Generating a Highlight Moments Summary Video of Apolitical Event using Ontological Analysis on Social Media Speech Sentiment

Abid Mehdi[1]
Laboratoire Génie Industriel, Traitement de l'Information et
Logistique
Faculty of Sciences Ain Chock, Hassan II University.
Casablanca Morocco

Yassine Serhane[3]
Laboratoire Recherche et Innovation Informatique Faculté
des Sciences Ain Chock
Faculty of Sciences Ain Chock, Hassan II University
Casablanca Morocco

Benayad Nsiri[2]
Research Center STIS, M2CS,
Higher School of Technical Education of Rabat (ENSET),
Mohammed V University
Rabat, Morocco

Miyara Mounia[4]
Laboratoire: Informatique & Aide à la Décision Faculté des
Sciences Ain Chock
Faculty of Sciences Ain Chock, Hassan II University
Casablanca Morocco

Abstract—**Numerous viewers choose to watch political or presidential debates highlights via TV or internet, rather than seeing the whole debate nowadays, which requires a lot of time. However, the task of making a debate summary, which can be considered neutral and does not give out a negative nor a positive image of the speaker, has never been an easy one, due to personal or political beliefs bias of the video maker. This study came up with a solution that generates highlights of a political event, based on twitter social network flow. Twitter streaming API is used to detect an event's tweets stream using specific hashtags, and detect on a timescale the extreme changes of volume of tweets, which will determine the highlight moments of our video summary at first, then a process is set up based on a group of ontologies that analyze each tweet of these moments to calculate the percentage of each sentiment's positivity, then classify those moments by category (positive, negative or neutral).**

Keywords—Debate summary; API; hashtags; twitter; highlights moment; ontologies; sentiment analysis

I. INTRODUCTION

In the 2017 Republicans primaries, CNN claimed that more than 84 million people have watched the republican candidates debating on its channel, breaking records for most events seen on CNN. FOX also cited that more than 83 million have seen the debate between the republican candidates, which made it the most watched event in the history of television. The majority of these audiences are social media users, who respond to every controversial moment [1] on various platforms in real time, such as Twitter, Facebook, Snap, Instagram.

The study uses Twitter as the main audience feedback source [2], [3], because of its worldwide use (Fig. 1), and people use it more than other social platforms to express their immediate feelings and opinions.

Several studies gave an interesting insights about the social network twitter evolution due to his dynamic nature with more than 400 million tweets posted everyday [4], using the hashtags (Significant continuation of characters without space beginning with the sign #, Which refers to a subject and inserted into a message by its author, in order to facilitate the location) helps to look for trending topics and look up thousands of tweet (Table 1).

Twitter users usually respond to political speaker statements or point of views during a political speech, which offers a fertile ground for sentiment analysis [5], due to the outrageous tweets against the opposite political speaker or the encouraging tweets from their supporters, those tweets usually come as a reaction to the big (Good or Bad) moments of the speech, which makes their reactions a good highlights' indicator for the event

In this article, the volume of these tweets was used in a preset amount of time as an indicator of an event highlight, gathering those highlights to wind up with a video summary generated only using a random sample of tweets which grant our summary the neutrality and avoid unwanted bias.

In the next chapter the KDD (knowledge discovery in databases) process will be discussed. This approach utilizes these tweets to score the sentiment's positivity percentage in them, in order to classify these tweets into positive, negative or neutral and able to determine the nature of the moment.

TABLE I. TABLEAU 1: NUMBER OF QUARTERLY ACTIVE TWITTER USERS IN MILLIONS

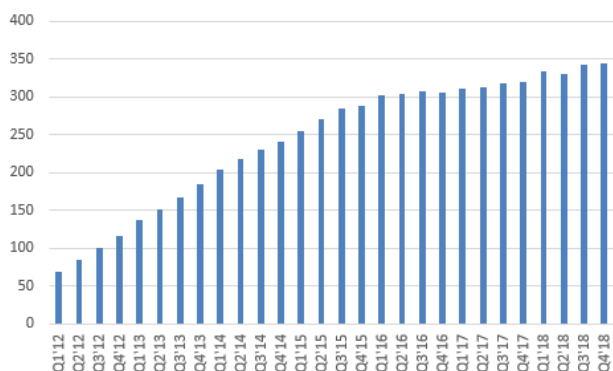| Topics (Hashtag)	# Tweets	Time span
Black Friday	1 085 365	2018.11.01 -- 2018.11.01
Trump	410 854	2018.11.01 – 2018.11.30
Iphone X	98 716	2018.06.30 – 2018.07.31
FC barcelona	302 523	2018.02.01 – 2018.02.28

Fig. 1.   Number of Quarterly Active Twitters users in Million.

II. Background

The construction of a summary video, which generates the highlight moments of a political, social or cultural event, is usually based on the image processing of the event, this process is basically established on objects detection [6], [7]. This detection of objects in the case of a speech, can be based on the change of the camera angle from speaker towards public, whenever the audience applauds or boo to show their disapproval or start chatting about controversial statements. However, this approach becomes delicate in cases where certain events or rallies take place where the spectators are seated behind the speaker.

Moreover, other studies rely on approaches that are based on the exploitation of audio by detecting any sudden variation of sound recording [8]–[10] caused by audience reaction, such as applauds or shouts out, this variation reflect the highlight moment during the event (Fig. 2), but unfortunately this approach can have several inconveniences such as special effects added during the event, or even the presence of noise during the whole event.

Due to the inconveniences of objects detection and audio recording analysis approaches to detect the highlight moments of a specific event, and the evolution of social media use, this massive quantity of data generated from these social media platforms especially twitter can be utilized, to generate a summary of this event. Similarly, several research works, such as predicting a movie success based on the reaction tweets from the trailer watching [11], and the prediction of the presidential elections established in several countries such as the USA, France and Pakistan [13],[14],[15].

Those approaches have shown a major success in their predictions, which proves the credibility of using the social network twitter as a source of information to figure out the public tendency.

Fig. 2.   Highlight Moment Detection based on Sound Frequency.

Including, the research [12] which use Twitter's data to predict the results of the Pakistan's elections in 2013, thanks to a model of classification developed in Machines Learning by using learning algorithms, in order to classify the tweets into two categories positive (Pro) or negative (Anti), through the sentiment analysis of every collected tweet, this classification is based on the contents of tweets (Hashtags and key word) e.g. The use of capital letters which means a person is shouting, words, emoticons etc. and then a comparison is done by attributing every tweet to the appropriate presidential candidate.

Furthermore, other studies have also been based on the sentiment analysis process of the tweets, i.e. due to the feelings polarization of their spectators during a soccer match [15], which can be identified thanks to the use of the standardized hashtag or the one made or official by their team. This approach creates a framework that handles various reactions from numerous Twitter users during a soccer match [16], and showed as expected positive results, the tweets from users are positive when their team scored a goal and negative if they concede one.

In addition, some researches were developed on fans swearing in tweets, while watching a soccer match and how they used it as a sentiment marker [15]. Their work concentrate heavily on the context of the tweet rather than the swearing itself, because not all swearing tweets reflect negative sentiment. They started by collecting tweets in relation with the English Premier League matches, then they linked these tweets to teams based on how many times a fan tweeted using his team hashtag the most, after that, they filtered these tweets by use of swearing, taking into consideration complication like fans using their opponent hashtags to get their attention. They conclude their work by showing that bad language is not always negative and some of the strongest sentiments expressed are self-critical.

Most of the studies described previously, have used in their approach various methods of data mining, such as KDD process [17], [18], which is used widely in the research field, or using process intended for the professional area such as CRISP-DM (Cross Industry Standard Process for Data Mining) which is considered as an iterative process, and strongly used to satisfy the industrial needs (Domain of engineering, medicine, sales and marketing).

In our study, we will be using the KDD process, because it is complete, precise and answers our needs, which is the search for the knowledge in big data.

Knowledge Discovery and Data Mining is a process that allows the extraction of the different information out of the massive data according to a predefined goal, in order to find oneself with a useful knowledge [19], [20] (Fig. 3).

Fig. 3.   KDD Processes.

This process is composed of 5 main steps (Selection, Pretreatment, Transformation, Data mining and Interpretation) [21].

- Selection: Consists of collecting and choosing the data which results in aggregating a variety of sources into a single target data.

- Pre-processing: Stage that contains the removal of noise and handles the missing values into clean target data.

- Transformation: In this phase, every data is transformed through the reduction of the database dimensions, and the transformation of the attributes, to wind up with a database that meets the requirements of our project objectives.

- Data Mining: This stage consists of choosing and adapting the algorithms of data mining, based on intelligent methods in order to extract data patterns.

- Interpretation: is the final stage of this process, which includes the evaluation and the interpretation of the patterns discovered in order to determine the useful information.

III. METHOD

The moment there is a broadcasted political event live on television, users begin to tweet about it using related Hashtags, in order to share their opinion and symbolize them in relation with the theme of this event.

Thanks to Twitter's streaming API, the contents was recovered as well as the volume of tweets by their Hashtags in real time via a request sent to the twitter's servers, which allows to obtain a stream of data $\{(x_i, y_i), i = 1,..., n\}$; Taking into example two features of the data it can be represented in the form of a cloud of points of data in a (x, y) plan (Fig. 4), where the x-axis represents speech time interval and the y-axis represents the number of tweets.

This research purpose is a summary video generation based on the highlight moment detection of an event and the analysis of the sentiment of these moments tweets, i.e. The detection of tweets volume extreme changes on a timescale at first, and then analyze the sentiment of each tweet belonging to the highlight moments in order to measure the percentage of its positivity.

To achieve that, the computing of a function f, that would allow to reflect the partner of points $(x_i, f(x_i))$ obtained on a graph, remains indispensable even though it was not explicitly known.

However, the mathematical approach used is the optimization of Lagrange polynomial obtained from the plot described previously [22], the existence of this polynomial is asserted by the following theorem:

There is a unique polynomial $p_n \in \mathbb{R}_n[x]$, ($\mathbb{R}_n[x]$ being the vector space of polynomials which degree is lower or equal to n) such as:

$$p_n(x_i) = f(x_i) \ \forall i \in \{1,2,...,n\}$$

and P_n is given by Lagrange formula:

$$P_n(x) = \sum_{i=0}^{n} f_i l_i(x)$$

with

$$f_i := f(x_i) \ \text{et} \ l_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \ , i \in \{1,2,...,n\}$$

P_n is called the polynomial of interpolation of Lagrange in points x_i for the measures f_i.

The theorem above allows to create a polynomial function passing by all the points obtained, e.g. as represented in Fig. 5.

However, the peaks of this polynomial function which varies according to the change of the tweets volume in a predefined period lead to the detection of the highlight moments. The latest can be determined through the spikes, which are the local maximums of the polynomial function p_n i.e. Points that satisfy the optimality conditions:

- Condition 1 (Stationarity)

$$\frac{\partial p_n}{\partial x}(x_i) = 0$$

- Condition 2

$$\frac{\partial^2 p_n}{\partial x^2}(x_i) < 0$$

First, the method of steepest descent for the stationary points of f shall be used, after that a simple selection of the points whose second order derivative are positive will take us to our objective (Peaks Detection) (Fig. 6).
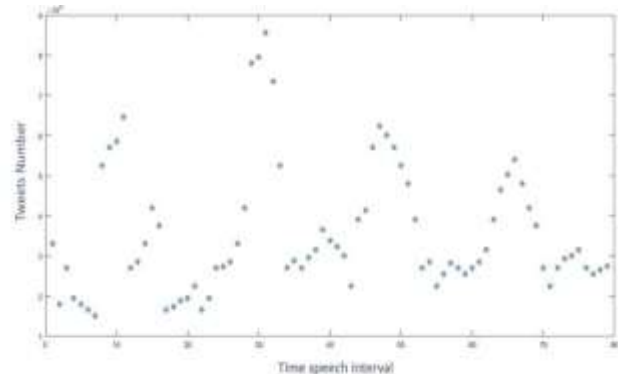

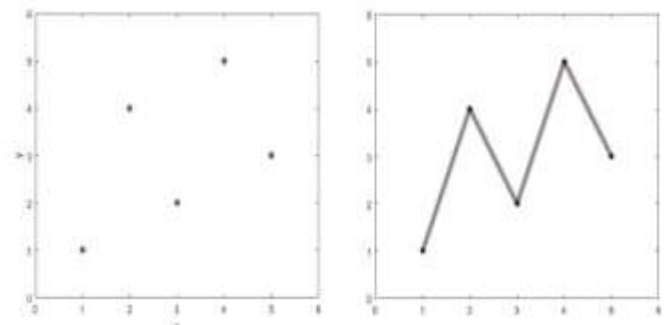Fig. 4. Tweet Data Volume Represented as a Scatter Graph.

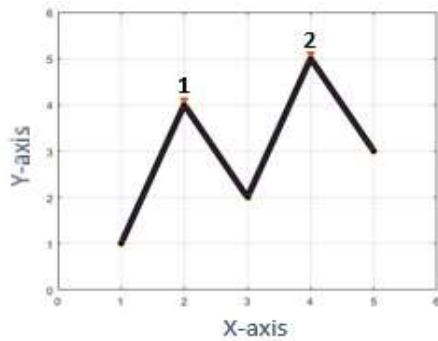Fig. 5. Lagrange Polynomial for the Interpolation.

Fig. 6.   Peaks Detection.

Once the highlight moments is defined by the generating the polynomial function and by detecting its peaks, the nature of these highlight moments was discovered by applying the sentiment analysis process on each tweets that belong to every peak, in order to measure the percentage of tweeter users' sentiment positivity toward the speaker.

The appropriate process of sentiment analysis comes down to developing a process that allows a classification of published tweet's sentiment, where data extracted from Twitter is analyzed in a granular way, by decomposing sentences into a group of words linked to a global ontology that includes various types of terminology. The aim of the sentiment analysis process is the ability to analyze a sentence and to measure the percentages of its positivity (Positive, negative and neutral) (Fig. 7). However the use of the ontologies in the analysis will have numerous advantages, in particular with regards to the cultural, linguistic and regional expressions... [23].

The global ontology used above allows to regroup different local ontologies, which describe their own local knowledge space in relation to a precise specification of each word or sentence category (positive, negative or neutral), in other words, each local ontology contains words and sentences that used to categorize the tweet components, at this level, this

those local ontology becomes a class that belongs to the global ontology. The process of the data creation or modification, within a local ontology, is based on a specific life cycle, which starts from draft mode to the published mode (Fig. 8).

In draft mode, system users can create or edit sentences or words samples, come back to them, save them and continue to work on them until they're ready to be submitted. Once the sample is submitted, it goes into the understood approved folder, where an ontology's local manager would review it.

During the review process, the sample can be either rejected, which would then put it back into the draft, or approved, in this case the sample is published.

At same point some sentences or words samples need to be updated and transferred into another local ontology, due to their meanings or their semantic change.

When the sample is selected to be revised, it goes back into the submitted stage, where the reviewer (manager) can either, once again, reject it or approve it to be revised. In case it's rejected, it goes all the way back to draft mode and starts the process all over again.

In conclusion, the work ends with measuring the sentiment percentage of each peak, that reflects the highlight moment, of tweets volume tweeted by a group of people in a specific moment. In the first stage, the calculation of each tweet sentiment percentage, which participated in generating this highlight, is done by measuring each sentiment category percentage, by using the process that allows the calculation of the sentiment classification percentage via the use of ontologies.

In the Second stage, to make the decision about the analyzed sentiment category of the highlight moment that was generated by the peak of tweets volume, the calculated average sentiment percentage after merging the sentiment classification of each tweet into three major sentiment categories. After that the sentiment with the maximum percentage to that highlight moment was assigned (Fig. 9).
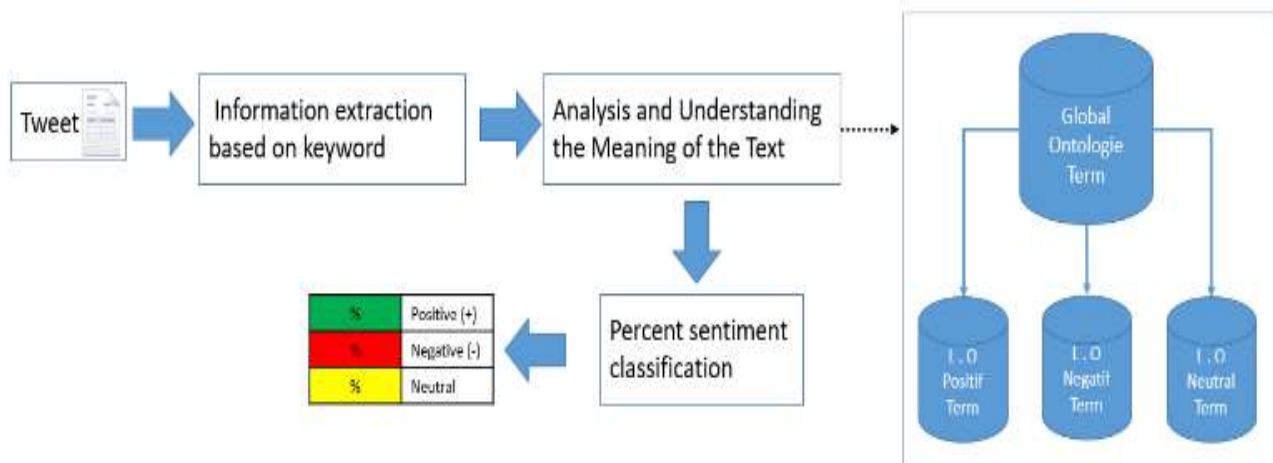

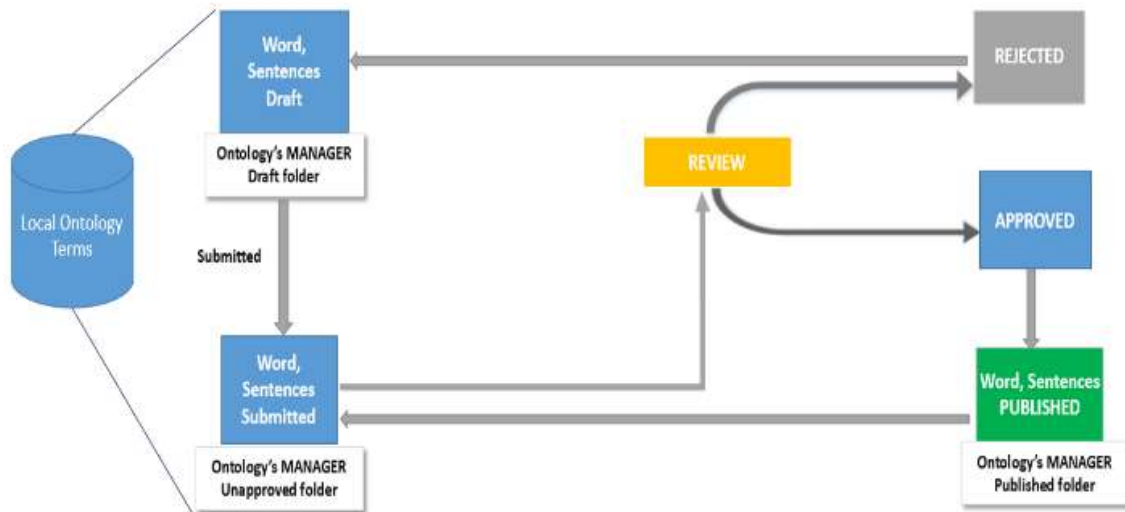Fig. 7.   Process of Classification Percentage Sentiment.

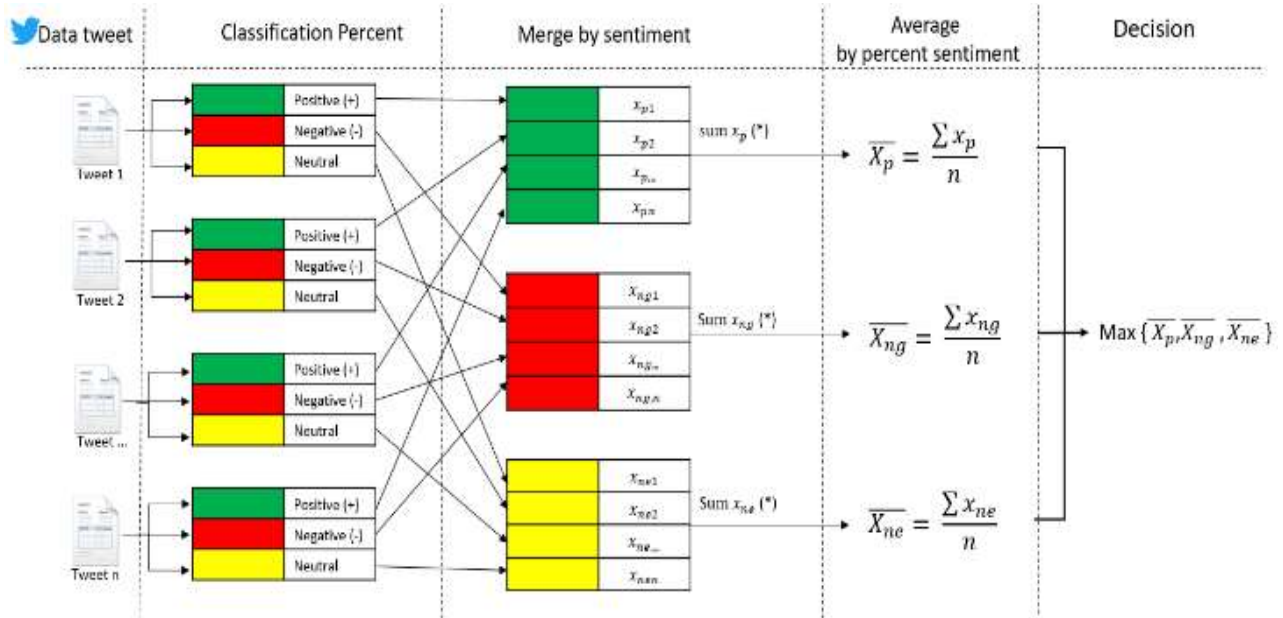Fig. 8. Process of Classification Percentage Sentiment.

Fig. 9. Highlight Moment Positivity Decision by the Calculating and Merging Process.

IV. IMPLEMENTATION

To realize our objectives, which are to generate a highlight moments summary video of a live broadcasted event, and to calculate the sentiment percentage by category of each one of the highlight moments, we used the Twitter Streaming API that allows us to query Tweeter databases and get only the tweets data in regard of a specific Hashtag in real time and which were generated in an exhaustive way, those hashtags have in general a relation with our speaker official account, such as #Donaldtrump, #Gop, #Maga, #Trump, #TinyTrump Furthermore, thanks to LaGrange mathematical approach which has been presented in the previous chapter, we project the collected data from the twitter streaming API as a polynomial function in terms of time speech (Fig. 10), in order to detect its local maximums (spike). The obtained peaks can

be considered as the highlight moments detection key of our video summary.

Certainly, one of the work main objectives is the capacity to analyze the sentiment of every highlight moment's tweets and this by measuring the positivity rate for each one of them by category (Positive, Negative and neutral). After using the KDD process that provided us with useful information of the big data recovered previously [24].

The measure of this sentiment positivity can arise many challenges, due to many obstacles e.g. linguistic, cultural, regional expressions, etc.

To cope with these challenges, we came up with a reliable approach that uses ontologies, which turns out to be reliable and robust at resolving the semantic problems of the sentence or the group of word that composed the tweets.
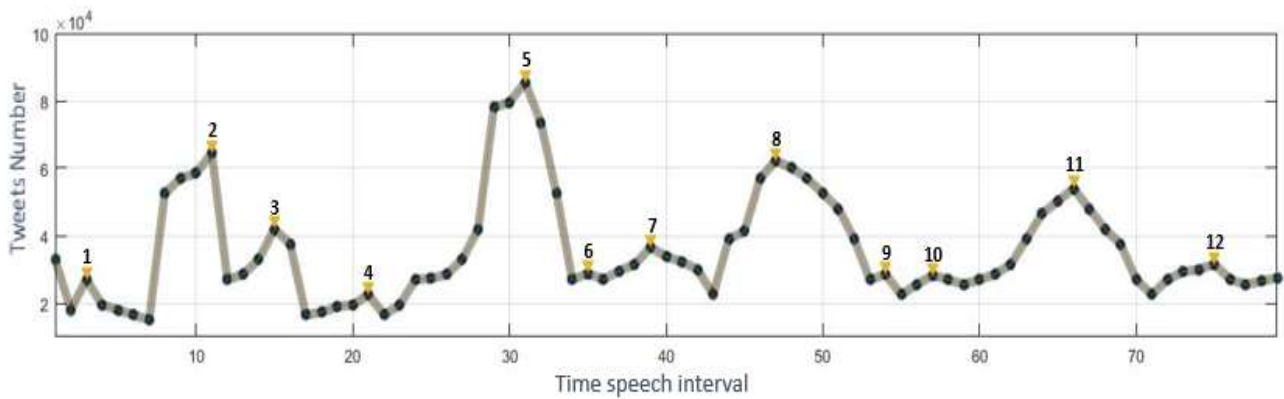
Fig. 10. Highlights Detection in Political Speech by Detection the Locals Maximums of the Obtain Polynomial Function that Reflect the Volume of Tweets During a Political Event.

To improve the interpretation of the sentiment analysis regarding the tweets data extracted at the semantic level, we created a global ontology that can be defined as a warehouse of generic knowledge; this global ontology is based on three types of local ontologies. These local ontologies are kept up to date regularly by adding, modifying or removing regularly the expressions or set of words, according to its category (Positive, Negative or Neutral sentiment).

The measurement of the sentiment positivity percentage of a tweet is based on the existence of the expression patterns that constitute the processed tweet, as well as their rate of occurrence within every local ontology (Fig. 11). Likewise, for a global sentiment classification of a single specific highlight moment. A simplified process was established, and this by merging all together each sentiment category percentage of each tweet that generates the highlight moment detected and by assigning the sentiment with the maximum percentage to the highlight moment global sentiment category (Fig. 12).

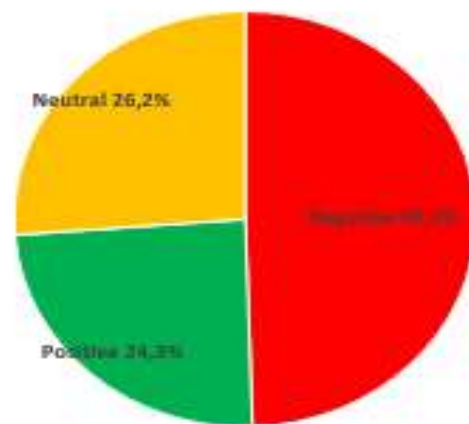Fig. 11. Percentage Measure of Tweeter users' Sentiment Positivity.

Fig. 12. Sentiments Classification Percent of the Fifth Highlight Moment.

V. CONCLUSION

In this article, a study was established on the generation of a video summary of an event, based on highlight moment detection using tweets volume changes, furthermore, a set-up of a process that allows measuring the sentiment positivity percentage of the tweets of these highlight moments, then classifying those tweets by category (Positive, Negative and Neutral) to wind up by classifying each highlight moment by category (Positive, Negative and Neutral) after merging the percentage of each tweet that composed that moment. From the results obtained, it was concluded that our proposed approach can play an important role on the detection of the citizen's sentiment in response to the speaker, which can open up a new perspective that will facilitate the voters to better choose their presidential candidate during an event of a future election and not rely on media.

REFERENCES

[1] J. E. M. de Oliveira, M. Cotacallapa, W. Seron, R. D. C. dos Santos, and M. G. Quiles, "Sentiment and Behavior Analysis of One Controversial American Individual on Twitter," Neural Inf. Process. -Springer Cham, pp. 509–518, Oct. 2016.

[2] C. Paris, H. Christensen, P. Batterham, and B. O'Dea, "Exploring Emotions in Social Media," IEEE Conf. Collab. Internet Comput., pp. 54–61, Oct. 2015.

[3] E. Lahuerta-Otero and R. Cordero-Gutiérrez, "Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter," Comput. Hum. Behav., vol. 64, pp. 575–583, Nov. 2016.

[4] F. Kalloubi, E. H. Nfaoui, and O. El Beqqali, "Graph based tweet entity linking using DBpedia," in Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on, 2014, pp. 501–506.

[5] M. Lai, C. Bosco, V. Patti, and D. Virone, "Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization," in Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, 2015, pp. 1–9.

[6] B. Vijayalaxmi, R. Putta, G. Shinde, and P. Lohani, "Object Detection Using Image Processing for an Industrial Robot," International Journal of Advanced Computational Engineering and Networking, 2013.

[7] J. Komala Lakshmi and M.Punithavalli, "A Survey on Performance Evaluation of Object Detection Techniques in Digital Image Processing," IJCSI Int. J. Comput. Scie Nce Issues, vol. 7, no. 6, 2010.

[8] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," in Proceedings of the Eighth ACM International Conference on Multimedia, New York, NY, USA, 2000, pp. 105–115.

[9] P. Suksai and P. Ratanaworabhan, "A new approach to extracting sport highlight," in Computer Science and Engineering Conference (ICSEC), 2016 International, 2016, pp. 1–6.

[10] K.-S. Lin, A. Lee, Y.-H. Yang, C.-T. Lee, and H. H. Chen, "Automatic highlights extraction for drama video using music emotion and human face features," Neurocomputing, vol. 119, pp. 111–117, Nov. 2013.

[11] V. Jain, "Prediction of movie success using sentiment analysis of tweets," Int. J. Soft Comput. Softw. Eng., vol. 3, no. 3, pp. 308–313, 2013.

[12] T. Mahmood, T. Iqbal, F. Amin, W. Lohanna, and A. Mustafa, "Mining Twitter big data to predict 2013 Pakistan election winner," in INMIC, 2013, pp. 49–54.

[13] K. Wegrzyn-Wolska and L. Bougueroua, "Tweets mining for french presidential election," in Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, 2012, pp. 138–143.

[14] J. Bachhuber, C. Koppeel, J. Morina, K. Rejström, and D. Steinschulte, "US Election Prediction: A Linguistic Analysis of US Twitter Users," in Designing Networks for Innovation and Improvisation, Springer, Cham, 2016, pp. 55–63.

[15] E. Byrne and D. Corney, "Sweet FA: Sentiment, Swearing and Soccer," in 1st International Workshop on Social Multimedia and Storytelling co-located with ACM International Conference on Multimedia Retrieval (ICMR 2014), 2014.

[16] P. C. Guerra, W. Meira Jr., and C. Cardie, "Sentiment Analysis on Evolving Social Streams: How Self-report Imbalances Can Help," in Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, NY, USA, 2014, pp. 443–452.

[17] J. Pérez, E. Iturbide, V. Olivares, M. Hidalgo, A. Martínez, and N. Almanza, "A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases," J. Med. Syst., vol. 39, no. 11, 2015.

[18] O. Marbán, J. Segovia, E. Menasalvas, and C. Fernández-Baizán, "Toward data mining engineering: A software engineering approach," Inf. Syst., vol. 34, no. 1, pp. 87–107, Mar. 2009.

[19] S. K. Gupta, V. Bhatnagar, and S. K. Wasan, "Architecture for knowledge discovery and knowledge management," Knowl. Inf. Syst., vol. 7, no. 3, pp. 310–336, Mar. 2005.

[20] A. Vedder, "KDD: The challenge to individualism," Ethics Inf. Technol., vol. 1, no. 4, pp. 275–281, Dec. 1999.

[21] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Commun ACM, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[22] J.-P. Berrut and L. N. Trefethen, "Barycentric Lagrange Interpolation," SIAM Rev., vol. 46, no. 3, pp. 501–517, Jan. 2004.

[23] M. Abid, B. Nsiri, and Y. Serhane, "Interoperability between different port information systems," Int. J. Math. Comput. Simul., vol. 8, pp. 156–161, 2014.

[24] S. Jai-Andaloussi, I. E. Mourabit, N. Madrane, S. B. Chaouni, and A. Sekkaki, "Soccer Events Summarization by Using Sentiment Analysis," in 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 2015, pp. 398–403.