A novel adaptable approach for sentiment analysis on big social data

Imane El Alaoui, Youssef Gahi, Rochdi Messoussi, Youness Chaabi, Alexis Todoskoff, Abdessamad Kobi

El Alaoui *et al. J Big Data (2018) 5:12*
https://doi.org/10.1186/s40537-018-0120-0

Journal of Big Data

METHODOLOGY Open Access

A novel adaptable approach for sentiment analysis on big social data

Imane El Alaoui[1,2*], Youssef Gahi[3], Rochdi Messoussi[1], Youness Chaabi[1], Alexis Todoskoff[2] and Abdessamad Kobi[2]

*Correspondence:
imane.el.alaoui@uit.ac.ma
[1] Laboratoire des Systèmes de Télécommunications et Ingénierie de la Décision, University of Ibn Tofail, Kenitra, Morocco
Full list of author information is available at the end of the article

Abstract

Gathering public opinion by analyzing big social data has attracted wide attention due to its interactive and real time nature. For this, recent studies have relied on both social media and sentiment analysis in order to accompany big events by tracking people's behavior. In this paper, we propose an adaptable sentiment analysis approach that analyzes social media posts and extracts user's opinion in real-time. The proposed approach consists of first constructing a dynamic dictionary of words' polarity based on a selected set of hashtags related to a given topic, then, classifying the tweets under several classes by introducing new features that strongly fine-tune the polarity degree of a post. To validate our approach, we classified the tweets related to the 2016 US election. The results of prototype tests have performed a good accuracy in detecting positive and negative classes and their sub-classes.

Introduction

Social media and its corresponding applications allow millions of users to express and spread their opinions about a topic and show their attitudes by liking or disliking content. All these constantly accumulating actions on social media generate high-volume, high-velocity, high-variety, high-value, high-variability data termed as big social data. In general, this kind of data refers to massive set of opinions that could be processed to determine people tendencies in the digital realm. Several researchers have shown a keen interest in the exploitation of big social data in order to describe, determine and predict human behaviors in several domains [10, 26]. Processing this kind involve various research avenues, particularly, text analysis. In fact, almost 80% of internet data is text [23], therefore, text analysis has become key element for public sentiment and opinion elicitation. Sentiment analysis, which is also called opinion mining, aims to determine people's sentiment about a topic by analyzing their posts and different actions on social media. Then, it consists of classifying the posts polarity into different opposite feelings such as positive, negative and so on.

Sentiment analysis could be divided into two main categories:

- *Lexicon analysis* aims to calculate the polarity of a document from the semantic orientation of words or phrases in the document. However, applications based on lexicon analysis do not consider the studied context.

- *Machine learning (ML)* involves building models from labeled training dataset (instances of texts or sentences) in order to determine the orientation of a document. Studies that used this type of methods have been carried out on a specific topic.

These two analysis methods have been widely used on big social data to gather public opinion in order to asssess internauts satisfaction of a subject (services, products, events, topics or persons) in several domains including politics [3], marketing [4] and health [7]. However, the results are varying, sometimes concluding with a reasonable degree of accuracy and sometimes are not. The failure is generally due to the opinion mining challenges such as the semantic orientation of a word which could change depending on the context. In this paper, we aim to tackle semantic analysis by introducing a novel adaptable approach that relies on social media posts and big data architecture to analyze internauts' behaviors and feelings toward a subject in real-time. The proposed approach is based on three stages as shown in Fig. 1.

In order to validate our proposed approach, we built a prototype and conducted a study on analyzing the 2016 US election related tweets to find out which candidate is the favorite.

The remainder of this paper is organized as follows. In the second section, work related to analyze social media data and its correlation with trends are explored. In the third section, we highlight the theoretical basis on textual analysis. Section four presents an overview of the proposed method. The experimental methodology and

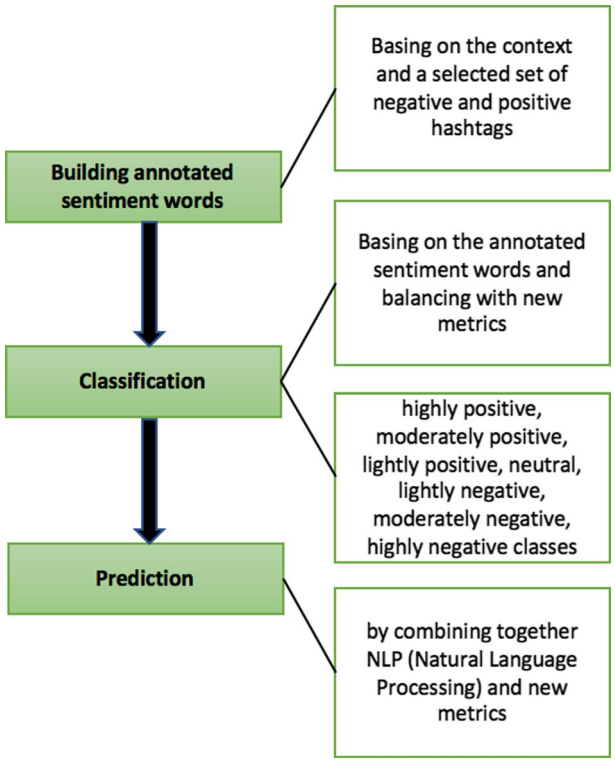

Fig. 1 Stages of the model

results are presented in "Methods and experience: US presidential elections" and "Results" sections respectively.

Related work

Several studies have focused on analyzing social media especially when it is related to some big events that attract a great attention like presidential elections. Social media became an important platform for both, candidates to share their programs and to get in direct touch with people, and voters to express themselves about each candidate. This intensive use of social media has attracted wide attention in academic research and many contributions have been conducted to follow that kind of events.

These contributions could be categorized into four categories:

- *Opinion-based approach* In which the authors have based their models on opinion mining methods, detailed in "Theoretical basis" section, in order to classify posts related to a candidate. There are two categories classes:
  - *Sentiment classes* researchers in [1, 22, 24, 34], have classified sentiments under two classes (positive and negative). Other authors [15, 16, 25, 28, 33, 37, 40] have added a third class (neutral), Wang et al. [35] have defined one more class (unsure).
  - *Context classes* Other researchers, such as [6, 19], have personalized classes depending on the context. Conover et al. [6] have classified the posts in three classes which are Left, Ambiguous and Right while authors in [19] have based their prediction model on two classes (Pros and Anti) for each party.
- *Volume-based approach* Here, researchers aim to predict the elected candidate basing on the number of tweets mentioning them (% mention) or retweet volume. In fact, researchers in [9, 18, 27, 29, 39] have discovered an interesting correlation between name mentions percentage or retweet volume and the vote. Finn et al. [11], have come up with a new approach in order to measure political polarization without using text. They have used a co-retweeted network, as well as the retweeting behavior of the users.
- *Opinion and volume (OV)-based approach* In which both opinion mining and volume approaches are combined. For instance, sentiment classes (positive, negative, mixed, neutral), three other classes (unannotable, non-relevant, unclear) and volume-based measures were combined in [15]. Wong et al. [38], have used positive, negative, neutral classes, retweets and tweeters' information. Authors in [6] have used context classes as well as the number of tweets mentioning each party and retweet volume. Khatua et al. [17] have used tweet volume as well as the two following opinion mining methods: Hu and Liu's opinion lexicon method [14], that classifies words into positive and negative categories, and list of AFINN-111 [13], that attributes score to words (from strongly negative -5 to strongly positive $+5$).
- *Emoji-based approach* Here, the classification of posts is based on the emoji. Researchers in [8] have selected relevant emoji and have categorized them into happy, sad, fear, laughter, and angry classes, then, have stripped the sentiment of the first emoji in the post.

There are conflicting views on the reliability of social media analysis, some of the previous studies have demonstrated the correlation between election outcomes and their related posts while others, such as [8, 12, 32], have shown the reverse. Although their former research has been based on efficient statistical methods or/and some well-known dictionaries as shown in Table 1, they failed. In general, this issue is due to opinion mining challenges such as the definition of word semantic orientation which could strongly change depending on the context. For instance, using the term sexual would be considered as negative word in a tweet related to trump, who has been accused of sexual assault by fifteen women, and by opposition, it will be considered as a positive word for Hillary. Another example, the semantic orientation of the word email, which is generally a neutral word but could change to negative in the context of Hillary, who contravenes the federal laws by using personal email account for government business.

In order to classify social media posts, previous contributions have either used generic dictionaries that do not consider the studied context, or have, used machine learning methods that use training data for a specific domain. However, in both cases, the stated methods are less efficient. In this paper, we aim to fill the gap of opinion mining in terms of context semantic orientation by proposing a novel adaptable approach that allows to automatically attribute positive or negative score to words depending on the context. In order to calculate the overall polarity of a post, we referred to the previous researches and added some value by introducing new metrics. Then, we classify posts basing on the final score under seven classes: highly positive, moderately positive, lightly positive, neutral, lightly negative, moderately negative, and highly negative. Finally, to gather public opinion, we assume that these different classes impact significantly and dissimilarly the overall public opinion.

The proposed method is applicable to diverse Twitter analyzing scenarios and rely on textual analysis that we overview in what follow.

Theoretical basis

Textual analysis consists in extracting hidden patterns from textual data and includes several techniques such as data mining, natural language processing and ML.

Text mining

Text mining is the process of deriving useful information from unstructured textual data. This process relies on two major phases: The analysis phase refers to the process of structuring text by using linguistic analysis techniques such as recognizing the words, sentences, their grammatical roles and relationships. It involves several methods:

- *Language identification* is the act of determining the natural language in which a given text is written.
- *Tokenization* is the process of segmenting a sequence of strings into words and sentences by discarding some characters like punctuation marks.
- *Filtering* consists of applying filters such as removing empty words.
- *Lemmatization* consists of grouping together the different inflected forms of a word by removing plurals, genders and conjugations. Then, we analyze them as a single item.

Table 1 Commonly used approach in related work

Approach	Lexicon-based		Learn-based				Statistics		
	Manually	Dictionary	SVM	K-means	NB/K-NN	Other	LR	OLS	MA
Opinion-based									
Sentiment classes	Razzaq et al. [25], Ramteke et al. [24], Tunggawan et al.[34], Jahanbakhsh et al. [15], Wang et al. [35], Šmailović et al. [28], Tumitan et al. [33]	Balasubramanyan et al. [1], Jose et al. [16], Wicaksono et al. [37], Peka et al. [22], [24, 33]	[24, 25, 28, 33, 40]	[15, 24, 25, 34, 35, 37]	Xing et al. [40]	[16, 25, 33]			[1, 15]
Context classes	Mahmood et al. [19]		[19]	[19]		[19]			
Volume-based				Finn et al. [11]	[11]	Xie et al. [39], Soler et al. [29]	Livne et al. [18]	Digrazia et al. [9]	[39]
OV-based	Conover et al. [6]	Wong et al. [38], Khatua et al. [17]	[6]					[17]	
Emojis-based			Delenn et al. [8]		[8]	[8]			

- *Named-entity recognition* is the process of searching text object that can be categorized in classes such as persons, dates, localization.

The interpretation phase evaluates and interprets the output of the first phase by using data mining methods [12]. Its purpose is to find patterns, relevance, novelty, and interestingness [32]. It is worth noting that text mining does not allow to extract opinion, so it must be combined with other techniques. Next, we present some of those techniques.

Opinion mining

Opinion mining is the science of using text analysis to determine the sentiment orientation of a text (positive, negative or neutral). It can be found under different umbrella terms: sentiment analysis subjectivity, analysis of stance. One of its promising application is to track and understand the mood of the public in social media about a particular topic in several domains such as marketing, health and politic. For instance, potential buyers can make their decision according to the product reviews.

Opinion mining is generally carried out by the following approaches:

Lexicon-based approach

The lexicon-based approach relies on a sentiment lexicon and a collection of known sentiment terms. It is roughly divided into dictionary-based approach and corpus-based approach. The first one finds opinion words in the text, then, finds their semantic orientation in the dictionary. There is several dictionaries such as SentiWordNet, however, they can be created manually.

The second one consists of finding opinion words in a context specific orientation. It starts with a seed list of opinion words and then find other opinion words in a large corpus. Most of the lexicon-based researches have used adjectives and verbs as indicators of the semantic orientation of text [2, 5, 41].

Learn-based approach

The learn-based approach relies on the famous ML algorithms (i.e. supervised and unsupervised methods). In the supervised methods, we train the model trough a large number of labeled documents. The most known ones for opinion mining are: Naïve Bayesian classification, maximal entropy principle, and the support vector machine. Nevertheless, although these methods reach quite a high accuracy in detecting the polarity in the domain that they are trained on, their performance fall precipitously when the same model is used in a different domain.

When it is difficult to find a labeled training documents, the unsupervised methods are used. However, due to their poor performances in this area, it is rarely adopted at present.

Hybrid approach

The hybrid approach employs both, the lexicon and the learn-based approaches. It uses the lexicon-based approach for sentiment scoring. Then, these scored documents will represent the training data for the learn-based part. Hybrid approach is widely used coz

of its high accuracy and its stability inherited from ML powerful and the lexicon based approach, respectively.

Lexicon-based, learn-based and hybrid approaches have been widely used in various domains, in different ways and improved by several searchers. More details are given in [20].

By referring to opinion mining approaches, we present in the following section a method that analyzes social media posts and extracts user's opinion.

The proposed method

In order to build a sentiment analysis model, we propose in this paper a three stages-based methodology that consists of, first building sentiment words, then classifying and balancing this set of words before executing the prediction algorithm.

The description of the three stages are explained bellow. Let: $Y_i, i = 1, \ldots n$ be a set of products, services or persons that we aim to compare in a specific context.

Let's consider $D = \{Y_1, Y_2, \ldots, Y_n\}$ as the targeted context.

First stage: constructing dictionaries

Various researches in social media analysis, such as [30, 31], have identified whether the intention behind a post is positive or negative based on hashtags description. However, they have used a very large set of manually annotated hashtags (which is time consuming) or they have combined these latter with dictionaries in order to enhance classification posts accuracy. At the first stage of this work, we use a small set of hashtags, in a new way, in order to build dictionaries of words, annotated with the word's semantic orientation for a given context as following:

We assume that every word in a tweet that contains a positive hashtag is positive and vice versa, then, we process it by different steps.

- *Step1* consists of collecting and storing posts related to Y_i. Since our approach aim to compare Y_i, we identify hashtags that have a high frequency as the most popular hashtags for each Y_i. Then, we classify a very small set (between two and three for each one) of them manually into positive and negative classes and collect related data for each class separately. In other words, collected tweets will be classified into a positive or negative class basing on the upper defined polarity of hashtags.
- *Step2* consists of preprocessing classified data from hashtags. Social data is informal and could contain misspelling and non-textual information, hence the need of a preprocessing step. For this, we apply various filters on tweets as following:

  - *Tokenization* this sub-step consists of identifying nouns, verbs, adverbs, adjectives, URLs, common emoticons, phone numbers, HTML tags, Twitter mentions hashtags, and repetition of symbols and Unicode characters.
  - *Conversion* all words will be converted to lowercase and replace more than two of the same consecutive letters in a word with only one occurrence of the letter (e.g., we replace fuuunny by funny and ANGRY by angry).
  - *Stemming* by removing plurals genders and conjugation (applying morphology stemming).

– *Filtering* Different researches [2, 5, 41] have proved that both adjectives and verbs are good indicators for positive and negative sentiment analysis. However, as social data could contain more information than a formal text, we enhance these indicators by applying other various filters and sentiment indicators such as hashtags.

The output of this step will represent the intermediate sentiment words of each: *inter-posSW* (Y_i) and *inter-negSW* (Y_i)

- *Step 3* The purpose of this step is to refine the annotated dictionary: positive *posSW*(), negative *negSW*()and neutral *neutSW*() dictionaries for each Y_i. As the task of neutral hashtags classification is difficult and could affect the result, we ignored them during the collect. In fact, a tweet that contains a neutral hashtag such as #Trump could be either negative or positive. Therefore, we construct neutral basing on the word occurrence $Occ(w_j)$ for all in the different classes. This allows us to construct the final dictionaries by Algorithm 1.

pw and *nw* are words in intermediate positive and negative dictionary, respectively.

Algorithm 1 Constructing the final dictionaries

Require: $inter\text{-}posSW(Y_i)$,$inter\text{-}negSW(Y_i)$
Ensure: $posSW(Y_i), negSW(Y_i), neutSW(Y_i)$
$posSW(Y_i) = inter\text{-}posSW(Y_i)$
$negSW(Y_i) = inter\text{-}negSW(Y_i)$
for pw in $inter\text{-}posSW(Y_i)$ **do**
  for nw in $inter\text{-}negSW(Y_i)$ **do**
    if $pw == nw$ and $occ(pw) > occ(nw)$ **then**
      $ratio = occ(nw)/occ(pw)$
    else if $pw == nw$ and $occ(nw) > occ(pw)$ **then**
      $ratio = occ(pw)/occ(nw)$
    end if
    if $pw == nw$ and $occ(pw) > occ(nw)$ and $ratio < 0.7$ **then**
      Delete ng from $negSW(Y_i)$
    else if $pw == nw$ and $occ(pw) < occ(nw)$ and $ratio < 0.7$ **then**
      Delete pw from $posSW(Y_i)$
    else if $ratio > 0.7$ **then**
      Add pw in $neut\text{-}SW(Y_i)$
      Delete pw from $posfromSW(Y_i)$
      Delete ng from $negSW(Y_i)$
    end if
  end for
end for

We employed empirical test, which consists of testing a number of values (between 0.5 and 0.8), in order to constitute the limit that allows to classify sentiment words with the smallest error rate. In our case 0.7 was the best value.

Finally, we assign a score to sentiment words: 1, 0, − 1 for positive, neutral and negative, respectively. Figure 2 illustrates the modules of the first stage:

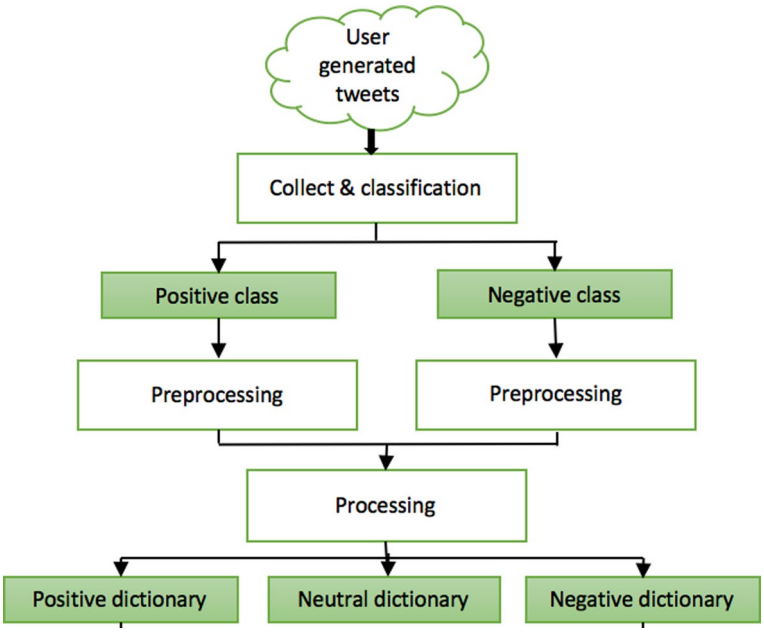

Fig. 2 First stage of the model

Second stage: classification

In this stage, we classify new tweets basing on the *SW* dictionary built in the previous stage. Classification steps are illustrated in Fig. 3.

- *Step 1* collect and store new tweets for each Y_i separately. Collected data goes through the following steps.
- *Step 2* Consist of preprocessing the data as following:
  - *Removing duplicated data* As a post could include more than one hashtag, it could be extracted for multiple times. Thus, we remove duplicated tweets in order to avoid misleading results.
  - *Tokenization* The same tokenization used in the first stage is applied to the new used tweets.
  - *Handling negation* As negation words (no, not, nothing and nonce) could significantly affect the overall polarity of a sentence, it is a very important criterion in sentiment classification. As recommended in [16, 21], we reverse the sentiment polarity of the words that come after a negation word until reaching a punctuation mark.
  - *Handling repetition* Detecting words that are written in uppercase or constitute more than two of the same consecutive letters.
  - *Applying morphology* by following the same rules as the first stage.
- *Step 3* In this step, we calculate polarity degree of the tweets basing on the semantic orientation of words assigned in stage 1. For this, we apply the two following actions:

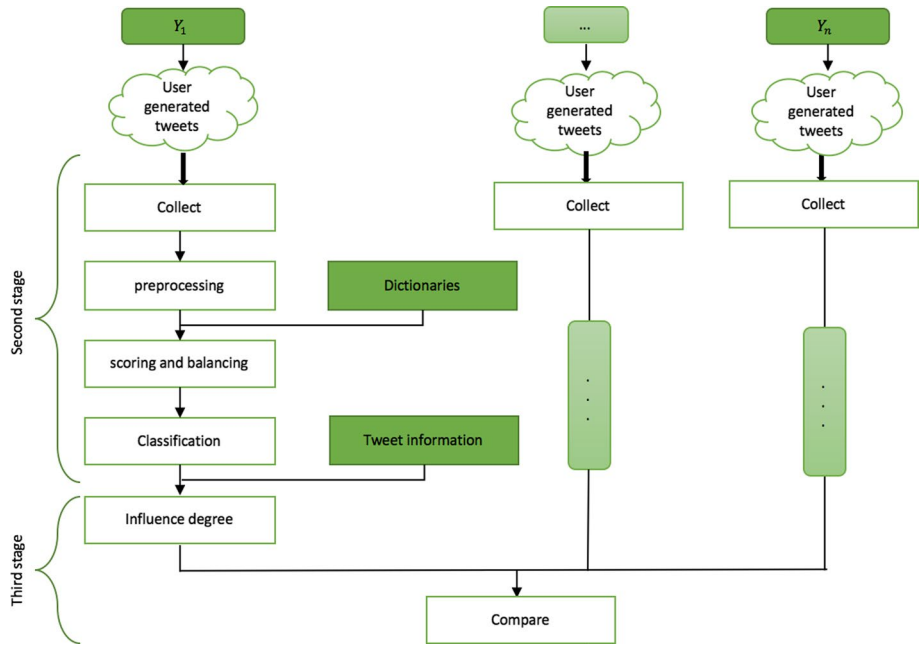

Fig. 3 Second and third stages of the model

– *Balancing* It is worth noting that the used language in social media posts is not conventional and could contain some special words such as those that are written in upper case or contain the repetition of more than two consecutive letter—"extended word"—. This kind of words are generally not fully exploited in most of sentimental analysis techniques [8, 12, 32]. However, extended words could intensify or reduce the polarity degree of the post. Therefore, we assign great attention to this kind of words in our proposed approach by balancing the words score while relying on these metrics. We have highlighted the desire to emphasize sentiment by using the uppercase and extended words. However, we have made the choice of not distinguish the different metrics in order to accentuate a sentiment. Therefore, we affect the same weight by adding $+1$ for positive words and -1 for negative words. For instance, if $score(\text{love}) = 1$, then $score(\text{looove}) = score(\text{LOVE}) = score(\text{LOOOVE}) = 2$.
– *Calculating polarity degree* The polarity of a tweet is calculated by adding up the independent score values of words stated in t. m is the length of t.

$$p(t) = \sum_{k=1}^{m} score(w) \tag{1}$$

- *Step 4* in order to classify tweets, we use $p(t)$ as following: We classify scored tweets into seven classes according to the polarity degree into $C_{+3}, C_{+2}, C_{+1}, C_0, C_{-1}, C_{-2}, C_{-3}$ (highly positive, moderately positive, lightly positive, neutral, lightly negative, moderately negative, highly negative classes, respectively). For this, we employed empirical test to determine the limit of each class.

If $0 < p(t) \leq 3$, then, the tweet is classified as lightly positive. If $4 \leq p(t) \leq 6$, it is moderately positive. If $p(t) \geq 7$, then, it is highly positive. If $-3 \leq p(t) < 0$, then, the tweet is classified as lightly negative. If $-6 \leq p(t) \leq -4$, it is moderately negative. If $p(t) \leq -7$, then, it is highly negative. There is a possibility for sentiment score to be equal to 0, if it is the case, then, the tweet is classified as neutral.

Third stage: prediction

Several researchers such as Wang et al. [36] have considered positive, negative and neutral classes to extract the sentiment of a document (based on words and/or emoticons) and only few ones, such Khatua et al. [17], have examined the polarity degree (i.e. highly, moderately, weakly positive and negative classes). However, to gather public opinion, authors have considered only strongly positive and strongly negative classes as indicator. Unlike the previous researches, we assume that each class impacts dissimilarly the strength of public opinion and the rate of Y_i for a given topic or domain.

- *Step 1* we attribute a weight for each class as following: $weightC_0 = 0$, $weightC_{+1} = 1$, $weightC_{+2} = 2$, $weightC_{+3} = 3$, $weightC_{-1} = -1$, $weightC_{-2} = -2$ and $weightC_{-3} = -3$.
- *Step 2* we assume that a tweet that have been retweeted or liked many time does not have the same weight (influence degree) than a tweet that have not been retweeted or liked by followers. A retweet is a way to support an opinion or share a post that users consider interesting. It could be done in two ways: through Retweet button or in the old way by writing RT, the user name and the post to retweet. In our case, we consider only retweet by button because the retweet in old way could be different from the original tweet. To measure the influence degree of a tweet $Inf(t)$, we add other metrics that balance the weight of a tweet in formula (2). We define these metrics as the number of retweets and the number of likes of a given tweet $(N_R(t), N_L(t) respectively)$.

$$Inf(t) = weight(t) * (N_R(t) + N_L(t)) \tag{2}$$

  $weight(t)$ is the weight of the class to which belongs.
- *Step 3* is the final step, in which we calculate the overall rating for each Y_i: $(Rate(Y_i))$. First, we calculate the influence degree for each tweet related to Y_i. $Rate(Y_i)$ is determined by both, the sum of $Inf(t_{Y_i})$ and the total volume tweets shared about $Y\ i$ as following:

$$Rate(Y_i) = \frac{\sum_{k=1}^{q} Inf(t_{Y_i})}{q} \tag{3}$$

  $Inf(t_{Y_i})$ positive tweets regardingY_i. q is the number of positive tweets related to Y_i. We consider the largest $Rate(Y_i)$ as the likelihood of a Y_i to be the most appreciated by internet users.

In what follow, we apply our proposed approach to the US presidential elections in order to show its robustness.

Methods and experience: US presidential elections

Social media, such as Facebook and Twitter, is increasingly used in the presidential elections by both, candidates to further their campaigns, to share their programs and communicate with people, and voters to express their opinion about each candidate.

In our case, we limit our study to Twitter, which is a micro-blogging service that allows users to post and interact with messages called "tweets", restricted to 140 characters. This information is widely accessible to developers and researchers due to REST API that allows to collect published tweets about a given hashtag.

Implementation

Big data analytics systems should have enough memory, bandwidth and the capacity to perform parallel processing of tasks in real-time. Therefore, we implemented the prototype system over a cluster of 3 servers. Each of these 3 servers has two Intel Xeon E5530 quad-core CPU 2.4 GHz processors that run on 64-bit Linux Ubuntu. Also, the servers are equipped with 24 Go DDR3 RAM and 1 TB hard disk.

For data gathering, we used apache Kafka which is a distributed streaming platform. Kafka uses publish-subscribe messaging and offer a distributed and replicated service. More specifically, we used the integrated Streams API library that allows building applications for stream processing. Then, the large volume of collected data is stored in HDFS (Hadoop Distributed File System).

For data processing (building dictionaries and data classification), Spark is employed. Spark allows scalable and fault-tolerant processing of data streams in near real time. In our case, Spark is used on Hadoop, managed by YARN and distribute processing across 3 nodes. Also, in order to compare the proposed method with other methods, we use Spark MLlib's implementation of Naïve Bayes classifier? for classifying the tweets in real-time. The model created by Naïve Bayes is applied in real-time to retrieved tweets.

All the algorithm for data gathering and processing were implemented using Python.

Data gathering

- *First stage* To focus on 2016 US presidential election, we only used two candidates' names: Hillary and Trump as keywords to retrieve tweets (i.e $D = \{Y_1, Y_2\}$ with $Y_1 =$ Trump and $Y_2 =$ Hillary). Data is collected from Twitter REST API in real time and tweets should be written in English. In order to identify the most popular hashtags, we generated a hashtag frequency list for the entire collected data. We selected most relevant hashtags from this word frequency list, then, we classified them manually as following: Positive Trump: #MAGA, #makeamerikagreatagain and #voteTrump Negative Trump: #NeverTrump #dumptrump Positive Hillary: #imwithher, #strongertogether and #voteHillary Negative Hillary: #podestaemails, #NeverHillary and #crokeedHillary. In order to validate the construction method of dynamic dictionary, we collected data related to chosen hashtags over 5 and 6 October 2016. Prototype data contains a total of 120,000 tweets (divided into 30,000 tweets for positive and negative classes)

- *Second stage* In order to predict the election outcomes, we collected all Twitter messages posted over 6 and 7 November 2016 that contains a total of 3,600,000 tweets for both Trump and Hillary.
- *Third stage* Here, we used the result of processed data in the second stage as well as tweets' information (i.e. the class, the weight of the class to which the tweet belongs, number of retweets and likes).

Data processing

- *First stage* Here, we constructed positive and negative dictionaries for the both "subject": Hillary and Trump. For this, we processed collected data related to chosen hashtags as mentioned in "First stage: constructing dictionaries" section.
- *Second stage* First, we preprocessed data by removing duplicated tweets, stop words and applying filters as mentioned in step 2, "Second stage: classification" section. Then, we classified tweets under seven classes and compared the performance for the three methods: Using $p(t)$ (the proposed method), Naïve Bayes method and Google prediction API. For instance, the original tweet in Tables 2, 3.
- *Third stage* First, we calculated the influence degree for collected data. This processing applied to the previous example in Table 4.

Then, we calculated and compared them in order to find the most appreciated candidate and predict the presidential election outcomes.

Table 2 Tweet example

Date	Retweet	Like	Tweet
06/11/2016 23:05:00	4	12	#ImWithHer #Hillary #Stronger-Together I SUPPORT her #DumpTrump

Table 3 Applying the second stage

Step	Action	Output
Step 2	Preprocessing	#ImWithHer #Hillary #Stronger-Together intensifier(support) #DumpTrump
Step 3	Balancing	Support
	Scoring	4
Step 4	Classification	C_{+2}

Table 4 Applying the third stage

Step	Output	Value
Step1	*Weight*	2
Step2	*Inf*(*t*)	$2 * (4 + 12) = 32$

In the following section, we evaluate the results of our proposed approach.

Results

In order to validate our approach, we first evaluated the post classification performance. Then, we compared it with other sentiment analysis tools such as Ibn Waston text analytics. Finally, we investigated whether it is possible to predict correctly the winner of 2016 US election.

Classification accuracy evaluation

To assess the ability of classifying tweets basing on the automatically constructing dynamic dictionary, we have randomly selected a subset of 600 tweets from the political Twitter corpora: 50 for each class. The tweets were carefully inspected and manually labeled as positive, moderately positive, highly positive, lightly negative, moderately negative, strongly negative or neutral for each candidate. Then, the same data was processed, as mentioned above, by removing stop words, applying tokenization, stemming and various filters. This step was done by TreeTagger, which is a tool for annotating text with part-of-speech and lemma information. TreeTagger was modified to handle negation, URLs, usernames, Twitter mentions and hashtags and intensifiers.

In order to test the proposed approach, we compared it with two following classifiers approaches:

- *Google cloud prediction API* that provides a RESTful API as a black box to build ML models basing on a training dataset. Prediction's cloud-based ML tools analyze data in several domains such as customer sentiment analysis, recommendation systems and spam detection.
- *Naïve Bayes* is simplest and most commonly used classifier. The model computes the posterior probability of a class, based on bag of words, and uses Bayes Theorem to predict the probability that a given feature set belongs to a particular class. In our case, the model classifies tweets into positive and negative sub-classes.

Four commonly used metrics accuracy, precision, recall, and F-score are for evaluating the performance of a classification method are used. The performance of classifiers (Table 5) is calculated by taking the average of the four metrics for each class of both candidates.

While considering accuracy and macro F-measure, it is observed that classification using our proposed method achieves a good accuracy (90.21, 89.98% respectively) against Naïve Bayes and Google prediction API.

Comparison with other sentiment analysis tools

In Table 6, we establish a comparison, basing on various metrics, between the proposed approach and other sentiment analysis tools such as IBM Watson text analytics, Rapidminer, Meaning cloud and StreamCrab.

Table 5 Classification accuracy

	Proposed method					Naïve Bayes					Google Prediction API			
	Accuracy	Mean Accuracy	Macro Precision	Macro Recall	Macro F-measure	Accuracy	Mean Accuracy	Macro Precision	Macro Recall	Macro F-measure	Mean Accuracy	Macro Precision	Macro Recall	Macro F-measure
Highly positive Trump	88.52%					85.71%								
Moderately positive Trump	88%					35.71%								
Lightly positive trump	88%					14.21%								
Highly positive Hillary	98%					21.42%								
Moderately positive Hillary	86%					14.28%								
Lightly positive Hillary	86%	90.21%				57.14%	34.45%	33.92%			66.66%			
Highly negative Trump	90%					0.70%								
Moderately negative Trump	96%					57.14%								
Lightly negative Trump	88%					42.28%								
Highly negative Hillary	98%					14.28%								
Moderately negative Hillary	88%					14.28%								
Lightly negative Hillary	88%					57.14%								

Table 6 Comparison with other sentiment analysis tools

	The proposed approach	Rapidminer	IBM Watson	Meaning cloud	StreamCrab
Automatically constructed dictionary	Yes	No	No	No	No
Context-based dictionary	Yes	No/customizable	No/customizable	No	No
Classification degree	(Highly, moderately, lightly) positive, negative and neutral	Positive, negative and neutral/customizable	Positive, negative and neutral	Positive, negative and neutral	Positive and negative
Intensifier	Yes	No	No	No	No
Big data tools	Yes	No (integrable with Hadoop)	Yes	No	No
Visualization	No	Yes	Yes	Yes	Yes (limited)

Prediction accuracy evaluation

We applied the last stage of the proposed approach on the collected data (posted over 6 and 7 November 2016 for both Trump and Hillary) to find the most appreciated candidate. The results are shown in Table 7.

Conclusion and discussion

Sentiment analysis or opinion polarity has been proven to be effective in predicting people attitude by analyzing big social data. In this contribution, we present a novel adaptable approach that aims to extract people opinion about a specific subject by relying on social media contents. The proposed technique consists to first building a dictionary of words' polarity based on a very small set of positive and negative hashtags related to a given subject, then, classifying posts into several classes and balancing the sentiment weight by using new metrics such as uppercase words and the repetition of more than two consecutive letter in a word. In order to test this model, a case study has been conducted for the 2016 US presidential election to go through our model step by step to guess which of candidates was the favorite. The performance results have shown some promising results compared to our model.

However, the proposed approach still suffers from some shortcomings. First, it does not distinguish the impact degree of the different metrics in order to accentuate a feeling. Second, we used only Twitter data. Third, the system is a prototype designed to assess the ability of automatically constructing dynamic dictionary using small samples. As future work, we intent to tackle these three limitations by proposing a more global and efficient model using larger volumes of data.

Table 7 US Presidential election result

	Donald trump	Hillary clinton
$Rate(Y_i)$	30.85	18.34

Abbreviations
OV: opinion and volume; ML: machine learning.

Authors' contributions
IE performed the primary literature review, experiments, proposed the method and also drafted the manuscript. YG supervised the programming of the application, wrote a part of the manuscript. YC worked with IE to develop the application. YG, RM, AT and AK provided reviews on the manuscript. All authors read and approved the final manuscript.

Author details
[1] Laboratoire des Systèmes de Télécommunications et Ingénierie de la Décision, University of Ibn Tofail, Kenitra, Morocco. [2] Laboratoire Angevin de Recherche en Ingénierie des Systèmes, University of Angers, Angers, France. [3] LGS, Ecole Nationale des Sciences Appliquéees, University of Ibn Tofail, Kenitra, Morocco.

Acknowledgements
A particular acknowledgement for the scientific and the editorial committee.

Competing interests
The authors declares that they have no competing interests.

Availability of data and materials
The data sets are available in (https://osf.io/kvxqs).

Ethics approval and consent to participate
Not applicable.

Funding
None.

Publisher's Note

Received: 19 December 2017 Accepted: 24 February 2018
Published online: 08 March 2018

References
1. Balasubramanyan R, Routledge BR, Smith NA. From tweets to polls: linking text sentiment to public opinion time series. Icwsm. 2010;11:1–2.
2. Benamara F, Cesarano C, Picariello A, Recupero DR, Subrahmanian VS. Sentiment analysis: adjectives and adverbs are better than adjectives alone. In: Proceedings of ICWSM conference. 2007.
3. Bermingham A, Smeaton A. On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the workshop on sentiment analysis where AI meets psychology. 2011.
4. Bhatt R, Chaoji V, Parekh R. Predicting product adoption in large-scale social networks. In: Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM; 2010. p. 1039–48.
5. Chesley P, Vincent B, Xu L, Srihari RK. Using verbs and adjectives to automatically classify blog sentiment. In: AAAI symposium on computational approaches to analysing weblogs (AAAI-CAAW). 2006. p. 27–9.
6. Conover MD, Goncalves B, Ratkiewicz J, Flammini A, Menczer F. Predicting the political alignment of twitter users. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. 2011. p. 192–9.
7. De Choudhury M. Predicting depression via social media. ICWSM. 2013;13:1.
8. Delenn C, Jessica Z, Zappone A. Analyzing Twitter sentiment of the 2016 presidential candidates. Stanford: Stanford University; 2016.
9. DiGrazia J, McKelvey K, Bollen J, Rojas F. More tweets, more votes: social media as a quantitative indicator of political behavior. PLOS ONE. 2013;8(11):e79449.
10. Ekaterina O, Jukka TO, Hannu K. Conceptualizing big social data. J Big Data. 2017;4:3.
11. Finn S, Mustafaraj E, Metaxas PT. The co-retweeted network and its applications for measuring the perceived political polarization. Faculty Research and Scholarship. 2014.
12. Gayo-Avello D. No, you cannot predict elections with Twitter. IEEE Internet Comput. 2012;16(6):91–4.
13. Hansen LK, Arvidsson A, Nielsen FA, Colleoni E, Etter M. Good friends, bad news-affect and virality in twitter. In: Future information technology, communications in computer and information science. Berlin: Springer; 2011. p. 34–43. https://doi.org/10.1007/978-3-642-22309-9_5.
14. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD'04. New York: ACM; 2004. p. 168–77.
15. Jahanbakhsh K, Moon Y. The predictive power of social media: on the predictability of US presidential elections using Twitter. arXiv:1407.0622 [physics]. 2014.
16. Jose R, Chooralil VS. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach. In: 2016 international conference on data mining and advanced computing (SAPIENCE). 2016. p. 64–7.
17. Khatua A, Khatua A, Ghosh K, Chaki N. Can #Twitter_trends predict election results? Evidence from 2014 Indian general election. In: 2015 48th Hawaii international conference on system sciences. 2015. p. 1676–85.

18. Livne A, Simmons M, Adar E, Adamic L. The party is over here: structure and content in the 2010 election. In: Fifth international AAAI conference on weblogs and social media. 2011.
19. Mahmood T, Iqbal T, Amin F, Lohanna W, Mustafa A. Mining Twitter big data to predict 2013 Pakistan election winner. In: INMIC. 2013. p. 49–54.
20. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5(4):1093–113.
21. Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, vol. 10. Stroudsburg: EMNLP'02, Association for Computational Linguistics; 2002. p. 79–86.
22. Pääkkönen P. Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream-based processing. J Big Data. 2016;3:6. https://doi.org/10.1186/s40537-016-0041-8.
23. Ramanathan V, Meyyappan T. Survey of text mining. In: International conference on technology and business and management. 2013. p. 508–14.
24. Ramteke J, Shah S, Godhia D, Shaikh A. Election result prediction using Twitter sentiment analysis. In: 2016 international conference on inventive computation technologies (ICICT), vol. 1. 2016. p. 1–5.
25. Razzaq MA, Qamar AM, Bilal HSM. Prediction and analysis of Pakistan election 2013 based on sentiment analysis. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014). 2014. p. 700–3.
26. Ruths D, Pfeffer J. Social media for large studies of behavior. Science. 2014;346(6213):1063–4.
27. Shi L, Agarwal N, Agrawal A, Garg R, Spoelstra J. Predicting US primary elections with Twitter. Stanford: Stanford University; 2012.
28. Smailović J, Kranjc J, Grčar M, Žnidaršič M, Mozetič I. Monitoring the Twitter sentiment during the Bulgarian elections. In: 2015 IEEE international conference on data science and advanced analytics (DSAA). 2015. p. 1–10.
29. Soler JM, Cuartero F, Roblizo M. Twitter as a tool for predicting elections results. In: 2012 IEEE/ACM international conference on advances in social networks analysis and mining. 2012. p. 1194–200.
30. Speriosu M, Sudan N, Upadhyay S, Baldridge J. Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the first workshop on unsupervised learning in NLP, EMNLP'11. Stroudsburg: Association for Computational Linguistics. p. 53–63.
31. Stavrianou A, Brun C, Silander T, Roux C. NLP-based feature extraction for automated tweet classification. In: Proceedings of the 1st international conference on interactions between data mining and natural language processing, vol. 1202, DMNLP'14. Aachen: CEUR-WS.org; 2011. p. 145–146.
32. Tumasjan A. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social media. 2010.
33. Tumitan D, Becker K. Sentiment-based features for predicting election polls: a case study on the Brazilian scenario. In: 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT), vol. 2. 2014. p. 126–33.
34. Tunggawan E, Soelistio YE. And the winner is...: Bayesian Twitter-based prediction on 2016 US presidential election. arXiv:1611.00440 [cs]. 2016.
35. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle. In: Proceedings of the ACL 2012 system demonstrations, ACL'12. Stroudsburg: Association for Computational Linguistics; 2012. p. 115–20.
36. Wang H, Castanon JA. Sentiment expression via emoticons on social media. In: 2015 IEEE international conference on Big Data (Big Data). 2015. p. 2404–8.
37. Wicaksono AJ, Suyoto P. A proposed method for predicting US presidential election by analyzing sentiment in social media. In: 2016 2nd international conference on science in information technology (ICSITech). 2016. p. 276–80.
38. Wong FMF, Tan CW, Sen S, Chiang M. Quantifying political leaning from tweets, retweets, and retweeters. IEEE Trans Knowl Data Eng. 2016;28(8):2158–72.
39. Xie Z, Liu G, Wu J, Wang L, Liu C. Wisdom of fusion: prediction of 2016 Taiwan election with heterogeneous big data. In: 2016 13th international conference on service systems and service management (ICSSSM). 2016. p. 1–6.
40. Xing F, Justin ZP. Sentiment analysis using product review data. J Big Data. 2015;2:5.
41. Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on empirical methods in natural language processing, EMNLP'03. Stroudsburg: Association for Computational Linguistics; 2003. p. 129–36.