Extracting Unique Discussions of Interests for Entrepreneurs and Managers in a Set of Business Tweets Without Any Human Bias

Jafar Mansouri, Fabrice Cavarretta, Wassim Swaileh, Dimitris Kotzinos

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2023, accepted 10 December 2023, date of publication 18 December 2023,
date of current version 26 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3343756

RESEARCH ARTICLE

Extracting Unique Discussions of Interests for Entrepreneurs and Managers in a Set of Business Tweets Without Any Human Bias

JAFAR MANSOURI[1,2], FABRICE CAVARRETTA[2], WASSIM SWAILEH[1,3], AND DIMITRIS KOTZINOS[1]

[1]ETIS Lab, UMR 8051, CY Cergy Paris University, ENSEA, CNRS, 95302 Pontoise, France
[2]ESSEC Business School, 95021 Cergy- Pontoise, France
[3]Cloud & AI Team, Huawei Technologies, 00180 Helsinki, Finland

Corresponding author: Jafar Mansouri (jafar.mansouri@gmail.com)

The work of Jafar Mansouri was supported by the Initiative of Excellence from CY Cergy Paris University. The work of Fabrice Cavarretta was supported by ESSEC Business School Research Center.

ABSTRACT This study proposes a framework for extracting unique discussions of the interests of managers and entrepreneurs on Twitter (X). By unique discussions of interests, we mean those that are more tweeted by these communities but rarely by public people. These discussions can be facts and/or sentiments related to some topics. Since this is a subjective problem, human intervention can lead to bias in the results. Therefore, we propose an unsupervised method with zero information about the context since prior knowledge stems from human intervention. Consequently, there is no real ground truth. To retrieve such discussions of interests, first, unique tweets (discussions) are identified in two stages. In the first stage, a scoring algorithm is proposed that gives a score to each tweet of a specific year and tweets are sorted based on their scores. Different sets of tweets are selected based on their scores and considered automatically created ground truths. In the next stage, an unsupervised convolutional neural network trained on the created ground truth is used for the classification of tweets of other years (whether they are unique to these communities). Finally, latent Dirichlet analysis is applied to the detected unique tweets to give the most common interest topics discussed by these communities. Experimental analysis is performed on tweets from 2017-2019. The results reveal these communities' attitudes and highlight interesting common and different topics discussed between managers and entrepreneurs; some of them can be difficult for humans to predict in advance. The proposed approach is applicable to any community.

INDEX TERMS Entrepreneurs, managers, opinion mining, topic extraction, unique discussions of interests, unsupervised classification.

I. INTRODUCTION

Social networks and microblogs such as Twitter (X), Facebook, LinkedIn, and industry-based applications such as Booking, TripAdvisor, and Expedia are becoming very popular among people. Social media is a rich source of information exchanged among people to share their opinions. People can share their opinions and stories in the form of text, images, videos, voices, and links. These opinions can be related to general topics or specific professions. Therefore, opinion mining from these platforms has drawn much attention in recent years. Companies, organizations, scientists, politicians, and many people receive feedback from these opinions [1], [2], [3], [4], [5]. For example, a family can choose their destinations for a trip based on the experiences of travelers that have been expressed in the related sites or applications [6]. Customers use the opinions of previous buyers to select the best service/goods from various brands. Companies use this information to improve their services and products or their marketing strategies [7]. Politicians can find the opinion of a nation on cultural, social, economic, and political issues, such as which candidate has the most chance to be elected in a presidential competition [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

Twitter (X) is a widely used platform for expressing the opinions of people related to various subjects, usually in the form of texts along with links, images, videos, or voices. It is logical to assume that some tweets of everybody are usually related to their profession or shared between individuals of a professional community. We name these opinions specific/unique opinions or specific/unique discussions of interests of a group that has a specific profession. However, in addition to unique interests for a community, these people can tweet about topics, namely, public opinions or public discussions of interests, that are shared by many others who do not have that profession, namely, public people. For example, if there is a society of users of mathematicians, it is expected that a considerable amount of their tweets is related to mathematics. However, it is unlikely that a subject related to mathematics (e.g., wavelets) is discussed by many public people. Note that this does not mean that public users do not discuss mathematics; engineers may tweet about mathematics too. However, if there are two large groups, one group of mathematicians (target group) and the other group is chosen randomly from people with other professions (public group), the number or density of tweets related to mathematics will be low in the public group in comparison with that of the target group. In addition, we do not seek just discussions related to the job of a professional community. We also want to find other interests, not (directly) related to their profession but common habits or interests, which are discussed frequently in their tweets, but public people rarely discuss them. As an example, mathematicians may tweet more about reading fiction books than public people. Therefore, this is considered a unique interest for mathematicians.

We want to find unique opinions or discussions of interests of entrepreneurs and managers (business people) that are not usually discussed by the people who are not entrepreneurs and managers (public people). For this task, since people can tweet about any topic, first for tweets of managers and entrepreneurs, their unique tweets and common tweets with public people should be discriminated against. For this purpose, tweets of public users are used as a catalyst or auxiliary set. These tweets help us divide tweets of target users into common (with the public) and unique ones.

Since this task is a subjective problem, for labeling tweets, individuals may have different inferences if a tweet is unique or common to the public, and the results can change from one person to another. To remove this bias, a method without human intervention and any prior knowledge is presented. The cost of these assumptions is that there is no ground truth, which is necessary for training and evaluating a machine learning algorithm.

In this study, *tweets* are used to find the interests of entrepreneurs and managers. The focus is on interests that are expressed in the text of tweets. We do not consider audios, images, videos, or links that are referred to in some tweets. There are several reasons why tweets are selected for this purpose. Twitter (X) is one of the most widely used social media platforms for sharing daily stories or ideas of people regarding different issues. Tweets can also show the change in the opinions of users over time. Due to its limitation in the number of maximum characters (at the time of writing this paper, it is 280 characters), people express their ideas directly, briefly, and sometimes with abstractions along with links, emojis, images, or videos. In many tweets, text is the main source of information, although these texts are usually short and unstructured. In contrast, books are written by few people, so there are no opinions of many target people. Additionally, papers are usually written by academic or well-known people. In this project, we target groups of entrepreneurs and managers around the world, regardless of their fame, academic background, or other factors. This goal is also achieved by tweets so that any person can express their ideas. To the best of our knowledge, this is the first work that aims to find unique discussions of interests (and not just sentiments) of a professional community from unstructured short texts.

This paper has the following contributions: We propose a fully unsupervised method with zero information about the context for identifying unique discussions of interests in tweets of entrepreneurs and managers. This is the first study that addresses the unique discussions of interests of these communities. Unique discussions (tweets) are identified using an unsupervised scoring algorithm and an unsupervised convolutional neural network (CNN). The scoring algorithm gives us an automatically created ground truth. This scoring algorithm is based on the comparison of the density of concepts used in some tweets of entrepreneurs/managers and public users. CNN is used to confirm that the created ground truth is suitable for the training and is mainly used for the classification of other tweets. Finally, LDA is applied to unique tweets for the extraction of unique interest topics. Entrepreneurs and managers (especially those new to these fields) can use these results to gain insight into the attitudes and thoughts of their colleagues and to identify what is important to them, particularly in their profession. This information can be used to improve their performance and enable their companies to better compete with others. Although the professional groups in this work are entrepreneurs and managers, the method can be applied to any other professional community.

The rest of the paper is organized as follows: Section II presents related works on opinion mining, especially in business, and states the differences between our work and others. Section III explains the proposed method. Experimental results are demonstrated in Section IV. Section V discusses some issues related to the algorithm and future works. Section VI concludes the paper.

II. RELATED WORKS

Usually, in opinion mining, there are five key elements *e, a, s, h, t* [9], [10], where *e* is the entity or the target, *a* is the aspect of the entity *e*, *h* is the opinion holder, *t* is the time when the opinion is stated, and *s* is the opinion that the opinion holder states about the aspect of *a* of the

entity e at t. For example, in ''The price of the laptop is good,'' the laptop is the entity, price is the aspect and good is the opinion or sentiment. In most applications of opinion mining, the aim is to perform sentiment analysis or aspect extraction. While sentiment analysis is more about the sentiment polarity s, aspect extraction focuses on finding the aspect a from the corresponding opinion s of the text [10]. Sentiment analysis can be performed at the document level, sentence level, or aspect level. Direct opinions give positive or negative opinions directly such as ''the sound quality of the phone is perfect.'' In contrast, comparison opinions compare opinions about some features between two or more entities, such as ''the quality of camera X is better than that of camera Y.''

Wang et al. [10] reviewed different techniques for extracting opinions and aspects at the document and sentence levels. Particularly, for aspect extraction, unsupervised methods such as topic modeling that could extract the product aspect from an unlabeled review corpus were discussed. As noted, most unsupervised methods for aspect detection use topic modeling or POS (part of speech) tagging. Standard topic models, such as LDA, could consider word co-occurrences at the document level.

Bouazizi and Ohtsuki [11] considered multi-class sentiment analysis in tweets with the assumption that each tweet may have more than one sentiment. A method (named quantification) is proposed to find all sentiments in a tweet and assign weights to them, representing how much they are relevant to the tweet. This is beyond finding overall sentiment polarity (positive, negative, and neutral). For example, two tweets may have negative sentiments about a company, but one tweet is about frustration and anger, while the other expresses sadness or bad luck. Their method identifies 11 different sentiment classes. This work is a continuation of previous authors' work [12] that classifies tweets into one of 7 different sentiment classes.

Wang et al. [13] considered the usage of social media for entrepreneurs to find both the level of engagement for their startup and the level of venture financing. The results show how differences in entrepreneurs' tweets (differences in the level of informativity, persuasiveness, and transformativity) are associated with different levels of startup engagement and venture financing, using activities related to the number of tweets, the number of mentions of other accounts, and the number of retweets.

A new strategic business partnership recommendation service was developed by Tsutsumi et al. [14]. They have explored approximately 280 million user reactions on Facebook by a similarity model between businesses that considers the opinions of users on content shared by businesses on Facebook. This model represents virtual relationships among businesses in the virtual world generated by users.

Obschonka et al. [15] used digital footprints on Twitter (X) and Facebook to identify the personality characteristics of superstar entrepreneurs and managers. Specifically, they compare the personality characteristics of 106 of the most influential business leaders based on the individuals' Twitter (X) messages (Receptivity). Their results show that superstar managers are more entrepreneurial in many personality characteristics than superstar entrepreneurs. Additionally, superstar entrepreneurs seem to show features of a classic ''Schumpeterian'' entrepreneurial personality with respect to being creative and independent rule-breakers.

Mukhopadhyay [16] studied sentiment analysis in management, and Chen et al. [17] considered opinion mining in the financial domain and for investors. Opinion mining in the financial sector for big data was demonstrated by Bach et al. [18], and applications of text and opinion mining in the financial domain were illustrated by Kumar and Ravi [19]. Kozinets [20] has discussed the importance and applications of opinion mining in social media in management and business. Many papers, [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], and [32] have addressed opinion mining based on various criteria, such as the type of domain, techniques, and different types of modalities, such as text, image, video, and audio.

In conventional methods for opinion mining, the aim is usually sentiment (feeling or emotion) analysis, aspect extraction, or entity recognition and linking. However, our work is different from other works since we want to find unique opinions for a community. Opinions can be expressed in facts, feelings, emotions, aspects, entities, or any other way. Sentences can be considered factual and sentimental. In sentimental sentences, a positive or negative (sometimes natural) sentiment regarding an entity or an aspect such as a service or a product is expressed. For example, ''AI is important in business.'' However, in factual sentences, facts are stated such as ''AI is used in business'' or ''Today is the inauguration of the World Cup.'' Therefore, in general, methods for extracting aspect or sentiment analysis are not suitable here since they look for specific features and try to find aspects and sentiments based on a relationship between positive/negative or emotional words and aspects, while in our case, we may have tweets without positive/negative words. Furthermore, we do not know what our entities or aspects are in advance. Additionally, there is no real ground truth in our work. Moreover, this is the first work that considers unique interests of specific communities of entrepreneurs and managers. Therefore, our work is different from the works in the literature.

III. PROPOSED METHOD

As reviewed in the previous section, in most applications of opinion mining, we face sentimental sentences. However, if we want to find a discussion related to a subject, both factual and sentimental sentences should be considered. The aim of this paper is to find the unique discussions of interests of entrepreneurs and managers that are expressed in their tweets. *By unique discussions of interests, we mean discussions and topics that are usually tweeted by the entrepreneurs and managers community but rarely (in comparison) by public people.* As stated earlier, by public people, we mean those who are not managers or entrepreneurs. These unique

interests can be related to their profession or not. For example, they may tweet frequently about marketing (a profession-related interest), or they may tweet more about reading books or creating weblogs than public people. However, people such as entrepreneurs can also tweet about *a topic that can usually be discussed by other people, namely, the public interest*. Therefore, we propose a method to discriminate between *unique and public discussions of interests*.

In our method, the *text of tweets* is used to identify the interests of entrepreneurs and managers. Note that for our goal, we do not know if a tweet is unique or common to the public in advance. Furthermore, in some applications, such as social sciences, it is important to guarantee that there will be no bias involved in the selection of tweets. Since our work is a subjective task, the results can change greatly from one person to another for annotation. Even it might be difficult to say whether a tweet is related to entrepreneurship; for example, consider this tweet (courtesy of Twitter or X): ''*The private sector could build roads, shopping centers build their own roads all the time. They'd be willing to front the cost because control of transport route is valuable and they wouldn't need tolls since the road would draw in business to offset cost*.'' One person can say this tweet is expressing an entrepreneurship concern while another one can say it is related to a social issue, or someone can say it is related to politics. Furthermore, it is possible that some tweets are ambiguous and nuance between different topics is not clear or it may need some pre-knowledge of the intention of the writer. Therefore, there are some constraints in our work:

- **There is no human intervention to avoid bias in the results.**
- **There is no prior knowledge (by humans) about the context**(it can be considered a special case of the previous condition since any implicit knowledge implies human intervention).

One consequence of the above conditions is that **there is not a real ground truth**. In other words, manual labeling of tweets is not authorized. The reason is that we do not know if the topic of a tweet is just discussed by entrepreneurs/managers or by many public people. Therefore, even crowdsourcing such as MTurk cannot be used. Furthermore, since there is no human intervention, **we should use unsupervised algorithms;** otherwise, we would predefine what the unique discussions of interests are by entrepreneurs/managers, and the interest system is selected by humans. Unsupervised methods can be problematic for evaluating classification results when there is no ground truth. Moreover, methods that require some initial knowledge are not applicable here. In the following, we explain the different stages of our algorithm. Fig. 1 shows the general block diagram of the proposed method.

A. TWEET COLLECTION

In our work, only English tweets are used, and the Tweepy [33] API is used for collecting tweets. There are different types for collecting tweets by this API. In the first phase, we collected two sets of tweets from users who had the keywords ''entrepreneur'' or ''manager'' on their profile description (in each set, users had only one of the keywords). The location could be anywhere in the world to have generality and diversity. For the set of entrepreneurs, we also tried other related keywords, such as ''founder'' and ''founded,'' but we noticed that many users with these keywords were not entrepreneurs and were mostly related to political fractions, liberty, or peace associations. Thus, only the keyword ''entrepreneur'' was selected for our purpose. We also collect tweets for public users. The public set plays the role of a control group, which helps us to find common public discussions of interests. Each API call is between 90-120 seconds. Then, there is exactly a 15-minute break. Since this method searches recent tweets from anybody in the world with the defined conditions, few tweets were collected from users of entrepreneurs and managers. At the end of the first phase, three groups of users are created:

FIGURE 1. General block diagram of the proposed method.

- *Entrepreneurs*: users who have the keyword ''entrepreneur'' but not ''manager'' in their profile
- *Managers*: users who have the keyword ''manager'' but not ''entrepreneur'' in their profile
- *Public*: users who do not have the keywords ''manager'' or ''entrepreneur'' in their profile

In the second phase of collecting tweets, we specified the condition ''screen name'' or ''user_id'' for the API by the collected users, and the API returned a maximum of approximately 3200 tweets per user (their recent tweets). In this phase, tweets from the above users for three groups of entrepreneurs, managers, and public people for the specific years 2017, 2018, and 2019 were collected. For each tweet, the following information provided by the API is stored:

- user screen name
- user name
- user_id
- tweet
- tweet time
- full name of the city for the tweet
- country code for the tweet
- location in profile
- profile description

It should be noted that there are some websites for finding users based on keywords of the profiles. These websites usually return users who have many followers; in other words, users who are ''famous.'' However, we need to be careful with introducing ''fame,'' as this is biasing the sample. Fame (mostly related to visibility) often relates to a specific kind of language and positioning. Thus, by following ''famous'' entrepreneurs, for example, we would end up with a biased

opinion system, one that is possibly very far from that of average entrepreneurs who are not famous at all. This study is about average entrepreneurship/management, not about how entrepreneurship or management is conceptualized in a media arena, and the two are suspected to be possibly very far apart.

It is possible that the profession of people changes or there are some fake users, so they are not real entrepreneurs or managers. Since LinkedIn is a social network that usually contains a history of jobs that a person has had, we tried to use LinkedIn by obtaining their career information to verify whether the collected users are real entrepreneurs/managers. However, there is no API for LinkedIn, and many users on Twitter (X) do not write their LinkedIn address in their account; similarly, many users do not write their Twitter (X) address in the contact section of their LinkedIn. In this case, finding a LinkedIn account corresponding to its Twitter (X) account is impossible. For the experiments, three sets of tweets for entrepreneurs, managers, and public users are created that contain user_id, tweets, and tweet time. Hereafter, the Ent, Mng, and Public sets represent entrepreneurs, managers, and public sets, respectively.

B. PREPROCESSING

Tweets are very short and generally colloquial. They usually contain targets, emojis, emoticons, interjections, abbreviations, links, images/videos/audios, typos, and even sometimes in the wrong grammatical form or strange format, such as ''I like football 4 fuuuuuuuun -;).'' Therefore, preprocessing is necessary for further operations. In our work, two types of cleaning are performed: light cleaning which is applied to tweets in all stages of the proposed algorithm, and full cleaning. For each algorithm in the proposed method, different kinds of cleaned tweets are utilized.

In the light cleaning, which is used in the unsupervised convolutional neural network stage, emojis, emoticons, targets, links, images, videos, audios, and non-ASCII characters are removed. Also lowercase conversion (for some words, they are converted to their equivalents before lowercase conversion, like US/U.S. to USA), conversion of currencies and their signs to ''money,'' and equivalent conversion for a few common words, such as ''united states'' to ''usa'' or ''can't'' to ''cannot,'' are performed.

For the stage of the unsupervised scoring algorithm and topic extraction, the full cleaning is performed. In this cleaning, in addition to the light cleaning, stop words (which are not related to the context) in libraries of NLTK [34], Spacy [35], PyPI [36], and scikit-learn [37] are removed. Furthermore, numbers, punctuations, special characters such as # and %, and English alphabets (many of them exist in Spacy and NLTK) are removed. Additionally, since the required memory and time increase significantly when there are more new words, interjections [38], [39], [40], which are not related to an interest, are removed. Additionally, in this stage, we need the root and meaning of the words so there is no difference between ''book'' and ''books'' or ''register'' and ''registered.'' Therefore, *lemmatization*is used. This process is done to convert the plural form to singular and to find the same form (infinitive) of the verbs. For example, the lemmatization of ''They have bought the books'' is ''They have buy the book.'' In other words, we want to know which concepts have been used in the tweets, not which forms of these concepts. This process is very useful in the scoring algorithm, which is described later. Note that the *stemming* technique is not used since for some words, this technique gives us incorrect answers for our purpose. For example, the stemming result for ''busy'' and ''business'' is the same, while their meaning or root is generally different (at least in our work).

Example: In the following case, the light and full cleanings are applied to a tweet:

Raw tweet: @John AI is used in business. https://abcde.com

Tweet after light cleaning: ai is used in business.

Tweet after full cleaning: ai use business

C. UNSUPERVISED SCORING ALGORITHM

The logic behind the scoring algorithm is that every community with a specific profession frequently uses some words or concepts related to their job or their common habits or interests more than other people. As an example, business people are more likely to tweet about the market or customers than public people. The idea of this algorithm is to give more weight to words and combinations of words that discriminate between entrepreneurs/managers and the public. These words and combinations are frequently used by entrepreneurs/managers and rarely by other people. Thus, discriminative or indicative tweets (discussions) between entrepreneurs/managers and the public can be found by these words and combinations. If a tweet contains more discriminative words or combinations compared to all the words or combinations in the tweet, it is more likely to be unique to the target communities. For example, a tweet with a total of 5 words and 4 discriminative words or combinations is more likely to be unique than a tweet with 5 words but only 1 discriminative word or combination.

We select tweets from the year 2018 for the scoring algorithm. These tweets are later used for creating the ground truth and training the CNN and tweets from 2017 and 2019 are used for the test and finding unique interests. To obtain more precise results, we chose the middle year for training. This selection ensures that the trends in the discussions of the training dataset are close to both 2017 and 2019. Additionally, it allows us to detect changes in the trends.

For each set of tweets for the Ent, Mng, and Public sets, words and combinations of two words in tweets and their frequencies are found. Note that each hashtag (without the ''#'' sign) is considered one word. A combination of two words is the coappearance of two words in a tweet, not in two tweets. Thus, it is more comprehensive than bi-grams. The order of words in the tweet is not considered in the combination. In the scoring algorithm, the aim is to give more weight to words/combinations that are used frequently by one

community and less by public people. Therefore, the criterion is the ''frequency of appearance by more users''. The goal is not to give more weight to hashtags or consider the order or position of words within a sentence. If a hashtag is related to the unique interests of entrepreneurs/managers and is utilized by many users, it is assigned more weight. The order of words in the combination does not necessarily improve the performance since

- There are active and passive sentences in which the order of words in the sentence can change. For example, the tweet ''This laptop is produced by our company'' can be written as ''Our company produces this laptop''. After applying cleaning, we have the cleaned tweets ''laptop produce company'' and ''company produce laptop''. For the algorithm, it is important that the same concepts appear together in both cleaned tweets. However, if we consider the order for the combination, we have combinations of {(laptop, produce), (produce, company), (laptop, company)} for the first cleaned tweet and {(company, produce), (produce, laptop), (company, laptop)} for the other one. In this case, these tweets do not have any common combination of two words, which leads to incorrect results.
- Many tweets have hashtags, and they are usually considered to have the same importance. For example, ''key factors in the market are #money #customers #stocks'' can be written as ''key factors in the market are #customers #stocks #money.'' The three factors can be written in different orders. We do not want to know the importance of each factor to consider their order. Our focus is solely on the concepts and whether they are used together in a tweet or not. Moreover, these factors can appear at the beginning of the tweet ''#customers #stocks #money are key factors in the market''. Considering the order can result in combinations such as (market, customer) for the second cleaned tweet and (customer, market) for the last cleaned tweet. However, for the algorithm, it is important that both ''customer'' and ''market'' have been used together in one tweet. The algorithm does not prioritize which word comes before another or which word has a more important role in a tweet. This is true for nonhashtag words; for example, ''A and B are necessary for this factory'' can be written as ''B and A are necessary for this factory.'' Emphasis can be done by the order, but it is not the focus of the algorithm.

Here, *frequency* for a word or a combination of a two-word means the *number of users* in each set (Ent, Mng, and Public sets) that have used that word or combination in their tweets. ''For each person, each word/combination in their tweets is counted only once.'' Note that it is very likely that there are some repetitive tweets for each user. This is because one person repeats a tweet (or very similar tweets) many times on some days for emphasis or other reasons, such as in reporting weather, such as these tweets: ''Today is sunny'' for one day and ''Today is almost sunny'' for the next day. The word ''sunny'' is just counted once for this user. Additionally, one user may mention the name of their company multiple times in their tweets. If all usages of a word for a user are counted, this will lead to the wrong inference that it is unique to a community. However, *the aim is to find concepts that are unique and common for a community, not just for one or a few users.* This means that repetitive tweets from a user are not considered, as repetitive tweets for ''a user'' do not have any new words/combinations. The algorithm does not differentiate if ''a user'' has used a word in one tweet or multiple tweets. Similarly, for the combination of two words, the algorithm checks if there is any tweet for the user that has both words together. If the user has used that combination in some tweets, it is just counted once for that user. If a user retweets another user's tweet, it is checked if there is a new word/combination for the user or not.

Since the number of users in the Ent/Mng set can differ from the number of users in the public set, we divide the frequency in each set by the total number of users in that set. If the *normalized frequency* or *density* (i.e. frequency of a word/combination divided by the total number of users in that set) of a word/combination in the Ent set is more in comparison with its normalized frequency in the Public set, we can say it is an indicative word/combination for entrepreneurs (i.e. that word/combination is usually used just by entrepreneurs and by no or few users by the public users). Similarly, indicative words/combinations for the Public set can be found. Thus, for each word/combination, a *weight* is assigned that is the difference of its normalized frequency in the two sets. This weight represents the relative normalized frequency or relative density of one word or combination of two words in the Ent/Mng sets with respect to the Public set.

For example, suppose that one Ent user has three tweets: {''Brand is important'', ''Our products show our quality'', ''Products of that brand are important''}. After cleaning the tweets, this user has the following words and combination of two words in their tweets {brand, important, product, show, quality, (brand, important), (product, show), (product, quality), (quality, show), (important, product), (brand, product)}. For every user in the Ent and Public sets we find their words and a combination of two words used in their tweets. As seen in the example, every word and combination of two words is counted just once for each user. Suppose that we have 10 users in the Ent set, and 8 of them have used ''brand'' in their tweets. In the Public set, there are 10 users, but only 1 user has used ''brand'' in their tweets. Therefore, the weight for the word ''brand'' is (8/10) – (1/10) = 0.7. This is a soft weighting scheme in which the weight can be a value between 1 (indicating that all users in the Ent set have used that word/combination and no user in the Public set has used it) and -1 (indicating that no user in the Ent set has used it but all users in the Public set have used it). This is in contrast to hard weighting, which assigns a weight for ''brand'' as 1 (indicative for the Ent), 0 (not indicative), or -1 (indicative for the Public). In other words, hard weighting only has three possible values for the weights.

A word/combination is more indicative if it has more weight. The indicative words/combinations for the Ent/Mng sets have positive weights, while for the Public set, they have negative weights. For combinations of words, we multiply weights by 2 since a combination of two words is more informative than one word. Combinations of more than two words were not exploited. This soft weighting scheme has some advantages. It does not need human intervention. Furthermore, it is less sensitive to an imbalance in the number of users between two sets and it is independent of the number of users. However, for better performance and more precise weights, there should be more users in each set and more tweets from any user.

For all words and combinations of two words in the Ent (Mng) set their weights are obtained. Next, for each tweet in the Ent set, all words and combinations of two words are found; their weights are added, and the sum of weights is divided by the number of words and combinations to find a *score* for that tweet. The tweets are sorted based on their scores from the most positive to the most negative.

These procedures are explained below. These formulas are for calculating weights and scores for the Ent set. For the Mng set, they are obtained in a similar manner. For each word in the Ent set, we calculate:

$$w\left(word_{Ent_x}\right) = \left(\frac{freq_{Ent_x}}{N_{Ent}}\right) - \left(\frac{freq_{Public_x}}{N_{Public}}\right) \quad (1)$$

where $w$is the weight for word x in the Ent set, representing its relative normalized frequency, $freq_{Ent_x}$ is the frequency of word x in the Ent set, $freq_{Public_x}$ is the frequency of word x in the Public set, and N_{Ent} and N_{Public} are the numbers of users in the Ent and Public sets, respectively.

For each combination of two words in the Ent set, the weight for the combination is calculated as follows:

$$c\left(word_{Ent_x}, word_{Ent_y}\right) = 2 * \left(\frac{freq_{Ent_{x,y}}}{N_{Ent}}\right) - 2 * \left(\frac{freq_{Public_{x,y}}}{N_{Public}}\right) \quad (2)$$

where $freq_{Ent_{x,y}}$ is the frequency of the combination of words x and y in the Ent set.

After finding weights for all words and combinations in the Ent set, for each tweet in the Ent set, the weights for words and combinations of two words in the tweet are aggregated to find a score for that tweet:

$$score\ (tweet) = \frac{1}{N} * \sum_{i=1}^{N} w_i + \frac{1}{M} * \sum_{j=1}^{M} c_j \quad (3)$$

where N and M are the number of words and combinations of two words in the tweet, respectively.

Tweets in the Ent set are sorted based on their scores, and the following hypothesis can be considered:

- A more positive score for a tweet means that it is more likely to hold a unique discussion of entrepreneurs.
- A more negative score for a tweet means that it is more likely to hold a public discussion.

Example:

- A tweet in the Ent set: "@John, AI is used in business. http://abc.com"
- After the full cleaning: "ai use business"
- Words and their weights: ai $= 0.2$, use $= -0.001$, business $= 0.3$
- Combinations of two words and their weights: (ai, business) $= 0.15$, (ai, use) $= 0.01$, (business, use) $= 0.0005$
- $N = 3$ and $M = 3$
- Score for the tweet $= \frac{0.2-0.001+0.3}{3} + \frac{0.15+0.01+0.0005}{3} = 0.2198$
- This tweet is more likely to have a unique discussion of entrepreneurs with a score of 0.2198

As stated before, before the scoring algorithm, full cleaning is applied to tweets since stop words, interjections, and common words/expressions are not related to any opinion or interest. Additionally, removing them leads to decreasing

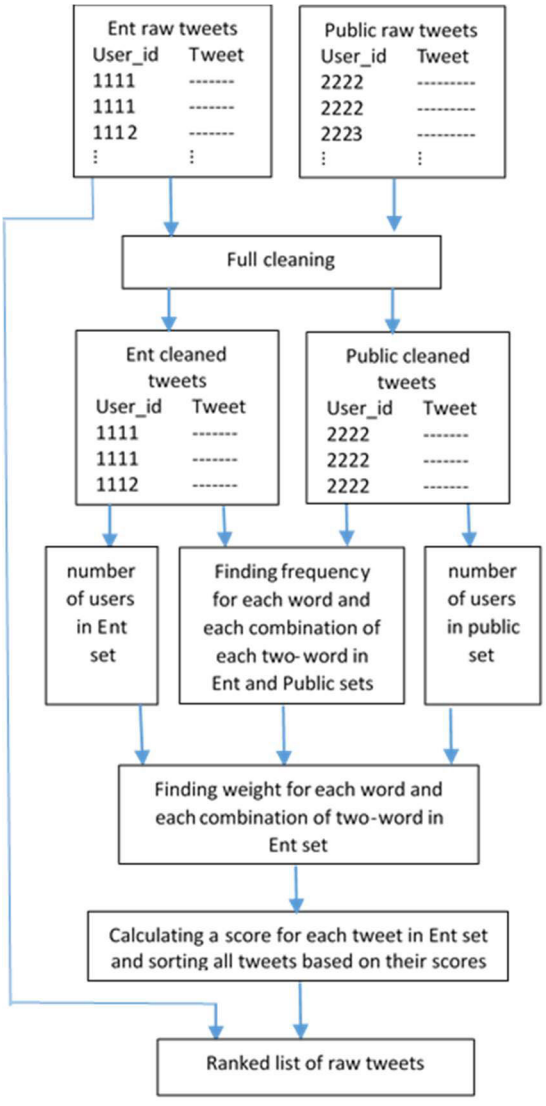

FIGURE 2. Block diagram of the scoring algorithm for the Ent set.

memory and increasing accuracy in calculating weights. The block diagram of the scoring algorithm for the Ent tweets is shown in Fig. 2. For the Mng tweets, it is similarly.

D. UNSUPERVISED CONVOLUTIONAL NEURAL NETWORK

Tweets are short, usually with one or a few sentences, probably with some hashtags. In the case of some sentences in tweets, sentences are mostly highly dependent on each other contextually. Therefore, after the light cleaning, it is rational to assume that each cleaned tweet looks like a long sentence. Therefore, Kim's strategy of convolutional neural network (CNN) for sentence classification [41] can be used for tweet classification. CNN is a well-known model that has been widely used for many text classification tasks, especially for short texts and sentences. A set of tweets with positive and negative scores found by the scoring algorithm can be used as an automatically created ground truth and is used for training a CNN. The CNN is used for two purposes:

- Testing the hypothesis that the scoring algorithm works well and consequently training samples are good.
- Determining the labels of the remaining tweets (tweets from 2017 and 2019) to determine whether they are unique to entrepreneurs or managers.

As stated in the previous subsection, whatever a tweet has a more positive score, it is more likely that it is unique to entrepreneurs or managers, and if a tweet has a more negative score, it is very likely that it is common to the public. The hypothesis that positive and negative samples found by the scoring algorithm are good samples (i.e. whether many labels found by the scoring algorithms are true) for training is tested by CNN. Of course, for the tweets with scores near zero, the degree of uncertainty is high, and they can be unique to entrepreneurs/managers or common with the public. For this purpose, different sets are selected (as will be explained in the experimental section). Each set is partitioned into training, validation, and test subsets. By training the CNN for the training subsets and evaluating the results for the validation and test subsets, this hypothesis can be checked. The block diagram of the tweet classification is shown in Fig. 3.

To feed the CNN, tweets are first converted to vectors by an embedding algorithm. For the word (token) embedding, FastText [42], [43] is used. FastText can consider synonyms or similar words. For example, the result for "team work" is close to "group work." Additionally, if there is a word that does not exist in its vocabulary (unknown words), FastText can find a vector for that word based on other words that share substantial substrings and use subword information. This is very useful in our case that tweets are used since there are many unknown words or typos in the tweets.

CNN is utilized to extract semantic and abstract features from the input data. CNN considers grams and context in the tweet. Some concepts, such as "business" or "bank," can have different meanings based on the context: "Many people are investing in the bank nowadays" and "The mountain is

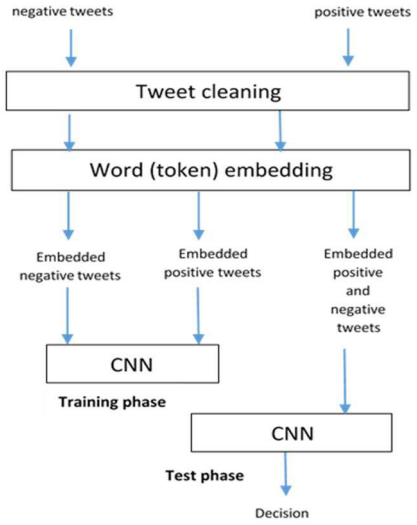

FIGURE 3. Block diagram of the tweet classification.

near the bank river." This difference can be detected in the context of a tweet by CNN.

For the tweet classification, the strategy for the sentence classification by CNN [41], [44], [45], as shown in Fig. 4, is used.

A sentence of m words (or generally tokens), with zero padding if necessary, is shown by $X_{1:m} = X_1 \blacksquare X_2 \blacksquare \ldots \blacksquare X_m$, where $X_i \in \mathbb{R}^k$ is the k-dimensional word vector for the i_th word in the sentence and $\blacksquare$ is the concatenation operator [41]. A convolution is performed by filter $V_l \in \mathbb{R}^{hk}$ that is applied to a window of h words (shown by $X_{i:i+h}$) and gives us a feature o_i:

$$o_{l,i} = f(V_l.X_{i:i+h} + b) \tag{4}$$

where $b \in \mathbb{R}$ and f is a nonlinear activation function. By applying a filter to all possible windows of words, we obtain a feature map:

$$\tilde{o}_l = [o_{l,1}, \ldots o_{l,m-h+1}] \tag{5}$$

Multiple filters with different window sizes are applied to the sentence and each of them gives a feature map. In the next layer, the max pooling layer operates on each feature map and gives us the most important feature in each feature map.

$$O_l = \max\{\tilde{o}_l\} \tag{6}$$

The output of the max pooling layers will be a vector consisting of the maximum value from each feature map:

$$M = [O_1, O_2, \ldots, O_L] \tag{7}$$

where L is the number of all filters. This vector is fed to a multilayer deep fully connected network in which the output layer is a sigmoid function (positive as unique discussion and negative as public discussion). Dropout layers are also used for regularization.

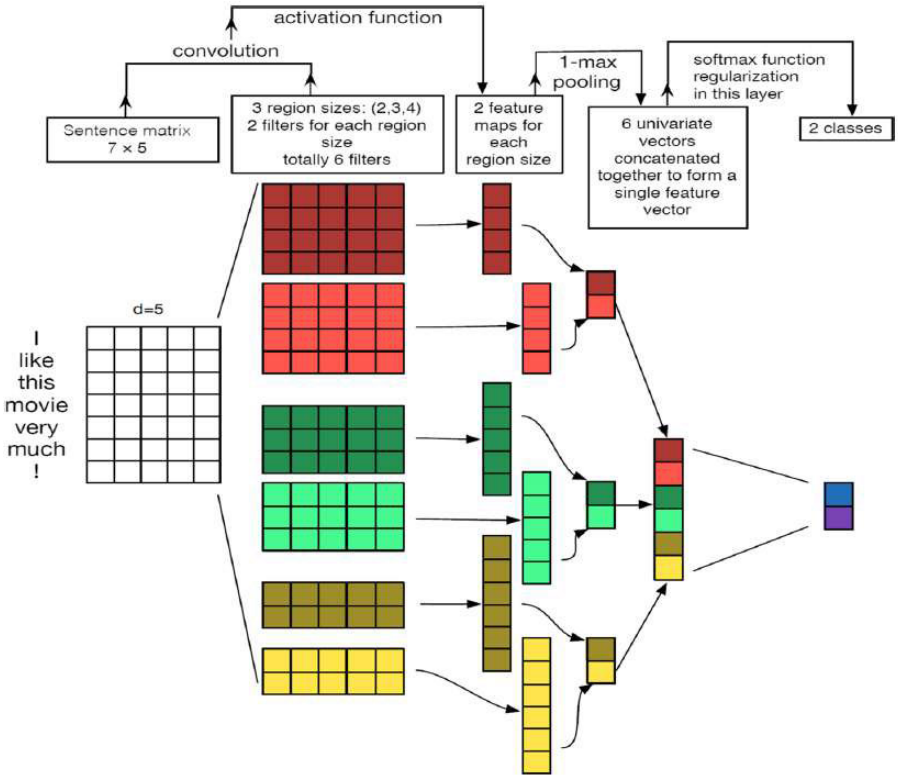

FIGURE 4. A CNN architecture for sentence classification. In this figure, filters with sizes of 2, 3, and 4 are shown [44].

E. TOPIC EXTRACTION

In this stage, the aim is to find which topics as unique interests are discussed in the unique tweets found by the unsupervised CNN. The well-known method of latent Dirichlet allocation (LDA) is used for this purpose. It is an unsupervised generative probabilistic model for a collection of data in which each data point is modeled by a finite mixture over an underlying set of topics [46]. LDA is a Bayesian hierarchical model that assumes that documents (in our case, sets of unique tweets) are produced from a mixture of topics. Similarly, topics are produced from a mixture of words or tokens. In other words, LDA is a matrix factorization method and converts the document-term matrix into a document-topic matrix (a matrix that represents the probability distribution of topics in documents) and a topic-word matrix (a matrix that represents the probability distribution of words in topics). Each topic is represented as a weighted list of words. In LDA there is a need to specify the number of topics in advance. We consider this to be an arbitrary parameter that can only change the resolutions of topics. Therefore, although we may choose a number of topics, there is no human intervention, and we can increase or decrease it to obtain more general or more detailed topics.

IV. EXPERIMENTAL RESULTS

For the experiments, since we consider general communities from managers and entrepreneurs, the more tweets from many users around the world are used, the more reliable results are obtained. The reason is that we do not know how much percentage of a person's tweets is related to their job and unique habits. Most tweets of one user can be related to the community's unique discussions, while for another user, most of their tweets can be related to public discussions. Additionally, many tweets may be repetitive or very similar to each other for one user. For example, a manager may have similar tweets about the teamwork in their company. This is generally true for most (not all) users since usually, every person has a limited vocabulary (in speaking and writing) and interests. For example, a person may tweet about management but not about mathematics, chemistry, art, or "all" jobs, sciences, and sports. They may like some sports such as tennis and football, but they do not tweet about many or all sports. Unless a tweet is related to a news TV or newspaper account or people who tweet about everything (which is very rare). Therefore, it is logical to say that people usually tweet about topics that are important or favorite to them but not every topic.

Furthermore, there may be robot users (fake accounts). They are usually in public tweets, and they mostly tweet about politics or famous people. Robot accounts may have the target keyword "entrepreneur" or "manager" on their profile. However, it is difficult to detect and remove these accounts. Additionally, there might be accounts for users who are entrepreneurs or managers for a very short time. These

accounts and robot accounts can be considered noise in each group.

Therefore, more users in a community give us better results for unique interests regarding a professional community. The experiments were performed for tweets from 2017-2019. Tweets from the years 2020 and after that were not used because many tweets are related to the pandemic issue, which makes the data very noisy.

For the scoring algorithm, for the year 2018, 2M tweets from 11440 entrepreneurs, 2M tweets from 7724 public users, and 2M tweets from 10368 managers are used. As stated before, the imbalance in the number of users does not affect the scoring algorithm significantly. This algorithm returns ranked lists of tweets for each Ent and Mng set. The number of positive tweets (assumed to be unique to the Ent/Mng communities) and negative tweets (assumed to be common to the public) are shown in Table 1. The selected tweets from this labeling (as will be described) are used to create ground truths and for training the CNN.

TABLE 1. Number of tweets with positive, negative, and zero scores for tweets in 2018 detected by the scoring algorithm. Most tweets with zero scores are empty after cleaning.

	Number of positive tweets	Number of negative tweets	Number of tweets with zero score
Ent set	963673	988183	48144
Mng set	1137835	808207	53958

For the tweet classification, tweets are first converted to vectors by the embedding algorithm. FastText is used for the embedding. Each embedded vector for a token has a dimensionality of 300. These embedded tweets are fed to the CNN.

For the CNN, we have used filter windows (h) of sizes of $s = 1, 2, \ldots, 7$. Here, even filter windows with small sizes of 1 and 2 have been used. The reason is that in our case, even words alone can be important, and especially each hashtag can convey a meaning or hashtags can be a concatenated form of some words (such as #socailmedia). For each size, there are 100 feature maps; thus, in total, there are 700 filters. For the convolutional layer, the activation function is ReLU. Then, maximum pooling is performed, and from each feature map, one feature is selected. These selected features are concatenated to form the input vector to a network of deep fully connected dense layers with dropout layers. The number of hidden layers is 7. The activation function for the hidden layers is linear, but for the output layer, it is sigmoid. The Adamax optimizer with binary cross entropy is used. For each model (as described), 70% of the samples are used for training the neural network, 10% are used for validation, and 20% of the samples are used as test samples. In each of the training, validation, and test subsets, the numbers of positive and negative labels are the same. The values for some hyperparameters such as CNN window sizes, the number of filters in each layer, the dropout rate, the batch size, and the number of hidden layers were manually selected from sets of various values. Other hyperparameters, such as the learning rate, were chosen based on the default setting of Keras. The configuration settings are reported in Table 2.

TABLE 2. Configuration settings for tweet classification by CNN.

Parameter	Value
Embedding dimensionality	300
Maximum number of tokens per tweet	5
CNN window sizes	7
Number of feature maps per window size	100
Number of hidden layers	7
Dropout rate	0.2
Number of epochs	10
Batch size	512
Learning rate	0.001

As stated before, CNN is utilized to verify the performance of the scoring algorithm and for the classification of tweets from 2017 and 2019 for entrepreneurs and managers. First, we want to check the hypothesis that the labeling by the scoring algorithm is good, i.e. most positive and negative labels are true. In other words, whatever a tweet has a more positive score (in Ent or Mng sets), it is more likely to be related to a unique discussion of entrepreneurs or managers; similarly, a tweet with a more negative score is more likely to represent a public discussion. Tweets with scores near zero can be unique (to entrepreneurs/managers) or common with the public. Since this work is an unsupervised task without a real ground truth, determining a threshold for scores is not possible. For this reason, different sets from the labeled samples are created so that each set acts as an "*automatically created ground truth" made without human intervention*. Each set is selected based on the ranks of tweets and their positive or negative labels. Three sets are selected from the most positive and most negative scores but with different total numbers of samples. Additionally, one set is created from

TABLE 3. Various cases for creating ground truths for training, validation, and testing the CNN. For each set, random shuffling is performed, and 70% of its samples are used for training, 10% of samples are assigned for validation, and 20% of samples are used as test samples. Each subset of training, validation, and test has an equal number of positive and negative samples.

Case 1	250K tweets with the most negative scores and 250K tweets with the most positive scores
Case 2	250K tweets with the least negative scores and 250K tweets with the least positive scores
Case 3	250K tweets chosen randomly from tweets with negative scores and 250K tweets chosen randomly from tweets with positive scores
Case 4	150K tweets with the most negative scores and 150K tweets with the most positive scores
Case 5	50K tweets with the most negative scores and 50K tweets with the most positive scores

the tweets with scores near zero. Another set is also created from uniform random sampling from tweets with positive and negative scores. These sets can be used to evaluate the effectiveness of labeling for tweets with the highest scores, the lowest scores, and in total. These sets are described in Table 3.

By training a CNN on the training subsets and considering the results for the corresponding validation and test subsets, we can verify how much a ground truth is good, and consequently, the effectiveness of the scoring algorithm can be assessed. These results are presented in Tables 4 and 5, in which each model is related to a CNN trained by one case of ground truth since training the network on different sets gives us different weights and hence different models. Note that if there is low accuracy for a model, the reason can be from an improper framework of classification (here, the structure of the CNN) or the lack of good ground truth. However, if there is high accuracy, it can be said that both the structure of the CNN and the ground truth are good. Receiver operating characteristic (ROC) curves and area under curves (AUCs) for five models are shown in Figs. 5 and 6. In these figures, the curves for models 1, 4, and 5 almost coincide with each other. These curves demonstrate the high performance of models 1, 4, and 5.

TABLE 4. Accuracy of CNN models trained by cases 1 to 5 for Ent tweets.

	accuracy for validation	accuracy for test
Model 1	0.9895	0.9890
Model 2	0.6238	0.6220
Model 3	0.8547	0.8544
Model 4	0.9812	0.9806
Model 5	0.9960	0.9955

TABLE 5. Accuracy of CNN models trained by cases 1 to 5 for Mng tweets.

	accuracy for validation	accuracy for test
Model 1	0.9759	0.9745
Model 2	0.6144	0.6143
Model 3	0.8610	0.8629
Model 4	0.9456	0.9439
Model 5	0.9826	0.9828

The results from models 1, 4, and 5 show very good accuracy (for both Ent and Mng), which indicates that tweets with the most positive scores are unique to entrepreneurs/managers and tweets with the most negative scores are related to public discussions and hence show the effectiveness of the scoring algorithm for these tweets. For model 2, tweets with scores near zero, the accuracy decreases significantly, but it is still relatively good. Note that for tweets with scores near zero, there are two cases; they are unique to entrepreneurs/managers (which are our target) or they are related to the public, but their labeling accuracy is more than

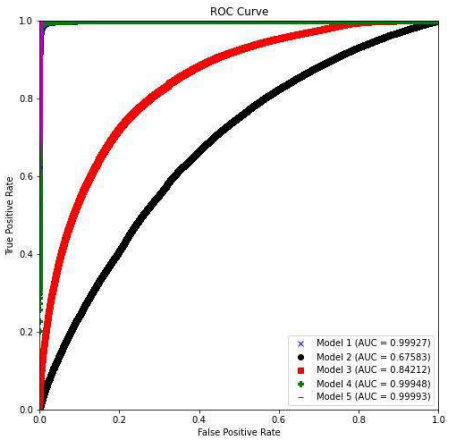

FIGURE 5. ROC curves and AUCs for models 1 to 5 for Ent tweets.

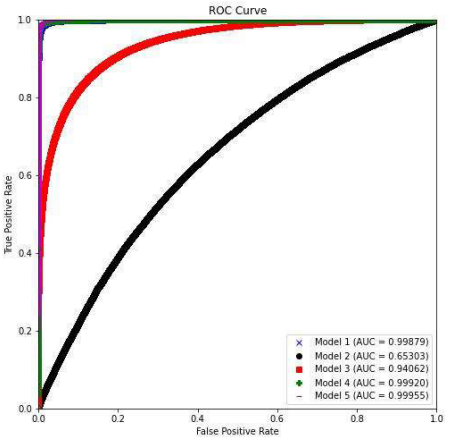

FIGURE 6. ROC curves and AUCs for models 1 to 5 for Mng tweets.

TABLE 6. Results of classification by fusion (maximum voting) of 5 models.

	Number of tweets with a positive label	Number of tweets with a negative label
Ent tweets 2017	1063213	936787
Ent tweets 2019	892698	1107302
Mng tweets 2017	756871	1243129
Mng tweets 2019	600143	1399857

60%. In general, the scoring algorithm works well from random samples of all positive and negative labels (model 3), and the accuracy is high. Therefore, in total, the scoring algorithm gives us good labeling, especially for tweets with very high positive and negative scores. Thus, when models with high positive and high negative scores are used, more accuracy is obtained for the prediction. However, if just models 1, 4, and 5 are used for the "*classification of unseen tweets*"(i.e. tweets from 2017 and 2019), the diversity of unique discussed interests *might* be limited. To improve the power of generalization, models 3 and even 2 are also used for the classification of unseen tweets. This issue is due to the intrinsic nature of our

TABLE 7. Topics and related words found by LDA for Ent unique tweets from 2017.

Topic	Number of assigned tweets	Words (concepts) related to the topic
1	179395	trump, money, bitcoin, president, tax, news, usa, people, country, good, need, vote, government, india, pay, new, know, time, state, year
2	39515	follow, thank, connect, link, recent, money, agree, happy, great, check, want, video, hi, click, free, appreciate, run, quickly, startup, libra
3	179367	thank, new, great, join, event, share, business, today, video, day, week, startup, entrepreneur, meet, post, come, team, excite, community, support
4	148022	market, business, social, medium, content, tip, seo, socialmedia, startup, new, brand, late, use, google, digital, sale, strategy, facebook, tool, thank
5	54373	book, new, ai, capricorn, machinelearning, today, available, read, write, iphone, good, money, free, author, datascience, apple, learn, suesparrow, mean, bigdata
6	78715	new, photo, post, city, energy, money, world, true, travel, car, facebook, startup, solar, late, york, year, oil, build, home, water
7	39935	thank, check, follow, late, thx, crowdfunding, pisces, people, automatically, aries, daily, work, twist, party, join, gemini, support, washington, lisa, wo
8	206749	business, work, entrepreneur, success, good, time, people, life, learn, need, way, thing, change, money, start, leadership, want, help, goal, great
9	90881	follower, new, nigeria, quote, problem, people, africa, need, today, day, stats, work, help, power, know, week, unfollowers, peace, nigerian, sir
10	46261	read, new, shoutout, check, growthhacking, money, solution, cybersecurity, intelligence, follow, artificial, business, time, aquarius, security, valley, tweet, stock, silicon, today

TABLE 8. Topics and related words found by LDA for Ent unique tweets from 2019.

Topic	Number of assigned tweets	Words (concepts) related to the topic
1	158348	work, people, good, life, time, success, thing, way, business, need, change, learn, goal, entrepreneur, know, want, help, world, think, great
2	130684	trump, agree, people, president, true, news, law, country, know, point, need, right, medium, good, fact, truth, money, usa, think, government
3	37786	post, new, video, check, add, design, fashion, link, shop, poshmark, shopmycloset, available, facebook, realestate, order, closet, construction, style, code, homes
4	78162	nigeria, vote, state, india, election, president, nigerian, minister, government, country, people, buhari, south, police, govt, know, party, indian, pakistan, leader
5	98345	join, event, new, day, today, meet, come, business, week, great, thank, excite, year, city, host, tomorrow, ticket, conference, open, free
6	82892	startup, ai, business, tech, technology, new, innovation, entrepreneur, data, company, need, read, problem, digital, late, help, learn, fund, work, money
7	47692	thank, share, great, job, check, idea, learn, support, student, podcast, good, work, course, opportunity, new, article, appreciate, need, apply, interview
8	139139	money, pay, tax, year, bitcoin, bank, business, price, time, new, need, buy, trade, market, company, people, good, cost, invest, work
9	90274	follow, market, business, thank, social, medium, content, brand, help, tip, late, instagram, check, online, facebook, email, free, blog, new, need
10	29376	book, team, congratulation, award, thank, congrats, new, winner, bos, win, great, player, read, league, write, champion, late, author, copy, preach

problem that there is no known answer in advance and also there is no real ground truth and human intervention.

We use a strategy similar to [47] and [48] (when there is no real ground truth) in which the maximum voting of some models is used as the final result. For this purpose, all of the above five models are applied for the binary classification of unseen Ent and Mng tweets. Then, for each tweet, maximum voting is used to assign the label of that tweet; i.e. if the number of positive classifications is more than negative ones, it is labeled positive (unique discussion to the community); otherwise, it is labeled negative (common discussions with the public). The 2M tweets of the Ent and Mng sets in 2017 and 2019 are classified by five CNN models. These results of maximum voting are presented in Table 6.

As it is observed from this table, while for entrepreneurs, approximately half of their tweets are related to their unique discussions about their profession and common habits and interests, for managers, there are fewer tweets with unique discussions (less than 40%) than common tweets with public discussions. This means that managers share more common ideas with public people in their tweets than entrepreneurs. To understand the unique interests discussed by the

TABLE 9. Topics and related words found by LDA for Mng unique tweets from 2017.

Topic	Number of assigned tweets	Words (concepts) related to the topic
1	142802	learn, business, data, new, cloud, work, need, customer, way, service, help, build, digital, technology, management, change, company, look, experience, http
2	72137	trump, usa, uk, new, work, president, sign, http, london, money, report, world, deal, country, brexit, news, time, problem, car, need
3	61032	work, job, enjoy, check, team, nice, hope, new, start, time, plan, glad, hard, join, meet, love, interview, club, help, set
4	61210	money, market, price, new, growth, rat, invest, tax, investment, home, course, work, rise, time, investor, increase, high, stock, help, mortgage
5	45603	luck, new, meet, work, view, team, travel, celebrate, step, xx, mile, support, event, award, join, weather, time, fitbit, area, start
6	98923	agree, work, share, people, vote, support, need, new, health, help, leader, team, leadership, http, research, nh, public, care, goal, service
7	83034	excite, team, help, look, work, support, idea, new, email, money, proud, time, christmas, run, people, raise, need, service, staff, project
8	84770	join, event, ticket, fun, new, open, http, book, conference, team, visit, excite, register, work, session, time, free, tour, host, meet
9	57913	follow, new, read, win, follower, book, chance, free, link, happy, giveaway, time, stats, enter, appreciate, money, share, copy, write, scorpio
10	49447	post, photo, facebook, news, new, market, content, medium, social, video, seo, blog, story, daily, list, digital, update, socialmedia, tip, advertise

TABLE 10. Topics and related words found by LDA for Mng unique tweets from 2019.

Topic	Number of assigned tweets	Words (concepts) related to the topic
1	45934	check, new, help, list, update, ticket, available, work, event, feature, book, free, release, excite, need, dm, site, time, announce, launch
2	81108	post, photo, new, market, book, email, work, business, content, medium, start, social, link, write, read, blog, need, tip, hope, video
3	29933	follow, win, join, time, giveaway, kindly, chance, money, enter, prize, incredible, new, follower, event, note, team, twitter, view, competition, winner
4	29836	look, enjoy, christmas, fun, money, plan, new, visit, work, earn, coffee, london, bite, xx, concern, invite, time, badge, disney, level
5	39735	work, brexit, support, vote, labour, party, uk, new, deal, tory, johnson, boris, mp, eu, time, people, corbyn, nh, leader, need
6	125915	agree, money, data, new, need, business, work, company, market, change, learn, ai, technology, help, people, time, problem, security, service, customer
7	62446	news, vote, trump, excite, president, usa, election, result, report, official, new, uk, state, work, country, government, support, time, sportlomo, office
8	42318	read, share, proud, award, work, team, new, winner, book, support, join, time, deserve, help, happy, celebrate, special, win, event, excite
9	46178	idea, new, service, work, travel, road, city, car, time, bring, station, home, train, set, run, area, flight, team, weather, join
10	96740	team, work, job, meet, new, look, people, support, help, learn, opportunity, join, manager, nice, need, time, want, community, project, staff

managers' and entrepreneurs' communities, LDA is applied to their unique tweets (after light cleaning) for topic extraction. LDA allocates topics to a set of tweets and words to topics. The number of topics is selected as 10, and for each topic, the most 20 keywords are found. These results are shown in Tables 7-10.

As seen from the tables, it is relatively difficult to ''name'' an exact topic for each assigned topic. However, by taking a closer look at the words (concepts) for topics, we can say that entrepreneurs mostly tweet about ''business, startup, leadership, social media, innovation, money, work, artificial intelligence, data, market (and marketing), time, (digital) technology, company, support, need, people, sale, brand, change, cryptocurrency, book, tax, job, goal, and learning.'' Managers mostly discuss ''business, money, work, people, team, office, change, support, customers, time, data, tax, market (and marketing), social media, artificial intelligence, leadership, (digital) technology, service, book, sale, social media, help, need, learning, and health.'' While entrepreneurs and managers have many common interests, such as ''artificial intelligence, social media, money, market, technology, and work,'' they have notable differences. For example, entrepreneurs pay more attention to ''innovation, new, strategy, brand, thinking, and cost.'' Managers are also more

interested in ''management, team, staff, service, health, and care.'' Additionally, it is seen that there are some interests for *both* communities, such as ''goal in work, social media, change, (cyber) security, and (reading) books,'' that these interests may not be detected by humans in advance, at least for one of the communities. Additionally, the tables show that nearly the same topics are discussed with changes in time (years 2017 and 2019).

However, it seems that some topics are common to the public, such as some concepts in politics. There are some reasons for these ''*noises.*'' These concepts are used with other concepts related to their professions, so these tweets may be difficult to be correctly classified by the CNN models. Nevertheless, these concepts are separated by LDA, as seen in the tables. Furthermore, the lack of diversity and balance in geographical location can be important and create noise in the results; however, due to the specific conditions of the API, there is not much control over it. Moreover, regarding the users, it should be noted that

- ''Entrepreneur'' is a specific word that *is used just in business,* so when we find users that have this keyword in their Twitter (X) profile description, there are three kinds of users:
  - Real entrepreneurs (our target)
  - Not real (i.e. for a very short time) entrepreneurs
  - Fake (robot) users
- ''Manager'' is not a unique word, and in addition to business, it*can be used in other domains*. For example, one user may be a manager of their lab at the university, a manager of a sports team, or a manager of a political party. Therefore, if the word ''manager'' appears in the Twitter (X) profile description, there are these types of users:
  - Real manager related to business (our target)
  - Not real (i.e. for a very short time) managers related to business
  - Managers in other domains
  - Fake (robot) users

Therefore, there are nontarget users in the sets collected by the API that we have little control over. Tweets from these undesirable users can decrease the performance of the scoring algorithm and the CNN models trained on the automatically created ground truths. Consequently, the result for the Ent set ''*seems*'' to be better than the Mng set, although it seems that both have non-unique (or public) interests. However, since this task is subjective and there is no human intervention and predefined answers, we cannot confirm it with certainty.

V. DISCUSSION

One important issue is the selection of the model for the tweet classification. It is crucial that the algorithm does not have any bias. Large Language Models (LLMs), such as GPT-4 [49] and Llama [50], have good performance when utilized with their pre-trained case. A pre-trained model indicates that it has been trained on a vast amount of texts. These texts include books and Wikipedia that have been written by humans, mostly scholars. However, this introduces bias in LLMs and possess prior knowledge. This contradicts our conditions of avoiding any human bias, prior knowledge, and famous authors.

In this study, just English tweets were considered. However, it is possible to use machine translation techniques to include tweets in other languages. This would greatly enhance our understanding of the unique discussions of interests of managers and entrepreneurs worldwide, as diversity has increased with users tweeting in their own languages.

In the experiments, the data covered a relatively short time period (2017-2019), and there were no significant changes in the trends of interests, as observed in the results. This is probably because there were no prominent events in the world during those years. If we have tweets from more years, the trends of interests may change significantly over the years. There are several hypotheses for potential changes in trends. One hypothesis is that changes can happen after a certain period, such as every 5 or 10 years. Another idea is that changes could be triggered by significant changes in situations, such as a pandemic, a major advancement or invention in technology, or significant shifts in global politics. Exploring these ideas would be an intriguing topic for future research.

VI. CONCLUSION

This study proposed a completely unsupervised method to identify unique discussions of the interests of business communities on Twitter or X (entrepreneurs and managers) that were rarely tweeted by other people. These interests could be in factual or sentimental sentences, related to their professions or not. This task was subjective, and the results could be changed by the intervention of humans. Furthermore, there was not a real ground truth. Therefore, conventional methods were not applicable here. Hence, an unsupervised method without human intervention and any prior knowledge was proposed. First, unique tweets (discussions) were identified in two stages. In the first stage, a scoring algorithm was designed to give a weight to each word or combination of two words based on the difference in its relative normalized frequency in two datasets (entrepreneurs/managers and public datasets). Frequency meant how many users in a dataset had used that word or combination of two words in their tweets. To be independent of the number of users, the frequencies were divided by the number of users in each set to give the normalized frequency. Based on these weights, a score was assigned to each tweet (from the year 2018) and the tweets were sorted based on their scores in the entrepreneur and manager sets. Five ground truths were created based on the sorted lists of tweets for each Ent and Mng set. In the next stage, the unsupervised CNN showed the effectiveness of the automatically created ground truths and consequently the scoring algorithm. Five CNN models trained on these ground truths were also used for the classification of other tweets (from 2017 and 2019) to determine unique tweets.

The final label for each tweet was obtained by maximum voting on the results of these five models. Finally, LDA was applied to the extracted unique tweets for topic extraction as unique interests. The feature of this work is that it is the first study for extracting unique discussions of interests for entrepreneurs and managers from their tweets. This is performed without any human bias or prior knowledge and using automatically created ground truth by an unsupervised method. Entrepreneurs and managers can use these results to understand the trends and thoughts of their colleagues, particularly in their profession to enhance their performance set appropriate strategies, and thus help their companies become more successful. Although this method was used for business communities, it can be generalized to any professional group.

REFERENCES

[1] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl.-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.
[2] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.
[3] F. Ji, Q. Cao, H. Li, H. Fujita, C. Liang, and J. Wu, "An online reviews-driven large-scale group decision making approach for evaluating user satisfaction of sharing accommodation," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118875.
[4] S. J. Lodha and M. Damle, "Sentiment and statistical analysis of customer reviews for strategic decision on positioning and marketing," in *Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA)*, Mar. 2022, pp. 100–107.
[5] A. Afaq, L. Gaur, and G. Singh, "A latent Dirichlet allocation technique for opinion mining of online reviews of global chain hotels," in *Proc. 3rd Int. Conf. Intell. Eng. Manage. (ICIEM)*, Apr. 2022, pp. 201–206.
[6] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A review of text corpus-based tourism big data mining," *Appl. Sci.*, vol. 9, no. 16, p. 3300, Aug. 2019.
[7] A. V. Mohan Kumar and A. N. Nandkumar, "A survey on challenges and research opportunities in opinion mining," *Social Netw. Comput. Sci.*, vol. 1, no. 3, p. 171, May 2020.
[8] J. S. Santos, F. Bernardini, and A. Paes, "A survey on the use of data and opinion mining in social media to political electoral outcomes prediction," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 103, Oct. 2021.
[9] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, C. C. Aggarwal C. Zhai, Eds. Boston, MA, USA: Springer, 2012, pp. 415–463.
[10] R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang, "A survey on opinion mining: From stance to product aspect," *IEEE Access*, vol. 7, pp. 41101–41124, 2019.
[11] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis in Twitter: What if classification is not the answer," *IEEE Access*, vol. 6, pp. 64486–64502, 2018.
[12] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in Twitter," *IEEE Access*, vol. 5, pp. 20617–20639, 2017.
[13] F. W. Wong, J. Kuruzovich, and Y. Lu, "Entrepreneurs' activities on social media and venture financing," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1–10.
[14] D. P. Tsutsumi, A. T. Fenerich, and T. H. Silva, "Towards business partnership recommendation using user opinion on Facebook," *J. Internet Services Appl.*, vol. 10, no. 1, p. 11, Jun. 2019.
[15] M. Obschonka, C. Fisch, and R. Boyd, "Using digital footprints in entrepreneurship research: A Twitter-based personality analysis of superstar entrepreneurs and managers," *J. Bus. Venturing Insights*, vol. 8, pp. 13–23, Nov. 2017.
[16] S. Mukhopadhyay, "Opinion mining in management research: The state of the art and the way forward," *Opsearch*, vol. 55, no. 2, pp. 221–250, Jan. 2018.
[17] C.-C. Chen, H.-H. Huang, and H.-H. Chen, "A research agenda for financial opinion mining," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 15, May 2021, pp. 1059–1063.
[18] M. P. Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, Feb. 2019.
[19] B. S. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowl.-Based Syst.*, vol. 114, pp. 128–147, Dec. 2016.
[20] R. V. Kozinets, "Netnography for management and business research," in *The SAGE Handbook of Qualitative Business and Management Research Methods: Methods and Challenges*. Newbury Park, CA, USA: Sage, 2018, pp. 384–397.
[21] C. Messaoudi, Z. Guessoum, and L. B. Romdhane, "Opinion mining in online social media: A survey," *Social Netw. Anal. Mining*, vol. 12, no. 1, p. 25, Jan. 2022.
[22] L.-C. Chen, C.-M. Lee, and M.-Y. Chen, "Exploration of social media for sentiment analysis using deep learning," *Soft Comput.*, vol. 24, no. 11, pp. 8187–8197, Jun. 2020.
[23] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022.
[24] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D., "Multimodal sentimental analysis for social media applications: A comprehensive review," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 5, Sep. 2021, Art. no. e1415.
[25] Y. Guo, F. Wang, C. Xing, and X. Lu, "Mining multi-brand characteristics from online reviews for competitive analysis: A brand joint model using latent Dirichlet allocation," *Electron. Commerce Res. Appl.*, vol. 53, May 2022, Art. no. 101141.
[26] C. Qian, N. Mathur, N. H. Zakaria, R. Arora, V. Gupta, and M. Ali, "Understanding public opinions on social media for financial sentiment analysis using AI-based techniques," *Inf. Process. Manage.*, vol. 59, no. 6, Nov. 2022, Art. no. 103098.
[27] S. Khan, "Business intelligence aspect for emotions and sentiments analysis," in *Proc. 1st Int. Conf. Electr., Electron., Inf. Commun. Technol. (ICEEICT)*, Feb. 2022, pp. 1–5.
[28] K. Arava, R. S. K. Chaitanya, S. Sikindar, and S. P. Praveen, "Sentiment analysis using deep learning for use in recommendation systems of various public media applications," in *Proc. 3rd Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Aug. 2022, pp. 739–744.
[29] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Analytics J.*, vol. 3, Jun. 2022, Art. no. 100073.
[30] R. Kim Amplayo, A. Brazinskas, Y. Suhara, X. Wang, and B. Liu, "Beyond opinion mining: Summarizing opinions of customer reviews," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 3447–3450.
[31] J. Lee, B. Jeong, J. Yoon, and C. H. Song, "Context-aware customer needs identification by linguistic pattern mining based on online product reviews," *IEEE Access*, vol. 11, pp. 71859–71872, 2023.
[32] V. Ganganwar and R. Rajalakshmi, "Implicit aspect extraction for sentiment analysis: A survey of recent approaches," *Proc. Comput. Sci.*, vol. 165, pp. 485–491, Jan. 2019.
[33] *Tweepy*. Accessed: Jan. 15, 2020. [Online]. Available: https://www.tweepy.org/
[34] *NLTK: Natural Language Toolkit*. Accessed: Mar. 3, 2020. [Online]. Available: https://www.nltk.org/
[35] *SpaCy Industrial-Strength Natural Language Processing in Python*. Accessed: Apr. 4, 2020. [Online]. Available: https://spacy.io/
[36] *PyPI the Python Package Index*. PyPI. Accessed: Apr. 5, 2020. [Online]. Available: https://pypi.org/
[37] *Scikit-Learn: Machine Learning in Python Scikit-Learn 1.0.2 Documentation*. Accessed: Mar. 3, 2020. [Online]. Available: https://scikit-learn.org/stable/
[38] *Dictionary of Interjections (AWW, OH, AH, EK, OOPS)*. Accessed: Nov. 24, 2020. [Online]. Available: https://www.vidarholen.net/contents/interjections/
[39] *Interjection Guide. Learn the Interjection Definition*. Accessed: Nov. 23, 2020. [Online]. Available: https://www.easybib.com/guides/grammar-guides/parts-of-speech/interjection/
[40] *Interjections Vocabulary Word List*. Enchanted Learning. Accessed: Nov. 23, 2020. [Online]. Available: https://www.EnchantedLearning.com/wordlist/interjections.shtml
[41] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1746–1751.

[42] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 427–431.
[43] *Word Representations FastText*. Accessed: Feb. 19, 2021. [Online]. Available: https://fasttext.cc/index.html
[44] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, vol. 1, Taipei, Taiwan, Nov. 2017, pp. 253–263.
[45] C.-X. Wan and B. Li, "Financial causal sentence recognition based on BERT-CNN text classification," *J. Supercomput.*, vol. 78, no. 5, pp. 6503–6527, Apr. 2022.
[46] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[47] J. Viinikka, R. Eggeling, and M. Koivisto, "Intersection-validation: A method for evaluating structure learning without ground truth," in *Proc. 21st Int. Conf. Artif. Intell. Statist.*, Mar. 2018, pp. 1570–1578. https://proceedings.mlr.press/v84/viinikka18a.html
[48] M. Fedorchuk and B. Lamiroy, "Binary classifier evaluation without ground truth," in *Proc. 9th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Bengaluru, India, Dec. 2017, pp. 1–6.
[49] *GPT-4*. Accessed: Nov. 12, 2023. [Online]. Available: https://openai.com/gpt-4
[50] H. Touvron, "LLaMA: Open and efficient foundation language models," Feb. 2023, *arXiv:2302.13971*.

JAFAR MANSOURI received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2005, 2008, and 2015, respectively. He was a joint Postdoctoral Researcher with CY Cergy Paris University and ESSEC Business School, Paris, France. His research interests include artificial intelligence, machine learning, natural language processing, data mining, and computer vision. He has served as a reviewer and a program committee member for several journals and conferences.

FABRICE CAVARRETTA received the B.A. degree in mathematics from École Polytechnique, the dual M.S. degree in CS from Stanford University and ENSTA, the M.B.A. degree from Harvard University, and the Ph.D. degree from INSEAD.

He is currently an Associate Professor in management with the ESSEC Business School. He combines management scholarly expertise with his quantitative background as a Software Engineer with Silicon Valley. His research interests include the logic managers use to develop new ventures, on the application of artificial intelligence to people and organization analytics; and on how organizational factors influence performance volatility and risk. His work has been published in the *Strategic Entrepreneurship Journal*, the *Journal of Organizational Behavior*, *Leadership Quarterly*, and *Industrial Corporate Change*. He mainly teaches leadership and entrepreneurial management in M.S. programs, coordinates the Ph.D. entrepreneurship seminar, and he has developed a corporate venturing/intrapreneurship course for executives. He has 12 years of operational management experience, including stints as a Division General Manager in a large media/telecom firm and the Founder of a social network start-up. He has published the book *Yes, France is a Paradise for Entrepreneurs* (Plon, 2016, French), in which he describes how to properly tackle each entrepreneurial ecosystem. It was widely featured in the French press. For more information visit the link: www.cavarretta.fr/fpe/en.

WASSIM SWAILEH received the B.Sc. degree in computer engineering from Sana'a University, Yemen, and the M.Sc. and Ph.D. degrees in computer science from the University of Rouen, Normandie, France, in 2008 and 2017, respectively. He is currently a Senior Researcher in computer vision with the Huawei Technologies Centre, Finland. He was a contracted Associate Professor with the Computer Science Department, CY Cergy Paris University. He was a member of the Multimedia Indexation and Data Integration (MIDI) Research Team, ETIS UMR-8051 Laboratory, CY Cergy Paris University, ENSEA, and CNRS, France. His main research interests include machine learning, pattern recognition, computer vision, and natural language processing. He was in charge of setting up the double degree project in data science and big data with the Zhejiang University of Science and Technology.

DIMITRIS KOTZINOS received the M.Sc. degree in transportation, in 1996, and the Ph.D. degree in computer science (in real-time web information systems), in 2001. During the master's and Ph.D. degrees, he studied networks and their applications in transportation systems. He is currently a Professor with the Department of Computer Science, CY Cergy Paris University, and a member of the ETIS Laboratory, where he is also a member of the MIDI Team. His main research interests include data management algorithms, techniques, and tools; the development of methodologies, algorithms, and tools for web-based information systems, portals, and web services; and the understanding of the meaning (semantics) of interoperable data and services on the web. Recently, he has started working on studying the formation and evolution of discussions in online social networks using machine learning (ML) and artificial intelligence (AI) techniques. Additionally, he has started working in the area of accountability, explainability, and fairness of the ML and AI algorithms, especially when applied in data engineering and analysis problems; this includes issues on data privacy and especially their intersection with the publication of linked open data. He has published more than 70 articles in various journals, books, conferences, and workshops and serves as a program committee member and a reviewer for various conferences and journals. He is also participating in nationally and internationally funded research programs around data analytics, data models, and networks and their integration into everyday life.