Decision Making over Multiple Criteria to Assess News Credibility in Microblogging Sites

Gabriella Pasi
Department of Informatics,
Systems, and Communication
University of Milano-Bicocca
Milan, Italy
0000-0002-6080-8170

Marco De Grandis
Department of Informatics,
Systems, and Communication
University of Milano-Bicocca
Milan, Italy
marcodegra@live.it

Marco Viviani
Department of Informatics,
Systems, and Communication
University of Milano-Bicocca
Milan, Italy
0000-0002-2274-9050

***Abstract*—Locating Web content useful to specific user needs and tasks concerns nowadays, in many circumstances, to assess the credibility of the content itself. With the diffusion of social media and the possibility for everyone to become a content generator, the problem of assessing information credibility has become a major research issue, in particular in microblogging sites, where fake news, hoaxes and other kinds of misinformation are diffused almost without any traditional form of trusted intermediation. In this paper, we propose an approach based on multiple criteria associated with news, on which the use of aggregation operators guided by linguistic quantifiers allow the modeling of the decision maker behavior into the news credibility assessment process. The operation and the evaluation of the approach are illustrated by considering the Twitter microblogging platform.**

***Index Terms*—Credibility, Microblogging, Fake News, Multi-Criteria Decision Making, Aggregation Operators, Linguistic Quantifiers.**

I. Introduction

The birth and progressive evolution of the World Wide Web (WWW) have made available to everyone a massive and distributed repository of heterogeneous data and *potential* information. This phenomenon has been further emphasised by the conception and implementation of Web 2.0 technologies, which allow every user to become a generator of content, the so-called *User-Generated Content* (UGC). In this scenario, UGC can be published and shared with peers through social media, without almost any traditional form of intermediated trusted control [1].

In this ‘disintermediation’ context, one of the big challenges related to the problem of locating on-line content useful to specific user needs and tasks, is how to assess the quality of the content itself and, consequently, how to retrieve only that at the highest standing. One important dimension of quality is *credibility*, which is a perceived characteristic of the information receiver, and which can be assessed by considering different aspects connected to the source that generated the content, the content itself, and the medium across which the content is diffused [2]. In the so-called *post-truth* era, a huge deal of research addresses the issue of discriminating in an automatic or semi-automatic way fake contents from genuine ones [3]. Tackling this issue is particularly important in microblogging platforms, which are nowadays a vehicle of fake news, hoaxes, conspiratorial theories, etc., thus strongly affecting the behavior of public opinion in real life.

In the literature, several solutions – mostly *data-driven* approaches based on *machine learning* (ML) – have been proposed to assess the credibility of news in microblogs [4]. In most cases, these approaches perform a binary classification of news items into genuine and fake. In a different perspective, in this article we focus on the presentation of a *model-driven* approach based on *Multi-Criteria Decision Making* (MCDM) to assess news credibility. The proposed solution relies on both several *criteria* (i.e., features providing evidence of credibility) related to the news items, and *prior domain knowledge*. The modeling of the above-mentioned aspects allows to compute an *overall credibility score* for each news item, which represents the satisfaction degree of a set of flexible constraints, obtained by aggregating the distinct credibility scores that represent how much each news item meets each criterion (constraint) from the point of view of credibility. Based on this overall score, either a *ranking* or a *binary classification* of news items can be provided. Compared to the solutions based on ML, the proposed MCDM-based approach allows incorporating the decision maker preferences in the credibility assessment process; as such, this can lead to a higher interpretability of the results, which is a benefit considering that credibility is a feature perceived by the information receiver [5].

The proposed approach takes inspiration from some prior works in the field of opinion spam detection [6], [7], but it addresses a different yet related problem, i.e., news credibility in microblogging sites, and explores the possibility of hybridizing the MCDM paradigm with some learning aspects. The approach is compared with multiple data-driven literature baselines to assess its effectiveness. To evaluation purposes, the Twitter microblogging site is taken into consideration.

II. Background and Related Works

On microblogging sites, users share information in the form of short textual messages, which can also contain figures, videos, and URLs. In this context, Twitter, in particular, has gained reputation as a prominent medium for *news* diffusion [8], given that the majority of *trending topics* on this platform

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

can be considered as *persistent news* [9]. Because of this aspect, and the impact that news diffused on microblogs has on the formation of the public opinion, being able to identify fake news on these platforms is of paramount importance. But what constitutes *fake news*? In the traditional journalism context, fake news refers to "articles that are intentionally and verifiably false, and could mislead readers" [10].

On-line, and in microblogs in particular, fake news refers to information related specifically to public news events that can be verified as false [4], [11]. In this context, fake news can be of various kinds: (i) *completely fake and large-scale hoaxes*, which is news deliberately fabricated or falsified in the mainstream or social media to deceive audience [12]; (ii) *humorous/satire news*, which relies on irony and humor, mimicking credible news stories [12]; (iii) *poorly written news articles*, which are constituted by statements presented as facts, without any verification of the sources and characterized by a mixture of subjective opinions and facts [4]; (iv) *conspiracy theories* [4]; (v) *misinformation*, which is constituted by news that the person diffusing it believes true, but which then turns out to be totally or partially fake [4]; (vi) *disinformation*, which is false information intentionally and deliberately spread by individuals [11]; (vii) *fake news automatically generated* by spam profiles, trolls and bots [13]. Several approaches have been proposed in the last years for detecting fake news (of type (i) in particular) in microblogging sites. In these works, a *news item* has been intended either as a *single post* (e.g., a tweet), or as a *thread of posts* (e.g., a set of tweets on the same topic), which represents a so-called *news event* [14].

Made the above premises, a general classification of approaches to news credibility assessment into two main categories can be done, namely: (i) *classification-based*, and (ii) *propagation-based* approaches. The latter are mainly concerned with studying the influence that *social bots* have on the dissemination of fake news [13], [15] and how low-credibility information spreads over the social network structure [16]–[19]. Classification-based approaches are either based on the use of external Knowledge Bases and Semantic Web technologies to represent news items as *facts*, focusing in particular on *automated fact checking* [20]–[23], or on the use of *multiple features* connected with news items to perform the classification task. The proposed approach follows this second strategy and, for this reason, only feature-based approaches belonging to category (i) will be detailed in the following.

Castillo *et al.* [24], [25] were among the first to tackle in a structured way the problem of news credibility on microblogging sites, Twitter in particular, by using classification-based approaches based on multiple features related to the textual content of tweets (linguistic features) and to their authors (behavioral features). In particular, they tested Bayesian methods, Logistic Regression, J48, Random Forests, and Meta Learning based on clustering, trained over labeled data obtained using crowdsourcing tools. Other classification-based approaches (mostly supervised or semi-supervised) are those described in [26]–[30]. Each of these solutions proposes different features (i.e., linguistic, behavioral, social, multimedia), machine learning algorithms, and evaluation datasets, depending on the considered problem, i.e., the assessment of the credibility of trending topics in Twitter [30], the identification of credible tweets during high-impact events [28], the detection of spammers [29] and troll profiles [27] in microblogging sites, the classification of credible versus non-credible multimedia tweets, i.e., accompanied by a multimedia item (image or video) from an event [26]. The work described in [14] considers a large set of credibility features (the same that are used in this paper) that are employed to automatically identify fake news in Twitter threads (disregarding multimedia content, which is out of the scope of this paper). The model proposed in [14] is trained over large-scale labeled datasets, including the one employed in this paper for evaluation purposes, i.e., CREDBANK [31].

III. MCDM and News Credibility Assessment

In this section, an approach to the assessment of news credibility is presented, which is based on the definition of a *model* based on *Multi-Criteria Decision Making* (MCDM). In an MCDM problem, there are usually a set of candidate solutions, i.e., *alternatives* that are available to a *decision maker* (DM), multiple *criteria* on which the alternatives are evaluated, and, possibly, distinct *importance weights* associated with each criterion. Solving an MCDM problem means to provide the decision maker with one or more optimal solutions (alternatives) complying to her/his preferences [32].

In the context considered in this paper, i.e., news credibility evaluation in microblogging sites, the alternatives are the considered *news items*, and the criteria to be assessed are related to the *credibility features* characterizing the news items. For each credibility feature, a numeric value is assessed, the so-called *performance score*, which can be interpreted as the *degree of credibility* of the news item with respect to that feature. These multiple *credibility scores* are subsequently *aggregated* to obtain an *overall credibility score* associated with the news item. Formally:

- $A = \{a_1, a_2, \dots, a_m\}$ is the set of *alternatives*, i.e., the *news items*;
- $C = \{c_1, c_2, \dots, c_n\}$ is the set of *criteria*, i.e., the *credibility features* characterizing each news item;
- s_i is the *satisfaction function* that, for a criterion c_i $(1 \leq i \leq n)$, returns the *performance score* $s_i(a_j) \in I$, $I = [0, 1]$, intended as the extent to which the alternative a_j $(1 \leq j \leq m)$ satisfies the criterion c_i (i.e., a *credibility score* in this context).

To obtain an overall performance score (i.e, an overall credibility score) σ_j for each alternative a_j, (i.e., each news item), the distinct performance scores (i.e., credibility scores) must be aggregated [33]. To this aim, an *aggregation operator* (AGOP) is applied; an AGOP is an n-ary function $\mathcal{A} : [0, 1]^n \rightarrow [0, 1]$, which is monotonic non decreasing with respect to each variable, and which satisfies the following boundary conditions: $\mathcal{A}(0, 0, \dots, 0) = 0$ and $\mathcal{A}(1, 1, \dots, 1) = 1$. Formally:

$$\sigma_j = \mathcal{A}(s_1(a_j), s_2(a_j), \dots, s_n(a_j)).$$

In the literature, several classes of aggregation operators have been employed to solve Multi-Criteria Decision Making problems, *averaging operators* in particular [32], [34]–[36]. A family of averaging aggregation operators that is of potential interest for the considered problem is that of *Ordered Weighted Averaging* (OWA) operators [37].

Definition 1: An *Ordered Weighted Averaging* (OWA) operator $\mathcal{A}_{\text{OWA}} : [0,1]^n \to [0,1]$ of dimension n has associated a weighting vector $W = [w_1, w_2, \ldots, w_n]$ such that $w_k \in [0,1]$ and $\sum_{k=1}^{n} w_k = 1$, where:

$$\mathcal{A}_{\text{OWA}}(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} w_k b_k, \tag{1}$$

in which b_k is the kth largest of the x_i.

OWA operators have the interesting possibility of allowing to guide the aggregation by the specification of *linguistic quantifiers* (e.g., *all*, *some*, *many*, etc.). This makes it possible to represent more easily the trade-off that the decision maker is leaning to accept among the satisfaction of the considered criteria, which lies between two borderline situations: (i) the situation in which the DM desires that *all* criteria are satisfied by the alternative, corresponding to the *min* operator, modeled by the following vector: $W_{min} = [0, 0, \ldots, 1]$, and (ii) the situation in which the satisfaction of *at least one* criterion is what the DM desires, corresponding to the *max* operator, modeled by the following vector: $W_{max} = [1, 0, \ldots, 0]$. Between these two extremes lie all averaging operators, among which the *arithmetic mean*, modeled by the following vector: $W_{am} = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]$.

OWA operators give the possibility of constructing the weighting vector based on the formal definition of a linguistic quantifier as a fuzzy subset. Specifically, the decision maker provides a linguistic quantifier Q indicating the number (*absolute* quantifier) or the proportion (*relative* quantifier) of criteria s/he believes should be satisfied to have a good solution. The procedure of generating the weighting vector W from a linguistic quantifier Q depends on its type. In this paper, *Regular Increasing Monotone* (RIM) relative quantifiers are considered,[1] such as *at least k%* and *most*. In the following, the formal procedure aimed at constructing the weighting vector associated with a RIM quantifier is shortly reported.

A. Equal Importance of Criteria

Starting form the definition of a RIM quantifier Q, the weights w_i of a weighting vector W of dimension n (n values to be aggregated) can be defined as follows:

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), \quad \text{for } i = 1, \ldots, n \tag{2}$$

Equation (2) allows to define the weighting vector W by assuming that all the considered criteria are *equally important* for the DM.

In real scenarios, it is often crucial to be able to discriminate the importance of the criteria that concur in a decision making process. For example, in the considered problem of assessing the credibility of news, not all the features connected with an information item are equally significant in terms of credibility. For this reason, is it necessary to consider unequal importance associated with features, as detailed in Section III-B.

B. Unequal Importance of Criteria

In [38], a way has been proposed for aggregating n scores by considering the *different importance* associated with the criteria that generated them. Let us consider an alternative a (i.e., a news item in the considered context) to be evaluated with respect to n criteria; the performance scores of a related to the n criteria are denoted by $x_1, x_2, \ldots, x_n$, each $x_i \in [0,1]$, while the numeric values denoting the importance of the n criteria are denoted by $V_1, V_2, \ldots, V_n$. In the reordering process of the x_i values, it is important to maintain the correct association between the values and the importance of the criteria that originated them. For this reason, u_j denotes the importance originally associated with the criterion that has the jth largest satisfaction degree. For example, assuming that x_5 is the highest value among the x_i values, thus $b_1 = x_5$ and $u_1 = V_5$. At this point, to obtain the weight w_j of the weighting vector with weighted criteria, it is possible to employ, for each alternative a, the following equation:

$$w_j = Q\left(\frac{\sum_{k=1}^{j} u_k}{T}\right) - Q\left(\frac{\sum_{k=1}^{j-1} u_k}{T}\right) \tag{3}$$

where $T = \sum_{k=1}^{n} u_k$ is the sum of the importance values u_js.

The weighting vector used in this aggregation *will generally be different* for each a, i.e., for each considered news item [37].

IV. Assessing News Credibility on Twitter

In this section, the proposed MCDM approach based on the use of OWA aggregation operators to perform news credibility assessment is illustrated on Twitter. After having identified and represented the *credibility features* to be employed in this context, different *aggregation functions* guided by distinct *linguistic quantifiers* are presented, which allow to vary the percentage of (important) features to be considered in the credibility assessment process, and to provide an overall credibility score associated with each news item. This allows to inject in the credibility assessment process the DM's preferences (i.e., depending on the way the aggregation function is defined and on which linguistic quantifier is based).

A. Features Identification and Representation

Several features have been proposed and employed in the literature for evaluating the credibility of news items on Twitter. In this paper, one of the most informative feature sets available today and illustrated in [14] is considered. It is composed of the following 15 features belonging to four macro-categories:

(i) *Structural features* [S], i.e., ***media count***: the number of tweets that contain media contents (images, videos,

[1] A linguistic quantifier is said to be a RIM quantifier if [38]: $Q(0) = 0$, $Q(1) = 1$, and $Q(r) \geq Q(s)$ if $r > s$ ($r, s \in [0,1]$).

etc.), ***mention count***: the number of tweets that contain mentions, ***URL count***: the number of tweets that contain URLs, ***retweet count***: the number of retweets for the news item, ***hashtag count***: the number of tweets that contain hashtags, ***status count***: the average number of tweets with respect to each user profile (in the thread), ***tweet count***: the number of tweets that contain only text (no media, mentions, hashtags or URLs);

(ii) *User-related features* [U], i.e., ***verified***: the number of verified profiles (in the thread), ***density***: the density of the network w.r.t. users (nodes) and their interactions (edges, i.e., mentions, replies, etc.), ***friends***, also known as *followees*: the average number of followees with respect to each user profile (in the thread), ***followers***: the average number of followers with respect to each user profile (in the thread);

(iii) *Content-related features* [C], i.e., ***polarity***: the average positive or negative feelings expressed by the tweets (in a thread), ***objectivity***: the score of whether a thread is objective or not;

(iv) *Temporal features* [T], i.e., ***ages***: the author account age w.r.t. to her/his first tweet, ***lifespan***: the minutes between the first and the last tweet of the thread.

Numerical values are associated with the above-mentioned features, but, since they represent different concepts, their values are expressed on different numerical scales. In the proposed MCDM approach, these values are transformed into meaningful performance scores (credibility scores) in the $[0, 1]$ interval, which will be subsequently aggregated in an overall performance score (overall credibility score) as explained in Section III. To do this, the *min-max* feature scaling function has been employed.[2] Formally:

$$s_i(a_j) = \frac{x_{i,j} - min(x_{i,h})}{max(x_{i,h}) - min(x_{i,h})} \tag{4}$$

where, for a news item a_j, $s_i(a_j)$ is the performance score associated with feature c_i, $x_{i,j}$ is the value of feature c_i for a_j, $h = 1, \ldots, m$, and m is the total number of news items. In the considered problem, the value '1' is assumed as the evidence of a full satisfaction in terms of credibility, and the value '0' as a complete dissatisfaction.[3]

B. Quantifier-guided Aggregation Functions

To aggregate the single credibility scores into an overall score for each news item, two distinct functions – denoted as (i) OWA_MORE, and (ii) OWA_MOST – have been defined. They are OWA operators guided by the *more than k%* and the *most* linguistic quantifiers, which are formally defined in the following. The choice of these quantifiers is motivated by the fact that, ideally, a decision maker would desire the fulfilment of *all* criteria. As illustrated in Section III, this corresponds to using the *min* operator. Usually, to tackle MCDM problems, aggregation functions lying between the minimum and the maximum are employed, since they allow to *compensate* low scores on some criteria by high scores on other criteria [42]. For this reason, we have considered those linguistic quantifiers that perform an aggregation on *the majority* of the criteria satisfied.[4]

According to [43], the *more than* $k\%$ quantifier, denoted as Q_{more}, can be defined as follows:

$$Q_{more}\left(r\right) = \begin{cases} 0 & \text{for } 0 < r \leq k \\ \frac{r-k}{1-k} & \text{for } k < r \leq 1 \end{cases} \tag{5}$$

In this paper, two configurations of this quantifier have been considered, i.e., with $k = 50$ and $k = 75$, representing different percentages of the required criteria to be satisfied. The shape of Q_{more} for both configurations is illustrated in Figure 1 (a) and (b).

Two definitions of the *most* quantifier are considered in this paper. According to [38], Q_{most} can be expressed as follows:

$$Q_{most}\left(r\right) = r^2 \tag{6}$$

According to [43], it can be defined as:

$$Q_{most}\left(r\right) = \begin{cases} 0 & \text{for } 0 < r \leq \alpha \\ \frac{r-\alpha}{\beta-\alpha} & \text{for } \epsilon < r < \beta \\ 1 & \text{for } r \geq \beta \end{cases} \tag{7}$$

The shape of Q_{most} under the two different definitions is illustrated in Figure 1 (c) and (d). In particular, Figure 1 (d) reports the case of $\alpha = 0.3$ and $\beta = 0.8$.

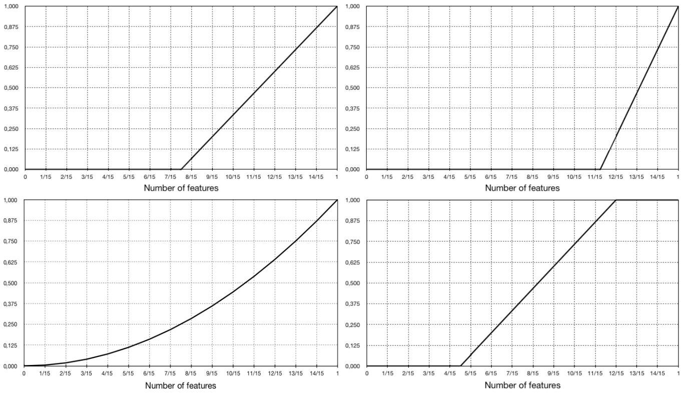


Fig. 1. Graphical representation of the Q function associated with the '*more than 50%*' (a), '*more than 75%*' (b), and '*most*' linguistic quantifiers, expressed according to Equations (6) and (7), in (c) and (d) respectively.

When considering all criteria as *equally important*, the weighting vector W for aggregation functions (i) OWA_MORE, and (ii) OWA_MOST, is obtained according to Equation (2), as illustrated in Section III-A. In this case, the

[2] The *min-max* normalization is the simplest method to rescale feature values on the unit interval in data processing [39]. This method was chosen for its immediacy. Other normalization functions can be tested; in the future, our intention is to use a different function for each criterion, based on its semantics.

[3] It has been empirically verified that, for all features, higher values can be interpreted as 'more credible'. Theoretical justifications are provided in the literature [40], [41].

[4] This choice does not prevent from employing and testing other linguistic quantifiers within the developed model.

above-defined linguistic quantifiers represent the proportion of criteria to be satisfied by the alternatives. To consider the proportion of the *important* criteria to be satisfied, two other aggregation functions have been defined, where the weighting vector W is built by employing Equation (3), as illustrated in Section III-B, together with the two linguistic quantifiers defined by Equations (5) – (7). The additional aggregation functions are denoted as: (iii) OWA_MORE_I, and (iv) OWA_MOST_I.

To assign *distinct importance values* to each credibility feature, two methods have been tested: 1) assigning them in a *heuristic* way, or 2) *learning* them from a subset of the available data. The second method illustrates that, in the presence of some labeled data (i.e., some news for which credibility is known), the proposed MCDM approach can be hybridized with a learning phase by considering a subset of the available training data. The two proposed methods are shortly explained in Sections IV-B1 and IV-B2 here below.

1) Importance values assigned in a heuristic way: In this case, the importance values associated with the considered features are based on a priori knowledge. In the literature, it has been highlighted that usually *temporal* and *user-related* features are particularly effective in assessing information credibility, more than *content-related* and *structural features* taken individually [3], [40], [44]. Based on these findings, with respect to the proposed categorization provided in Section IV-A, discrete importance values in the set $\{1, 2, 3, 4\}$ have been assigned to each category of features: in particular, to temporal features, an importance value equal to 4 has been assigned; to user- and content-related features an importance value equal to 3 and 2 has been respectively assigned; to structural features an importance value equal to 1 has been assigned. It is useful to notice that importance values could also take values on continuous intervals, like $[0, 1]$.

2) Importance values learned from data: Numerous are the ways that could be adopted to learn importance values from data in an MCDM scenario, as illustrated in the literature [45], [46]. In this work, a simple solution has been employed to learn the importance weights of categories of features. The labeled dataset described in Section V-A and employed for evaluation purposes has been split into three parts: 1/3 has been employed as the training set (balancing fake and genuine news items), and the residual part as the test set. After that, a 100-tree Random Forest classifier (one of the baselines described in Section V) has been trained and tested by excluding one feature at a time from the initial feature set, to assess the influence of that feature on the final classification results. The importance V_i of each single feature c_i has been obtained by evaluating the Area Under the ROC Curve (AUC) value [39], when each feature is removed from the classifier (the lower the result, the higher the importance of the removed feature), by complementing and normalizing this value as follows:

$$V_i = 1 - \left[(\beta - \alpha) \frac{\text{AUC}_i - min(\text{AUC}_k)}{max(\text{AUC}_k) - min(\text{AUC}_k)} + \alpha \right] \quad (8)$$

where AUC_i represents the AUC value obtained by excluding the feature c_i, $k = 1, \ldots, n$, n is the total number of features, and $\alpha = 0.1$, $\beta = 0.9$ are constant values set to obtain normalized values in the $[0.1, 0.9]$ range (to exclude the 'extreme' values 0 and 1). By observing the features reordered according to the obtained importance values in Table I, it is interesting to notice that the learning process confirms, for the most part, the heuristic assignment of importance by category.

TABLE I
FEATURES ORDERED ACCORDING TO THEIR IMPORTANCE VALUES, COMPUTED ACCORDING TO EQUATION (8).

Category – Feature	*AUC*	*Importance value*
[T] – ***ages***	0.734	0.9
[U] – ***friends***	0.742	0.836
[S] – ***media count***	0.756	0.724
[U] – ***density***	0.770	0.612
[T] – ***lifespan***	0.776	0.564
[S] – ***tweet count***	0.776	0.564
[C] – ***objectivity***	0.779	0.550
[C] – ***polarity***	0.779	0.550
[S] – ***retweet count***	0.780	0.532
[S] – ***mention count***	0.797	0.396
[U] – ***verified***	0.801	0.364
[S] – ***hashtag count***	0.802	0.356
[U] – ***followers***	0.809	0.300
[S] – ***status count***	0.822	0.196
[S] – ***URL count***	0.834	0.1

To sum up, let us consider a news item a characterized by the 15 features illustrated in Section IV-A, whose numerical values are denoted as $x_1, x_2, \ldots, x_{15}$. The performance scores $s_1(a), s_2(a), \ldots, s_{15}(a)$ are obtained after the normalization of $x_1, x_2, \ldots, x_{15}$ according to Equation (4), and the final credibility score σ_a is computed as:

$$\sigma_a = \mathcal{A}_{\text{OWA}}(s_1(a), s_2(a), \ldots, s_{15}(a)) = \sum_{k=1}^{15} w_k b_k, \quad (9)$$

where b_k is the kth largest of the $s_i(a)$, and the values w_k are the elements of the weighting vector W that characterizes the employed OWA operator; in particular the weights are computed according to Definition 1, Section III, which allows to incorporate in the weighting vector the importance weights associated with the considered criteria.

Example 1: Let us consider 4 credibility features. By considering an OWA operator with weighting vector $W = [0.2, 0.4, 0.1, 0.3]$ and a news item a, whose normalized credibility feature values are $(0.6, 0.4, 0.9, 0.5)$, the overall credibility score for the news item is: $\sigma_a = \mathcal{A}_{\text{OWA}}(0.6, 0.4, 0.9, 0.5) = (0.2)(0.9) + (0.4)(0.6) + (0.1)(0.5) + (0.3)(0.4) = 0.59$.

When features are considered as *equally important*, the value of the w_k weights is computed according to Equation (2), where Q is expressed according to Equation (5) for aggregation function (i), i.e., OWA_MORE, and to Equation (6) or Equation (7) for aggregation function (ii), i.e., OWA_MOST.

When features have *distinct importance* associated with them, the value of the w_k weights is computed according to Equation (3), where importance values can be computed according to both methods described in Sections IV-B1 and IV-B2, and where Q is expressed according to Equation (5) for aggregation function (iii), i.e., OWA_MORE_I, and to Equation (6) or Equation (7) for aggregation function (iv), i.e., OWA_MOST_I.

V. Evaluations

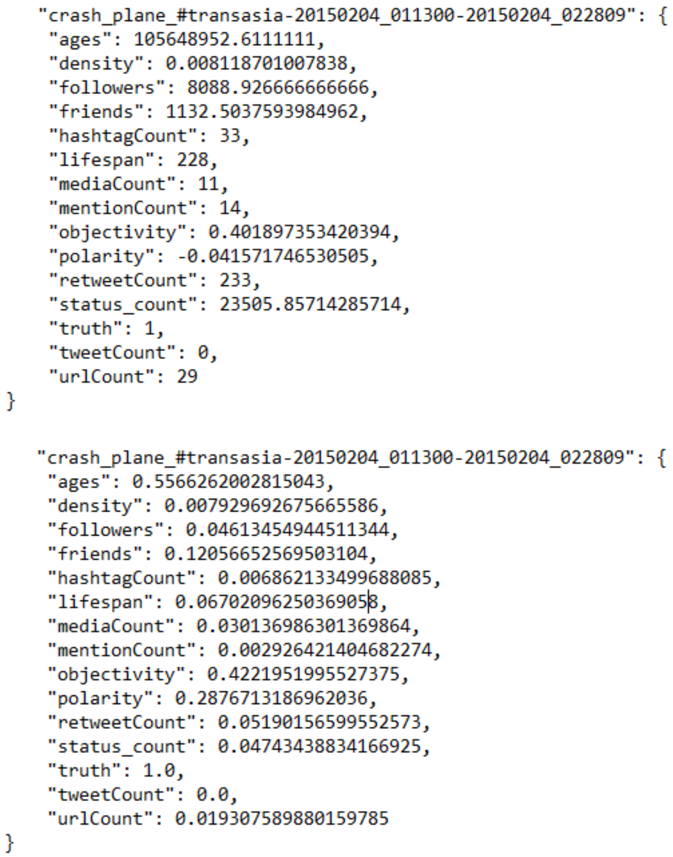
To evaluate the proposed model, the CREDBANK dataset detailed in Section V-A has been employed. On this dataset, first, a binary classification task has been performed by employing the aggregation functions (i)–(iv) defined in Section IV-B, by assigning importance values both heuristically (Section IV-B1) and based on a subset of training data (Section IV-B2). Also, various well-known machine learning algorithms employed successfully in the literature for news credibility assessment have been implemented (i.e., SVM, kNN, Decision Trees, Naive Bayes, and Random Forests [3], [47]). The effectiveness of the different classifiers has been evaluated by considering the following metrics: *accuracy* (*Acc*), *precision* (*Prec*), *recall* (*Rec*), $F1$*-score* ($F1$), and *Area Under the ROC Curve* (AUC) [39].

A. The CREDBANK Dataset

The dataset has been defined in [31] as "*a unique dataset compiled to link social media event streams with human credibility judgments in a systematic and comprehensive way*". It is composed of about 80 millions of tweets, grouped into 1,376 news events (about 60,000 tweets per event). With each news event, a 30-element vector of *credibility labels* (called *accuracy labels* in [31]) is associated, provided by 30 distinct experts. Each credibility label is expressed on a 5-point Likert scale ranging from -2 (*certainly false*) to 2 (*certainly true*).

In this article, a 'reduced' version of the CREDBANK dataset is employed, i.e., the one described and provided in [14], where the authors have considered the most retweeted tweets in order to discard, among the 1,376 events, those provoking less reactions. To have an overall score associated with each news event, the authors have computed the *mean accuracy rating* based on the 30 accuracy labels provided by experts. This led the authors to finally select 156 news events, of which 99 are labeled as true and 57 as fake. It is worth to be underlined that in the reduced version, news events represent only the most *significant news* (in terms of reactions), and each news event is made up of *thousands* of individual tweets, for a total of more than 9 million tweets. In fact, in this evaluation section, *news events* are considered as news items to be classified in terms of credibility.

In Figure 2, two fragments of the JSON file representing a news event in the CREDBANK dataset with the associated features and feature values (not normalized and normalized according to Equation (4), respectively) are illustrated.



```
   "crash_plane_#transasia-20150204_011300-20150204_022809": {
    "ages": 105648952.6111111,
    "density": 0.008118701007838,
    "followers": 8088.926666666666,
    "friends": 1132.5037593984962,
    "hashtagCount": 33,
    "lifespan": 228,
    "mediaCount": 11,
    "mentionCount": 14,
    "objectivity": 0.401897353420394,
    "polarity": -0.041571746530505,
    "retweetCount": 233,
    "status_count": 23505.85714285714,
    "truth": 1,
    "tweetCount": 0,
    "urlCount": 29
}

   "crash_plane_#transasia-20150204_011300-20150204_022809": {
    "ages": 0.5566262002815043,
    "density": 0.007929692675665586,
    "followers": 0.04613454944511344,
    "friends": 0.12056652569503104,
    "hashtagCount": 0.006862133499688085,
    "lifespan": 0.06702096250369058,
    "mediaCount": 0.030136986301369864,
    "mentionCount": 0.002926421404682274,
    "objectivity": 0.4221951995527375,
    "polarity": 0.2876713186962036,
    "retweetCount": 0.05190156599552573,
    "status_count": 0.04743438834166925,
    "truth": 1.0,
    "tweetCount": 0.0,
    "urlCount": 0.019307589880159785
}
```

Fig. 2. A news event extracted from the CREDBANK dataset, and its connected features, before and after the normalization phase.

B. Implementation Details

The classification and experimental phases have been conducted by employing the *Python* programming language. To manage data, the *pandas* library has been used;[5] to make numerical computations on data, such as the development of the proposed aggregation functions, the *NumPy* library has been used;[6] finally, the *scikit-learn* library has been employed to implement and evaluate the baseline classifiers.[7] In particular, as regards the parameter setting, the *linear kernel* of the `sklearn.svm.SVC` function has been used to implement the SVM classifier. Concerning the other classifiers, they have been implemented by keeping the default parameters of the following `scikit-learn` functions:

- `sklearn.neighbors.KNeighborsClassifier` for the kNN classifier;
- `sklearn.tree.DecisionTreeClassifier` for the Decision Tree classifier;
- `sklearn.naive_bayes.GaussianNB` for the Naive Bayes classifier;
- `sklearn.ensemble.RandomForestClassifier` for the Random Forest classifier.

To prevent possible overfitting, *5-fold cross validation* has been performed to evaluate the above-mentioned machine learning baselines, as done in the related literature [14], [23].

[5] https://pandas.pydata.org

[6] https://www.numpy.org

[7] https://scikit-learn.org/stable/index.html

For the proposed MCDM approach, aggregation functions (i)–(iv) have been implemented to aggregate the performance scores associated with the considered credibility features of news items. For each news item, an overall credibility score in the [0,1] interval has been obtained. Then, news items have been classified as *genuine* or *fake* by selecting an optimal *threshold* over these overall scores. The threshold has been set, for each distinct method, by selecting the value that maximizes the classification effectiveness, as discussed in [48]. The selection of a classification threshold was necessary to comparatively evaluate the proposed model with respect to those in the literature that apply a *binary classification*. The model proposed in this paper generates credibility values associated with news items, and, as such, it is also suitable for proposing a *ranking* of news items to the decision maker on the basis of their possible level of credibility, so that the DM can directly take a final decision of which news to trust more.

C. Summarization of Results and Discussion

In this section, the values of the considered evaluation metrics for each aggregation function, and for each considered baseline, are reported. All classifiers have been tested over the CREDBANK dataset described in Section V-A. Table II summarizes the obtained results.

TABLE II
SUMMARIZATION OF RESULTS OF ALL THE EXPERIMENTS.

	AUC	*Acc*	*Prec*	*Rec*	*F*1
SVM	0.80	66%	66%	99%	80%
kNN	0.62	68%	70%	87%	77%
Decision Trees	0.75	76%	89%	82%	81%
Naive Bayes	0.78	71%	71%	93%	80%
Random Forests	**0.87**	79%	79%	90%	84%
OWA_MORE$_{(50\%)}$	0.79	76%	82%	81%	81%
OWA_MORE$_{(75\%)}$	0.81	79%	**87**%	79%	83%
OWA_MOST$_{(6)}$	0.68	65%	77%	65%	70%
OWA_MOST$_{(7)}$	0.79	78%	79%	89%	84%
OWA_MORE_I$_{(50\%)}$ (a)	**0.84**	**83**%	83%	91%	**87**%
OWA_MORE_I$_{(75\%)}$ (a)	0.83	**83**%	85%	89%	**87**%
OWA_MOST_I$_{(6)}$ (a)	0.75	73%	78%	80%	79%
OWA_MOST_I$_{(7)}$ (a)	0.83	82%	82%	91%	86%
OWA_MORE_I$_{(50\%)}$ (b)	0.80	78%	80%	86%	83%
OWA_MORE_I$_{(75\%)}$ (b)	0.82	77%	85%	77%	81%
OWA_MOST_I$_{(6)}$ (b)	0.64	63%	74%	65%	69%
OWA_MOST_I$_{(7)}$ (b)	0.78	77%	85%	77%	81%

It can be noticed that, for the aggregation functions guided by the *more than k%* and *most* linguistic quantifiers, several configurations have been tested. In particular, the functions denoted as OWA_MORE$_{(50\%)}$ and OWA_MORE$_{(75\%)}$ have been considered (i.e., more than 50% and more than 75% of criteria have to be satisfied). Furthermore, also the functions denoted as OWA_MOST$_{(6)}$ and OWA_MOST$_{(7)}$ have been tested, where in the first case the *most* quantifier guiding the aggregation is expressed according to Equation (6), while in the second case it is expressed by means of Equation (7), with $\alpha = 0.5$ and $\beta = 0.6$ (these parameters are those that provided the best results for the considered aggregation function).

Different configurations of aggregation functions considering *different importance* associated with the considered criteria have also been tested: OWA_MORE_I$_{(50\%)}$, OWA_MORE_I$_{(75\%)}$, OWA_MOST_I$_{(6)}$, OWA_MOST_I$_{(7)}$, both when importance values have been obtained heuristically (a), and when they have been learned from a subset of the available data (b) (see Sections IV-B1 and IV-B2).

Based on the results reported in Table II, a first consideration is that aggregation functions based on OWA operators guided by the *more than k%* quantifier perform better with respect to those based on the *most* quantifier, in any case, at least regarding the percentages of criteria to be satisfied that have been selected (remember that these represent examples of the decision maker preferences, which are modifiable parameters in the proposed model). Furthermore, as it was reasonable to expect, the aggregation functions considering a different importance associated with criteria perform better than those considering all criteria as equally important.

With respect to the aspect of how defining importance values, it is interesting, in particular, to notice that the aggregation functions for which the importance values have been defined heuristically based on a prior knowledge, have similar (and even better) performance of those where the importance values have been learned from a subset of the available data. This confirms the feasibility and the effectiveness of the use of a completely model-driven approach to tackle the considered fake news detection problem, not forgetting, at the same time, that more complex solutions to learn the values of importance from some given training data, together with the MCDM approach, could provide better results.

Globally, for the considered aggregation functions and ways of setting importance values, the best results are obtained by the aggregation functions OWA_MORE_I$_{(50\%)}$ (a) and OWA_MORE_I$_{(75\%)}$ (a), which exceed all baselines with respect to accuracy, precision and F1 score, with, on average, comparable AUC values.

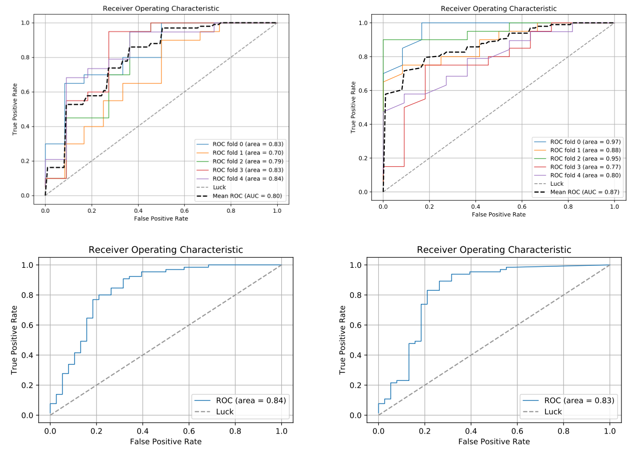


Fig. 3. From left to right, the ROC curves and AUC values for SVM and RF baselines, and for OWA_MORE_I$_{(50\%)}$ (a), and OWA_MORE_I$_{(75\%)}$ (a).

Figure 3 illustrates the ROC curves for the baselines that in the literature have been mostly employed in the specific context of news credibility assessment (i.e., SVM and Random Forests), and for the most effective aggregation functions proposed in this paper.

VI. Conclusions

In this article, the problem of assessing the credibility of information spreading on social media was considered, by defining a Multi-Criteria Decision Making (MCDM) approach for news credibility assessment in microblogging sites (Twitter in particular). With respect to the several approaches based on machine learning that have been proposed in last years in the literature, the proposed approach, by exploiting aggregation operators and prior domain knowledge, allows to design in a flexible way the credibility assessment model. In this way, the 'black-box' effect that characterizes some previous works can be avoided and, at the same time, this approach is not fully dataset-dependent.

Despite this, it is possible and desirable hybridizing the MCDM model with some learning aspects if unbiased datasets labeled with respect to the credibility of information are available, as illustrated in the article. In future research, other families of aggregation operators (for example fuzzy integrals) and formal approaches to learn the different importance weights to be associated with credibility features will be analyzed and evaluated.

References

[1] E. Ferrari and M. Viviani. Privacy in social collaboration. In *Handbook of Human Computation*, pp. 857–878. Springer, 2013.

[2] C. S. Self. Credibility. In M. B. Salwen and D. W. Stacks, editors, *An Integrated Approach to Communication Theory and Research, 2nd Edition*, pp. 435–456. Routledge, Taylor and Francis Group, 2008.

[3] M. Viviani and G. Pasi. Credibility in Social Media: Opinions, News, and Health Information-A Survey. *WIREs Data Mining and Knowledge Discovery*, 7(5), 2017.

[4] K. Shu et al. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[5] B. J. Fogg and H. Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 80–87. ACM, 1999.

[6] M. Viviani and G. Pasi. Quantifier guided aggregation for the veracity assessment of online reviews. *International Journal of Intelligent Systems*, 32(5):481–501, 2017.

[7] G. Pasi and M. Viviani. Application of aggregation operators to assess the credibility of user-generated content in social media. In *Proc. of IPMU'18*, pp. 342–353. Springer, 2018.

[8] L. M. Aiello et al. Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

[9] H. Kwak et al. What is Twitter, a social network or a news media? In *Proc. of the 19th Int. Conf. on WWW*, pp. 591–600. ACM, 2010.

[10] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[11] D. M. Lazer et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[12] V. L. Rubin et al. Deception detection for news: three types of fakes. *Proc. of the Assoc. for Inf. Sci. and Technology*, 52(1):1–4, 2015.

[13] E. Ferrara et al. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[14] C. Buntain and J. Golbeck. Automatically identifying fake news in popular Twitter threads. In *IEEE Smart Cloud (SmartCloud) 2017*, pp. 208–215. IEEE, 2017.

[15] C. Shao et al. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[16] M. Gupta et al. Evaluating event credibility on Twitter. In *SDM*, pp. 153–164. SIAM, 2012.

[17] Z. Jin et al. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE Int. Conf. on*, pp. 230–239. IEEE, 2014.

[18] N. Vo et al. Revealing and detecting malicious retweeter groups. In *Proc. of the 2017 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining 2017*, pp. 363–368. ACM, 2017.

[19] L. Zhao et al. A topic-focused trust model for Twitter. *Computer Communications*, 76:1 – 11, 2016.

[20] S. Cazalens et al. A content management perspective on fact-checking. In *The Web Conference 2018-alternate paper tracks" Journalism, Misinformation and Fact Checking"*, pp. 565–574, 2018.

[21] T. Mihaylova et al. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] K. Popat et al. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1003–1012, 2017.

[23] H. Rashkin et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, pp. 2931–2937, 2017.

[24] C. Castillo et al. Information credibility on Twitter. In *Proc. of the 20th Int. Conf. on World Wide Web*, pp. 675–684. ACM, 2011.

[25] C. Castillo et al. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2012.

[26] C. Boididou et al. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[27] P. Galán-García et al. Supervised machine learning for the detection of troll profiles in Twitter social network. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[28] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, pp. 2. ACM, 2012.

[29] A. Gupta and R. Kaushal. Improving spam detection in online social networks. In *Cognitive Computing and Information Processing (CCIP), 2015 International Conference on*, pp. 1–6. IEEE, 2015.

[30] B. Kang et al. Modeling topic specific credibility on Twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 179–188. ACM, 2012.

[31] T. Mitra and E. Gilbert. CREDBANK: A large-scale social media corpus with associated credibility annotations. In *Proc. of the Ninth International AAAI Conference on Web and Social Media*, 2015.

[32] S. Greco et al. *Multiple criteria decision analysis*. Springer, 2016.

[33] T. Calvo et al. *Aggregation operators: new trends and applications*, volume 97. Physica, 2012.

[34] P. Ceravolo et al. Adding a peer-to-peer trust layer to metadata generators. In *OTM Confederated International Conferences*, pp. 809–815. Springer, 2005.

[35] E. Damiani and M. Viviani. Trading anonymity for influence in open communities voting schemata. In *2009 International Conference on Social Informatics (SOCINFO'09)*, pp. 63–67. IEEE, 2009.

[36] S. Marrara et al. Aggregation operators in information retrieval. *Fuzzy Sets and Systems*, 324:3–19, 2017.

[37] R. R. Yager and J. Kacprzyk. *The ordered weighted averaging operators: theory and applications*. Springer Science, 2012.

[38] R. R. Yager. Quantifier guided aggregation using OWA operators. *International Journal of Intelligent Systems*, 11(1):49–73, 1996.

[39] M. Kubat. *An Introduction to Machine Learning*. Springer, 2016.

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

[41] J. Fontanarava et al. Feature analysis for fake review detection through supervised classification. In *2017 IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA)*, pp. 658–666. IEEE, 2017.

[42] D. Dubois and H. Prade. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142(1):143 – 161, 2004. Aggregation Techniques.

[43] D. Ben-Arieh. Sensitivity of multi-criteria decision making to linguistic quantifiers and aggregation means. *Computers & Industrial Engineering*, 48(2):289–309, 2005.

[44] A. Mukherjee et al. Fake review detection: Classification and analysis of real and pseudo reviews. Technical report, UIC-CS-03-2013, 2013.

[45] D. Filev and R. R. Yager. Learning owa operator weights from data. In *Proceedings of 1994 IEEE 3rd Int. Fuzzy Systems Conference*, pp. 468–473. IEEE, 1994.

[46] V. Torra. Learning weights for the quasi-weighted means. *IEEE Transactions on Fuzzy Systems*, 10(5):653–666, 2002.

[47] M. Crawford et al. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):23, 2015.

[48] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.