data from Amazon and proves its effectiveness. Moreover, authors in [**?**] proposed a method to detect the brust pattern in reviews given to some specific products or services. This approach generates five new spammer behavior indicators to enhance the review spammer detection. The authors used the Markov random fields to model the reviewers in brust and a hidden node to model the reviewer spamicity. Then, they rely on the loopy belief propagation framework to spot spammers. This method achieves 83.7% of precision thanks to the spammers behaviors indicators. Since then, behavioral indicators have become an important basis for spammer detection task. These indicators are used in several recent researches [**?**]. Nevertheless, we believe that the information or the reviewers' history can be imprecise or uncertain. Also, the deceptive behavior of users might be due to some coincidence which make the spammer detection issue full of uncertainty. For these reasons, ignoring such uncertainty may deeply affect the quality of the detection. To manage these concerns, we propose a novel method aims to classify reviewers into spammer and genuine ones based on K-nearest neighbors' algorithm within the Belief function theory to deal with the uncertainty involved by the spammer behaviors indicators which are considered as features. It is known as the richest theory in dealing with all the levels of imperfections from total ignorance to full certainty. In addition, it allows us to manage different pieces of evidence, not only to combine them but also to make decision while facing imprecision and imperfections. This theory prove its robustness in this field through our previous methods which achieve significant results [**?**], [**?**], [**?**], [**?**]. Furthermore, the use of the Evidential K-NN has been based on its robustness in the real world classification problems under uncertainty. We seek to involve imprecision in the spammers behaviors indicators which are considered as the fundamental interest in our approach since they are used as features for the Evidential K-NN. In such way, our method distinguishes between spammers and innocents reviewers while offering an uncertain output which is the spamcity degree related to each user.

This paper is structured as follows: In the first section, we present the basic concepts of the belief function theory and the evidential K-nearest neighbors, then we elucidate the proposed method in section 2. Section 3 is consacred for the experimental results and we finish with a conclusion and some future work.

II. Belief function theory

In section, we elucidate the fundamentals of the belief function theory as well as the Evidential K-nearest neighbors classifier.

A. Basics

The belief function theory, called also the Dempster Shafer theory, is one of the powerful theories that handles uncertainty in different tasks. It was introduced by Shafer [**?**] as a model to manage beliefs.

1) Basic concepts: In this theory, a given problem is represented by a finite and exhaustive set of different events called the frame of discernment Ω. 2^Ω is the power set of Ω that includes all possible hypotheses and it is defined by: $2^\Omega = \{A : A \subseteq \Omega\}$.

A basic belief assignment (bba) or (a belief mass) represents the degree of belief given to an element A. It is defined as a function $m^\Omega$from 2^Ω to $[0, 1]$ such that:

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{1}$$

A focal element A is a set of hypotheses with positive mass value $m^\Omega(A) > 0$.

Several types of bba's have been proposed [**?**] in order to model special situations of uncertainty. Here, we present some special cases of bba's:

- The certain bba represents the state of total certainty and it is defined as follows: $m^\Omega(\{\omega_i\}) = 1$ and $\omega_i \in \Omega$.
- The categorical bba has a unique focal element A different from the frame of discernment defined by: $m^\Omega(A) = 1$, $\forall A \subset \Omega$ and $m^\Omega(B) = 0$, $\forall B \subseteq \Omega\ B \neq A$.
- Simple support function: In this case, the bba focal elements are $\{A, \Omega\}$. A simple support function is defined as the following equation:

$$m^\Omega(X) = \begin{cases} w & \text{if } X = \Omega \\ 1 - w & \text{if } X = A \text{ for some } A \subset \Omega \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where A is the focus and $w \in$ [0,1].

2) Belief function: The belief function, denoted bel, includes all the basic belief masses given to the subsets of A. It quantifies the total belief committed to an event A by assigning to every subset A of Ω the sum of belief masses committed to every subset of A.

bel is represented as follows:

$$bel(A) = \sum_{\varnothing \neq B \subseteq \Omega} m^\Omega(B) \tag{3}$$

$$bel(\emptyset) = 0 \tag{4}$$

3) Plausibility function: The plausibility function, denoted pl, calculates the maximum amount of belief that could be provided to a subset A of the frame of discernment Ω. Otherwise, it is equal to the sum of the bbm's relative to subsets B compatible with A.

$$pl(A) = \sum_{A \cap B \neq \varnothing} m^{\Omega}(B) \tag{5}$$

4) Combination Rules: Various numbers of combination rules have been proposed in the framework of belief functions to aggregate a set of bba's provided by pieces of evidence from different experts. Let m_1^{Ω} and m_2^{Ω} two bba's modeling two distinct sources of information defined on the same frame of discernment Ω. In what follows, we elucidate the combination rules related to our approach.

1) $Conjunctive\ rule$: It was settled in [**?**], denoted by $\bigcirc\!\!\!\!\cap$ and defined as:

$$m_1^{\Omega} \bigcirc\!\!\!\!\cap m_2^{\Omega}(A) = \sum_{B \cap C = A} m_1^{\Omega}(B) m_2^{\Omega}(C) \tag{6}$$

2) $Dempster's\ rule\ of\ combination$: This combination rule is a normalized version of the conjunctive rule [**?**]. It is denoted by $\oplus$ and defined as:

$$m_1^{\Omega} \oplus m_2^{\Omega}(A) = \begin{cases} \frac{m_1^{\Omega} \bigcirc\!\!\!\!\cap m_2^{\Omega}(A)}{1 - m_1^{\Omega} \bigcirc\!\!\!\!\cap m_2^{\Omega}(\emptyset)} & \text{if} \quad A \neq \emptyset, \forall A \subseteq \Omega, \\ 0 & otherwise. \end{cases} \tag{7}$$

5) Decision process: The belief function framework provides numerous solutions to make decision. Within the Transferable Belief Model TBM [**?**], the decision process is performed at the pignistic level where $bba's$ are transformed into the pignistic probabilities denoted by $BetP$ and defined as:

$$BetP(B) = \sum_{A \subseteq \Omega} \frac{|A \cap B|}{|A|} \frac{m^{\Omega}(A)}{(1 - m^{\Omega}(\emptyset))} \quad \forall\ B \in \Omega \tag{8}$$

B. Evidential K-Nearest neighbors

The Evidential K-Nearest Neighbors (EKNN) [**?**] is one of the best known classification methods based in the belief function framework. It performs the classification over the basic crisp KNN method thanks to its ability to offer a credal classification of the different objects. This credal partition provides a richer information content of the classifier's output.

Notations

- $\Omega = \{C_1, C_2, ..., C_N\}$: The frame of discernment containing the N possible classes of the problem.
- $X_i = \{X_1, X_2, ..., X_m\}$: The object X_i belonging to the set of m distinct instances in the problem.
- A new instance X to be classified.
- $N_K(X)$: The set of the K-Nearest Neighbors of X.

EKNN method

The main objective of the EKNN is to classify a new object X based on the information given by the training set. A new instance X to be classified must be allocated to one class of the $N_K(X)$ founded on the selected neighbors. Nevertheless, the knowledge that a neighbor X belongs to class C_q may be deemed d as a piece of evidence that raises the belief that the object X to be classified belongs to the class C_q. For this reason, the EKNN technique deals with this fact and treats each neighbor as a piece of evidence that support some hypotheses about the class of the pattern X to be classified. .In fact, the more the distance between X and X_i is reduces, the more the evidence is strong. This evidence can be illustrated by a simple support function with a bba such that:

$$m_{X,X_i}(\{C_q\}) = \alpha_0 \exp^{-(\gamma_q^2 d(X,X_i)^2)} \tag{9}$$

$$m_{X,X_i}(\Omega) = 1 - \alpha_0 \exp^{-(\gamma_q^2 d(X,X_i)^2)} \tag{10}$$

Where;

- α_0 is a constant that has been fixed in 0.95.
- $d(X, X_i)$ represents the Euclidean distance between the instance to be classified and the other instances in the training set.
- γ_q assigned to each class C_q has been defined as a positive parameter. It represents the inverse of the mean distance between all the training instances belonging to the class C_q.

After the generation of the different $bba's$ by the K-nearest neighbors, they can be combined through the Dempster combination rule as follows:

$$m_X = m_{X,X_1} \oplus ... \oplus m_{X,X_K} \tag{11}$$

where $\{1, ..., K\}$ is the set including the indexes of the K-Nearest Neighbors.

III. Proposed Method

The idea behind our method is to take into account the uncertain aspect in order to improve detecting the spammer reviewers. For that, we propose a novel approach based on different spammers indicators and we rely on the Evidential K-nearest neighbors which is famous classifier under the belief function framework. In the remainder of this section we will elucidate the different steps of our proposed approach; in the first step we model and calculate the spammers' indicators through the reviewers' behaviors. In the second step, we present the initialization phase. Moreover, the learning phase is detailed in the third step. Finally, we distinguish between the spammers and the innocent reviewers through the classification phase in which we also offer an uncertain input to report the spamicity degree of each reviewer. Figure **??** illustrates our method steps.

A. Step1: Pre-processing phase

As mentioned before, the spammers indicators become one of the most powerful tool in the spammers detection field used in several researches. In this part, we propose to control the reviewers behaviors if they are linked with the spamming activities and thus can be used as features to learn the Evidential KNN classifier in order to distinguish between the two classes spammer and innocent reviewers. We select the significant features used in the previous work [**?**]. Here, we detail them in two lists; in the first list we elucidate the author

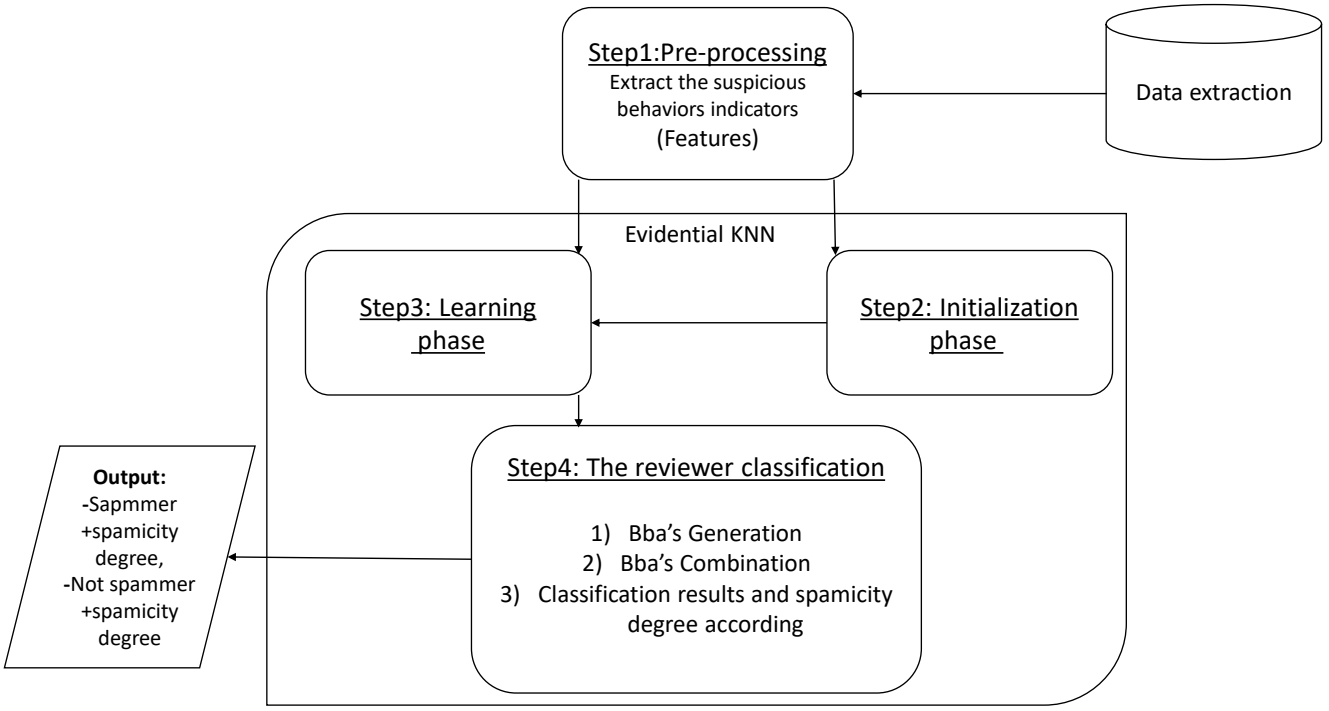

Fig. 1. Our method illustration

features and the second one presents the review features. To make the equations more comprehensible we present the different notations in the table **??**.

Reviewers features: The values of these features are into the interval [0,1]. The more the value is close to 1 the higher the spamicity degree is indicated.

1) Content similarity (CS): Generally, spammers choose to copy reviews from other similar products because, for them, creating a new review is considered as an action that required time. That's why, we assume that it is very useful to detect the reviews' content similarity (using cosine similarity) of the same reviewer. From this perspective and in order to pick up the most unpleasing behavior of spammers, we use the maximum similarity.

$$f_{CS}(R_i) = max_{R_i(r_j),R_i(r_k)\in R_i(Tr)} cosine(r_j, r_k) \quad (12)$$

Where $R_i(r_j)$ and $R_i(r_k)$ are the reviews written by the reviewer R_i, and $R_i(Tr)$ represents all the reviews written by the reviewer R.

2) Maximum Number of Reviews (MNR): Creating reviews and posting them successively in one day display an indication of a deviant behavior. This indicator calculates the maximum number of reviews per day for a reviewer normalized by the maximum value for our full data.

$$f_{MNR}(R_i) = \frac{MaxRev(R_i)}{Max_{R_i\in R_i(Tr)} MaxRev(R_i)} \quad (13)$$

3) Reviewing Burstiness (BST): Although authentic reviewers publish their reviews from their accounts occasionally, the opinion spammers represent a non-old-time membership in the site. To this point, it makes us able to take advantage of the account's activity in order to capture the spamming behavior. The activity window, which is the dissimilarity between the first and last dates of the review creation, is used as a definition of the reviewing burstiness. Consequently, if the time-frame of a posted reviews was reasonable, it could mention a typical activity. Nevertheless, posting reviews in a short and nearby burst (τ= 28 days, estimated in [**?**]), shows an emergence of a spam behavior.

$$f_{BST}(R_i) = \begin{cases} 0 & L(R_i(r)) - F(R_i(r)) > \tau \\ \frac{L(R_i(r)) - F(R_i(r))}{\tau} & Otherwise \end{cases} \quad (14)$$

Where $L(R_i(r))$ represents the last posting date of the review r given by the reviewer R_i and $F(R_i(r))$ is first posting date of the review.

4) Ratio of First Reviews (RFR): To take advantage of the reviews, people lean on the first posted reviews. For this reason, spammers tend to create them at an early stage in order to affect the elementary sales. Therefore, spammers believe that managing the first reviews of each product could empower them to govern the people's sentiments. For every single author, we calculate the ratio between the first reviews and the total reviews. We mean by the first reviews those posted by the author as the first to evaluate the product.

$$f_{RFR}(R) = \frac{|R_i(r_f) \in R_i(Tr)|}{R_i(Tr)} \quad (15)$$

TABLE I
LIST OF NOTATION

R_i	A reviewer
r	A review
p	A product
Tr	Total number of reviews
$R_i(r)$	Review written by the reviewer R_i
$R_i(Tr)$	Total number of reviews written by the reviewer R_i
$Tr(p)$	Total number of reviews on product or service p
$r(p)$	Review on product p
$R_i(r(p))$	Review given by the reviewer R_i to the same product p
$R_i(Tr(p))$	Total number of reviews given by the reviewer R_i to the same product p
$R_i(Tr_*(p))$	Total number of rating reviews given by the reviewer R_i to the same product p
$L(R_i(r))$	Last posting date of the review written by the reviewer R_i
$F(R_i))$	First posting date of the review written by the reviewer R_i
$A(p)$	The date of the product launch
I_i	Spamming indicator
S_{mean}	The mean score of a given product
S	The reviewing score of the reviews given to one product p by the same reviewer R_i.
$\Omega = \{S, \bar{S}\}$	The frame of discernment including the spammer and not spammer class

Where $R_i(r_f)$ represents the first review of the reviewer R_i.
Review features: These features have a binary values. If the feature value is equal to 1, then it indicates the sapmming. If not, it represents the non-spamming.

5) Duplicate/Near Duplicate Reviews (DUP): As far as they want to enhance the ratings, spammers frequently publish multiple reviews. They tend to use a duplicate/near-duplicate kind of preceding reviews about the same product. We could spotlight this activity by calculating the duplicate reviews on the same product. The calculation proceeding is as following:

$$f_{DUP}R_i(r)) = \begin{cases} 1 & r \in R_i(Tr(p)) = cosine(R_i(r), r) > \beta_1 \\ 0 & otherwise \end{cases} \tag{16}$$

For a review r each author R_i on a product p acquires as value 1 if it is in analogy (using cosine similarity based on some threshold, $\beta_1 = 0.7$) with another review is estimated in [**?**].

6) Extreme Rating (EXT): In favor of bumping or boosting a product, spammers often review it while using extreme ratings (1* or 5*). We have a rating scale composed by 5 stars (*).

$$f_{EXT}(R) = \begin{cases} 1 & R_i(Tr_*(p)) \in \{1, 5\} \\ 0 & R_i(Tr_*(p)) \in \{2, 3, 4\} \end{cases} \tag{17}$$

Where $R_i(Tr_*(p))$ represents all the reviews (ratings) given by the reviewer R_i to the same product p.

7) Rating Deviation: Spammers aim to promote or demote some target products or services to this point they generate reviews or rating values according the situation. In order to deviate the overall rating of a product, they have to contradict the given opinion by posting deceptive ratings strongly deviating the overall mean.

If the rating deviation of a review exceeds some threshold $\beta 2 = 0.63$ estimated in [**?**], this features achieves the value of 1. The maximum deviation is normalized to 4 on a 5-star scale.

$$f_{Dev}(R) = \begin{cases} 1 & \frac{|S - S_{mean}|}{4} > \beta_2 \\ 0 & otherwise \end{cases} \tag{18}$$

Where S_{mean} represents the mean score of a given product and S represents the reviewing score of the reviews given to one product p by the same reviewer R_i.

8) Early Time Frame (ETF): Since the first review is considered as a meaningful tool to hit the sentiment of people on a product, spammers set to review at an early level in order to press the spam behavior. The feature below is proposed as a way to detect the spamming characteristic:

$$ETF(r, p) = \begin{cases} 0 & L(R_i, p) - A(p) > \delta \\ 1 - \frac{L(R_i,p) - A(p)}{\delta} & otherwise \end{cases} \tag{19}$$

$$f_{ETF}(r) = \begin{cases} 1 & ETF(R_i, R_i(r(p))) > \beta_3 \\ 0 & otherwise \end{cases} \tag{20}$$

Where $L(R_i, p)$ represents the last review posting date by the reviewer R_i on the product p and $A(p)$ is the date of the product launch. The degree of earliness of an author R_i who had reviewed a product p is captured by $ETF(R_i, R_i(r(p)))$ the threshold symbolizing earliness is about $\delta = 7$ months (estimated in [**?**]). According the presented definition, we cannot consider the last review as an early one if it has been posted beyond 7 months since the product's launch. On the other hand, the display of a review following the launch of the product allows this feature to reach the value of 1. $\beta_3 = 0.69$

is considered as the threshold mentioning spamming and is estimated in [**?**].

9) Rating Abuse (RA): To bring up the wrongly use generated from the multiple ratings we adopt the feature of Rating Abuse (RA). Obtaining Multiple rating on a unique product is considered as a weird behavior. Despite the fact that this feature is alike to DUP, it does not focus on the content but rather it targets the rating dimension. As definition, the Rating Abuse, the similarity of the donated ratings by an author for a product beyond multiple ratings by the same author blended by the full reviews on this product.

$$RA(R_i, R_i(r(p)) = |R_i(Tr(p))|(1 - \frac{1}{4} max_{r \in R_i(Tr(p))}(r,p) - min_{r \in R_i(Tr(p))}(r,p)) \quad (21)$$

$$f_{RA} = \begin{cases} 1 & RA(R_i, R_i(r(p)) > \beta_4 \\ 0 & otherwise \end{cases} \quad (22)$$

We should calculate the difference between the two extremes (maximum/minimum) on 5-star scale rating to catch the coherence of high/low rating and to determine the similarity of multiple star rating. The maximum difference between ratings attains as normalized constant 4. Lower values are reached by this feature if, in authentic cases, the multiple ratings where in change (as a result of a healthy use). $\beta_4 = 2.01$ is considered as the threshold mentioning spamming and is estimated in [**?**].

B. Step2: Initialization phase

In order to apply the Evidential K-NN classifier, we should firstly assign values to parameters α_0 et γ_0 to be used in the learning phase. We will start by initializing the parameter α_0 and then computing the second parameter γ_{Ii} while exploiting the reviewer-item matrix. As mentioned in the EKNN procedure [**?**], the α_0 is initialized to 0.95. The value of the parameter α_0 is assigned only one time while the γ_{Ii} value change each time according to the current items' reviewers. In order ensure the γ_{Ii} computation performance, first of all we must find reviewers having separately exclusive spammers indicators. Based on the selected reviewers, we assign a parameter γ_{Ii} to each indicators I_i corresponding to the reviewer R_i which will be measured as the inverse of the average distance between each pair of reviewers R_i and R_j having the same spammers' indicators values. This calculation is based on the Euclidean distance denoted $d(R_i, R_j)$ such that:

$$d(R_i, R_j) = \sqrt{\sum_{i,j=1}^{n} (I_{(R,i)} - I_{(R,j)})^2} \quad (23)$$

Where $I_{(R,i)}$ and $I_{(R,j)}$ correspond to the value of the spammer indicators of the reviewer R to the indicators i and j.

C. Step3: Learning phase

Once the spammers indicators are calculated and the two parameters α_0 and γ_{Ii} have been assigned, we must select a set of reviewers. Then, we compute for each reviewer R_j in the database, its distance with the target reviewer R_i. Given a target reviewer, we have to spot its K-most similar neighbors, by selecting only the K reviewers having the smallest distances values that is calculated using the Euclidean distance and denoted by $dist(R_i, R_j)$.

D. Step4: Classification phase

In this step, we aim to classify a new reviewer into spammer or innocent reviewer. Let $\Omega = \{S, \bar{S}\}$ where S represents the class of the spammers reviewers and $\bar{S}$ includes the class of the not spammers (genuine) reviewers.

1) The bba's generation: Each reviewer R_I induces a piece of evidence that builds up our belief about the class that he belongs. However, this information does not supply certain knowledge about the class. In the belief function framework, this case is shaped by simple support functions, where only a part of belief is committed to $\omega_i \in \Omega$ and the rest is assigned to Ω. Thus, we obtain the following bba:

$$m_{R_i,R_j}(\{\omega_i\}) = \alpha_{R_i} \quad (24)$$

$$m_{R_i,R_j}(\Omega) = 1 - \alpha_{R_i} \quad (25)$$

Where R_i is the new reviewers and R_j is its similar reviewer that $j = \{1..K\}$, $\alpha_{R_i} = \alpha_0 \exp^{(-\gamma_{Ii} dist(R_i,R_j))}$, α_0 and γ_{Ii} are two parameters assigned in the initialization phase and $dist(R_i, R_j)$ is the distance between the two reviewers R_i and R_j computed in the learning phase.
In our case, each neighbor of the new reviewer has two possible hypotheses. It can be similar to a spammer reviewer in which his the committed belief is allocated to the spammer class S and the rest to the frame of discernment Ω. In the other case, it can be near to an innocent reviewer where the committed belief is given to the not spammer class $\bar{S}$ and the rest of is assigned to Ω. We treat the K-most similar reviewers as independent sources of evidence where each one is modeled by a basic belief assignment. Hence, K different bba's can be generated for each reviewer.

2) The bba's combination: After the generation of the bba's for each reviewer R_i, we describe how to aggregate these bba's in order to get the final belief about the reviewer classification. Under the belief function framework, such bba's can be combined using the Dempster combination rule. Therefore, the obtained bba represent the evidence of the K-nearest Neighbors regarding the class of the reviewer. Hence, this global mass function m is obtained as such:

$$m_{R_i} = m_{R_i,R_1} \oplus m_{R_i,R_2} \oplus .... \oplus m_{R_i,R_K} \quad (26)$$

3) Final classification result and the spamicity degree according: We apply the pignistic probability $BetP$ in order to select the membership of the reviewer R_i to one of the classes of Ω and to accord him a spamicity degree. Then, the classification decision is made either the reviewer is a spammer or not. For this, we select the $BetP$ with the grater value. Moreover, we assign to each reviewer even he is not a spammer the spamicity degree which consists on the $BetP$ value of the spammer class.

TABLE II
DATASETS DESCRIPTION

Datasets	Reviews (filtered %)	Reviewers (Spammer %)	Services (Restaurant or hotel)
YelpZip	608,598 (13.22%)	260,277 (23.91%)	5,044
YelpNYC	359,052 (10.27%)	160,225 (17.79%)	923

TABLE III
COMPARATIVE RESULTS

Evaluation Criteria	Accuracy				Precision				Recall			
Methods	NB	SVM	UCS	Our method	NB	SVM	UCS	Our method	NB	SVM	UCS	Our method
YelpZip	60%	65%	78%	**84%**	57%	66%	76%	**85%**	63%	68%	74%	**86%**
YelpNYC	61%	68%	79%	**85%**	62%	69%	79%	**86%**	61.8%	67.8%	76.7%	**83.6%**

IV. EXPERIMENTATION AND RESULTS

The evaluation in the fake reviews detection problem was always a challenging issue due to the unavailability of the true real world growth data and variability of the features also the classification methods used by the different related work which can lead to unsafe comparison in this field.

Data description

In order to test our method performance, we use two datasets collected from yelp.com. These datasets represent the more complete, largest, the more diversified and general purpose labeled datasets that are available today for the spam review detection field. They are labeled through the classification based on the yelp filter which has been used in various previous works [**?**], [**?**], [**?**], [**?**], [**?**] as ground truth in favor of its efficient detection algorithm based on experts judgment and on various behavioral features. Table **??** introduces the datasets content where the percentages indicate the filtered fake reviews (not recommended) also the spammers reviewers.

The YelpNYC dataset contains reviews of restaurants located in New York City; the Zip dataset is bigger than the YelpNYC datasets, since it includes businesses in various regions of the U.S., such that New York, New Jersey, Vermont, Connecticut and Pennsylvania. The strong points of these datasets are:

- The high number of reviews per user, which facilities to modeling of the behavioral features of each reviewer.
- The miscellaneous kinds of entities reviewed, i.e., hotels and restaurants
- Above all, the datasets hold just fundamental information, such as the content, label, rating, and date of each review, connected to the reviewer who generated them. With regard to considering over-specific information, this allows to generalize the proposed method to different review sites.

Evaluation Criteria

We rely on these three following criteria to evaluate our method: Accuracy, precision and recall and they can be defined as Eqs.**??**, **??**, **??** respectively where TP, TN, FP, FN denote True Positive, True Negative, False Positive and False Negative respectively:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{27}$$

$$Precision = \frac{TP}{(TP + FN)} \tag{28}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{29}$$

Experimental results

As our method relies on the Evidential KNN classifier to classify the reviewer into spammer and genuine ones. We propose to compare our method with the Support Vector Machine (SVM) and the Naive Bayes (NB) used by most of spammer detection method [**?**], [**?**], [**?**]in this field. Moreover, we propose to compare also with our previous proposed Uncertain Classifier to detect Spammers (UCS) in [**?**]. Table **??** reports the different results.

Our method achieves the best performance detection according to accuracy, precision and recall over-passing the baseline classifier. We record at best an accuracy improvement over 24% in both yelpZip and yelpNYC data-sets compared to NB and over 19% compared to SVM. Moreover, the improvement records between our two uncertain methods (over 10%) at best, shows the importance of the variety of the features used in our proposed approach.

Our method can be used in several fields by different reviews websites. In fact, these websites must block the detected spammers in order to stop the appearance of the fake reviews. Moreover and thanks to our uncertain output which represent the spamicity degree for each reviewer, they can control the behavior of the genuine ones with a high spamicity degree to prevent their tendency to turn into spammers.

V. Conclusion

In this work, we tackle the spammer review detection problem and we propose a novel approach that aims to distinguish between the spammer and the innocent reviewers while taking into account the uncertainty in the different suspicious behavioral indicators. Our method shows its performance in detecting the spammers reviewers while according a spamicity degree to each reviewer. Our proposed approach can be useful for different reviews sites in various fields. Moreover, our uncertain input can be used by other methods to model the reliability each reviewer. As future work, we aim to tackle the group spammer aspect in the interest of improving the detection in this field.

References

[1] Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 13, 2-11 (2013)

[2] Bandakkanavar, RV., Ramesh, M., Geeta, H.: A survey on detection of reviews using sentiment classification of methods. IJRITCC, 2(2):310–314 (2014) Advanced Intelligent System and Informatics (AISI), 395-404 (2018)

[3] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Multiple criteria fake reviews detection based on spammers' indicators within the belief function theory. The 19th International Conference on Hybrid Intelligent Systems (HIS'2019). Springer International Publishing. (To appear)

[4] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Fake reviews detection based on both the review and the reviewer features under belief function theory. The 16th international conference Applied Computing (AC'2019), 123-130 (2019)

[5] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Spammers detection based on reviewers' behaviors under belief function theory. The 32nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE'2019). Springer International Publishing, 642-653 (2019)

[6] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Multiple criteria fake reviews detection using belief function theory. The 18th International Conference on intelligent systems design and applications (ISDA'2018). Springer International Publishing, 315-324 (2018)

[7] Deng, X., Chen, R.: Sentiment analysis based online restaurants fake reviews hype detection. Web Technologies and Applications, 1-10 (2014)

[8] Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat.38, 325-339 (1967)

[9] Denoeux, T.: A K-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. Syst. Man Cybern. 25(5), 804–813 (1995)

[10] Lefèvre, E., Elouedi, Z.: How to preserve the confict as an alarm in the combination of belief functions? Decis. Support Syst.56, 326-333 (2013)

[11] Fayazbakhsh, S., Sinha, J.: Review spam detection: A network-based approach. Final Project Report: CSE 590 (Data Mining and Networks) (2012)

[12] Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 13, 175-184 (2013)

[13] Fontanarava, J., Pasi, G., Viviani, M.: Feature Analysis for Fake Review Detection through Supervised Classification. Proceedings of the International Conference on Data Science and Advanced Analytics, 658-666 (2017).

[14] Heydari, A., Tavakoli, M., Ismail, Z., Salim, N.: Leveraging quality metrics in voting model based thread retrieval. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10 (1), 117-123 (2016)

[15] Jindal, N., Liu, B.: Opinion spam and analysis. Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 219-230 (2008).

[16] Jousselme, A.-L., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. Inf. Fusion 2(2), 91-101 (2001)

[17] Liu, P., Xu, Z., Ai, J., Wang, F.: Identifying Indicators of Fake Reviews Based on Spammers Behavior Features," IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 396–403 (2017)

[18] Lim, P., Nguyen, V., Jindal, N., Liu, B., Lauw, H. : Detecting product review spammers using rating behaviors. Proceedings of the 19th ACM international conference on information and knowledge management, 939-948 (2010)

[19] Ling, X., Rudd, W.: Combining opinions from several experts. Applied Artificial Intelligence an International Journal, 3 (4), 439-452 (1989)

[20] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M.: Spotting opinion spammers using behavioral footprints. Proceedings of the ACM international conference on knowledge discovery and data mining, 632-640 (2013)

[21] Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What Yelp Fake Review Filter Might Be Doing. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 409-418 (2013)

[22] Ong, T., Mannino, M., Gregg, D.: Linguistic characteristics of shill reviews. Electronic Commerce Research and Applications, 13 (2), 69-78 (2014)

[23] Pan, L., Zhenning, X., Jun, A., Fei, W.: Identifying indicators of fake reviews based on spammer's behavior features. Proceedings of the IEEE International Conference on Software Quality, Reliability and Security Companion, QRS-C, 396-403 (2017)

[24] Savage, D., Zhang, X., Yu, X., Chou, P., Wang, Q.: Detection of opinion spam based on anomalous rating deviation. Expert Systems with Applications, 42 (22), 8650-8657 (2015)

[25] Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD, 985-994 (2015)

[26] Shafer, G.: A Mathematical Theory of Evidence, vol. 1. Princeton University Press (1976)

[27] Smets, P.: The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Mach. Intell. 12(5), 447-458 (1990)

[28] Smets, P.: The transferable belief model for expert judgement and reliability problem. Reliability Engineering and system safety, 38, 59-66 (1992)

[29] Smets, P.: The canonical decomposition of a weighted belief. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1896-1901 (1995)

[30] Smets, P.: The transferable belief model for quantified belief representation. In: Smets, P. (ed.) Quantified Representation of Uncertainty and Imprecision, 267-301. Springer, Dordrecht (1998)

[31] Wang, G., Xie, S., Liu, B., Yu, P. S.: Review graph based online store review spammer detection. Proceedings of 11th international conference on data mining, ICDM, 1242-1247 (2011)