Cluster Computing (2020) 23:2375–2388
https://doi.org/10.1007/s10586-019-03009-w

Expertise-aware news feed updates recommendation: a random forest approach

Sami Belkacem[1] · Kamel Boukhalfa[1] · Omar Boussaid[2]

Received: 26 February 2019 / Revised: 10 July 2019 / Accepted: 22 October 2019 / Published online: 1 November 2019

Abstract
With social media being widely used around the world, and because of the large amount of data, users are overcome by updates displayed chronologically in their news feed. Furthermore, most updates are considered irrelevant. To help beneficiary users quickly catch up with the relevant content, ranking news feed updates in descending relevance order has been achieved based on the prediction of a relevance score between a beneficiary and a new update in the news feed. Four types of features are generally used to predict the relevance: (1) the relevance of the update content to the beneficiary's interests; (2) the social tie strength between the beneficiary and the update's author; (3) the author's authority; and (4) the update quality. In this work, from the biography and the textual content posted, we propose an approach that infers and uses another type of feature which is the expertise of the update's author for the corresponding topic. Following extensive experiments on a real dataset crawled from *Twitter*, the results show that infer the author's expertise is critical for identifying relevant updates in news feeds.

Keywords Social media · News feed updates · Relevance · Ranking · Expertise · *Twitter*

1 Introduction

Social media such as *Facebook*, *Twitter*, and *LinkedIn* are used by hundreds of millions of users worldwide and contribute to the concept of *Big data* [1]. Social data are known for their large volumes that can reach petabytes (10^{15} bytes), their variety (text, images, videos, music, etc.), and their velocity (arriving near real-time) [2]. Due to the large amount of data posted and shared [3] on social media [3], users are overcome by a flow of updates displayed chronologically in their news feed [4]. For example, a survey of 587 *Twitter* users showed that 66.3% of them feel they cannot keep up with the large volume of updates in their news feed [5]. Moreover, most of these updates are considered irrelevant [6]. For example, a survey of 56 *Twitter* users indicated that users lose the most relevant tweets in a news feed of thousands of less useful tweets [7]. Therefore, large data volume and irrelevance make it difficult for users to catch up with the relevant updates in their news feed [8].

In several research approaches, ranking news feed updates in descending relevance order has been achieved based on the prediction of a relevance score between a beneficiary user and a new update in the news feed [9]. These approaches generally use four types of features that may influence relevance [10]: (1) the relevance of the update content to the beneficiary's interests; (2) the social tie strength between the beneficiary and the update's author; (3) the author's authority; and (4) the update quality. We believe that using these features is necessary, but not sufficient. For example, updates on a specific topic authored by a novice user may not attract the attention as much as those posted by a recognized expert in his field. Indeed, updates posted by experts, also known as topical influencers [11] or topical authorities [12], are considered credible, important, and interesting [13]. Hence, leverage

✉ Sami Belkacem
s.belkacem@usthb.dz

Kamel Boukhalfa
kboukhalfa@usthb.dz

Omar Boussaid
omar.boussaid@univ-lyon2.fr

[1] LSI laboratory, USTHB, Algiers, Algeria

[2] ERIC laboratory, University Lyon 2, Lyon, France

the author's expertise could be necessary to enable users to catch up with the valuable and trustworthy updates on specific topics.

Expertise is usually not explicitly provided by users [14]. Therefore, existing methods rely on implicit expert finding, which aims at identifying users with the relevant knowledge or experience on a given topic [15]. The main techniques used to infer a user's expertise leverage other users' interactions with the textual content he posted [16] as well as his social behaviors, including: user-generated content, biographical information, social relationships, etc. [14]. In our previous work [17], we proposed an approach that predicts the relevance of news feed updates using supervised prediction models based on *Decision Trees*. This approach leverages the author's expertise that we inferred from the textual content posted, in addition to other features used in related work. The experimental results have shown that judging expertise is crucial for maximizing the relevance in news feeds. However, we believe that the previous approach can be enhanced. In this paper, to improve the latter and better study the contribution of expertise to rank news feed updates, we first extend it by using other features that may influence relevance, and especially those that infer the author's expertise from the biography and the textual content posted. Moreover, we use prediction models based on *Random Forest* to address the limitations of the decision trees. Finally, we make a deeper analysis with a greater number of users.

This work focuses on *Twitter* for the followings reasons: (1) the large flow of tweets encountered by users; (2) the irrelevance of a large part of tweets; (3) the fact that social data are public by default unlike most of the other social media platforms; and (4) the availability of API for easy crawling [18]. However, it would be possible to adapt this work to other social platforms. *Twitter* is a widely used microblogging social media that allows users to communicate using short messages of 280 characters called "tweets" [19]. Each tweet has (see Fig. 1): (1) an author; (2) a set of beneficiaries users who can read and interact with it; (3) a textual and/or multimedia content; (4) a publication date; (5) mentions which represent links to other users; (6) hashtags which identify tweets on specific topics; and (7) URLs to websites or online articles [10]. Following is the only type of social relationship such that users who follow a user $\boldsymbol{u}$ are called *followers* of $\boldsymbol{u}$ and users that $\boldsymbol{u}$ follows are called *followings* of $\boldsymbol{u}$ [20]. If a user $\boldsymbol{u}$ follows another user $\boldsymbol{u}'$, $\boldsymbol{u}$ will receive in the news feed the tweets posted by $\boldsymbol{u}'$.

The paper is structured as follows: Sect. 2 provides background on ranking news feed updates, Sect. 3 discusses work carried out in this area, Sect. 4 describes the proposed approach which uses the author's expertise in addition to other features, Sect. 5 presents the experiments

Fig. 1 Tweet posted by *Elon Musk*

we performed to evaluate our approach, and Sect. 6 concludes the paper.

2 Background

A user's news feed on *Twitter*, or *home timeline*, is a list of tweets where are displayed from most recent to least recent tweets posted by his followings and tweets they retweeted and/or liked [21]. The drawback of the chronological feed is that a user must browse a large number of tweets to not miss those that are relevant [22]. A user can perform three actions to interact with a tweet: (1) *Retweet*: when the user finds the tweet interesting and wants to share it with his followers; (2) *Reply*: when the user wants to answer or comment on the tweet; and (3) *Like*: when the user finds the tweet interesting and wants to save it in the "Likes" section.

Ranking news feed updates on *Twitter* involves ranking and displaying the tweets in descending relevance order [23]. We note that other terms can be used to refer to the ranking process, e.g.: reordering, recommendation, personalization, etc. Unlike the chronological news feed, the ranking process is done in such a way that the most relevant tweets are found at the top of the news feed and the least relevant at the bottom [21].

Berkovsky and Freyne [18] propose the following formalization of the problem of ranking news feed updates: "Let $\boldsymbol{F(u)}$ denotes tweets unread by the beneficiary user $\boldsymbol{u}$

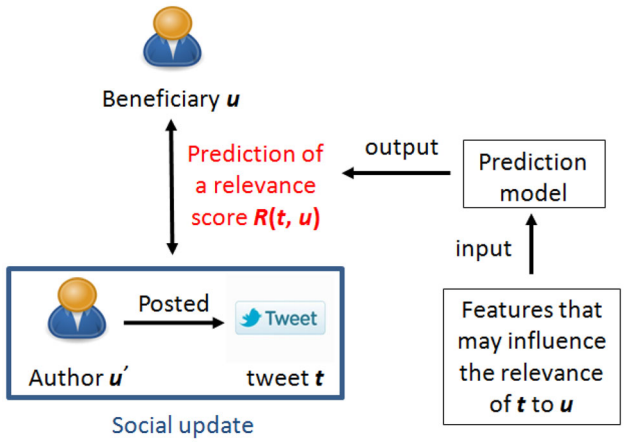


Fig. 2 Prediction of a relevance score

that can potentially be included in the news feed. The ranking process implies selecting and displaying a subset $\boldsymbol{K}(\boldsymbol{u}) \in \boldsymbol{F}(\boldsymbol{u})$, such that $|\boldsymbol{K}(\boldsymbol{u})| \ll |\boldsymbol{F}(\boldsymbol{u})|$, that corresponds to the most relevant tweets to $\boldsymbol{u}$. This ranking involves three steps: (1) predict and assign a relevance score to each tweet $\boldsymbol{t} \in \boldsymbol{F}(\boldsymbol{u})$; (2) select and display, in descending relevance order, the $|\boldsymbol{K}(\boldsymbol{u})|$ tweets with the highest relevance scores at the top of the news feed; and (3) save or delete the remaining $\boldsymbol{F}(\boldsymbol{u})\backslash\boldsymbol{K}(\boldsymbol{u})$ tweets". The rest of this paper focuses on the first step, which is the most important. Figure 2 describes the primary technique used to predict and assign a relevance score to a tweet $\boldsymbol{t} \in \boldsymbol{F}(\boldsymbol{u})$. This technique is based on a relevance prediction model that uses as input a set of features that may influence the relevance of $\boldsymbol{t}$ to the beneficiary user $\boldsymbol{u}$, to output a corresponding relevance score $\boldsymbol{R(t,u)}$ that measures the relevance of $\boldsymbol{t}$ to $\boldsymbol{u}$. We point out that $\boldsymbol{t}$ is posted by an author user $\boldsymbol{u}' \in \boldsymbol{A}(\boldsymbol{u})$, such that $\boldsymbol{A(u)}$ is the set of users that $\boldsymbol{u}$ follows.

3 Related work

Ranking and predicting the relevance of news feed updates has been studied in both industrial and academic communities. In the industrial community, *Facebook*, *Twitter*, and *LinkedIn* are making efforts to rank news feed updates. However, their approaches are most often undisclosed due to commercial sensitivity and competition between companies [18]. These companies further claim that their algorithms have several limitations[1] [9, 22]. For example, *Will Oremus*, a journalist at *Slate.com*, had the rare privilege of meeting the *Facebook* team in charge of the news feed[1]. He stated that the *Facebook* ranking algorithm combines hundreds of features to predict the relevance of news feed updates, but is still likely to provide updates that users find irrelevant. Indeed, according to the journalist, the *Facebook* team has been running a test in which it shows users the top update in their news feed alongside one lower-ranked update, asking them to pick the one they would prefer to read. *Facebook* acknowledged that the results correspond to users' preferences "sometimes", declining to be more specific. When the results do not match, *Facebook* says that points to an area for improvement.

In the academic community, ranking and predicting the relevance of news feed updates has drawn much attention from researchers over the past few years [24]. Due to lack of space, we present and discuss the most representative as well as the most recent works. For the purpose of assigning continuous relevance scores to news feed updates on the Chinese social media *Sina Weibo*, Kuang et al. [8] proposed a ranking model based on weighted linear combinations with static weights. The model uses three types of features: (1) social tie strength between $\boldsymbol{u}$ and $\boldsymbol{u}'$; (2) relevance of the content of $\boldsymbol{t}$ to the interests of $\boldsymbol{u}$; and (3) quality of $\boldsymbol{t}$. To evaluate their approach, the authors asked 1048 volunteers to explicitly assign boolean values to updates (True for relevant and False for irrelevant). The model's *MAE* (Mean Average Precision) was 75%, outperformed several baseline methods, and was improved by 57% as compared to the results of the chronological model.

In an effort to reorder and recommend relevant tweets to *Twitter* users, Shen et al. [21] and Chen et al. [25] proposed prediction models that use five types of features: (1) social tie strength between $\boldsymbol{u}$ and $\boldsymbol{u}'$; (2) relevance of the content of $\boldsymbol{t}$, its hashtags, and mentions to the interests of $\boldsymbol{u}$; (3) quality of $\boldsymbol{t}$; (4) authority of $\boldsymbol{u}'$; and (5) the activity of $\boldsymbol{u}'$ from the number of tweets posted. In [21], the authors first modeled the relevance of tweets by minimizing the pairwise loss of relevant and irrelevant tweets. Then, they used a supervised regression model based on a *Gradient Boosted ranking* algorithm (GBrank) [26] to predict continuous relevance scores for tweets. The model average pairwise reordering accuracy (*ACC*) was improved by 34.5% when comparing the results of the chronological model. On the other hand, in [25], the authors used a probabilistic collaborative ranking model based on *latent factors* to capture users' interests and predict binary rating scores for tweets. The model's *MAE* was 76% and outperformed several baseline methods including the chronological model. The results also indicated that recommended tweets attracted more attention than unrecommended tweets.

In an attempt to rank tweets by relevance on *Twitter*, Feng and Wang [23], De Maio et al. [22], and Vougioukas et al. [6] proposed supervised binary classifier models that use five types of features: (1) social tie strength between $\boldsymbol{u}$ and $\boldsymbol{u}'$; (2) relevance of the content of $\boldsymbol{t}$, its hashtags, and

[1] https://longform.org/posts/who-controls-your-facebook-feed.

mentions to the interests of u; (3) quality of t; (4) authority of u'; and (5) activity of u'. First, in [23], the tweet ranking model is based on matrix factorization and predict the likelihood that a beneficiary retweet a tweet from the news feed. The model's *MAE* was 76.27% and outperformed several baseline methods including the chronological model. In contrast, in [22], the relevance prediction model is based on a *deep learning* method attempting to re-adapt the ranking by preferring the tweets that are interesting to the beneficiary user. The model's *MAE* and *NDCG* (Normalized Discounted Cumulative Gain) outperformed several baseline methods including the chronological model. Lastly, in [6], the prediction model is based on *logistic regression* and predicts if the beneficiary user may find a tweet interesting enough to retweet it. In experiments with a collection of 130K tweets received by 122 journalists, the model average *F1 score* was 90% using the *Pearson* correlation of the top ten features.

In all previous work on *Twitter*, to obtain training and evaluation data, users' interactions with tweets in terms of retweets and replies were used as implicit indicators relevance [6, 21–23, 25]. This implicit method has been widely used for its ease of use since it does not require much effort from users who naturally tend to interact with others. We also notice that supervised learning models have been commonly used and seem to be suitable to rank news feed updates [6, 21, 22]. Indeed, using labeled training data, these models analyze users past behaviors to make predictive assumptions about future outcomes by inferring a function that maps an input tweet to an output relevance score [27]. Finally, we note that four types of features that may influence relevance were widely used (see Fig. 2):

- Features between u and t that measure the relevance of the textual content of t, its hashtags, and mentions to the interests of u [6, 8, 21–23, 25]. Certainly, the features that match between the tweet content and the beneficiary's interests (inner product, cosine similarity, common words, etc.) may serve as direct predictors of relevance [8].
- Features between u and u' that measure social tie strength between them: interaction rate, number of common friends, etc. [6, 8, 21–23, 25]. The assumption is that t could be relevant to u if he has a strong social relationship with u' [22]. Indeed, close friends tend to have common interests and want to catch up with the latest updates from each other [6].
- Features of u' that measure his authority: followers count, followings count, seniority, etc. [6, 21–23, 25]. The assumption is that t could be relevant to u if u' has authority on the social media [23]. Nagmoti et al. [28] state that if a user is important, i.e. has authority, then his tweets are also important.
- Features of t that measure its quality: length, popularity, the presence of a URL, hashtags, multimedia content, etc. [6, 8, 21–23, 25]. Note that these features (length, popularity, etc.) were used by related work to assess quality, but others may be used. The assumption is that t could be relevant to u if it is of high quality (long, popular, formal, informative, etc.) independently of the personal interests [25].

Based on existing work and to the best of our knowledge, the features that measure the expertise of u' in the topics of t (features between u' and t) have not been used by others when predicting the relevance. The assumption is that t could be relevant to the beneficiary u if the author u' is an expert in the topics of t. Indeed, according to Wagner et al. [13], tweets posted by experts, i.e. users with the relevant skill or knowledge on a given topic [29], are considered credible, important and interesting. For example, *Elon Musk*, one of the most famous heroes of the tech culture, is known for his warnings about the risks of Artificial Intelligence and his tweets on this subject often attract users' attention[2]. Therefore, unlike novice users, experts know what they are talking about when it comes to topics they master [30] and identifying these experts could be crucial to allow users to catch up with the valuable tweets on specific topics. In the next section, we introduce our approach that uses the author's expertise in addition to other features considered in related work.

4 Proposed approach

The proposed approach takes as an input a set of tweets $F(u)$ unread by the beneficiary user u that can be included in the news feed, to predict and output a relevance score to each tweet $t \in F(u)$. This approach uses *Random Forest* classifiers [31] as relevance prediction models and, in addition to other features used in related work, leverages the author's expertise that we infer from the biography and tweets posted.

Let denote by S the set of beneficiary users for whom we apply the proposed approach and $D(u)$ a subset of tweets previously read by the user u. In order to use supervised prediction models based on random forests, we first create a training database for each beneficiary user $u \in S$. The training database is a set of input-output pairs, such that an input represents the features that may influence the relevance of a tweet $t \in D(u)$ to the user u, and the corresponding output represents the implicit relevance score $R(t,u)$ that measures the relevance of t to u. The proposed approach involves three steps: (1) assign implicit relevance

[2] www.wired.co.uk/article/elon-musk-artificial-intelligence-world-war-3.

scores to tweets; (2) extract the features that may influence relevance; and (3) train the relevance prediction model. In this section, we describe each of the steps.

4.1 Relevance scores

We assume that a previously read tweet $t \in D(u)$ is relevant to a beneficiary user $u \in S$ if u interacted with t. As given by Eq. 1, predicting implicit relevance scores results in a binary classification problem.

$$R(t,u) = \begin{cases} 1 & \text{if } u \text{ interacted with } t \text{ (retweet, reply, like)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We use the implicit method, which has been used by most related work [6, 21–23, 25], due to its simplicity and the fact that the explicit method used by Kuang et al. [8] has several limitations. It is not related to users' interactions on the one hand (users' feedbacks were obtained via a survey or specifically developed tools), and on the other, it is binding as it asks users to assign relevance scores to a large number of tweets [10]. We also split relevance scores into two bins, relevant and irrelevant, for several reasons. First, we intend to employ the relevance prediction model as a type of spam filter which is binary. Second, train a finer-grained classifier (e.g., t is very relevant to u if he retweeted, liked, and replied to it) would have been difficult since users' multiple interactions with the same tweet are not common [6]. For example, out of 569 tweets, we found that only 5% and 0% of tweets get respectively two and three types of interaction from the same user.

4.2 Feature extraction

We define and extract 16 features that may influence the relevance score $R(t,u)$ that measures the relevance of a tweet t, posted by an author u', to the beneficiary u. These features are summarized in Table 1 and divided into five categories. We use for each category the most relevant features according to related work. The features 1, 2, 4, 5, 6, 7, and 10 are gradually updated as tweets are injected into the news feed from least recent to most recent. We thus simulate an evolution of the social media platform over time. In the rest of the section, we provide a detailed description of each feature.

4.2.1 Relevance of the keywords of t to u

According to Shen et al. [21], keywords of tweets posted by a user and/or with which he interacted implicitly reflect his topics of interest and may serve as direct predictors of relevance. First, we represent the topics of t by keywords extracted using *DBpedia*[3] *Spotlight* annotation service[4]. Each keyword is represented by the URI[5] (Uniform Resource Identifier) corresponding to the annotated resource. Note that this method is not compromised by the short and informal nature of tweets [32], unlike methods used in related work which are mainly based on TF-IDF [6, 23] or a topic model [21, 25]. Using a *support* = 20 and a *confidence* = 0.32 for example, *DBpedia Spotlight* outputs the following keywords to the tweet in Fig. 1: *2060s, National_Geographic_Channel, Human, Mars*. After that step, we use the following proposed equation to calculate this feature:

$$f_1(u,t) = \sum_{i=1}^{nbk(t)} P(u, k_i(t)) \quad (2)$$

where $k_i(t)$ is the i^{th} keyword of t, $nbk(t)$ is the number of keywords of t, $P(u, k_i(t))$ is the number of times u has previously posted and/or interacted with $k_i(t)$.

For example, if t has 2 keywords k_1 and k_2, and u has previously posted and/or interacted 10 times with k_1 and 5 times with k_2, then the sum of the two values, i.e. 15, will be assigned to this feature.

4.2.2 Relevance of the hashtags of t to u

According to Chen et al. [25], hashtags of tweets posted by a user and/or with which he interacted implicitly reflect his topics of interest and may serve as key predictors of relevance. We first lowercase hashtags and then use the following proposed equation to calculate this feature:

$$f_2(u,t) = \sum_{i=1}^{nbh(t)} P(u, h_i(t)) \quad (3)$$

where $h_i(t)$ is the i^{th} hashtag of t, $nbh(t)$ is the number of hashtags of t, $P(u, h_i(t))$ is the number of times u has previously posted and/or interacted with $h_i(t)$.

4.2.3 Relevance of the mentions of t to u

A tweet that mentions u is likely to match his interests and be relevant. Feng and Wang [23] state that a mention draws a user's attention to the corresponding tweet. The authors propose to calculate this feature with a boolean variable.

4.2.4 Interaction rate of u with u'

As reported by Vougioukas et al. [6], if u interacted (retweet, reply, like) frequently with tweets posted by u' in

[3] A project that extracts structured content from *Wikipedia*.

[4] http://demo.dbpedia-spotlight.org/.

[5] A string of characters used to identify a resource.

Table 1 Features that may influence relevance

Features that may influence relevance		N°
Relevance of the content of t, its hashtags, and mentions to u	Relevance of the keywords of t to u	*f1*
	Relevance of the hashtags of t to u	*f2*
	Relevance of the mentions of t to u	*f3*
Social tie strength between u and u'	Interaction rate of u with u'	*f4*
	Number of times u mentioned u'	*f5*
Expertise of u' in the topics of t	Publishing rate of u' for keywords of t	*f6*
	Interaction rate with keywords of t posted by u'	*f7*
	Keywords of u' biography and keywords of t	*f8*
Authority of u'	Followers count / Followings count	*f9*
	Seniority	*f10*
	Listed count	*f11*
Quality of t	Length	*f12*
	Presence of hashtags	*f13*
	Presence of a URL	*f14*
	Presence of an image or a video	*f15*
	Popularity	*f16*

the past, then they tend to have a strong social relationship, and u may find the tweets of u' relevant in the future. We propose to calculate this feature with the following equation:

$$f_4(u,u') = \frac{|\text{Tweets posted by } u' \text{ with which } u \text{ interacted}|}{|\text{Tweets posted by } u' \text{ that } u \text{ has previously read}|} \quad (4)$$

4.2.5 Number of times u mentioned u'

According to Chen et al. [25], if u mentioned u' in the past, then they tend to have a strong social relationship, and u may find the tweets of u' relevant in the future. The authors propose to calculate this feature by counting the number of times u mentioned u' in tweets he posted.

4.2.6 Publishing rate of u' for keywords of t

Xu et al. [33] assert that a user is likely to have expertise in the topics in which he frequently expresses his opinion. We propose to calculate this feature using Eq. 5 since we represent tweets' topics by keywords extracted using *DBpedia Spotlight*. We recall that our aim is not to propose a method that infers expertise, but rather to study its contribution to rank news feed updates.

$$f_6(u',t) = \frac{\sum_{i=1}^{nbk(t)} Post(u', k_i(t))}{nbk(t) \times nbp(u')} \quad (5)$$

where $k_i(t)$ is the i^{th} keyword of t, $nbk(t)$ is the number of keywords of t, $nbp(u')$ is the number of tweets previously posted by u', $Post(u', k_i(t))$ is the number of times u' has previously posted $k_i(t)$.

For example, if t has 3 keywords k_1, k_2, k_4 and u' has previously posted 2 tweets t_1 and t_2, that have respectively the keywords k_1, k_2 and k_1, k_3, then the following value will be assigned to this feature:

$$f_6(u',t) = \frac{2+1+0}{3 \times 2} = 0.5 \quad (6)$$

4.2.7 Interaction rate with keywords of t posted by u'

Li et al. [16] state that tweets on specific topics authored by a recognized expert may attract users' attention and get more interactions. We propose to calculate this feature using Eq. 7. The assumption is that the author u' is an expert in the topics with which his followers interacted most (retweet, reply, like).

$$f_7(u',t) = \frac{\sum_{i=1}^{nbk(t)} Interaction(u', k_i(t))}{\sum_{i=1}^{nbk(t)} Post(u', k_i(t)) \times nbf(u') \times 3} \quad (7)$$

where $k_i(t)$ is the i^{th} keyword of t, $nbk(t)$ is the number of keywords of t, $Post(u', k_i(t))$ is the number of times u' has previously posted $k_i(t)$, $nbf(u')$ is the number of followers of u', 3 is the maximum number of interactions by a user, $Interaction(u', k_i(t))$ is the number of interactions with tweets previously posted by u' that have $k_i(t)$.

For example, if t has 3 keywords k_1, k_2, k_4, and u' has 20 followers and has previously posted 2 tweets t_1 and t_2, that have respectively the keywords k_1, k_2 and k_1, k_3, and have get respectively 35 and 15 interactions, then the following value will be assigned to this feature:

$$f_7(u',t) = \frac{50+35+0}{(2+1+0)\times 20\times 3} = 0.47 \quad (8)$$

4.2.8 Keywords of u' biography and keywords of t

As reported by Wagner et al. [13], self-reported biographies often contain information that indicates users' expertise such as education, skills, and career information. We propose to calculate this feature as follows:

$$f_8(u',t) = \begin{cases} 1 & \text{if } |kb(u') \cap k(t)| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where $k(t)$ is the set of keywords of t, $kb(u')$ is the set of keywords of the biography of u', extracted in the same way as tweets' keywords.

4.2.9 Followers count/followings count

According to Pan et al. [34], users who have authority on social media tend to have more followers than followings. The authors propose to calculate this feature by dividing the followers count of u' by the followings count.

4.2.10 Seniority

Shen et al. [21] state that senior users, i.e. users whose accounts were created early, tend to have authority on social media. We propose to calculate this feature with the following equation:

$$f_{10}(u',t) = \text{Year in which } t \text{ was posted} - \text{Year in which the account of } u' \text{ was created} \quad (10)$$

4.2.11 Listed count

Twitter lists allow users to organize people they follow into labeled groups [15]. According to Duan et al. [35], the number of lists to which a user has been added is an adequate representation of the authority on social media. Uysal and Croft [36] propose to assign the list count of u' to this feature.

4.2.12 Length

Chen et al. [25] assert that a long tweet is likely to be more formal, more informative, and of better quality. Vougioukas et al. [6] propose to calculate the length of the tweet t by counting its number of characters.

4.2.13 Presence of hashtags

A tweet with hashtags can provide more information and be of better quality. Indeed, Chen et al. [25] state that the author spent time on tagging the tweet thinking it might be useful. We propose to calculate this feature with a boolean variable.

4.2.14 Presence of a URL

As a tweet is limited to 280 characters, De Maio et al. [22] assert that users tend to include a URL to a website containing more details. According to the authors, a tweet with a URL can provide more information and be of better quality. They propose to calculate this feature with a boolean variable.

4.2.15 Presence of an image or a video

Feng and Wang [23] state that a tweet with multimedia content can give more details and be of good quality. The authors propose to calculate this feature with a boolean variable.

4.2.16 Popularity

Kuang et al. [8] assert that the more users interact with a tweet, the better its quality. We propose to calculate this feature with the following equation:

$$f_{16}(t) = \text{Number of retweets of } t + \text{Number of replies to } t + \text{Number of likes of } t \quad (11)$$

4.3 Relevance prediction model

First, considering each previously read tweet $t \in D(u)$ from least recent to most recent, we create the training database instances of each beneficiary user $u \in S$ in the form of input-output pairs. An input represents the features that may influence the relevance of t to u, and the output represents the implicit relevance score $R(t,u)$ that measures the relevance of t to u.

Second, we divide the training database of each beneficiary user $u \in S$ into two sets: a training set of the

relevance prediction model for 70% of the first instances (the least recent ones) and a test set for 30% of the remaining instances (the most recent ones). The latter set will be used to evaluate the prediction model.

Finally, to create a supervised prediction model for each user $\boldsymbol{u} \in \boldsymbol{S}$, we use the corresponding training set to train a *Random Forest* classifier [31] which fits a number of *Decision Tree* estimators [37]. The purpose is to predict relevance scores for tweets unread by $\boldsymbol{u}$ using multiple decision trees learned from previously read tweets in the training set. The random forest algorithm has been extremely successful as a classification and regression method in a wide range of prediction problems such as data science, bioinformatics, 3D object recognition, etc. [38]. We choose random forests as prediction models because they [39]: (1) tend to result in powerful prediction models, especially for binary classification problems; (2) do not need data preprocessing; (3) do not require parameterization; (4) are fast to train; (5) can handle a large number of features; (6) implicitly perform feature selection; (7) seldom overfit; and (8) allow to compute feature importance in judging the relevance of tweets by users and study the importance of expertise (given by equations 5, 7, and 9) compared to other features used in related work. Note that other supervised learning algorithms are also applicable and that it is out of the scope of this paper to compare them.

Random Forest is an *ensemble learning* method[6] for classification and regression problems that operate by constructing a multitude of *Decision Trees* [39]. Each tree is distinguished by the sub-sample vector on which it is trained, and which is randomly selected with the same distribution from the training set [31]. This method uses averaging to improve the predictive performance and correct the decision trees' habit of overfitting [40]. As discussed in Sect. 4.1, we recall that predicting relevance scores of tweets is a binary classification problem. The rest of the section therefore focuses on decision tree and random forest classifiers. As shown in Fig. 3, a decision tree classifier for a beneficiary user $\boldsymbol{u} \in \boldsymbol{S}$ is a flowchart-like structure in which a node represents a test on a feature that may influence relevance, a branch represents the outcome of the test, a leaf node represents the class label (relevance score) of a tweet $\boldsymbol{t}$. Note that more details about decision trees are provided in [37]. As regards random forest classifiers, Breiman [31] proposes the following definition:

Definition 1 A random forest classifier consists of a collection of tree-based classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed

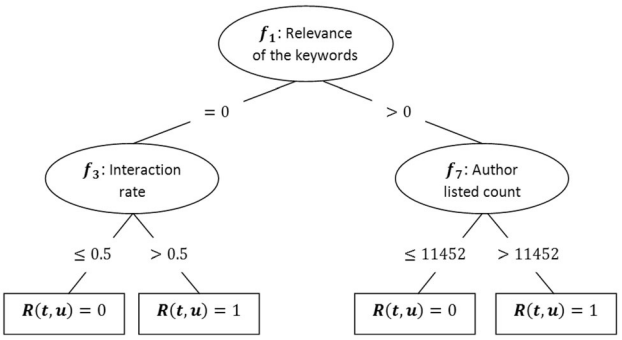


Fig. 3 *Decision Tree* for the *Twitter* user *Medium*

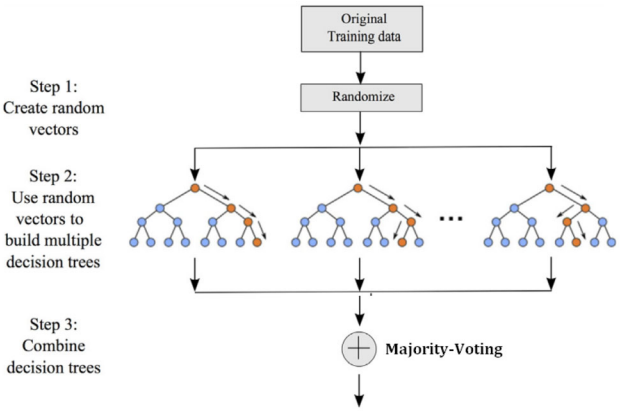


Fig. 4 The random forest methodology [41]

random vectors. Each tree in the forest casts a unit vote for the most popular class at input $\mathbf{x}$.

From the previous definition, and as reported by Malekipirbazari and Aksakalli [41], the random forest classifier methodology can be formally described in the following three steps (see Fig. 4):

1. a random sub-sample vector Θ_k is created from the training set for the k^{th} decision tree. The independence property enforces that the random vector Θ_k is independent of the past random vectors $\Theta_1 \ldots \Theta_{k-1}$, but with the same distribution;
2. multiple decision trees are build so that the k^{th} tree is build using the random vector Θ_k, resulting in a classifier $h(\mathbf{x}, \Theta_k)$ where $\mathbf{x}$ is an input vector.
3. to classify a new tweet from an input vector $\mathbf{x}$, the tweet is presented to each of the trees in the forest. Each tree first gives a classification output (relevant tweet, or not) and then the forest chooses the classification having the most votes. To combine tree classifiers $\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_k(\mathbf{x})\}$ using the training set drawn randomly from the distribution of the

[6] A method that use multiple machine learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms alone.

random vector $\mathbf{X}$, Y, the margin function mg can be defined with the following equation [31]:

$$mg(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j) \quad (12)$$

where av is the average and I the indicator function [42]. The margin measures the extent to which the average number of votes at $\mathbf{X}$, Y for the right class exceeds the average vote for any other class.

5 Experimentation and results

To evaluate our approach and study the contribution of the author's expertise to rank news feed updates, we describe in this section: (1) the dataset used in the experiments we performed; (2) the measures used to evaluate the performance; and (3) the obtained results.

5.1 Dataset

First, we randomly selected a set $\boldsymbol{S}$ of 46 beneficiary users for whom we apply the proposed approach. Each user $\boldsymbol{u} \in \boldsymbol{S}$ meets the following criteria: (1) has a great number of social relationships (at least 20 followings, which we believe is the minimum number). Ranking news feed updates is proposed to help especially this kind of user to catch up with the relevant updates [34]. We note that the average number of followings is 72 followings per beneficiary user; (2) interacts (retweet, reply, like) frequently with tweets from the news feed (interaction rate greater than or equal to 10%, which we believe is the minimum rate). We apply this criterion due to the use of the implicit training and evaluation method, which assumes that user interaction with a tweet involves its relevance; and (3) English-speaking in order to use the English version of *DBpedia Spotlight* which is the most efficient [43]. Then, we collected using *Twitter Rest API*[7] over a period of 10 months all data needed for the proposed approach. As regards tweets' keywords, they were extracted using the English version of *DBpedia Spotlight* with a *support* = 20 and a *confidence* = 0.32 (values determined through experiments).

We report that it is impossible to directly retrieve a user's news feed on *Twitter* [25] and say in case of non-interaction if a particular user has read a given tweet [36]. Therefore, to simulate the news feed of each beneficiary user $\boldsymbol{u} \in \boldsymbol{S}$, we used a variant of the principle proposed by Feng and Wang [23] to select, $\boldsymbol{D(u)}$, the subset of tweets posted by the followings of $\boldsymbol{u}$ that he may have read. We recall that the tweets read by $\boldsymbol{u}$ with which he did not interact are considered irrelevant according to the implicit training and evaluation method (see Sect. 4.1). The variant is as follows:

1. sort all the tweets posted by the followings of $\boldsymbol{u}$ in chronological order from least recent to most recent;
2. for each tweet $\boldsymbol{t}$ with which $\boldsymbol{u}$ interacted, keep the chronological session defined by the tweet $\boldsymbol{t}$, the tweet before $\boldsymbol{t}$ and the tweet after $\boldsymbol{t}$;
3. after deleting duplicates, sort the selected tweets again from least recent to most recent.

Note that the use of our variant resulted in an interaction rate with tweets of approximately 35% for each beneficiary user and an average number of instances of 569 instances in the training database of each user.

5.2 Measures

First, we train a random forest classifier for each user $\boldsymbol{u} \in \boldsymbol{S}$ using the corresponding training set (70% of the least recent instances in the training database, see Sect. 4.3). Then, we define the following concepts to evaluate our approach using the corresponding test set (30% of the most recent instances) [44]:

- TP (True Positive): number of relevant tweets correctly predicted relevant to $\boldsymbol{u}$
- TN (True Negative): number of irrelevant tweets correctly predicted irrelevant to $\boldsymbol{u}$
- FP (False Positive): number of irrelevant tweets incorrectly predicted relevant to $\boldsymbol{u}$
- FN (False Negative): number of relevant tweets incorrectly predicted irrelevant to $\boldsymbol{u}$

After that, we use the weighted *F1 score* measure denoted by F and given by Eq. 13 [44]. This equation calculates the standard *F1 score* for each class and finds their average weighted by support (the number of true instances for each class). We use this measure to evaluate the performance since classes are slightly unbalanced with an interaction rate with tweets of approximately 35% for each beneficiary user. Moreover, we are interested in measuring the performance of predicting both relevant and irrelevant tweets classes.

$$F = \frac{(F_r \times (TP + FN)) + (F_i \times (TN + FP))}{TP + TN + FP + FN} \quad (13)$$

where F_r is the *F1 score* for the relevant tweets class, F_i is the *F1 score* for the irrelevant tweets class.

Lastly, for each beneficiary user $\boldsymbol{u} \in \boldsymbol{S}$, we perform experiments with the corresponding test set using the F *score* and compare two approaches: our approach that uses

[7] https://dev.twitter.com/rest/public.

Table 2 Experimental results

User's screen name	Number of instances	F with expertis	F without expertise	C
Astro0Glen	1245	83,26	81,61	1,65
Astro_Pam	837	72,83	73,77	− 0,94
bamwxcom	123	86,62	86,62	0
Baronatrix	220	76,43	77,79	− 1,36
BethStamper6	760	72,57	71,93	0,64
byudkowsky	185	70,51	70,51	0
ch402	497	86,12	79,66	6,46
demishassabis	319	86,58	88,5	− 1,92
eevil_abby	148	77,41	78,94	− 1,53
elonmusk	303	91,13	92,34	− 1,21
GeorgeHarrison	177	85,49	79,76	5,73
GilmoreGuysShow	105	90,2	64,84	25,36
gwern	771	69,54	70,26	− 0,72
homebrew	1115	81,89	80,21	1,68
HybridZizi	1039	73,87	73,31	0,56
jadelgador	1401	84,46	81,82	2,64
JHUBME	516	90,45	89,18	1,27
JohnDawsonFox26	1649	79,69	80,04	− 0,35
john_walsh	154	97,89	93,73	4,16
kilcherfrontier	708	85,77	81,69	4,08
LKrauss1	115	68,32	54,29	14,03
mastenspace	135	90,38	80,76	9,62
Medium	56	87,55	80,47	7,08
microphilosophy	112	65,04	67,3	− 2,26
MIRIBerkeley	195	87,87	87,87	0
NASAKepler	86	83,23	74,84	8,39
NASA_Wallops	116	80,99	70,69	10,3
newscientist	281	78,66	75,74	2,92
PattiPiatt	1173	80,05	80,11	− 0,06
peterboghossian	1388	74,83	69,96	4,87
rafat	782	81,87	81,53	0,34
realDonaldTrump	140	85,89	88,17	− 2,28
Red_or_MC1R	306	72,09	69,26	2,83
renormalized	543	73,77	74,9	− 1,13
RossTuckerNFL	772	89,21	87,11	2,1
RoxanneDawn	366	83,24	84,37	− 1,13
scimichael	2940	79,87	78,45	1,42
SfNtweets	198	96,67	96,67	0
slatestarcodex	271	65,85	64,77	1,08
SLSingh	560	73,34	68,26	5,08
sxbegle	382	79,25	77,07	2,18
TeslaRoadTrip	1692	86,71	84,34	2,37
TheMuslimReform	27	100	100	0
TheRickDore	202	86,51	82,81	3,7
USDISA	522	76,19	71,48	4,71
WestWingWeekly	548	82,83	80,7	2,13
Average	569	81,59	78,88	2,71

C = F with expertise − F without expertise

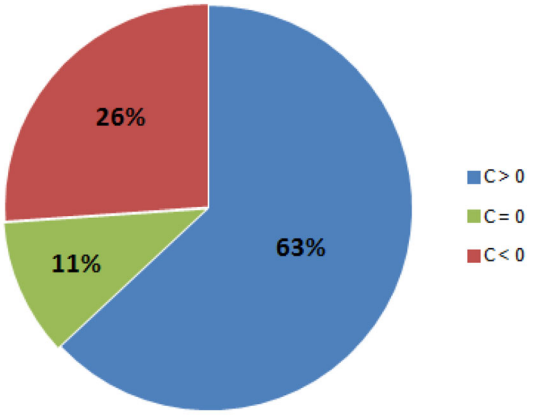


Fig. 5 The contribution of expertise for all users

the author's topical expertise and a classical approach that is the same as ours except that it does not use it. Moreover, to study the importance of the author's expertise compared to other features used in related work, we compute the importance scores of features when judging the relevance of tweets by users. As described by Breiman and Cutler [39], the importance of a feature is calculated as the normalized total reduction of the criterion brought by that feature and is known as the *Gini importance*. Note that the higher the value, the more important the feature. More details about the importance of features are provided in [39].

5.3 Results

In all experiments, the results were obtained using 50 decision tree estimators in each random forest classifier. This choice was made following other experiments.

First, Table 2 presents the experimental results of the comparison between the proposed approach and the classical one, and Fig. 5 summarizes the contribution of the author's expertise for all beneficiary users. The contribution of expertise for a beneficiary user, denoted *C*, is computed by subtracting the *F score* with expertise from the corresponding *F score* without it.

The results of Table 2 show that our approach most often succeeds in predicting relevance scores of tweets with an average *F score* of 81.59%. Note that, considering the current dataset, the latter score has been improved by +12.34%, from 69.25% to 81.59%, when comparing to the previous approach which uses decision tree models with fewer features [17]. Furthermore, we point out that the proposed approach gives remarkable results for several beneficiary users with an *F score* of more than 90%: 100% for *TheMuslimReform*, 97.89% for *john_walsh*, 96.67% for *SfNtweets*, etc. Certainly, this proves that the features and random forest models we used are perfectly suitable to rank news feed updates for these users. The results of Table 2 and Fig. 5 also indicate that the gain brought by the author's expertise is positive for 63% of beneficiary users when predicting relevance scores of tweets. Undoubtedly, this highlights that judging expertise is critical for

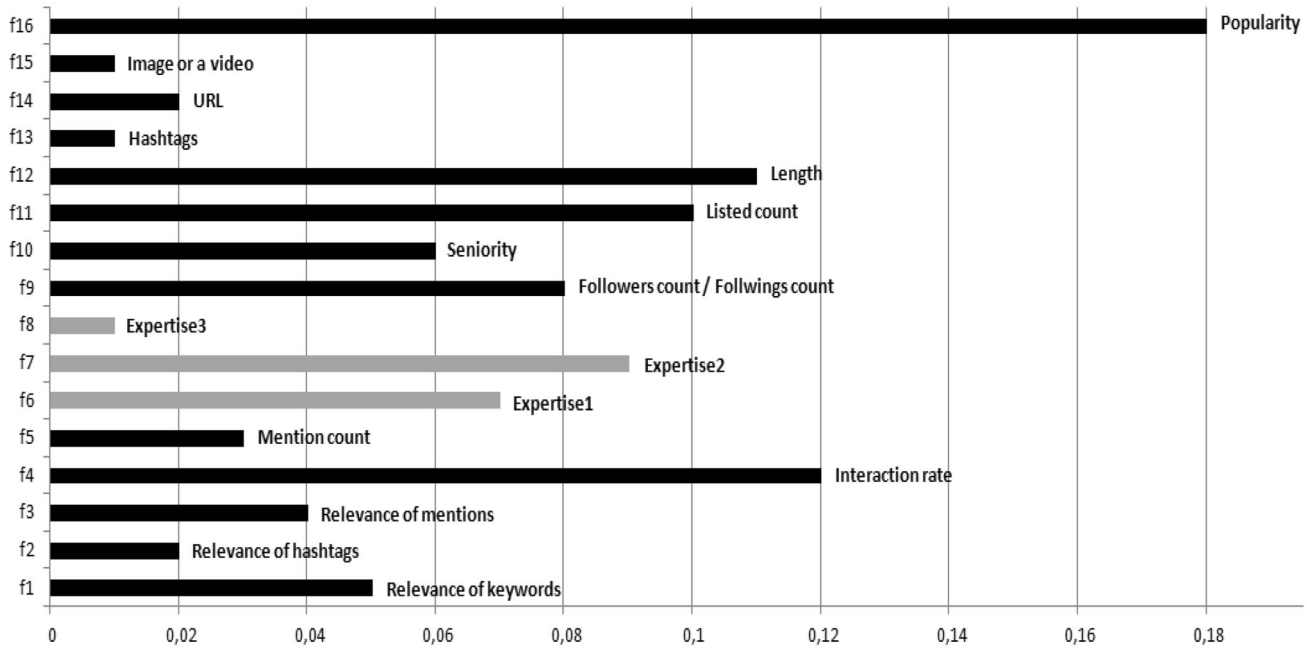


Fig. 6 Average importance of features for users

identifying relevant updates in news feeds. Moreover, we notice that infer the author's expertise made a highly significant contribution to several users, e.g.: +25.36% for *GilmoreGuysShow*, +14.03% for *LKrauss1*, +10.3% for *NASA_Wallops*, etc. Indeed, this confirms that infer expertise enables beneficiary users, and especially those who value it, to catch up with the valuable, up-to-date, and reliable tweets on specific topics.

Second, we computed the average feature importance scores for all beneficiary users which are presented in Fig. 6. The results show that the top feature is the feature f_{16} (0.18) which measures the tweet's popularity. Indeed, according to Kuang et al. [8], the more users interact with a tweet, the more likely it is to be of high quality. The second most important feature is the feature f_4 (0.12), the interaction rate of $\boldsymbol{u}$ with tweets posted by $\boldsymbol{u}'$. Certainly, if $\boldsymbol{u}$ found tweets posted by $\boldsymbol{u}'$ relevant in the past, then they tend to have a strong social relationship, and $\boldsymbol{u}$ may find the tweets of $\boldsymbol{u}'$ relevant in the future [23]. The third most important feature is the feature f_{12} (0.11) which measures the tweet length. Chen et al. [25] claim that a long tweet is likely to be more formal, more informative, and of better quality. The results also indicate that the features f_6 and f_7 (0.07 and 0.09 respectively) which we have introduced to infer the author's expertise from the tweets posted are extremely important. Undoubtedly, tweets posted by a user have been used in several works to infer expertise [14] and leveraging the latter is efficient in judging the relevance of tweets. Moreover, the results show that the features f_9, f_{10}, and f_{11} (0.08, 0.06, and 0.1 respectively) which measure the author's authority are very important. These features are consistent with the observations in [28] that state that if a user is important, then his updates are also important. We further notice that the features f_1, f_2, and f_3 (0.05, 0.02, and 0.04 respectively) which measure respectively the relevance of the content of $\boldsymbol{t}$, its hashtags, and mentions to $\boldsymbol{u}$, are not the most important features. This proves that predicting relevance scores is a difficult task because the most important features are not the most intuitive. Furthermore, we note that the features f_5 (0.03) which represents the number of times $\boldsymbol{u}$ mentioned $\boldsymbol{u}'$ in tweets he posted is not very important. Indeed, it is not very common for users to mention other users, except for dialogues [6]. The results also reveal that the feature f_8 (0.01) which we have used to infer the author's expertise from the biography is surprisingly not important. This is probably due to the fact that users do not frequently change their biography and that this feature is sparse since not all users provide a full description [12]. Finally, we notice that the features f_{13}, f_{14}, and f_{15} (0.01, 0.02, and 0.01 respectively) which measure tweet quality are not really important since they are non-personalized features which do not take into consideration the preferences of each user [23].

Despite the improvements we have made, we observe that the proposed approach has limitations for a small number of users. First, from Table 2, the experimental results show that our approach gives modest results for a few users with an *F score* of less than 70%: 69.54% for *gwern*, 68.32% for *LKrauss1*, 65.85% for *slatestarcodex*, and 65.04% for *microphilosophy*. The number of instances in the training database does not seem to affect the results obtained (higher number of instances does not necessarily involves better results and vice-versa), e.g., results for *gwern* and *Medium*. Therefore, we believe that these results are due to: (1) the fact that we do not have the browsing information of users which is crucial to our approach. More specifically, we are not sure whether a beneficiary user has read a tweet or not in case of non-interaction [36]. This certainly affects the results for both training and evaluation; and (2) the implicit training and evaluation method and more precisely the FP measure (which was on average 13 irrelevant tweets incorrectly predicted relevant) may wrongly penalize our approach. Indeed, non-interaction is not always synonym of irrelevance. A user can find a tweet relevant and choose not to interact with it [10].

Second, from Table 2 and Fig. 5, the results show that infer the author's expertise has no contribution to 11% of beneficiary users when predicting relevance scores of tweets, e.g. *byudkowsky* and *SfNtweets*. The results also indicate that the gain brought by this feature is negative for 26% of users, e.g.: − 1.21% for *elonmusk*, − 1.53 % for *eevil_abby*, and − 1.13% for *renormalized*. We believe that these results are due to: (1) the fact that *Twitter* is a general public social media and not a specialized, professional, or academic one [45]. Indeed, different types of tweets are encountered by users in their news feed. Some tweets are related to highly specialized areas where expertise is important, e.g.: technology, science, sports, etc. and some tweets are not, e.g.: tweets about friends, routine activities, personal experiences, etc. [15]; and (2) users information needs which are different on social media [30]. Certainly, a minority of beneficiary users may not give importance to the author's expertise when judging the relevance of tweets.

6 Conclusion and future work

In this work, we proposed an approach that predicts the relevance of news feed updates using supervised prediction models based on *Random Forests*. In addition to the four types of features used in related work: (1) the relevance of the update content to the beneficiary's interests; (2) the social tie strength between the beneficiary and the update's author; (3) the author's authority; and (4) the update quality, our approach uses the author's expertise that we

inferred from the biography and the textual content posted. Following extensive experiments on a real dataset crawled from *Twitter*, the results show that our approach most often succeeds in predicting relevance and that infer expertise is critical for identifying valuable updates in news feeds.

For now, we only evaluated our approach implicitly using users' interactions. For further work, we first intend to get explicit users' feedback by asking their opinion on the predicted relevance scores. Moreover, it would be interesting to integrate freshness of tweets (date and time) and conduct a study to identify the characteristics of beneficiary users who value expertise. Finally, we plan to use other relevance prediction models and compare them with random forests.

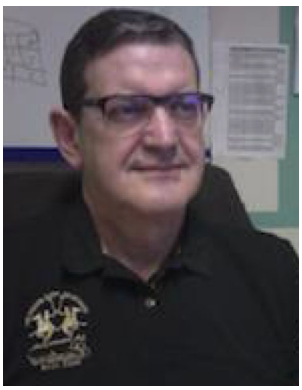
References

1. Jin, S., Lin, W., Yin, H., Yang, S., Li, A., Deng, B.: Community structure mining in big data social media networks with MapReduce. Clust. Comput. **18**(3), 999–1010 (2015)
2. Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., Hu, C.: Crowdsourcing based description of Urban emergency events using social media big data. In: IEEE Transactions on Cloud Computing, p. 1 (2016)
3. Srividya, M., Irfan Ahmed, M.S.: A filtering of message in online social network using hybrid classifier. Clust. Comput. (2017). https://doi.org/10.1007/s10586-017-1300-y
4. Arularasan, A.N., Suresh, A., Seerangan, K.: Identification and classification of best spreader in the domain of interest over the social networks. Clust. Comput. **22**, 4035–4045 (2018)
5. Bontcheva, K., Gorrell, G., Wessels, B.: Social media and information overload: survey results. arXiv preprint arXiv:1306.0813 (2013)
6. Vougioukas, M., Androutsopoulos, I., Paliouras, G.: Identifying Retweetable Tweets with a Personalized Global Classifier. CoRR. abs/1709.06518 (2017)
7. Ramage, D., Dumais, S.T., Liebling, D.J.: Characterizing microblogs with topic models. ICWSM **10**(1), 16 (2010)
8. Kuang, L., Tang, X., Yu, M., Huang, Y., Guo, K.: A comprehensive ranking model for tweets big data in online social network. EURASIP J. Wirel. Commun. Netw. **2016**(1), 46 (2016)
9. Agarwal, D., Zhang, L., Chen, B.-C., He, Q., Hua, Z., Lebanon, G., Ma, Y., Shivaswamy, P., Tseng, H.-P., Yang, J.: Personalizing LinkedIn Feed, pp. 1651–1660. ACM Press, New York (2015)
10. Belkacem, S., Boukhalfa, K., Boussaid, O.: News feeds triage on social networks: a survey. In: Proceedings of The 2nd International Conference on Computing Systems and Applications (CSA), pp. 34–43 (2016)
11. Alp, Z.Z., Öğüdücü, G.: Identifying topical influencers on twitter based on user behavior and network topology. Knowl. Based Syst. **141**, 211–221 (2018)
12. Pal, A., Herdagdelen, A., Chatterji, A., Taank, S., Chakrabarti, D.: Discovery of Topical Authorities in Instagram, pp. 1203–1213. ACM Press, New York (2016)
13. Wagner, C., Liao, V., Pirolli, P., Nelson, L., Strohmaier, M.: It's Not in their Tweets: Modeling Topical Expertise of Twitter Users, pp. 91–100. IEEE, Washington (2012)
14. Yu, X., Dong, Z., Lawless, S.: Inferring Your Expertise from Twitter: Combining Multiple Types of User Activity, pp. 589–598. ACM Press, New York (2017)
15. Wei, W., Cong, G., Miao, C., Zhu, F., Li, G.: Learning to find topic experts in Twitter via different relations. IEEE Trans. Knowl. Data Eng. **28**(7), 1764–1778 (2016)
16. Li, X., Cheng, S., Chen, W., Jiang, F.: Novel user influence measurement based on user interaction in microblog. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 615–619. IEEE (2013)
17. Belkacem, S., Boukhalfa, K., Boussaid, O.: Leveraging expertise in news feeds: a Twitter case study. In: Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA), vol. 14, pp. 1–16. Revue des Nouvelles Technologies de l'Information (2018)
18. Berkovsky, S., Freyne, J.: Personalised network activity feeds: finding needles in the haystacks. In: Atzmueller, M., Chin, A., Scholz, C., Trattner, C. (eds.) Mining. Modeling, and Recommending 'Things' in Social Media, vol. 8940, pp. 21–34. Springer, Cham (2015)
19. Gligoric, K., Anderson, A., West, R.: How Constraints Affect Content: The Case of Twitter's Switch from 140 to 280 Characters. arXiv preprint arXiv:1804.02318 (2018)
20. Nguyen, T.-T., Nguyen, T.-T., Ha, Q.-T.: Applying Hidden Topics in Ranking Social Update Streams on Twitter, pp. 180–185. IEEE, Washington (2013)
21. Shen, K., Jianmin, W., Zhang, Y., Han, Y., Yang, X., Song, L., Xiao, G.: Reorder user's tweets. ACM Trans. Intell. Syst. Technol. **4**(1), 1–17 (2013)
22. De Maio, C., Fenza, G., Gallo, M., Loia, V., Parente, M.: Time-aware adaptive tweets ranking through deep learning. Future Gen. Comput. Syst. (2017). https://doi.org/10.1016/j.future.2017.07.039
23. Feng, W., Wang, J.: Retweet or Not?: Personalized Tweet Re-ranking, p. 577. ACM Press, New York (2013)
24. Belkacem, S., Boukhalfa, K., Boussaid, O.: Tri des actualités sociales: Etat de l'art et Pistes de recherche. In: Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA), vol. 13, pp. 85–100. Revue des Nouvelles Technologies de l'Information (2017)
25. Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., Yu, Y.: Collaborative Personalized Tweet Recommendation, p. 661. ACM Press, New York (2012)
26. Zheng, Z., Chen, K., Sun, G., Zha, H.: A Regression Framework for Learning Ranking Functions Using Relative Relevance Judgments, p. 287. ACM Press, New York (2007)
27. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press, Cambridge (2012)
28. Nagmoti, R., Teredesai, A., De Cock, M.: Ranking Approaches for Microblog Search, pp. 153–157. IEEE, Washington (2010)
29. Martín-Vicente, M.I., Gil-Solla, A., Ramos-Cabrer, M., Blanco-Fernàndez, Y., López-Nores, M.: Semantic inference of user's reputation and expertise to improve collaborative recommendations. Expert Syst. Appl. **39**(9), 8248–8258 (2012)
30. Liao, Q.V., Wagner, C., Pirolli, P., Fu, W.-T.: Understanding Experts' and Novices' Expertise Judgment of Twitter Users, p. 2461. ACM Press, New York (2012)
31. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
32. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on Twitter using a hierarchical knowledge base. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Terzopoulos, D., Tygar, D., Weikum, G., Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M.,

Staab, S., Tordai, A. (eds.) The Semantic Web: Trends and Challenges, vol. 8465, pp. 99–113. Springer, Cham (2014)
33. Xu, Y., Zhou, D., Lawless, S.: Inferring Your Expertise from Twitter: Integrating Sentiment and Topic Relatedness, pp. 121–128. IEEE, Washington (2016)
34. Pan, Y., Cong, F., Chen, K., Yu, Y.: Diffusion-Aware Personalized Social Update Recommendation, pp. 69–76. ACM Press, New York (2013)
35. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.-Y.: An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 295–303. Association for Computational Linguistics (2010)
36. Uysal, I., Croft, W.B.: User Oriented Tweet Ranking: A Filtering Approach to Microblogs, p. 2261. ACM Press, New York (2011)
37. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
38. Biau, G., Scornet, E.: A random forest guided tour. TEST **25**(2), 197–227 (2016)
39. Breiman, L., Cutler, A.: Random forests-classification description: random forests. www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (2007). Accessed 20 Sept 2017
40. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
41. Malekipirbazari, M., Aksakalli, V.: Risk assessment in social lending via random forests. Expert Syst. Appl. **42**(10), 4621–4631 (2015)
42. Khan, R., Hanbury, A., Stoettinger, J.: Skin Detection: A Random Forest Approach, pp. 4613–4616. IEEE, Washington (2010)
43. Färber, M., Ell, B., Menne, C., Rettinger, A.: A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. Semant. Web J. **1**, 1–5 (2015)
44. Sammut, C., Webb, G.I.: Encyclopedia of Machine Learning. Springer, Berlin (2011)
45. Ester, M.: Recommendation in social networks. In: RecSys, pp. 491–492 (2013)

Sami Belkacem is a PhD candidate in Computer Science at the University of Sciences and Technology Houari Boumediene (USTHB) and currently a member of the Computer Systems Laboratory. He received a Bachelor's degree in Computer Science in 2013 and a Master's degree in Artificial Intelligence from USTHB in 2015. Sami has published several papers in international conferences, workshops, and journals. His research interests center on Social Network Analysis, Recommender systems, and Machine learning. His current research work focuses on Ranking news feed updates on social media.

Kamel Boukhalfa is a full Professor and researcher in Computer Science at the University of Sciences and Technology Houari Boumediene (USTHB) and currently a member of the Computer Systems Laboratory. He received an Engineering degree in 1997 and a Magister degree in Computer Science from USTHB in 2002. In 2009, Kamel obtained a PhD degree in Computer Science from Poitiers (France) and USTHB (Algeria) Universities. He has published more than 70 papers in international conferences, workshops, and journals. His main research interests are databases, data warehousing, cloud computing, data mining, and optimization.



Omar Boussaid is a full Professor and researcher in Computer Science at the University of Lyon 2 and currently a member of the ERIC Laboratory. In 1987, he obtained a PhD degree in Computer Science from the University of Lyon 1. He is head of Computer Science and Statistics Department and director of the Master Business Intelligence and Big Data. Omar is a member of several program committees of international journals and conferences and a founding member of the EDA conference on Data warehousing and OLAPing. He is also cofounder of the Maghrebian conference on Advanced Decision making Systems (ASD). He has published more than 200 papers in international conferences, workshops, and journals. His main research interests include business intelligence, business analytics, Big Data, data warehousing, text, social and graph OLAP, and data mining.