A Crowdsourcing Based Framework for Sentiment Analysis: A Product Reputation

F. Z. Ennaji, A. El Fazziki, H. El Alaoui El Abdallaoui, and H. El Kabtane

Original scientific paper

***Abstract*—As social media has spread, people started sharing their personal opinions and thoughts widely via these online platforms. This valuable data represents a rich data source for companies to deduct their products' reputation from both social media and crowds' judgments. To exploit this data, a framework was proposed to collect opinions and rating scores respectively from social media and a crowdsourcing platform to perform sentiment analysis, provide insights about a product and give consumers' tendencies. During the analysis process, a consumer category (strict) is excluded from the process of reaching a majority consensus. To overcome this, a fuzzy clustering is used to compute consumers' credibility. The key novelty of our approach is the new layer of validity check using a crowdsourcing component which collects the opinions expressed by the participants of the crowd. Finally, experiments are carried out to validate this model (Twitter and Facebook were used as data sources). The results show that this approach is more efficient and accurate than existing solutions thanks to our two-layer validity check design.**

***Index Terms*—Crowdsourcing, Product Reputation, Sentiment analysis, Social Media, Subjectivity Classification.**

I. INTRODUCTION

The emergence of the web has dramatically changed how people express their ideas and has allowed internet users to disclose their views and opinions openly [1]. Thanks to social media, several opinions on products and services are widely shared with the whole community [2][3].

In order to gather these opinions, a subjectivity and sentiment classification is required. This classification treats two sub-topics: i) determine whether the extracted reviews represent an opinionated text (subjective sentences) or not (objective sentences) and then ii) determine the polarity of the subjective sentences [4]. Unfortunately, this separation (subjectivity classification) cannot be deducted from the source or the type of the review. For example, newspaper articles are typically thought to be relatively objective, but [5] reported that 44% of sentences in their corpus (in articles that are not editorials or reviews) are subjective.

As a result, the sentiment and subjectivity analysis have become an important topic of data mining. Thus, an automatic tool can be considered in order to gather valuable opinions from social media [6]. The resulting information (users' opinion) might have great importance and usefulness for the development cycle of products and services.

The opinion extraction and analysis are not simple tasks, owing to the fact that the data is heterogeneous and different from traditional ones used in data mining. This contrast is illustrated not only in the size of extracted data, but also in its noisiness and formlessness. However, the main characteristic of social data is the integration of social relations, which lead us to use data analysis approaches that can combine social theories with statistical and data mining methods.

Several applications were created with the intention to collect public opinions from social media such as Facebook, Twitter, etc. Although collecting data can be very useful for decision makers, they are still not sufficient enough, seeing that consumers need to be involved in a much deeper level. This is due to the emergence of new information technologies and web platforms [7]. Users, now, have within their reach, and at all times, the necessary means to share their opinions about a product. Companies start understanding this fact by exploiting social media and crowdsourcing in order to build a social software to achieve powerful results, which provide a multidimensional participation based conversation mode.

Integrating crowdsourcing technologies, as a collaborative information sharing mechanism based on the principle of collective wisdom, will allow us to [8]:

- Get the consumers talking: This will help to be closer to the consumers and consequently know more about their expectations and feedbacks.
- Turn consumers into brand defenders by getting them involved and engaged: each brand has its own defendants; those people who remain attached to it regardless the quality of the produced model. Consequently, the product will tempt other visitors by reading their evaluations.

Social media data are very different from the traditional data that we know, not only in terms of the size of the data extracted, but also in terms of its noisiness and difficulties related to their forms. The inability of machine learning techniques to determine with certainty the expressed opinion in social media lead us to propose a product reputation tool that performs two main tasks, the opinion analysis on social media data and the

Manuscript received October 15, 2019; revised February 7, 2020. Date of publication October 23, 2020. Date of current version October 23, 2020. The associate editor prof. Mladen Russo has been coordinating the review of this manuscript and approved it for publication.

Authors are with the Computing Systems Engineering Laboratory (LISI), Cadi Ayyad University, Marrakech, Morocco (e-mails: f.ennaji@edu.uca.ma, elfazziki@uca.ma, {hasnaelalaoui57, elkabtanehamada}@gmail.com).

Digital Object Identifier (DOI): 10.24138/jcomss.v16i4.935

crowd rating collection. The first part includes a sentiment analysis tool that performs sentiment and subjectivity analysis. It comprises three features: collecting reviews from social media, finding positive, negative and neutral reviews and identify who has posted the most reliable information using the Fuzzy C-means algorithm that classifies opinions into majority's and leaders' opinion, since some social media users are more worth listening than others and in most cases they are overlooked.

As a new information source, a crowdsourcing component was built to collect crowd's rating. All this will enable us to provide a way to find out trends in the future. In this work, we focus on analyzing Twitter and Facebook reviews at the experimental stage.

The rest of the paper is organized as follows. Section II provides related studies. Section III describes an overview of the proposed framework and its main components, while section IV and section V presents in details, respectively, the opinion extraction and analysis component, and the crowdsourcing component. Sections VI is devoted to the development process. In section VII, we present experimental results of the proposed framework followed by discussion and conclusion in section VIII and IX.

II. Related Works

A. Opinion mining in Social Media

A considerable part of web browsing involves the use of social media. Assuming the large amount of content stored and disseminated quickly, companies have started to exploit it for competitive advantage [9]. Thus, Social Media Analytics (SMA) has been recognized as a distinct subcategory in the field of data analysis. It applies analytical skills appropriate to the social media content in order to generate specific types of knowledge [10]. Over time, several types of analysis have been applied such as topic modeling, network analysis and opinion mining to different social media (see Table I).

Table I
Examples of Social Media used in Opinion Mining

Article	Social Media
Kordonis et al. 2016 [11]	Twitter
Li et al. 2016 [12]	Twitter, Facebook
Ahn and Spangler 2014 [13]	Twitter, Facebook, Blogs
Kaur et al. 2019 [14]	Facebook

Several techniques and tools have been used for opinion mining in social media such as:

- Lexical-based: These techniques are essentially based on the lexicon of feelings (a collection of known and precompiled terms), that can be either based on a dictionary or on a corpus which use statistical or semantic methods to find the polarity of meaning.
- Machine learning-based: It trains a text classifier on a human-labeled training dataset. The machine learning techniques that have been proposed for classifying emotions be supervised or unsupervised learning techniques.
- Hybrid: it is based on the integration of two or more classification techniques. It helps to resolve their limits and to complement each other.

Many research-oriented applications have been proposed in this area like [15] where the authors tried to predict the stock market using Twitter moods and [16] that proposes a solution to fix the lack of suitable data sets, complicating the comparison between different approaches. Hamdan et al. [17] experimented with a different low-level features such as an adapted logistic regression classifier powered by n-grams, lexicons, Z-score and semantic features while taking into account negation expressions. Tweets are classified into two classes (negative and positive). The authors concluded that the lexicon-based approach provided the best results. However, the neutral class has been excluded, which could lead to misleading conclusions as they, generally, constitute a high percentage of short social media posts. Authors in [18] approached the problem of classification of tweets. They explored the influence of filter function selection techniques on the classification of tweets, using ten subsets of variable-length entities and four machine learning methods. The results prove that the choice of the features and the length of the subset of entities improve significantly the performance of the classification.

B. Crowdsourcing

The significant development experienced by the web and consequently the social media, has allowed a remarkable increase of the applicability and usefulness of crowdsourcing techniques that have been adopted in software development.

Crowdsourcing aims to outsource complex tasks by fragmenting them to several sub-tasks, to be carried out by the members of the selected crowd. Basically, the crowdsourcing techniques have been used in three main topics: image annotation (in [19] and [20]), language processing [21] and information retrieval (such as [22] and [23]). One of the subtopics of information retrieval is sentiment analysis, which is the area of this work.

Several platforms have exploited human intelligence to perform sentiment analysis. The crowd participants, in this case, can be either paid or volunteer. Amazon Mechanical Turk, one of the famous crowdsourcing platforms, was created in 2005. It makes people work and perform sub-tasks, called HITs (Human Intelligence Tasks), in exchange of money. CrowdSource aims to gather customers' feelings about an entity (brand, product or service) in real time by monitoring social media, blogs, media and entity's reviews. Unlike these two platforms, CrowdCrafting is an open source web-based service that invites volunteers to work on different scientific projects in different fields (such as sentiment analysis) that require the integration of human intelligence.

Several automatic tools have been created to perform sentiment analysis. However, combining them with crowdsourcing technologies will give more accurate results since the evaluation is based on human intelligence. Unfortunately, relying on manual work affects negatively and directly the execution time. [24] discusses the lack of the crowd competence in the sentiment analysis field. However, in our case, the crowd intelligence will be just used to gather the general opinion about a product. Thus, we are not interested in training the crowd, so a participant can extract the feeling expressed by a sentence.

Furthermore, in the work carried out by Tsapatsoulis N. and Djouvas C. [9], a comparison in terms of effectiveness between

the features indicated by humans with those extracted by deep learning was carried out. Their work focused on the classification of feelings for short texts (tweets and Facebook comments). This study has shown that crowdtagging (human intelligence based), can be effectively used to form sentiment classification models for short texts and that these models are at least as effective as those using deep learning or even better.

III. The Proposed Framework

In this work, we aim to propose a framework that can be used for distributed processing in social CRM applications (Customer Relationship Management). This framework will allow a better interpretation of topics of interest recorded in social media.

The proposed framework is structured into components. These components will be defined according to the classic principle of software engineering: strong cohesion and weak coupling. The strongly linked tasks will be grouped in the same component. As a result, two components showed up to perform two different tasks: extracting the opinion expressed by social media users and gathering the rating scores from the crowd. Fig. 2 shows an overview of the architectural components, which are described in the subsequent sections.

The proposed framework is divided into three main parts:

- Opinion Extraction and Analysis Component (OEAC): takes the responsibility to extract reviews from social media that concern a certain product (chosen by the administrator), to build the related social network and to analyze them in order to gather the public opinions.
- Crowdsourcing Component (CC): collects product assessments from the crowd with the intention of comparing and consolidates its results with those gotten from the OEAC.

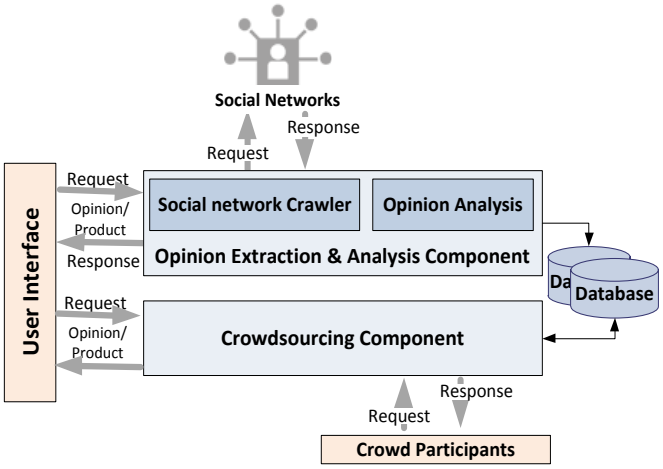


Fig. 2. The architecture of the product reputation framework

- Simple user front-end interfaces: ensure a communication between the users (the company and the crowd) and the framework.

In the following, we will give more details about the opinion extraction and analysis component (OEAC) and the crowdsourcing component (CC).

IV. Opinion Extraction and Analysis Component

In order to increase sales, companies have to improve their products and services continuously and compare them to the competitors' ones. This can be achieved by collecting and analyzing public opinions, which might be about the current, earlier, or the competitors' products. We then conclude the importance of integrating the social intelligence mechanism into firms' information systems.

To do so, an opinion extraction process has been proposed. The process starts by defining a keyword (e.g. the product's brand name). Based on that keyword, the related reviews and the social relationship between their authors, are both extracted in order to build a customized social network for that specific keyword. The collected online data are reorganized following a data pattern that will be detailed later (see data modelling section). The gathered data are refined to improve their pertinence and significance. This step is based on removing unnecessary reviews (objective reviews, duplicated reviews by the same user, repeated letters, URLs, emotion icons, etc.). Then, the authors who have no more reviews to be analyzed are removed. Moving to the next step, data analysis will allow concluding the expressed opinions. The components included in the framework are described as follows (see Fig. 1). The opinion extraction and analysis is a three step process:

- Data Extraction: It aims to extract reviews and the users' personal information from several social media in order to preserve the input data and build the related social network.
- Data Refinement: It was observed that the existing posts in social media are often trivial or off-topic even if it contains the wanted keyword. Thus, the extracted reviews have to be refined in order to maintain the opinionated reviews. This step may favorably affect the precision of the results. The performed quality

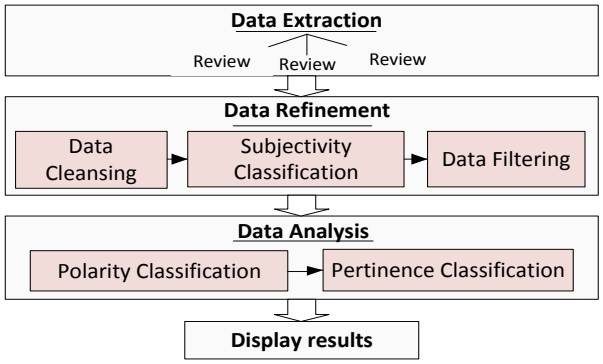


Fig. 1. The Opinion Extraction Process

verification has three main steps: data cleansing, Subjectivity Classification and data filtering

- Data Analysis: At this stage, the framework analyses the filtered reviews. The sentiment keywords are collected to detect their polarity. Thus, each polarity score is assembled with its user, to be used at the fuzzy clustering level with the objective of performing a pertinence classification.

A. Data Extraction

Social data collection refers to all the methods that have as a goal collecting social interactions between entities. Considering that people are socially connected, the task of extracting social media users is important to have more details about their profile information and their network. Members of the same network are often sharing similar interests and they even influence each other. Thus, several social networks are created for each keyword depending on the targeted social media (Facebook, Twitter, etc.). The social network is created by linking users with implicit and explicit bridges. The implicit links are built if users share, like or comment on a post of a user with whom they do not share any social relationship. The explicit link expresses the social link between two users (friends, followers, etc.). We suppose that we have "n" users, so the resulting social network for a certain keyword and social media will be defined as follows:

$$SN_{(KW,SN)} = (V,E), \text{ where} \begin{Bmatrix} V = (u_i)_n \text{ a set of vertices} \\ E \text{ a set of implicite/explicite directed edges} \end{Bmatrix} \quad (1)$$

The data are extracted using the suitable API for each social media and reassembled as one single huge storage. The data integration in the database should follow a standardized and comprehensive template so that the data can be analyzed easier and alike.

B. Data Refinement

The data refinement is achieved by performing three steps: data cleansing, Subjectivity Classification and data filtering.

1) Data Cleansing

This step aims to rectify data anomalies, irregularities and repetitions before exchanging them and to allow the production of a cleaned repository.

Each social media had its spam words/sentences. Twitter, for example, permits only a message with a maximum length of 140 characters. Thus, a tweet that contains a long URL or too many tags and hashtags can be considered as a spam content. URLs and tags/hashtags are detected using regular expression functions.

These are some examples of spam words/sentences:

TABLE II
SPAM TYPES ON TWITTER

Type of spam	Example
Long URL	http://car-us.com/car/volkswagen-touran/39883 …#volkswagen 2010 VW Touran 2.0 …
Hashtags and username	#nouveau#Mini#Clubman long commeune#Volkswagen#Golf@MINI_FR@MyMiniParis@lookatmyminihttp://urlc.fr/AdTaEM

At the end of data cleansing, the users, who only have empty reviews left, are automatically removed.

2) Subjectivity classification

Not everything posted in social media contains an opinion. They can be advertisements and other irrelevant texts containing no opinions. To overcome this problem, a subjectivity classification is required to determine whether a certain review is subjective or objective (contains an opinion or not).

The textual information is divided into two main types: facts and opinions. Facts are something that can be checked and backed up with evidences. Opinions are usually based on a belief or view but not on evidence. Thus, we can classify textual expressions into two types: objective and subjective sentences.

In this work, we are concerned with analyzing subjective reviews that describe people's sentiments, appraisals or feelings. Therefore, the task is to classify the posts as opinionated and not opinionated (subjectivity classification), and then to pick up the sentiment expressions that reveal the user's opinion about the targeted product from the subjective sentences. To do so, some tools allow establishing the subjectivity classification such as SentiWordNet [25]. This tool gives and enhances lexical resource for sentiment analysis and opinion mining. It assigns for each word several synonyms called "synsets". Each "synset" is associated with three scores: objective, the positive and negative score for the different terms in the "synset". The subjectivity of each "synset" is then evaluated as shown in (2):

$$Sub(t) = \frac{\sum_{s \in t} neg(s) + pos(s)}{|t|}, \text{ where t=}\{s_1,...,s_n\} \quad (2)$$

in which si represents a "synset" and t the word found in the review.

To Evaluate the subjectivity of a review, we use (3) in which $|\{t \cap t' \mid t \in d, t' \in SWN\}|$ means the total number of terms found in the review belonging to SWN:

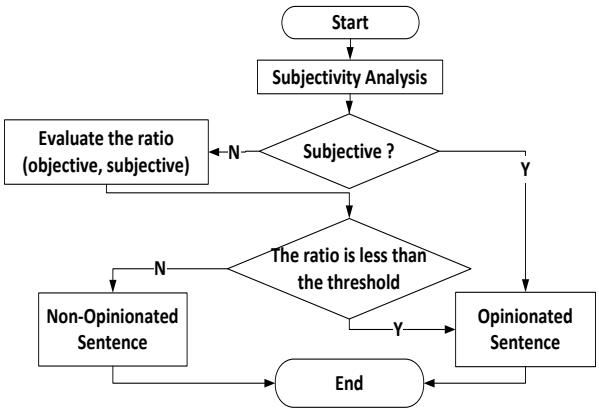


Fig. 3. Opinionated sentence identification algorithm

$$Sub(r) = \frac{\sum_{t \in r} Sub(t)}{\left|\{t \cap t' \mid t \in d,\ t' \in SWN\}\right|} \quad (3)$$

Using SentiWordNet, an algorithm was developed (as shown in Fig. 3) to classify sentences into two classes: subjective and objective. This algorithm is based on the idea that each sentence contains a percentage of subjectivity and objectivity. As a result, after assigning an assessment of subjectivity, the ratio of objective and subjective scores is evaluated to determine the amount of subjectivity infiltrated in a sentence that seems, at first sight, objective. If the percentage of subjectivity is significant then the post is considered as subjective (opinionated).

3) Data Filtering

Data filtering focuses on selecting the significant reviews among those extracted from social media. This can be done by determining whether the opinion of a node (user) in the constructed network is worth listening. To do so, two scores were established:

a) Knowledge score (KS):

Lately, a new category of social media's users has appeared. This group of people is categorized of their level of knowledge on a certain subject. As a result, a knowledge analysis is required in order to determine if the shared opinion was based on a high level of expertise in the field or simply based on personal experience. The knowledge score is evaluated as shown in the equation (4):

$$KS(u) = \frac{\sum_{s_i \in S} s_i \times pertinence(s_i, u)}{|S|} \quad (4)$$

in which |S| is the number of posts, "si" a post and "pertinence(si,u)" is a score that define the quality of the post and how much this review is interesting. This can be done by counting the number of times that the keyword and its features are repeated.

In other words, a social media user with the highest number of posts related to the subject of the research (product, brand, category of the product) is going to have the highest knowledge score. This proves that this user has more knowledge than others do and that his opinion will be more objective and interesting.

b) Authority score (AS):

The importance and the impact of an opinion in social media is not evaluated only based on the expertise level of its author, but also on his authority in his network. A person with a huge number of friends, followers, etc. means that he might have a high visibility [26]. This dominant factor have been taken into consideration in several works [27]. To make a use of this important parameter, an authority score is evaluated based on his influence and the rating score of each of the published reviews.

The first score measure the influence of a user in his network (5). This score can be determined using the implicit and explicit links between users of a given social media. Whether the relationship exists (followers, friends, subscribers, etc.) or not (commenting or liking a post without the existence of any relationship), these bridges can be explicit in the first case or implicit in the second.

$$IS(u) = \frac{\sum_{u \in SN} (E_{aij} + I_{aij})}{|SN - \{u\}|} \quad (5)$$

in which, Eaij (Iaij) stands for the explicit (implicit) link between two users. Eaij=1 (Iaij=1) if the link exists; Eaij=0 (Iaij=0), otherwise. |SN − {u}| represents the total number of users on the social network (SN) except user u.

On the other hand, most of the existent social media offer rating tools on publications that are used in order to generate a Rate Score (RS).

C. Data Analysis

Data analysis aims to extract the polarity of the extracted opinions. This task is done in two main steps: polarity classification and pertinence classification.

1) Polarity Classification

The sentence-level sentiment classification is the task of performing the subjectivity classification followed by the polarity classification. Therefore, considering the "subjectivity classification" have already been done in the previous section, the next step is to apply the polarity classification (classifying opinionated text into three groups: positive, neutral and negative opinion).

Sentiment analysis was a subject of several studies. The aim of sentiment analysis is to determine the attitude of a person toward a specific topic, an event, or an object. In our case, the main goal is to collect user's opinion about a product or a service from social media. To do so, a classification and quantification of the collected sentences is made using SentiWordNet (proposes a subjectivity classification and a polarity score), although there are some alternatives like SentiGem and SentiStrenght [28].

2) Pertinence Classification: Fuzzy C-means

At data refinement level, the credibility score for each user (node) has been evaluated. In order to reach a consensus among consumers' ratings, the framework uses the majority opinion.

The problem here is that an important segment of society, that can be in most cases a minority, are neglected even though they have both strong expertise and strong trustworthiness (leaders, experts or strict consumers [29]).

To separate the majority from the leaders, a fuzzy clustering is performed. Many benefits carry consensus clustering (K-means and fuzzy C-means algorithms); it helps to generate robust clusters, find "unusual" ones, and even handle noise and outliers.

V. The Crowdsourcing Component (CC)

Crowdsourcing describes a new form of outsourcing tasks. The term itself is a neologism that combines "crowd" and "outsourcing". In recent years, crowdsourcing [30] has been successfully used to deal with problems that are difficult to solve using only computer algorithms. However, crowdsourcing needs more costs in order to encourage people to participate; low efficiency is also a challenging problem.

In order to evaluate the relevancy of our framework, we are going to integrate a crowdsourcing component. In our case, we are interested in crowd rating type since there are different types of crowdsourcing, among them we can mention:

- Crowd Processing: It is based on dividing the process into micro-tasks and delegating them to the crowd.
- Crowd Solving: The contributions in this kind of crowdsourcing are independent from each other and represent alternative or complementary solutions to a given task or problem.
- Crowd Rating: Implementing an effective rating mechanism that sufficiently catches crowdsourcees' (individuals from the crowd who perform the tasks) perceptions and opinions is the central problem within crowd rating systems.

The assessments are collected from the crowd in order to

calculate their scores regarding the product and finally to compare them to the results extracted from social media.

Fig. 4 shows the general view of the proposed crowd rating process.

The main modules of CC are the Process Generator, the Process Engine, the Task Manager and the Data Mining and statistics module.

- Query: The request sent by the company in order to start the crowd rating process. This query is written as four components (Object, Context, Assessments, and Strategy) defined by the process generator.
- Process Generator: Receives the query Q = (object, context, assessments, strategy) and transform it into a processing plan. In our case, the query will be expressed as follows: Q = (product name, "collecting crowd rating", [very bad, bad, neutral, good, very good], Buffer strategy).
- Process Engine: prepares the product details in order to select the appropriate crowd participants (name, image …) and collects assessments and participants' location.
- Task Manager: receives progressively the tasks from the Process Engine and communicates with the crowd via a crowd platform to post tasks and retrieve the answers, then send them to the Process Engine.
- Data Mining/stats: Provides an integral dashboard as an easy way to display the found results.
- Crowd Participant Rating Component: Evaluates, after the end of each crowd rating process, the scores of the contributors. Thus, the crowd participants will be rated based on the quality of the result, the participation rate and the response time that should not exceed a specified deadline. So, at the end of the rating process, the more the crowd participant opinion is close to the reality and submitted in a short time, the more his score is high and vice versa. This will help to select the participants with the highest scores for the next survey, in order to improve the collected results from the crowd and consequently, to have a correct idea about the product reputation.

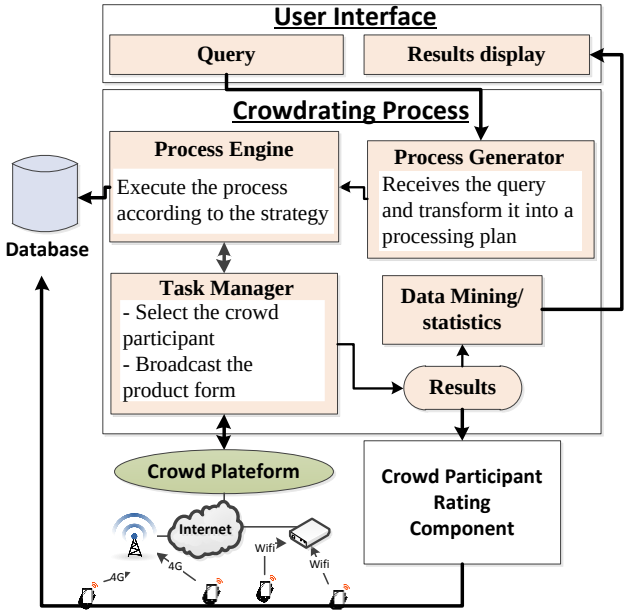


Fig. 4. The Crowdsourcing Component Architecture

The crowd sourcing data are composed of the crowd participants' personal information (age, gender, interests …) and their assessments. On the first hand, the crowd participants' details are used to select the appropriate crowd participants. On the other hand, the assessments are used as another input data in addition to the opinions extracted from social media.

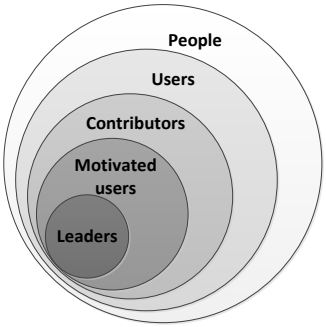


Fig. 5. People types

Obviously, the crowdsourcing component needs a crowd in order to collect knowledge. To do so, people should be motivated. There are several ways to attract people to make them collaborate with a crowdsourcing application, such as the financial reward, the opportunity to develop creative skills, to have fun and entertained, to share knowledge, the love of the community, etc. Considering this, people will love spending time on the application.

Maslow tried to organize people's motivations in a pyramid [31] (Maslow's pyramid) to order what people need. Realizing this will help us to satisfy one or more of the individual needs. The more the person is motivated, the more he is involved to the application, and consequently will have a higher score. As a result, five types of people arise (Fig. 5):

- Normal people: All the people in the world.
- Users: People registered in the application.
- Contributors: Users who accept to evaluate at least one product.
- Motivated users: Users who evaluate at least one product.
- Leaders: Motivated users who evaluate several products.

Assigning a score to each user is steadily done. The more the user participates and gives assessments close to the reality, the more his score is higher. Therefore, once a person is registered, his score will be equal to "0", and after each evaluation, this score is revaluated to determine the new one. This new score will help to select the new crowd in the case of a new investigation, and to determine the user's credibility.

VI. The Development Process

The component development process consists of analyzing, designing and implementing the software applications. To do so, a development modelling language such as UML (Unified Modelling Language) is used as a standard way to visualize the design of the component system.

In the following, we are going to present the chosen steps of the development process for a better planning and managing the framework functionalities and their interactions:

- The framework views: UML Use case Diagram and

Sequence Diagrams
- Data ware house Modelling: No SQL
- Data ware house implementing: HBase
- Data Processing: MapReduce

A. *UML Diagrams*

With the goal of describing the functionalities of the Opinion Extraction, analysis component (OEAC) and the crowdsourcing component (CC), we present several resulting UML diagrams based on an incremental iterative process driven by UML use case (Fig. 6).

The framework performs two main tasks: extracting opinion from social media and making a use of the crowd approach to collect more points of view about the targeted product. As a result, two UML sequence diagrams are made.

Figures 7 and 8 show, respectively the UML sequence diagrams for the assessment collection from the crowd and the opinion extraction from social media.

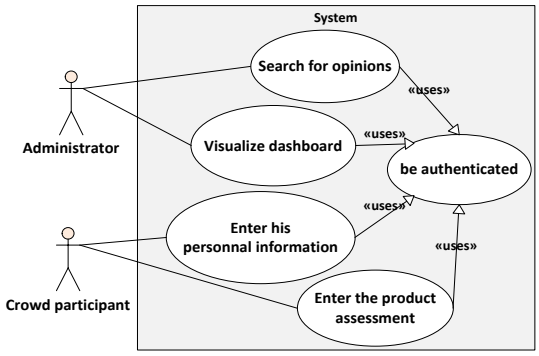


Fig. 6. Use Case Diagram

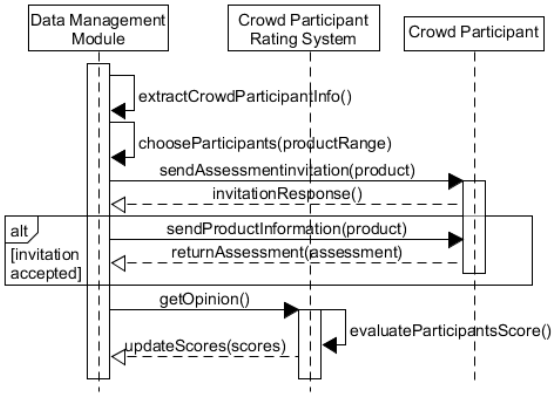


Fig. 7. Sequence Diagram for the Crowdsourcing Component

B. *Data warehouse modelling and implementation*

In this work, Apache Hadoop HBase was used for the data warehouse modelling and implementing in order to store data at the end of each step (data extraction, data refinement and data analysis). The most significant advantages of using this non-relational database is that it can handle storing and analyzing billions of rows and can be integrated with MapReduce.

Fig. 9 represents a part of data warehouse conceptual schema and Fig. 10 illustrates the corresponding HBase table structuring. In our case, the fact is the "Sentiment" entity (sentimentSM for the OAEC and sentimentCr for the CC) that has two dimensions: "Product" and "User" (crowd participant or social media users). Each dimension is mapped into a column family in the each HBase table.

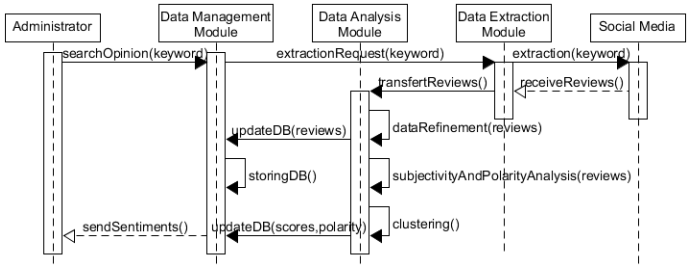


Fig. 8. A part of the sequence diagram for the Opinion Extraction process performed by the Opinion Extraction and Analysis Component (OEAC)

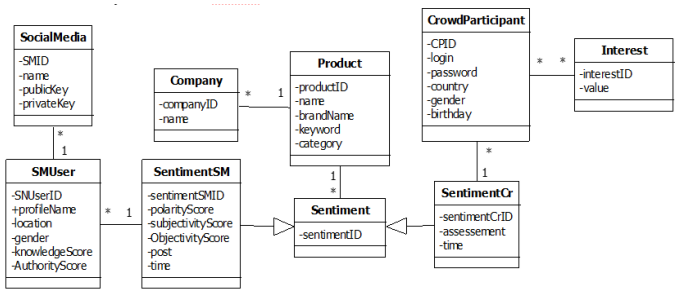


Fig. 9. A part of Social Media database schema

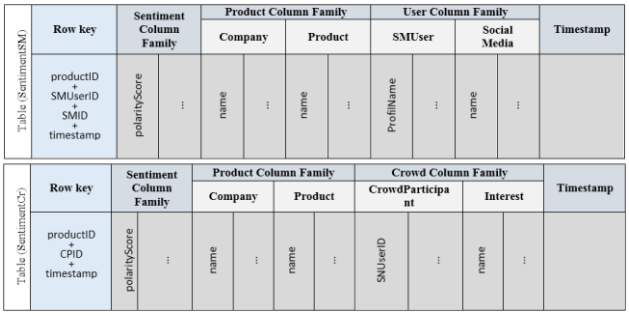


Fig. 10. The corresponding HBase tables

C. *Data Processing*

Processing vast amounts of data sets and easily writing applications within a reasonable time can only be done using a parallel computing programming model such as MapReduce.

The reviews are gathered and then partitioned according to the network's source (location) so it can be processed using MapReduce in order to perform a sentence-level sentiment classification.

Fig. 11 shows how MapReduce divides the input data set into independent fragments. The framework sorts the outputs of the map jobs and then inputs them to the reduce method.

At the Map stage, each review is refined in order to remove the invaluable reviews. The remaining reviews are then analyzed to get the sentiment scores (sentiment classification) using SentiWordNet. Each result is generated independently, comprising the review identifier, and the associated positive,

negative and neutral sentiment scores. Using this architecture, the Map algorithm can be easily adapted to perform different analyses on individual reviews by replacing "SentiWordNet" with another analysis package. The Reduce stage outputs the results obtained by the Mappers.

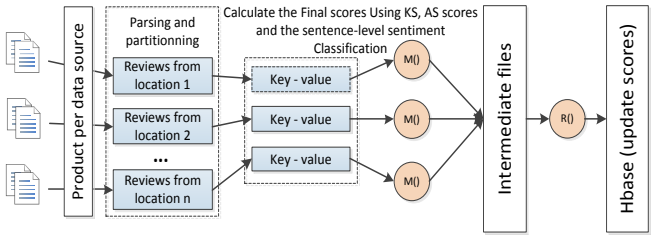


Fig. 11. The MapReduce data processing steps

VII. Case Study and Experimental Results

To evaluate our framework, a case study involving data from Twitter and Facebook have been performed. These two social media are used by a large number of people (approximately 332 million active users for Twitter and 1,59 billion active users of Facebook). These users are spread around the world and connected through computers, tablets and smart phones. We choose a car brand as our target of analysis.

Throughout this study, the data extracted from Twitter will be mainly used by detailing the process carried out during each step. Whereas Facebook data will only be used at the end to evaluate the maintenance of opinion in these platforms.

A. The Opinion extraction and analysis Component

1) Data Extraction

At this stage, the data is collected exploiting Twitter4J [32] and Restfb [33]. These unofficial Java libraries have allowed integrating respectively Twitter service and Facebook Graph API to our application. To test the proposed framework, 9500 reviews were collected from Twitter and Facebook, according to the chosen keyword (in our case a car brand name).

2) Data Refinement

We assuming that the extracted data size is 100%. After running the data cleansing (removing the RTtweets, URLs special symbols, etc.), the size of the full extracted reviews has gradually decreased to 78,3%. Performing the subjectivity classification has allowed to only retain the opinionated sentences, so the size of our repository has decreased again to 67,6%. Then, at the end of the data refinement (processing the data filtering stage), we are left with 63,59% of the totality of the reviews. These results (see Table III) have allowed us to conclude that refining data before analyzing them help gaining time and accuracy, since 36,41% of the reviews were detected as invaluable and been removed.

TABLE III
THE SIZE EVOLUTION AFTER PROCESSING DATA REFINEMENT

Pre-processing	Total %
Data Extraction	100%
Data Cleansing	78,3%
Subjectivity Classification	67,6%
Data Filtering	63,59%

3) Data Analysis

This step comes directly after executing data refinement module, in which unnecessary posts have to remove. In the experimental results, we have noticed how much reviews were removed. To illustrate more its impact on the results (the opinion polarity), we proceed by executing the data analysis module on the extracted and the refined data.

Fig. 12 shows the polarity dispersion before and after the execution of the pre-processing module (data refinement).

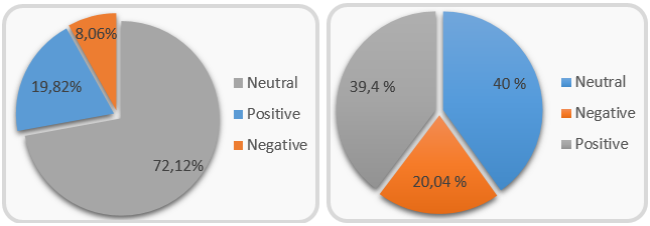


Fig. 12. Reviews polarity before and after the pre-processing

The experimental results show that 72,12% of neutral opinions were detected, but it turned out that only 40% are left after the pre-processing module. On the other hand, positive reviews have outstandingly increased from 19,82% to 39,6%. Contrariwise, negative opinions did not undergo any changes. To conclude, the results for each case were greatly different and this is up to the great number of the unwanted reviews hidden in the extracted data from Twitter.

Each social media has its own strategy and philosophy that characterizes it compared to the others. Facebook and Twitter, for example, are the most known and used social media in the world. The first one is usually used in a personal way. It allows creating a virtual community based on IRL (In Real Life) friends. On the other hand, Twitter can be used in a professional way, in most of the times, where users make create their community based on their interests. That leads us to search for opinions obtained from Facebook and compare them to Twitter results. To do so,

Fig. 13 shows the positive and negative scores gathered from Facebook and Twitter.

Fig. 13. Positive and Negative Opinions from Twitter and Facebook

The results show that both Twitter and Facebook users like the chosen car brand. For instance, 66,4% of the reviews extracted from Twitter expressed a positive opinion. In addition, the results a similar percentage have been found from Facebook.

To illustrate the effectiveness of our work, we collected the number of selling units of the chosen product over the year 2016 all over the world from the product's official website in order

to compare it with the results found via the proposed framework.

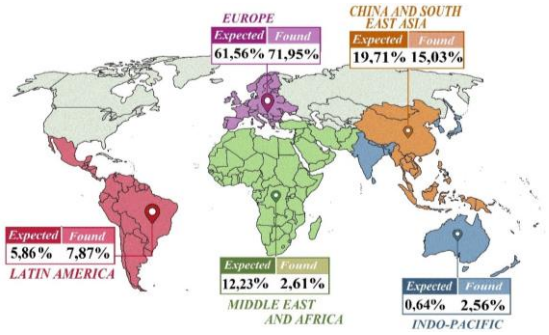


Fig. 14. Worldwide sales and experimental results in 2016

After extracting, refining and analyzing opinions from social media, we classify the results according to the user continent (

Fig. 14). Some results, especially those related to "Eurasia" and "North America", were not available for many reasons:

- These countries have initiated a major industrial take-off. Consequently, they produce their own brands in different product categories (e.g. cars).
- People in these regions publish in social networks with their tongues. However, data are usually extracted from the most known language (English).
- The official sites of the chosen product's brand themselves do not have statistics about these regions.

Classifying the opinions of the majority and leaders allows not only to retrieve the general opinion, but also to differentiate how common people points of view are spread among those who have great expertise and influence their network.

Fig. 15 represents the two clusters found after running the Fuzzy C-means in the filtered Twitter data. Therefore, the framework can return the majority opinion separated from the leaders' opinion.

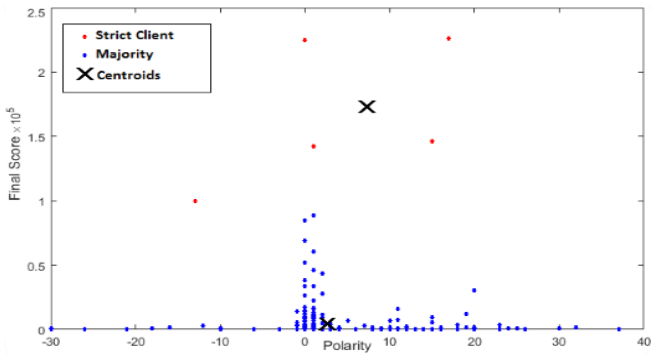


Fig. 15. Fuzzy C-means clustering

B. The Crowdsourcing Component

The crowdsourcing component aims to collect the rating scores of the targeted product, and as a result, it can be considered as a supportive data source. To reach our goal, a mobile application, called "Opinion Crowd Analytics (OC Analytics)", was developed so the crowd participants can easily submit their rating scores about the product.

This mobile application offers several interfaces (see Fig. 16) that allow the crowd to register, authenticate, visualize a list of products that have to be evaluated. In addition, a sheet that contains more details about the product is also provided. Finally, the application gives an assessment to the products that the "OCAnalytics" user accept to evaluate.

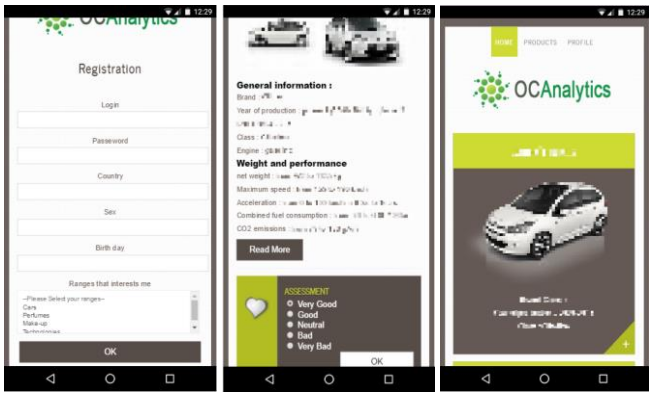


Fig. 16. Screenshots of the crowdsourcing component

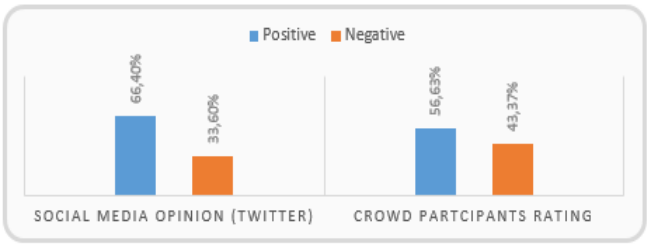


Fig. 17. Results obtained from Twitter and the crowd participants

A sample people was involved in order to get the experimental results for the same car brand. The crowd ratings will be compared to the extracted opinions from Twitter (see

Fig. 17).

The results show that 66,40% of positive opinions were gathered using the OEAC and, on the other side, 55,63% of the crowd have announced that they like that product.

VIII. Discussion

The proposed framework is divided into two main components:

- OEAC (Opinion Extraction and Analysis Component): that uses social media data (reviews and social relationships) as input data sets.
- CC (Crowdsourcing Component): that uses the crowd data to complete and support OEAC's results.

This combination is the strength of our approach (the previous researches rely only on social media results), and the two components complement each other. On one hand, the OEAC processes a large volume of data, which makes it able to reach a wide range of people. However, it lacks of human intelligence. On the other hand, the CC uses human intelligence to support the results obtained by the OEAC and interact with people who have interacted or are interested by the product.

Collecting these data (OEAC and CC data) makes this application suitable for big data analytics. The storage and the processing time are the two main constraints that have to be

considered while processing large data sets such as social media data. As a result, the Hadoop MapReduce approach was used in order to guarantee that the large volumes of collected, structured and unstructured data are handled with more efficiency than the traditional enterprise data warehouse, the distributed processing, the flexibility and the ability to adapt it to other technologies.

Twitter, one of the most famous social media, has been used just to attest the effectiveness of the proposed framework. However, the proposed approach is social media agnostic. In order to interface with other social media, we only need to change the initial configuration file.

IX. CONCLUSION

Initially, social media have been created to let their users to maintain contact with their friends, strengthen friendships, entertain alone or with others, plan trips, post photos, etc. However, people starts also exchanging information, feelings and opinions about entities (events, products or services). Consequently, these platforms are considered as an open source of valuable information for companies. Thus, gathering and analyzing these opinions seems to be a beneficial for companies and has become an essential trend. In this paper, we presented a social framework that predicts reputations via opinions expressed in social media and rating scores given from the crowd. In the first part, we focused on strict consumers who are generally excluded, so we argued about the use of fuzzy clustering to determine a consumer's credibility. In the second part, a crowdsourcing component was created in order to gather the crowd rating scores. The selection of the crowd is based on the crowd participants' rating component that evaluates the crowd participant's credibility. In order to evaluate the effectiveness of our framework, we placed it under probation using Twitter as a data source. The experimental results showed that the proposed framework gives results close to the reality and helps to classify both the general opinion and the leaders' opinion using Fuzzy C-means clustering (whether those collected from social media or the crowd). This review illuminates the need for a further study with the possibility to extend the extraction of general opinion from social media by gathering opinions about each feature of a certain entity (product or service). Using the opinions about the product's features instead of general opinions, will help to evaluate the product without prejudice and to specify the features that the clients hate or like. At the crowdsourcing component level, it was seen that the crowd need to be involved much deeper in the product development cycle. For that reason, as a future work, we propose to harness the creativity of the crowd in order to unleash their imagination to suggest new models for products' future generations. The goal is to make customers feel that they are the owners and the decision-makers because they are the main users of these products.

REFERENCES

[1] P. Gundecha, H. Liu: "*Mining Social Media: A Brief Introduction",* in 2012 Tutorials in Operations Research: New Directions in Informatics, Optimization, Logistics, and Production, pp. 1-17, 2012. doi: 10.1287/educ.1120.0105.

[2] W. Medhat, A. Hassan, H. Korashy: *Sentiment analysis algorithms and applications: A survey*, Ain Shams Eng. J., Vol. 5, No. 4, pp. 1093-1113, 2014. doi: 10.1016/j.asej.2014.04.011.

[3] J. Ge, M. A. Vazquez, and U. Gretzel: *Sentiment analysis: A review*, in Advances in Social Media for Travel, Tourism and Hospitality: New Perspectives, Practice and Cases, New York, USA, pp. 243-261, 2018. doi: 10.4324/9781315565736.

[4] X. Wang, C. Ding, W. Zheng, and M. Wu: *Sentiment Analysis based on Specific Dictionary and Sentence Analysis*, in Proc. of the 2017 International Conference on Economics and Management, Education, Humanities and Social Sciences (EMEHSS 2017), Hangzhou, China, 2017. doi: 10.2991/emehss-17.2017.2.

[5] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane: *Current state of text sentiment analysis from opinion to emotion mining*, ACM Comput. Surv., Vol. 50, No. 2, pp. 1–33, 2017. doi: 10.1145/3057270.

[6] D. M. E.-D. M. Hussein: *A survey on sentiment analysis challenges*, J. King Saud Univ. - Eng. Sci., Vol. 30, No. 4, pp. 330–338, 2018. doi: 10.1016/j.jksues.2016.04.002

[7] A. E. Green, M. De Hoyos, S.-A. Barnes, B. Baldauf, H. Behle: "CrowdEmploy Crowdsourcing Case Studies : An Empirical Investigation into the Impact of Crowdsourcing on Employability", Publications Office of the European Union, 2013. doi: 10.2791/53978

[8] K. J. Stol and B. Fitzgerald: *Two's company, three's a crowd: A case study of crowdsourcing software development*, in Proc. of the 36th International Conference on Software Engineering, pp. 187-198, 2014. doi: 10.1145/2568225.2568249

[9] N. Tsapatsoulis and C. Djouvas: *Opinion mining from social media short texts: Does collective intelligence beat deep learning?*, Front. Robot. AI, Vol. 5, No. JAN, pp. 1-14, 2019. doi: 10.3389/frobt.2018.00138

[10] C. W. Holsapple, S. H. Hsiao, and R. Pakath, "Business social media analytics: Characterization and conceptual framework," *Decis. Support Syst.*, no. March, pp. 1–14, 2018. doi: 10.1016/j.dss.2018.03.004

[11] J. Kordonis, S. Symeonidis, and A. Arampatzis: *Stock price forecasting via sentiment analysis on Twitter*, in PCI '16: Proceedings of the 20th Pan-Hellenic Conference on Informatics, pp. 1–6, 2016. doi: 10.1145/3003733.3003787

[12] Y. Li, V. Rakesh, and C. K. Reddy: *Project success prediction in crowdfunding environments*, in WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining, pp. 247–256, 2016. doi: 10.1145/2835776.2835791

[13] H. Ahn and W. S. Spangler: *Sales prediction with social media analysis*, in Annual SRII Global Conference, SRII, pp. 213-222, 2014. doi: 10.1109/SRII.2014.37

[14] R. Kaur, H. Singh, and G. Gupta: *Sentimental Analysis on Facebook Comments using Data Mining Technique*, Int. J. Comput. Sci. Mob. Comput., Vol. 8, No. 8, pp. 17–21, 2019.

[15] J. Bollen, H. Mao, and X. Zeng: *Twitter mood predicts the stock market*, J. Comput. Sci., Vol. 2, No. 1, pp. 1–8, 2011. doi: 10.1016/j.jocs.2010.12.007

[16] P. Nakov, S. Rosenthal, A. Ritter, and T. Wilson: *SemEval-2013 Task 2: Sentiment Analysis in Twitter*, in Proceedings of the International Workshop on Semantic Evaluation (SemEval-2013), Vol. 2, No. SemEval, pp. 312–320, 2013.

[17] H. Hamdan, P. Bellot, and F. Bechet: Lsislif: *Feature Extraction and Label Weighting for Sentiment Analysis in Twitter*, in Proceedings of the 9th International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp. 568–573, 2015. doi: 10.18653/v1/s15-2095

[18] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman: *Impact of feature selection techniques for tweet sentiment classification*, in Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, pp. 299–304, 2015.

[19] M. Cartwright, G. Dove, A. E. M. Méndez, J. P. Bello, and O. Nov: *Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists*, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11, 2019. doi: 10.1145/3290605.3300522

[20] S. Nowak, S. Rüger: *How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation*, in Proc. of MIR conf., pp. 557–566, 2010. doi: 10.1145/1743384.1743478

[21] R. Q. Wang, H. Mao, Y. Wang, C. Rae, and W. Shaw: *Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data*, Comput. Geosci., Vol. 111, pp. 139–147, 2018. doi: 10.1016/j.cageo.2017.11.008

[22] R. K. Qarout, A. Checco, and K. Bontcheva: *Investigating stability and*

reliability of crowdsourcing output, in CEUR Workshop Proceedings, pp. 83–87, 2018.

[23] C. Grady, M. Lease: *Crowdsourcing document relevance assessment with Mechanical Turk*, in Proc. of NAACL HLT 2010 Work. Creat. Speech Lang. Data with Amaz. Mech. Turk, pp. 172–179, 2010.

[24] P. Hsueh, P. Melville, V. Sindhwani: *Data quality from crowdsourcing: a study of annotation selection criteria*, in Proc. of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, pp. 27–35, June 2009. doi: 10.1.1.157.5154

[25] S. Baccianella, A. Esuli, F. Sebastiani: *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, in Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 2200–2204, 2010. doi: citeulike-article-id:9238846

[26] N. C. Rathore, S. Tripathy, and J. Chandra: *Predicting user visibility in online social networks using local connectivity properties*, in Distributed Computing and Internet Technology: 11th International Conference, ICDCIT 2015, Proceedings, Vol. 8956, pp. 419–430, 2015. doi: 10.1007/978-3-319-14977-6_46

[27] N. C. Rathore and S. Tripathy: *Epidemic model based visibility estimation in Online Social Networks*, in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI 2018), pp. 2161–2168, 2018. doi: 10.1109/ICACCI.2018.8554671

[28] M. Thelwall, K. Buckley, G. Paltoglou: Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol., Vol. 63, No. 1, pp. 163–173, 2012. doi: 10.1002/asi.21662

[29] Z. Saoud, N. Faci, Z. Maamar, D. Benslimane: A Fuzzy Clustering-based Credibility Model for Trust Assessment in a Service-oriented Architecture, in Proc. of the 23rd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE-2014), pp. 56–61, 2014. doi: 10.1109/WETICE.2014.35

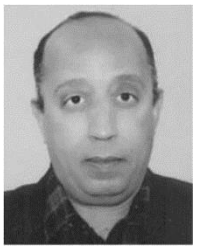
[30] M. Vukovic: *Crowdsourcing for enterprises*, in 2009 Congress on Services - I, pp. 686–692, 2009. doi: 10.1109/SERVICES-I.2009.56

[31] E. Estellés-arolas, F. González-ladrón-de-guevara: *Towards an integrated crowdsourcing definition*, Vol. 38, No. 2, pp. 189–200, 2012. doi: 10.1177/016555150000000

[32] "Twitter4J." [Online]. Available: http://twitter4j.org/.

[33] "Restfb." [Online]. Available: http://restfb.com/.

F. Z. Ennaji is a computer science engineer, graduated from the National School of Applied Sciences Marrakesh/Morocco at 2014. She joined in 2015 the Computing Systems Engineering Laboratory (LISI) of CADI AYYAD University. Her thesis includes many research interests like social media, sentiment analysis, data mining, Big Data Analytics, crowdsourcing, social CRM, etc. She participated in many international conferences and published many articles in international journals. Recently, and as a PhD, she is working on the same research areas on top of Machine Learning and Internet Of Things.



A. El Fazziki Received the M.S. degree from the University of Nancy, France, in 1985, and the Ph.D. degree in computer science from CADI AYYAD University in 2002. He has been with CADI AYYAD University since 1985, where he is currently a Professor of computer science. He has been responsible for the engineering since 2006. He was the Director of the Computer Systems Engineering Laboratory between 2011 and 2015. He has co-authored several papers on agent-based image processing. In addition, he is the main author of over 20 papers in software engineering and data analytics field. His research interests are related to software engineering, decision support, big data, data analytics, crowdsourcing, and e-government. In the MDA field, he has been involved in agent-based systems, service-oriented systems, and decisional systems.

H. El Alaoui El Abdallaoui is a computer science engineer, graduated from the National School of Applied Sciences Marrakesh/Morocco year 2014. After acquiring scientific and technical knowledge in the computer and information systems field; especially in designing, modeling and implementing software solutions from both the architectural and administrative perspectives, she then integrated the Computing Systems Engineering Laboratory of CADI AYYAD University since January 2016 to prepare her PhD thesis. Her research interests include e-government applications, crowdsourcing, image processing, mobile applications, etc.

H. El Kabtane received the M.S. degree from the Faculty of Sciences Ibn Tofail of Kenitra, Morocco, in 2012 and then he holds a PhD degree in the Information Systems Engineering Laboratory (LISI) in the Faculty of Sciences, Cadi Ayyad University of Marrakech, Morocco. He worked on the Virtual Learning Environments, the 3D Image Processing, Virtual Reality and Augmented Reality. Now, as a PhD and in addition to what have been said before, he is also working on Machine Learning and Internet of Things.