Multi-Criteria Decision Making and Supervised Learning for Fake News Detection in Microblogging

Marco De Grandis, Gabriella Pasi, and Marco Viviani
marcodegra@live.it,gabriella.pasi@unimib.it,marco.viviani@unimib.it
University of Milano-Bicocca / DISCo
Milan, Italy

ABSTRACT

In the last years, the success of social media platforms has lead to the spread of big volumes of the so-called User-Generated Content (UGC) across virtual communities. In microblogging sites, UCG can be often generated in the form of 'newsworthy' posts, i.e., related to information that has a public utility for the people. In this scenario, being the UGC diffused without almost any traditional form of trusted external control, the possibility of incurring in possible fake news is far from remote. For this reason, several approaches for fake news detection in microblogging have been proposed up to now. Many solutions that aim at classifying genuine news with respect to fake ones, deal with supervised machine learning techniques. In this paper, as an alternative to purely data-driven solutions, the use of the Multi-Criteria Decision Making (MCDM) paradigm is proposed, and in particular of aggregation operators. In this context, the decision maker can be involved in the process of fake news detection, by taking advantage of both some prior knowledge of the domain, and some potentially available learning data.

CCS CONCEPTS

• **Information systems** → **Decision support systems**; **Social networks**; • **Computing methodologies** → **Artificial intelligence**; • **Mathematics of computing** → *Numerical analysis.*

KEYWORDS

Credibility, Fake News, Multi-Criteria Decision Making, Aggregation Operators, Social Media, User-Generated Content, CREDBANK

1 INTRODUCTION

Nowadays, the interaction between users is promoted by a number of Web 2.0 technologies that facilitate the establishment of multiple social relationships [7], and the diffusion of information in the form of *User-Generated Content* (UGC). Often, UGC is referred to the so-called *conversation* posts, which usually have an interest only for friends, or people sharing the same interests, of the person who generated the content. In other cases, however, *newsworthy* or *news* posts can be diffused, which have a more general interest for the public. Considering social media, news posts are diffused in particular by microblogging platforms, such as Twitter and Sina Weibo,[1] where millions of users act as real-time news diffusers [19]. Through the phenomenon of 'disintermediation' that characterizes social media [12], the possibility of incurring in fake news is always higher. Due to the social consequences that relying on fake news can have (think, for example, of the possibility of directing political elections, or of spreading conspiracy theories or generating controversy [15, 35] at the level of public opinion) several approaches have been proposed in the last years to analyze and combat this phenomenon [11, 37, 44, 45]. Recent literature is addressing the problem of *automated fact checking* [34]; other approaches for fake news detection can be categorized into two main classes [37]: (*i*) *classification-based* and (*ii*) *propagation-based*. In the first category fall all those methods that are based on machine learning (ML) (mostly supervised) algorithms to classify in a binary way news credibility. The second category includes studies on the propagation of low-credibility content across virtual communities, often through *social bots* [32].

In this paper, considering the context of classification-based approaches, the *Multi-Criteria Decision Making* (MCDM) paradigm is proposed, and in particular, the use of a numerical solution based on aggregation operators. The proposed approach aims at considering the decision maker as part of the fake news detection process, by giving to her/him a certain control over the classification phase, both employing prior knowledge about the domain, and potentially available learning data in building the model.

To illustrate the approach and for evaluation purposes, the proposed solution has been instantiated over the CREDBANK dataset [26], constituted by microblogging posts gathered from Twitter; nonetheless, it can be generalized to other microblogging sites, and, potentially, to other kinds of social media.

2 BACKGROUND AND RELATED WORK

Addressing the fake news detection problem in microblogging sites requires, first of all, to identify the so-called *news posts* briefly introduced in Section 1, which are statements about a fact or an actual event of interest to a larger community, not only to the friends of the author of the message [8]. In the last years, Twitter has gained reputation as a prominent news medium [1], given that the majority of *trending topics* on this platform can be considered *headline news* or *persistent news* [22]. But what constitutes *fake news*? Fake news does not represent a 'new' problem emerged with social media; it always existed because of the human nature, and it has psychological and social foundations [33]. In the traditional journalism context, a definition of fake news is the following: "articles that are intentionally and verifiably false, and could mislead readers" [2].

[1] https://twitter.com, https://weibo.com

ROME 2019, July 25, 2019, Paris, France

On-line, and in microblogs in particular, things are more complex. There are the so-called *rumors*, which can usually be defined as pieces of "circulating information whose veracity status is yet to be verified at the time of posting" [46], and fake news that, differently from rumors, refers to information related specifically to public news events that can be verified as false [23, 33]. It may be interesting to briefly mention that fake news can in turn be of various kinds: (*i*) *completely fake and large-scale hoaxes* [30], which is news deliberately fabricated or falsified in the mainstream or social media to deceive audience; (*ii*) *humorous/satire news* [30], which relies on irony and humor, mimicking credible news stories (e.g., the Italian website *Lercio* or the American *The Onion*)[2]; (*iii*) *poorly written news articles* [33], which are constituted by statements presented as facts, without any verification of the sources and characterized by a mixture of subjective opinions and facts; (*iv*) *conspiracy theories* [33]; (*v*) *misinformation* [33], which is constituted by news that the person diffusing it believes true, but which then turn out to be totally or partially fake; (*vi*) *disinformation* [23], which is false information intentionally and deliberately spread by individuals; (*vii*) *fake news automatically generated* by spam profiles, trolls and bots [14, 16].

Several approaches have been proposed in the last years for detecting both rumors and fake news (of type (*i*) in particular) in microblogging sites. Among the works in the literature, some of them have considered as an information unit (to be evaluated in terms of credibility) a *single post* (e.g., a tweet); other works have considered a *thread* (e.g., a set of tweets on the same topic) that represents a single *news event*. Some approaches focus on *automated fact checking* [10, 25, 28, 29], often based on the use of external knowledge bases; other approaches belong to two main categories, namely (*i*) *classification-based*, and (*ii*) *propagation-based*; the latter are mainly concerned with studying the influence that *social bots* have on the dissemination of fake news [13, 32] and how low-credibility information spreads over the social network structure [18, 19, 39, 43].

The solution proposed in this paper for fake news detection falls in the *classification-based* category, i.e., including those approaches where multiple features connected with news are considered to classify them with respect to credibility; for this reason, in the following, only the main approaches belonging to category (*i*) will be detailed. Castillo *et al.* [8, 9] were among the first to tackle in a structured way the problem of information credibility on microblogging sites, Twitter in particular, by using classification-based approaches. The authors focus on automatic methods for assessing the credibility of a given 'time-sensitive' set of tweets, i.e., a *trending topic*, based on multiple features extracted from tweets (linguistic features) and their authors. In [9], the authors extend the model presented in [8], and evaluate it on the scenario of the use of Twitter during a crisis event. The approaches are based on the use of 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 [4, 14, 16, 17, 20], each of which 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 target-topics in Twitter [20], the identification of credible tweets during high impact events [17], the detection of spammers [16] and troll profiles [14] in microblogging sites, the classification of credible versus not credible multimedia tweets, i.e., accompanied by a multimedia item (image or video) from an event [4]. The work described in [5] aims at considering a large set of credibility features (the same that are used in this paper), which are employed to automatically identify fake news in Twitter threads (disregarding multimedia content, which is out of the scope of this paper). The model is trained over large-scale labeled datasets, CREDBANK in particular [26], which is a large-scale set of Twitter threads about news events and corresponding crowdsourced credibility assessments for each event. At the time of writing, the CREDBANK dataset represents a suitable solution to the scarcity of labeled datasets to develop and evaluate effective supervised classifiers for fake news detection.

With respect to the above-mentioned data-driven approaches, in the next section a classification-based solution focusing on Multi-Criteria Decision Making is described. It is based on a numerical solution exploiting aggregation operators, prior domain knowledge, and (potentially) some learning data to define the proposed model.

3 MCDM AND FAKE NEWS DETECTION

In this paper, the fake news detection problem is seen as a *Multi-Criteria Decision Making* (MCDM) problem, where there is: (*i*) a set of *alternatives*, i.e., *news* to be evaluated in terms of credibility, which are available to a *decision maker* (DM); (*ii*) multiple (potentially conflicting) *criteria*, i.e., credibility *features* associated with news, which are satisfied in a different way by each alternative (each piece of news); different *importance weights* associated with each criterion (each credibility feature). In the literature, a solution often employed to solve MDCM problems consists in assigning distinct scores, namely *performance scores*, to each alternative with respect to each criterion.

Each score represents a *degree of satisfaction* that expresses to what extent an alternative is satisfactory with respect to a criterion. In the considered context, where the alternatives are the *news* to be evaluated, and the criteria are the *features* characterizing the news, the aim is to develop a system automatically assisting the decision maker, by interpreting these performance scores as *degrees of credibility* of the news with respect to each feature. These multiple *credibility scores* can then be subsequently *aggregated* to obtain an *overall credibility score*.

Formally, let us assume that: $A = \{a_1, a_2, \ldots, a_m\}$ is the set of *alternatives*, i.e., the *news*; $C = \{c_1, c_2, \ldots, c_n\}$ is the set of *criteria*, i.e., the *features* characterizing the news; s_i is the *satisfaction function* that, for a criterion c_i ($1 \leq i \leq n$), returns the *performance score* $s_i(a_j) \in I$, $I = [0, 1]$, to which the alternative a_j ($1 \leq j \leq m$) satisfies the criterion c_i (i.e., the *credibility score* in the considered context). A solution to obtain an overall performance score σ_j for each alternative a_j, i.e., an overall credibility score for each piece of news, is to employ an n-ary function $\mathcal{A}$, called *aggregation operator* (or *aggregation function*), which is a mapping $\mathcal{A} : [0, 1]^n \rightarrow [0, 1]$ acting on a finite number n of performance scores to be aggregated (for $n \in \mathbb{N}_0$). Formally: $\sigma_j = \mathcal{A}(s_1(a_j), s_2(a_j), \ldots, s_n(a_j))$ [6].

[2] https://www.lercio.it, https://www.theonion.com

3.1 Quantifier-guided Aggregation

Depending on the aggregation operator (or family of aggregation operators) considered, it is possible to guide the aggregation by *linguistic quantifiers* (LQ) (e.g., *all*, *some*, *many*, ...). This allows to choose the best alternative(s) (i.e., news) based on the satisfaction of a 'certain amount' of the criteria (i.e., credibility features) by the alternative(s). In the literature, a family of aggregation operators that allows *quantifier-guided aggregation* is that of *Ordered Weighted Averaging* (OWA) operators [40], extensively studied in the literature [42].

Definition 3.1. An aggregation operator $\mathcal{A}_{\text{OWA}} : [0,1]^n \rightarrow [0,1]$ is called an *Ordered Weighted Averaging* (OWA) operator of dimension n if it has associated a weighting vector $W = [w_1, w_2, \ldots, w_n]$ such that $w_k \in [0,1]$ and $\sum_{k=1}^{n} w_k = 1$, and 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 allow to represent the trade-off that the decision maker is leaning to accept among the considered criteria, which lies between two extreme situations: (i) the situation in which s/he desires that *all* criteria are satisfied by the alternative, represented by the *min* operator, when $w_n = 1$, i.e., $W = [0, 0, \ldots, 1]$; (ii) the situation in which the satisfaction of *at least one* criterion is what the decision maker desires, represented by the *max* operator, when $w_1 = 1$, i.e., $W = [1, 0, \ldots, 0]$. An intermediate situation between these two extremes is represented by the use of the *arithmetic mean* operator, when $w_i = \frac{1}{n}$, i.e., $W = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]$.

3.2 Equal and Unequal Importance of Criteria

In general, in the process of quantifier-guided aggregation, the decision maker provides a linguistic quantifier Q indicating the number (*absolute* quantifier) or the proportion (*relative* quantifier) of criteria s/he believes is sufficient 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, such as *at least k%* and *most*. A linguistic quantifier is said to be a RIM quantifier if [41]: $Q(0) = 0$, $Q(1) = 1$, and $Q(r) \geq Q(s)$ if $r > s$ $(r, s \in [0,1])$.

3.2.1 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:

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

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, as detailed in section below (e.g., in fake news detection, not all the features connected with a piece of news are equally significant in terms of credibility assessment).

3.2.2 Unequal Importance of Criteria. In [41], a way has been proposed for aggregating n scores with *distinct importance* associated with the criteria that generated them. Let us consider an alternative a (i.e., a piece of news in the considered context) to be evaluated with respect to n criteria; the performance scores by a of 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 piece of news.

4 FAKE NEWS DETECTION ON TWITTER

In this section, an MCDM approach based on the use of OWA aggregation operators for fake news detection is proposed; different *aggregation functions* guided by distinct *linguistic quantifiers* are presented, which allow to tune the number of (important) features to be considered, and to provide an overall credibility score associated with each considered piece of news. Based on these overall scores, it is possible to identify which news has to be considered fake news and which not. In particular, as illustrated in Section 2, it is possible to consider as news either single posts, or threads representing *news events*. If every single post is considered as an alternative, credibility features (i.e., criteria) are those associated with the considered post and/or with the user who generated it. In the case of a news event, the features describe 'global properties' of the event, i.e., of the posts that compose the thread and their authors. Since prior classification-based solutions have referred to Twitter and employed (labeled) datasets from this social media platform to prove their effectiveness, the same context has been taken into account in this paper, to instantiate and evaluate the proposed approach, by focusing in particular on the assessment of the credibility of *news events* extracted from the CREDBANK dataset. It makes it possible to consider a large number of different types of features in Twitter, and can be employed for learning purposes.

4.1 The CREDBANK Dataset

The dataset, defined in [26], is composed of about 80 millions of tweets, grouped into 1,376 news events (about 60,000 tweets per event). To each news event, it is associated a 30-element vector of *credibility labels* (called *accuracy labels* in [26]) 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 [5], where the authors have considered the most retweeted tweets in order to discard, among the 1,376 events, those provoking less reaction. To have an overall score associated with each news event, they 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.[3] 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.

4.2 Features Identification and Representation

Several features have been used in the literature for evaluating the credibility of Twitter threads (i.e., representing news events); they belong to the following macro-categories: *structural features* [S], i.e., specific to the structure of each Twitter thread; *user-related features* [U], i.e., representing attributes related to the users, their profiles, their connections and interactions; *content-related features* [C], i.e., based on textual properties extracted from the content of the tweets; *temporal features* [T], i.e., allowing to take into consideration how the values of the other types of features change over time.

In this paper, only the most informative feature set is considered, as illustrated in [5], which is composed of the following 15 features: [S] ***media count***: the frequency of tweets that contain media contents (images, videos, etc.); [S] ***mention count***: the frequency of tweets that contain mentions; [S] ***URL count***: the frequency of tweets that contain URLs; [S] ***retweet count***: the number of retweets for the event; [S] ***hashtag count***: the frequency of tweets that contain hashtags; [S] ***status count***: the average number of tweets with respect to each user profile (in the thread); [S] ***tweet count***: the frequency of tweets that contain only text (no media, mentions, hashtags or URLs); [U] ***verified***: the number of verified profiles (in the thread); [U] ***density***: the density of the network w.r.t. users (nodes) and their interactions (edges, i.e., mentions, replies, etc.); [U] ***followers***: the average number of followers with respect to each user profile (in the thread); [S] ***friends*** also known as *followees*: the average number of followees with respect to each user profile (in the thread); [C] ***polarity***: the average positive or negative feelings expressed by the tweets (in a thread); [C] ***objectivity***: the score of whether a thread is objective or not; [T] ***ages***: the author account age relative to a tweet creation; [T] ***lifespan***: the minutes between the first and the last tweet of the thread.

The above-mentioned features are of a different nature, refer to distinct concepts and, therefore, are expressed on different numerical scales. In the proposed MCDM approach, starting from the values associated with features, it is necessary to select a suitable satisfaction function that is able to transform them into suitable performance scores in the $[0, 1]$ interval to be aggregated, as illustrated in Section 3. To do this, the *min-max* normalization function has been employed:

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

where, for a news event a_j, $s_i(a_j)$ is the performance score normalized in the $[0, 1]$ interval with respect to the feature c_i, $x_{i,j}$ is the value of the feature c_i for a_j, $h = 1, \ldots, m$, and m is the total number of news events. The performance scores obtained this way are considered as the degrees of satisfaction of each news event with respect to each feature in terms of credibility. The value '1' is assumed as the evidence of a full satisfaction in terms of credibility, and the value '0' as a complete dissatisfaction.[4]

4.3 Quantifier-guided Aggregation Functions

Two initial functions have been developed, based on OWA operators guided by the following linguistic quantifiers: (i) the *more than k%* quantifier – OWA_MORE; (ii) the *most* quantifier – OWA_MOST.

More than k%. According to [3], the *more than k%* (*more*) quantifier can be defined as:

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

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

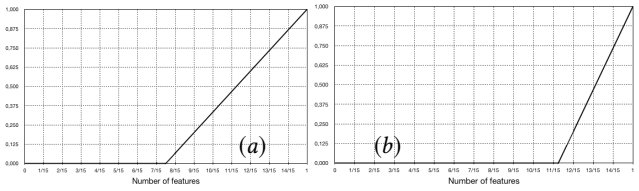

Figure 1: Graphical representation of the Q function for the '*more than 50%*' (a), and the '*more than 75%*' (b) LQ.

Most. Two definitions of the *most* quantifier are considered in this paper. According to [41], Q_{most} can be expressed as:

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

According to [3]:

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

The shape of Q_{most} under the two different definitions is illustrated in Figure 2 (a) and (b). In particular, Figure 2 (b) reports as an example the case of $\alpha = 0.3$ and $\beta = 0.8$.

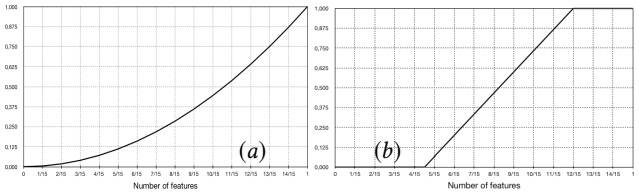

Figure 2: Graphical representation of the Q function for the '*most*' LQ, expressed according to Equations (6) and (7).

When considering all criteria as *equally important*, the weighting vector W for aggregation functions (i) and (ii) can be obtained

[3] https://github.com/cbuntain/CREDBANK-data

[4] It has been empirically verified that, for all the features, higher values can be interpreted as 'more credible'. Some theoretical justifications about the type of features and the values associated with them in the assessment of the credibility of information are provided in [24].

according to Equation (2), as illustrated in Section. 3.2.1. In this case, the 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, other aggregation functions have been considered, where the weighting vector W is built by employing Equation (3) illustrated in Section, 3.2.2, together with linguistic quantifiers defined by Equations (5)−(7). These additional funchions are denoted as: (*iii*) OWA_MORE_I; (*iv*) OWA_MOST_I.

In the proposed solution, to assign distinct importance values to each credibility feature, two methods have been proposed: (1) assigning them in a heuristic way, or (2) learning them from a subset of the available data. The first method represents a complete unsupervised way of implementing the proposed OWA-based approach; the second method illustrates that, in the presence of some labeled data (i.e., some news of which credibility is known), the proposed MCDM approach can also be hybridized with some data-driven aspects by considering a subset of the available training data.

4.3.1 Importance values assigned in a heuristic way. The importance values associated with features are based on a priori knowledge of the domain. In the literature [8, 24, 27, 37], 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. Therefore, with respect to the proposed categorization provided in Section, 4.2, discrete importance values in the set $\{1, 2, 3, 4\}$ have been assigned to each category of features: in particular, to temporal features it has been assigned an importance value equal to 4; to user- and content-related features it has been assigned an importance value equal to 3 and 2 respectively; to structural features it has been assigned an importance value equal to 1. It is important to notice that also continuous values, for example in the [0,1] interval, could have been employed.

4.3.2 Importance values learned from data. The dataset illustrated in Section 4.1 has been split into three parts: 1/3 of it has been employed as the training set, by considering the balancing between fake and genuine news events, and the residual part as the test set. After that, a 100-tree Random Forest classifier (the same employed by the reference baseline, as it will be illustrated in Section 5) has been trained and tested by excluding one feature at a time from the initial feature set, to asses the influence of that each feature has 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 [21], when each feature is removed from the classifier (the lower the result, the higher the importance of the removed feature), and by complementing and normalizing this value as follows:

$$V_i = 1 - \left[(b-a) \frac{\mathrm{AUC}_i - min(\mathrm{AUC}_k)}{max(\mathrm{AUC}_k) - min(\mathrm{AUC}_k)} + a \right] \qquad (8)$$

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 $a = 0.1$, $b = 0.9$ are constant values used to obtain normalized values in the range 0.1–0.9 (to exclude the 'extreme' values 0 and 1). The features reordered according to their importance values are shown in Table 1. As it emerges from the table, the learning process largely confirms the assignment of importance by category, as performed in the method described in Section 4.3.1.

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

Table 1: Features ordered according to their importance values, computed according to Equation (8).

To sum up, let us consider a news event e with 15 values denoted as $x_1, x_2, \ldots, x_{15}$ associated with the features $1 - 15$ previously described. The performance scores $s_1(e), s_2(e), \ldots, s_{15}(e)$ are obtained after normalization according to Equation (4), and the final credibility score σ_e is computed as:

$$\sigma_e = \mathcal{A}_{\mathrm{OWA}}(s_1(e), s_2(e), \ldots, s_{15}(e)) = \sum_{k=1}^{n} w_k b_k, \qquad (9)$$

in which b_k is the kth largest of the $s_i(e)$, as illustrated in Section 3.1. 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*), and to Equation (6) or Equation (7) for aggregation function (*ii*).

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 4.3.1 and 4.3.2, and where Q is expressed according to Equation (5) for aggregation function (*iii*), and to Equation (6) or Equation (7) for aggregation function (*iv*).

5 EVALUATION

For evaluation purposes, the CREDBANK dataset illustrated in Section 4.1 has been considered. On this dataset, first, a binary classification task has been performed by employing the aggregation functions (*i*)–(*iv*) proposed in this paper, by implementing both the totally unsupervised MCDM approach described in Section 4.3.1, and the hybrid approach described in Section 4.3.2. Then, some well-known machine learning algorithms employed in the literature have been tested: SVM, kNN, Decision Trees, Naive Bayes, and Random Forests [37], to comparatively evaluate them with respect to the proposed approach. The effectiveness of the approaches have been evaluated by considering the *accuracy* (*Acc*), *precision* (*Prec*), *recall* (*Rec*), *F1-score* (*F1*), and ROC-AUC metrics [21].

5.1 Implementation Details

The classification and experimental phases have been conducted by employing the *Python* programming language. To manage data, the *pandas* library[5] has been used; to make numerical computations on data, such as the development of the proposed aggregation functions, the *NumPy*[6] library has been used; finally, the *scikit-learn* library[7] has been employed to implement and evaluate the baseline classifiers. It is worth to be underlined that, in particular, the original code provided by Buntain and Golbeck[8] has been employed to perform the 5-fold cross-validation using the 100-tree Random Forest classifier at the basis of their approach.

With respect to the proposed MCDM approach, by applying aggregation functions (i)–(iv) to the performance scores of criteria associated with news events, for each news event an overall credibility score in the [0,1] interval has been obtained. Then, news events 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, in an experimental way, by selecting the one that maximizes classification effectiveness [31].

5.2 Summarization of Results and Discussion

In this section, the results of the above-mentioned evaluation metrics over the considered classification task with respect to the proposed aggregation functions, and the considered data-driven baselines are illustrated. All the approaches have considered the same CREDBANK dataset described in Section 4.1. Table 2 summarizes the obtained results.

	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 [5]	**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 (exp)	0.68	65%	77%	65%	70%
OWA_MOST $(0.5-0.6)$	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 (exp) (a)	0.75	73%	78%	80%	79%
OWA_MOST_I $(0.5-0.6)$ (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 (exp) (b)	0.64	63%	74%	65%	69%
OWA_MOST_I $(0.5-0.6)$ (b)	0.78	77%	85%	77%	81%

Table 2: Summarization of results of all the experiments.

It may be useful to emphasize that naive aggregation functions, such as those guided by the *all* or the *at least one* linguistic quantifiers (corresponding to the *min* and the *max* aggregation operators), were initially considered, but not presented in this work since they do not apply a trade-off among criteria (features), leading to unsatisfactory results. Instead, we considered linguistic quantifiers able to represent more 'flexible' requests of the decision maker, allowing her/him to consider a 'certain amount' of (important) features to be satisfied in terms of credibility. Some of them was already investigated in prior works in the field of fake review detection [36, 38], providing good results. With respect to aggregation functions guided by the *more than k%* and *most* linguistic quantifiers, several configurations have been tested for each of them, i.e., the OWA_MORE (50%) and the OWA_MORE (75%), i.e., more than 50% and more than 75% of criteria satisfied, and the OWA_MOST (exp) and OWA_MOST $(0.5-0.6)$, where in the first case the *most* quantifier is expressed according to Equation (6), while in the second case it is expressed by means of Equation (7) where $\alpha = 0.5$ and $\beta = 0.6$ (these parameters provided the best results for the considered aggregation function).

Aggregation functions considering *different importance* associated with criteria have also been tested, i.e., OWA_MORE_I (50%), OWA_MORE_I (75%), OWA_MOST_I (exp), and OWA_MOST_I $(0.5-0.6)$, both when importance values have been obtained heuristically (a), as described in Section 4.3.1, and when they are learned from a subset of the available data (b), according to Section 4.3.2.

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. Furthermore, as expected, those aggregation functions considering different importance associated with criteria perform better than those considering all criteria as equal.

With respect to the aspect of how defining importance values, in particular, it is interesting to notice that the aggregation functions for which importance values have been defined heuristically based on a prior knowledge, have similar (and even better) performance of those where 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.

In fact, globally, the best results are obtained by the aggregation functions OWA_MORE_I (50%) (a) and OWA_MORE_I (75%) (a), which exceed the baselines with respect to accuracy, precision and F1 score, with a comparable AUC. In Figure 3, are illustrated the ROC curves for the baselines that in the literature have been mostly employed in the specific context of fake news detection (i.e., SVM and Random Forests), and for the most effective aggregation functions proposed in this paper.

6 CONCLUSIONS

In this paper, an approach for fake news detection in microblogging based on the Multi-Criteria Decision Making (MCDM) paradigm has been proposed. In particular, it is a model-driven classification-based approach employing aggregation operators guided by linguistic quantifiers. This solution aims at providing the decision maker with a certain control over the fake news detection process, by exploiting some prior domain knowledge, and, virtually, some learning data. In the proposed model, news represent alternatives to be evaluated in terms of credibility; to each alternative are associated distinct credibility features that are satisfied by the alternative

[5] https://pandas.pydata.org
[6] http://www.numpy.org
[7] http://scikit-learn.org/stable/index.html
[8] https://github.com/cbuntain/CREDBANK-data/tree/master/src/main/python/Labeling

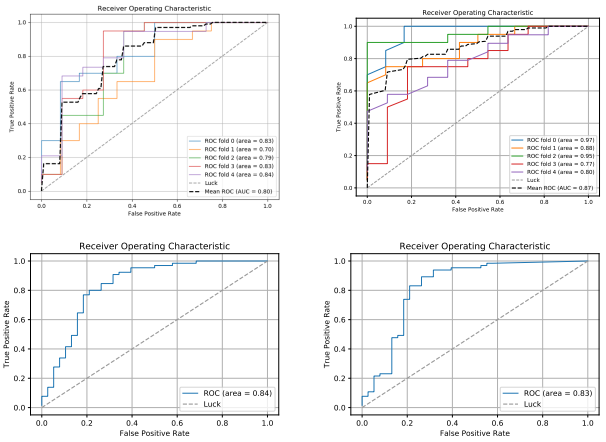

Figure 3: From left to right, the ROC curves and AUC values for SVM and Random Forest baselines, and for OWA_MORE_I (50%) (a), and OWA_MORE_I (75%) (a) aggregation functions.

to a certain credibility extent, which can be expressed as a numerical credibility score. The overall credibility score of an alternative, i.e., a piece of news, is therefore obtained as the aggregation of the distinct credibility scores associated with the alternative. In this scenario, the decision maker can act: (i) on the choice of the aggregation operator (or the family of aggregation operators) to be used, (ii) on the choice of the number or the proportion of the features s/he estimates sufficient to have a good solution (by choosing absolute or relative linguistic quantifiers and their formal representations), and (iii) on the assignment of different importance values to different credibility features, for example exploiting prior knowledge s/he has of the considered domain, or using a subset of the available learning data.

To illustrate the approach and for evaluation purposes, the proposed solution has been instantiated and tested over the Twitter scenario, by exploiting the publicly available CREDBANK dataset, which represents, at present, one the largest labeled dataset of news events. This dataset has been used in [5], which has been taken as baseline due to the high number of credibility features considered, and to its effectiveness with respect to other data-driven solutions, which have also been considered. The proposed model-driven approach, under different configurations of aggregation functions and linguistic quantifiers, has produced promising results.

In the future, further aggregation functions will be considered, taking into account, for example, the interaction among features.

7 ONLINE RESOURCES

Online resources, i.e., the code and the employed datasets, are available at the following link: https://github.com/ir-laboratory/fake-news-detection.

REFERENCES

[1] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

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

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

[4] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

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

[6] T. Calvo, G. Mayor, and R. Mesiar, editors. *Aggregation Operators: New Trends and Applications*. Physica-Verlag GmbH, Heidelberg, Germany, Germany, 2002.

[7] B. Carminati, E. Ferrari, and M. Viviani. A multi-dimensional and event-based model for trust computation in the social web. In *International Conference on Social Informatics*, pages 323–336. Springer, 2012.

[8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proc. of the 20th Int. Conf. on World Wide Web*, pages 675–684. ACM, 2011.

[9] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2012.

[10] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, and X. Tannier. A content management perspective on fact-checking. In *The Web Conference 2018-alternate paper tracks" Journalism, Misinformation and Fact Checking"*, pages 565–574, 2018.

[11] N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[12] G. Eysenbach. Credibility of health information and digital media: New perspectives and implications for youth. In *Digital Media, Youth, and Credibility*, pages 123–154. The MIT Press, 2008.

[13] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[14] P. Galán-García, J. Gaviria de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. Supervised machine learning for the detection of troll profiles in twitter social network. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[15] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3, 2018.

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

[17] 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*, page 2. ACM, 2012.

[18] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM, 2012.

[19] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 230–239. IEEE, 2014.

[20] B. Kang, J. O'Donovan, and T. Höllerer. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 179–188. ACM, 2012.

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

[22] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of the 19th Int. Conf. on WWW*, pages 591–600. ACM, 2010.

[23] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

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

[25] T. Mihaylova, P. Nakov, L. Marquez, A. Barron-Cedeno, M. Mohtarami, G. Karadzhov, and J. Glass. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] T. Mitra and E. Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267, 2015.

[27] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. Fake review detection: Classification and analysis of real and pseudo reviews. Technical report, UIC-CS-03-2013. Technical Report, 2013.

[28] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. 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*, pages 1003–1012. International World Wide Web Conferences Steering Committee, 2017.

[29] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

[30] V. L. Rubin, Y. Chen, and N. J. Conroy. Deception detection for news: three types of fakes. *Proc.s of the Assoc. for Inf. Sci. and Technology*, 52(1):1–4, 2015.

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

[32] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*,

9(1):4787, 2018.

[33] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Exp. News.*, 19(1):22–36, 2017.

[34] J. Thorne and A. Vlachos. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*, 2018.

[35] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):10, 2019.

[36] M. Viviani and G. Pasi. Quantifier guided aggregation for the veracity assessment of online reviews. *International Journal of Intelligent Systems*, 32(5):481–501, 2016.

[37] 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.

[38] M. Viviani and G. Pasi. *A Multi-criteria Decision Making Approach for the Assessment of Information Credibility in Social Media*, pages 197–207. Springer International Publishing, 2017.

[39] N. Vo, K. Lee, C. Cao, T. Tran, and H. Choi. Revealing and detecting malicious retweeter groups. In *Proc. of the 2017 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining 2017*, pages 363–368. ACM, 2017.

[40] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, Jan. 1988.

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

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

[43] L. Zhao, T. Hua, C.-T. Lu, and I.-R. Chen. A topic-focused trust model for twitter. *Computer Communications*, 76:1 – 11, 2016.

[44] X. Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.

[45] X. Zhou, R. Zafarani, K. Shu, and H. Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 836–837. ACM, 2019.

[46] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32, 2018.