WordNet based Implicit Aspect Sentiment Analysis for Crime Identification from Twitter

Hajar El Hannach[1], Mohammed Benkhalifa[2]
ANISSE research team, Computer science department
Faculty of science, Mohammed V University
Rabat, Morocco

***Abstract*—Crime analysis has become an interesting field that deals with serious public safety issues recognized around the world. Today, investigating Twitter Sentiment Analysis (SA) is a continuing concern within this field. Aspect based SA, the process by which information can be extracted, analyzed and classified, is applied to tweet datasets for sentiment polarity classification to predict crimes. This paper addresses the aspect identification task involving implicit aspect implied by adjectives and verbs for crime tweets. The proposed hybrid model is based on WordNet semantic relations and Term-Weighting scheme, to enhance training data for (1) Crime Implicit Aspect sentences detection (IASD) and (2) Crime Implicit Aspect Identification (IAI). The performance is evaluated using three classifiers Multinomial Naïve Bayes, Support Vector Machine and Random Forest on three Twitter crime datasets. The obtained results demonstrate the effectiveness of WN synonym and definition relations and prove the importance of verbs in training data enhancement for crime IASD and IAI.**

Keywords*—Implicit aspect based sentiment analysis; information retrieval; machine learning; supervised approaches; frequency model; WordNet; crime detection; hate crime twitter sentiment (HCTS)

I. INTRODUCTION

Sentiment Analysis (SA) has become one of the most active topics in information retrieval and text mining due to the large expansion of the World Wide Web. SA is the field of study that deals with automatic analysis of people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text [1]. The entities can be products, services, organizations, individuals, events, or topics such as crimes. SA research has been mainly carried out at three levels of granularity: document, sentence or aspect level. Aspect level SA is the most fine-grained model, which extracts opinions expressed against different aspects/features of the entity.

Classifying opinion text at the document level or at the sentence level as positive or negative is insufficient for most applications. These classifications do not tell what each opinion is about, that is, the target of opinion. Indeed, when a document or a sentence evaluates a single entity, it does not mean that this evaluation is true for all aspects of the entity [2]. For a more complete analysis, aspects need to be discovered before to determine whether the sentiment is positive, negative, or neutral about each aspect. To obtain this level of fine-grained results, Aspect-based Sentiment Analysis (ABSA) is applied [3]. This latter considers relations between the aspects of the object of the opinion and the document polarity (positive or negative feeling expressed in the opinion). An aspect is a concept on which the author expresses his/her opinion in the document. The aspects can be of two types: explicit aspects and implicit aspects. Explicit aspects correspond to specific terms that explicitly appear in the document. In contrast, an implicit aspect is not specified explicitly in the document. The implicit aspects (which can be indicated by adjectives, adverbs, verbs or phrasal verbs) are very important that they can convey the opinions and help in improving the performance of SA systems.

Within the next few years, SA and more particularly IASA is set to become a promising approach for crime prediction [4]–[6]. Nowadays, IASA is applied for crime prevention systems such as neighborhood crime rating systems and safety of school platforms that are developed to support crime prevention and fear reducing. The Most challenging task in crime prediction area is identifying the set of committed crimes according to their types, locations and individuals, especially when this information is implicitly implied and not mentioned explicitly in data. In this scenario, Implicit Aspect based Sentiment Analysis (IASA) can be used to highlight the patterns of crimes.

When applied to crime prediction, IASA operates in three steps: (1) implicit aspect sentences detection (IASD), (2) implicit aspect identification (IAI) and (3) sentiment classification.

For crime datasets, Twitter is a defensible and logical source of data widely used in crime prevention and pattern detection approaches[7]–[9]. When gathering implicit aspect sentences from this popular social networking site, the main issue is the huge number of tweets returned with poor grammar and spelling, hashtags, URL, and irrelevant sentences. Thus, the construction of implicit aspect crime datasets requires preprocessing treatment and information retrieval techniques in order to classify relevant and irrelevant sentences. This process is known as "implicit aspect tweets or sentences detection".

After building crime datasets, Implicit Aspect Identification (IAI) is performed. IAI encompasses implicit aspect term (IAT) extraction and IAT aggregation. For each implicit aspect sentence, IAT extraction aims at extracting adjectives, verbs implying aspects. Afterward, extracted terms suggesting the same aspect are assembled into one implicit aspect in IAT aggregation.

After the implicit aspect identification, sentiment classification can be applied to classify opinions, toward each aspect, into positive or negative classes.

In this paper, the focus is made on Implicit aspect sentence detection and Implicit Aspect identification. A hybrid model, coupling WordNet Synonym and Definition semantic relations and Term Weighting scheme, is proposed for training data improvement to support both IASD and IAI steps. The proposed hybrid model is empirically evaluated using three classifiers Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Random Forest(RF) on three Twitter crime datasets. The study shows that our approach helps the three classifiers achieve good performance for IASD and IAI tasks.

The remainder of the paper is organized as follows. In section 2, related works are reviewed. In section 3, the proposed hybrid model based approach is presented in details. In section 4, the experimental setting adopted is exposed. In section 5, obtained results are presented and discussed. Finally, conclusions and future work are presented.

II. Related Works

A considerable amount of works have been published in aspect based sentiment analysis [10], [11], while few have attempted to address the implicit aspect identification. The methods applied for this task are based on two major methods: lexical based and supervised learning approaches. Among the lexical based approaches, the semantic orientation methods are used to supports binary classification [12]. Dictionary based techniques are one of the most popular lexical approach used in this field. In [13], authors try a new approach based lexical method, Part of speech tagging, SentiWordNet and WordNet combined with a weighted model provided by Natural language processing NLP(weight assignment policies) in sentiment classification. Their results outperform the basic use of WEKA Naïve Bayes Classifier and prove the effectiveness and contribution of the lexical approach in opinion mining.

Several studies investigating machine learning have been carried out on sentiment analysis. Machine learning algorithms have been used to solve the sentiment analysis as a regular text classification problem. In [14] performed a comparative study involving different machine learning algorithms. Naïve Bayes, Support Vector Machine and maximum-entropy-based classifiers are applied for sentiment polarity classification for movies reviews. Compared to the human generated baselines, the ML techniques achieve the better performances. Data representation is also among factors that impact ML performances. In [15], authors aim at investigating the effectiveness of vector representation for explicit aspect extraction. Their approach is hybrid based on Semantic Role Labelling, Conditional Random Fields and Structural Support Vector Machines (SVM-HMM). The evidence presented in their work suggests that the vector space approach support explicit aspect extraction and SA classification.

Much of the current studies on SA pays particular attention to Twitter trends and opinions. A lot of research has been done in this field by researchers and scholars all around the world [16]–[18]. Sentiment analysis in tweets is done according to major steps, identifying opinion target, explicit or implicit aspect, and classifying the sentiment polarity of tweets. To perform Sentiment classification in twitter, most of the research applied the followed process: data collection, information retrieval and sentiment classification [19], [20]. For information retrieval, Term Frequency-Inverse Document Frequency (TF-IDF) is among the most popular technique used for text categorization and tweets selection [21]. This term weighting scheme is easy to compute, implement, and understand. However, its shortcoming is very well recognized. For imbalanced datasets, the TF-IDF need to be enhanced to allows better performances[22].

Sentiment analysis is fast becoming a key instrument in Crime prevention and data detection. In [4], authors elaborate a sentiment analysis approach based on lexicon methods and combined with kernel density estimation based on historical crime incidents to predict the time and location in which a specific crime will occur. Their approach provides a significant achievement comparing to the benchmark model. Others in [8], addressed the aspect-based sentiment analysis for crime tweets through the use of hybrid model. Based on Natural Language Processing techniques and SentiWordNet, the hybrid model detects the subjectivity of crime and then predicts the hate crime tweets polarity.

III. Proposed Framework

The proposed study is motivated by considering WordNet extracted terms according to Synonym and Definition subsets for adjectives and verbs coupled with a new Term Weighting model to represent implicit aspects and improve training data. This motivation is driven by two curiosities: (1) How these WN extracted terms can be exploited and combined with their corpus adjectives and verbs to best represent implicit aspects and (2) How this combination can be made optimally informative to both tasks: implicit aspect crime sentence detection and implicit aspect identification.

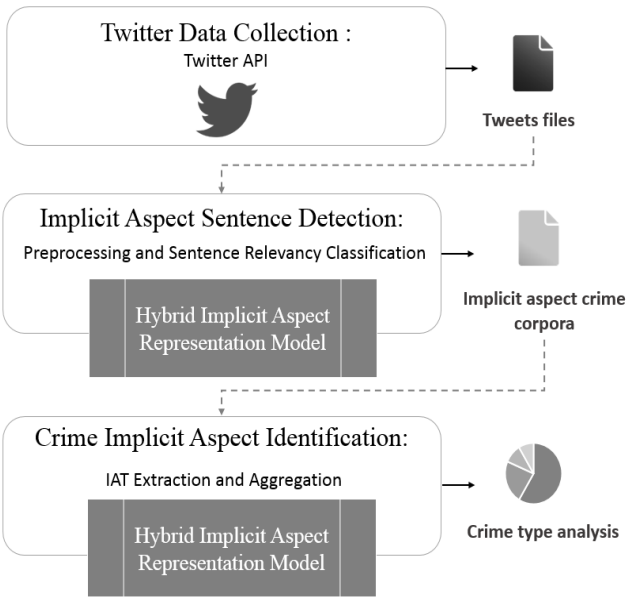

Fig. 1. Abstract Process of the Proposed Framework.

The proposed approach is supported by a hybrid representation model. It operates in three phases according to the schema shown in Fig. 1. The first phase collects tweet datasets using the official Twitter Search API v1.1. The second phase proceeds into 2 steps: The preprocessing that prepares tweet datasets and the sentence relevancy classification that detects implicit aspect crime sentences. The third phase performs IAT extraction and IAT aggregation for crime implicit aspect identification.

Before presenting the exhaustive outline of the proposed approach, the Hybrid Implicit Aspect representation model (as shown in Fig. 2) is explained in details in the following section.

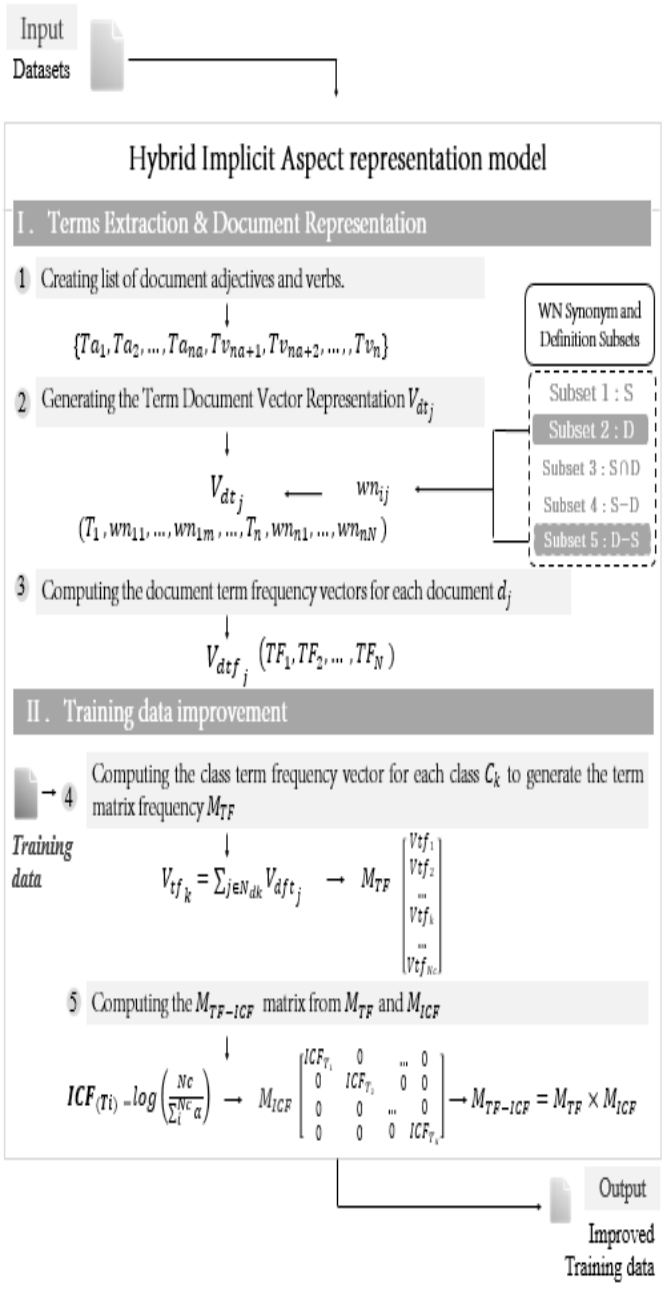

Fig. 2. Summary of the Proposed Hybrid Implicit Aspect Representation Model.

A. Hybrid Implicit Aspect Representation Model

To represent crime implicit aspects, our hybrid model proceeds in five steps. Steps 1, 2, and 3 deal with extracting implicit aspect terms for document representation whereas step 4 and 5 bring improvements to training data.

Step 1 creates a list of extracted adjectives and verbs called terms T_i.

$$\{Ta_1, Ta_2, \dots, Ta_{na}, Tv_{na+1}, Tv_{na+2}, \dots, Tv_n\}$$

Where Ta_i and Tv_i denotes adjective and verb term respectively, and n represents the number of terms T_i.

To represent dataset documents, step 2 generates a document term vector V_{dt_j} from WN extracted terms vectors V_{T_i}. The V_{T_i} vectors are generated using the appropriate WN semantic relation subsets according to adjectives and verbs. Indeed, V_{T_i} (as shown in (1)) are constructed from the best supportive subsets of WordNet semantic relations, empirically identified in [23]. In this latter work, five WN subsets are considered for adjectives and verbs:

- Subset 1: S which contains all words extracted from synonym relation.
- Subset 2: D containing all synonyms words and nouns appearing in phrases describing a given word from definition relation.
- Subset 3: S ∩ D that contains words appearing in synonym and definition relations.
- Subset 4: S-D, composed of words appearing in synonym relation and not in definition relation.
- Subset 5: D-S, representing words appearing in definition relation and not in synonym relation.

For Ta_i , V_{T_i} vectors are constructed from Subset 2 (D) containing synonyms and nouns appearing in Ta_i description. For Tv_i, V_{T_i} vectors are generated from subset 5 (D-S) that is composed of nouns appearing in verb definition and not in synonym relation.

$$V_{T_i} = (wn_{i1}, wn_{i2}, \dots, wn_{im}) \quad (1)$$

The document term vector V_{dt_j} representing a document j is generated as follows:

$$V_{dt_j} = (T_1, wn_{11}, \dots, wn_{1m}, \dots, T_n, wn_{n1}, \dots, wn_{nN} \quad (2)$$

Where wn_{in} is the n-th WN related word extracted for Term T_i and N denotes the number of terms and theirs WN extracted terms.

After the term document vector generation, step 3 computes the document term vector frequency V_{dtf_j} for each document j. TF is calculated for T_i and their WN extracted terms wn_{im} . The wn_{im} term frequency is equal to the number of times term Ti occurs in document d_j.

$$V_{dtf_j} = (TF_1, TF_2, \dots, TF_N) \quad (3)$$

Where TF_i is the document term frequency of term T_i.

Instead of using Term Frequency-Inverse Document Frequency (TF-IDF) the hybrid model uses TF-ICF which brings class information from training data.

$$TF - IDF(T_i, d_j) = tf(T_i, d_j) \times log\left(\frac{N}{N(T_i)}\right) \quad (4)$$

Where $tf(T_i, d_j)$ represents the number of times term T_i occurs in documents d_j, N denotes the number of documents, and $N(T_i)$ stands for the number of documents in which term T_i occurs at least once.

In fact, TF-IDF, that computes term weighting scores regardless the class information of documents, can't effectively deal with crime datasets which are imbalanced.

The next steps aim at including class category information from training data to provide the new term weighting ICF (inverse class frequency). This basically implies that the new ICF is class category specific and is computed using the class terms frequency vector V_{tf_k} (5) based on document term frequency vectors V_{dtf_j}

In step 4, the class terms frequency vector V_{tf_k} is generated. For each class C_k, V_{tf_k} presents the number of times that term T_i occurs in training data of class C_k. The class term frequency is obtained from V_{dtf_j} as follows:

$$V_{tf_k} = \sum_{j \in N_{dk}} V_{dft_j} \quad (5)$$

Where V_{tf_k} is the class term frequency of class C_k, and N_{dk} denotes the number of training document of C_k and V_{dtf_j} is the document frequency term for document j computed in (3). $M_{TF}(N_C, N)$ is defined as terms frequency matrix representing all V_{tf_k} vectors where N_C stands for the number of classes and N is the number of terms T_i.

Finally, in step 5, the ICF is computed for each term T_i as follows:

$$ICF(T_i) = log\left(\frac{Nc}{\sum_i^{Nc} \alpha}\right) \quad (6)$$

Where α takes 0 if term T_i does not appear in class C_k, and 1 in otherwise. The new ICF boosts the importance of terms appearing only at one class and penalizes irrelevant terms.

The final $M_{TF-ICF}(N_C, N)$ matrix is obtained by

$$M_{TF-ICF} = M_{TF} \times M_{ICF} \quad (7)$$

Where the $M_{ICF}(N, N)$ is the diagonal matrix of ICF.

As mentioned earlier, our approach proceeds in three phases (shown in Fig.1) as follows:

Phase 1: Twitter Data Collection

The data collection is done from twitter through the use of the official Twitter Search API v1.1. The Twitter API allows real time access and extraction of tweets according to a specific query. With more than 50 requests, we create three crime different datasets. The two first datasets consider the major crime types (Homicide, Rape, Robbery, Assault, Kidnapping). whereas the third one is a Hate Crime Twitter sentiment (HCTS) dataset with different aspects of Hate Crime as racism, terrorism, religious tolerance… The obtained datasets contain two types of tweets: (1) irrelevant tweets which refer to contexts not related to crimes (i.e., movies, games) or tweets without implicit aspects and (2) implicit aspect Crime tweets. Furthermore, certain tweets contain grammatical and spelling mistakes, abbreviations, URLs, sources of data, hashtags… These hurdles are addressed by the preprocessing step of the IASD phase to ensure better crime implicit aspect identification.

Phase 2: Implicit Aspect Sentence Detection

IASD phase, as shown in figure 3, consists of preprocessing and sentence relevancy classification process:

1) Preprocessing

The first step of the preprocessing is the removal of noisy data. The process begins with the removal of URL, @usernames and #hashtags. Then, the Part of speech tagger (POS) is used to parse tweets to extract adjectives and verbs as they represent potential implicit aspect terms implying crimes. For the elongate extracted terms, with more than three following occurrence of the same letter, we applied the compression words process commonly used for tweets. It's used to obtain the right form of word acceptable by the WordNet dictionary. At last, the stop words are removed from tweet datasets.

2) Sentence relevancy classification

Sentence Relevancy Classification, which encompasses two sub-steps, focuses on classifying relevant/irrelevant tweets in order to create an implicit aspect crime corpus from each dataset. The first sub-step preprocesses tweet datasets and uses the proposed hybrid model to enhance training data. The second sub-step employs the improved training data to build a classification model for crime implicit aspect sentences and then generate crime implicit aspect corpora.

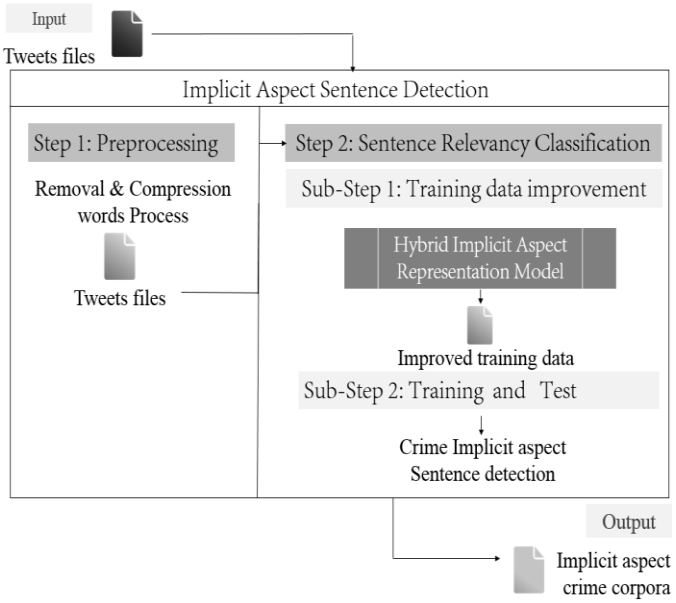

Fig. 3. Crime Implicit Aspect Sentences Detection using Hybrid Model.

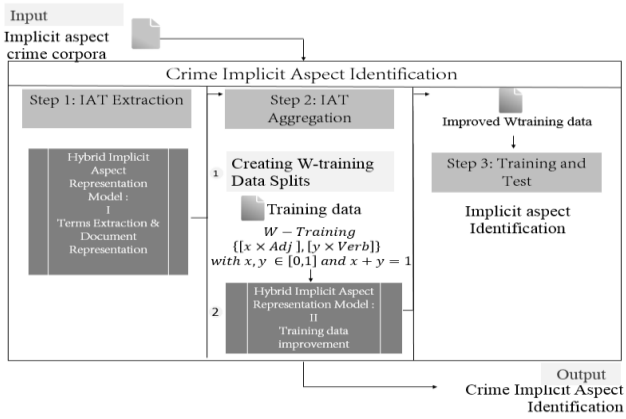

Fig. 4. Crime Implicit Aspect Identification using Hybrid Model.

Phase 3: Crime implicit aspect identification

As shown in Fig.4, the task aims at extracting crime implicit aspects from corpora prepared in phase 2. IAT extraction and aggregation are addressed using the two steps of the proposed hybrid model.

In IAT Extraction, the Terms Extraction & Document Representation steps of the hybrid model are applied to extract potential implicit aspect implied by adjectives and verbs. Then, for each dataset document, the hybrid model provides the document term frequency vectors V_{dtf_j}, which represents the contribution of adjectives and verbs and their WN extracted terms for a given document.

In IAT aggregation, the Training data improvement steps of the hybrid model are applied using several W-Training data splits. These splits are obtained using weighting schema assigning different weights for adjectives and verbs. This weighting schema is used to evaluate the impact of using different proportions of adjectives and verbs on the improvement of training data for crime datasets. Each W-Training data split is computed by equation 9 as follows:

$$W - Training\ split = \{[x \times Adj\ ], [y \times Verb]\} \quad (8)$$

$$where\ \ x, y\ \in [0,1]\ and\ x + y = 1$$

IAT aggregation task aims at identifying the implicit aspect for each document. To this end, IAT aggregation uses weighting model that measures the document terms reliability according to a given implicit aspect (class). Thus, IAT aggregation computes term matrix frequency M_{TF-ICF} , that reflects the term's strength of representing a specific class.

IV. Experiments and Evaluation

The experiments conducted to validate the proposed approach are presented in this section with the experimental design adopted, i.e. the pre-processing techniques utilized, classifies used, datasets chosen, the performance evaluation metrics used, and the results obtained based on those measures with the discussion.

A. Experimental Setup

1) Preprocessing: After gathering data from Twitter by means of the Twitter API within data collection phase, the preprocessing is done. it applied filtering text techniques to obtain a clear text without irrelevant content. At last, the POS tagger is used for parsing data and extracting a list of adjectives and verbs used at sub-step 1 of Sentence Relevancy Classification process.

2) Classifiers used: Three supervised classifiers are used to validate the proposed approach: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Random Forest (RF).

Multinomial Naïve Bayes is the most variation of NB that is mostly used in text categorization and sentiment analysis [24]. MNB is a probabilistic model based on the Bayes theorem. It uses the joint probabilities of features and categories to estimate the probabilities of classes given a document and makes the assumption that features are conditionally independent of each other to make the computation of joint probabilities simple.

In text categorization and sentiment analysis, *Support Vector Machine* is often considered as the best classifier providing the greatest performances for those tasks [25]. It's among the class of classifiers based on kernel substitution [26].In this work, the version Sequential Minimal Optimization (SMO) developed in [27] is used.

Random Forest is a popular tree classifier based on many classification trees, used for text categorization and sentiment analysis for Twitter [19], [28]. The forest construction is the base step in this classification. Each individual tree is constructed based on two procedures proposed by [29]*:* (1) to create decision tree nodes, subspace of features is randomly chosen, then (2) to generate training data subsets for building individual trees, the classifier relies on bagging method and finally (3) to obtain the random forest classifier all individual trees created are combined.

3) Datasets: The proposed approach is assessed using crime datasets collected and prepared in this work. The three crime datasets are extracted from twitter with different size and aspects.

The first crime dataset contains 2k tweets, of which 357 include implicit aspect sentences involving adjectives and verbs. The dataset covers the four major crime types namely, homicide, rape, robbery and aggravated assault.

The second dataset considers more specific type of crime as shooting, kidnapping, vehicle theft, violent crime, rape and homicide. It contains more than 600 implicit aspect sentences extracted from 3k tweets.

The hate crime dataset involves 6k tweets of which 648 include implicit aspect sentences and cover different predefined aspect racism, disability abuse, religious tolerance, terrorism and rape.

4) Evaluation measures: To evaluate the performance obtained after using the proposed approach, we use the standard metric F1-score which is commonly used to evaluate the classification task. F1, introduced by Van Rijsbergen [30] is the equally weighted average of recall and precision as

stated in (9). The Recall is defined to be the ratio of correct assignments by the system divided by the total correct assignments. The Precision is the proportion of correct assignments by the system within the total number of the system's assignments. All experiments are carried out using Weka platform [31]. We use the 10 Fold cross validation to reduce the uncertainty of data split between training and test data.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

B. Experimental Protocols

Experiments aimed at evaluating the effectiveness of the proposed approach for Crime IASD and crime IAI.

1) Crime implicit aspect sentence detection IASD: Experiments have been conducted according to three points relating to performance of the three classifiers for Crime IASD:

a) The use of TF-IDF versus TF-ICF for document representation

The first point pertains to evaluating and comparing the impacts of using TF-IDF and TF-ICF on the performances for the classifiers for IASD. Two categories of experiments are defined:

- TF-IDF (denoted baseline): it refers to the use of the three classifiers with hybrid model and without considering WordNet semantic relations. TF-IDF uses only the terms extracted from datasets and presents documents using the Term Frequency-Inverse Document Frequency vector model.
- TF-ICF: it represents the use of the three classifiers with hybrid model and without considering WordNet semantic relations. TF-ICF uses only the terms extracted from datasets and presents documents using the Term Frequency-Inverse Class Frequency vector model.

b) The integration of WordNet Synonym relations of adjectives and verbs in the document representation model

The second point relates to the comparison of the impacts on the performances of the classifiers for IASD using TF-IDF and TF-ICF with the integration of WordNet synonym relations for adjectives and verbs. Synonyms are considered here due to their wide use in SA. Two types of experiments are defined:

- TF-IDF+ Synonyms: it concerns the use of the three classifiers with the hybrid model using TF-IDF and integrating synonyms of adjectives and verbs.
- TF-ICF+ Synonyms: it refers to using the three classifiers with the hybrid model using TF-ICF and integrating Synonyms of adjectives and verbs.

c) The integration of the best WN subsets of adjectives and verbs in the document representation

The third point is similar to the second point except here the integration of WordNet relations concerns the best WN subsets for adjectives and verbs. Two types of experiments are defined:

- TF-IDF + Best-WN-subsets represents the use of the three classifiers with the hybrid model using TF-IDF and integrating the best WN subsets (subsets D and D-S for adjectives and verbs respectively).
- TF-ICF + Best-WN-subsets represents the use of the three classifiers using TF-ICF and integrating the best WN subsets (subsets D and D-S for adjectives and verbs respectively).

Experiments have been conducted according to two points that evaluate the effectiveness of the proposed approach for Crime implicit aspect sentence detection IASD and crime implicit aspect identification IAI.

2) Crime implicit aspect identification IAI: Experiments have been conducted according to two points relating to performance of the three classifiers for Crime IAI:

a) The use of adjectives and verbs for training data enhancement

The first point concerns the comparison of the impacts of adjectives and verbs on training data improvement. Experiments are done here using several W-Training data splits. These splits are obtained using weighting schema assigning different weights for adjectives and verbs. For each dataset, six testing datasets are prepared where each set combines adjectives and verbs with different weighting. These weightings are defined as follows: $(1, 0)$, $(0.8, 0.2)$, $(0.6, 0.4)$, $(0.4, 0.6)$, $(0.2, 0.8)$ and $(0,1)$. Three experiments are defined: MNB, SVM and RF that respectively refers to MBN, SVM and RF classifier using hybrid model, W-training data and integrating the best WN subsets of adjectives and verbs.

b) The Absence of WN terms of adjectives and verbs in the document representation

The second point deals with the effects on classifiers performances of the absence of WN terms of adjectives and verbs in the hybrid model. Three experiments are defined: MNB_{NoWn}, SVM_{NoWn} and RF_{NoWn} that respectively represents MNB, SVM and RF using the hybrid model with W-training data and without the best WN subsets of adjectives and verbs.

C. Results and Discussion

In this section, experiments results are presented according to the points mentioned in the experimental protocols section.

1) Crime implicit aspect sentence detection IASD: The three classifiers are assessed for crime Implicit aspect sentence detection using three crime datasets with varying sizes presented in table 1. Table 2 shows the performances with the Average Improvement Rates (AVG.Imp.R) of the three classifiers obtained from experiments related to the three points above mentioned in the experimental protocols for this phase.

TABLE I. SIZE OF DATASETS

	Crime dataset 1	*Crime dataset 2*	*Hate crime dataset*
Number of sentences	2k	3k	6k
Number of implicit aspect sentences	357	641	648
Number of irrelevant sentences	1643	2359	5352
Number of Training data for implicit aspect	180	350	300
Number of Training data for implicit aspect	670	1500	3500

TABLE II. MNB, SVM AND RF FOR RELEVANT / IRRELEVANT CLASSIFICATION

	Crime dataset 1			*Crime dataset 2*			*Hate crime dataset*		
	MNB	***SVM***	***RF***	***MNB***	***SVM***	***RF***	***MNB***	***SVM***	***RF***
(1)	0.51	0.63	0.63	0.52	0.65	0.63	0.63	0.69	0.68
(2)	0.57	0.65	0.65	0.59	0.68	0.68	0.74	0.77	0.75
(1)/(2)	11.6 %	3.1%	3.1%	13.4 %	4.6 %	7.9 %	17.4 %	11.5 %	10.2 %
(3)	0.71	0.68	0.68	0.69	0.71	0.69	0.76	0.78	0.78
(4)	0.74	0.71	0.71	0.74	0.74	0.72	0.74	0.78	0.78
(5)	0.78	0.71	0.71	0.74	0.73	0.73	0.80	0.81	0.80
(6)	**0.83**	**0.88**	**0.87**	**0.79**	**0.80**	**0.79**	**0.82**	**0.89**	**0.87**
(3)/(5)	9.8%	4.4%	4.4%	7.2%	2.8 %	5.7 %	5.2%	3.8%	2.5%
(4)/(6)	12.1 %	23.9 %	22.5 %	6.7%	8.1 %	9.7 %	10.8 %	14.1 %	11.5 %

(1) Baseline, (3) TF-IDF+Synonyms, (5) TF-IDF + Best WN Subsets
(2) TF-ICF, (4) TF-ICF+Synonyms, (6) TF-ICF + Best WN Subsets (The proposed Hybrid Model)

Firstly, it can be seen, from table 2, that the use of TF-ICF helps better the three classifiers deal with IASD than using TF-IDF for the three datasets. In fact, TF-IDF does need cope effectively with document representation for the three datasets because these latter are class imbalanced. Normally, terms with low TF-IDF are considered irrelevant terms since they appear in large part of documents. This is not definitely true for imbalanced datasets, because although these terms occur more often in one class than others they are relevant and important to distinguish between classes. On the contrary of TF-IDF, TF-ICF takes advantage of those unevenly distributed words by considering term contribution in class representation rather than document representation.

Secondly, table 2 shows that the integration of WN synonyms helps the three classifiers improve their performances for IASD when using TF-IDF and TF-ICF. Moreover, the use of TF-ICF is proven to be consistently more helpful for the three classifiers than using TF-IDF even with the integration of WordNet synonyms.

Thirdly, table 2 proves that the integration of the best WN subsets allows the three classifiers achieve their best performances for both TF-IDF and TF-ICF cases. Also, using TF-ICF is shown to help the three classifiers achieve better performances than using TF-IDF.

In fact, the integration of WN semantic relations promotes training data vocabulary by creating a large set of relevant terms that support system to learn better from data. However, the selection of WN semantic relation is crucial. The integration of synonym relation allows classifiers to achieve better scores, yet, it induces more noisy terms than definition subsets. WN semantic relations for adjectives and verbs must be appropriately selected (subsets D for adjectives and subsets D-S for verbs) so that they can help the classifiers achieve their best performances for IASD.

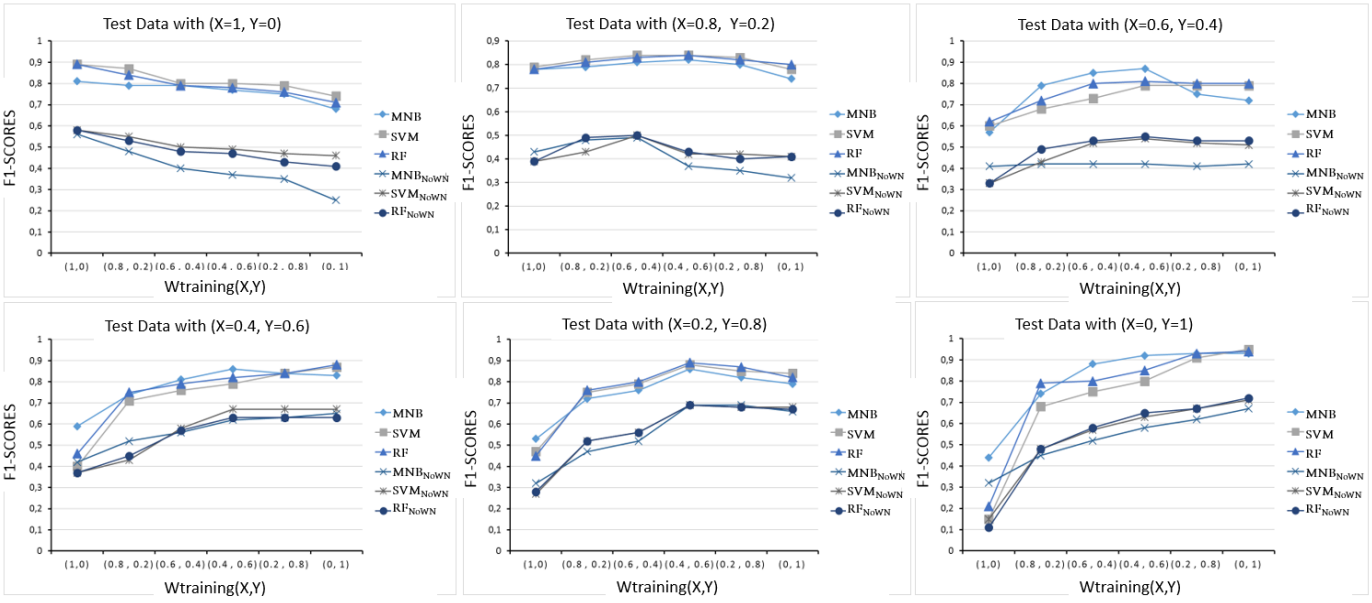

Fig. 5. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on **CRIME CORPUS 1** for IAI Phase.

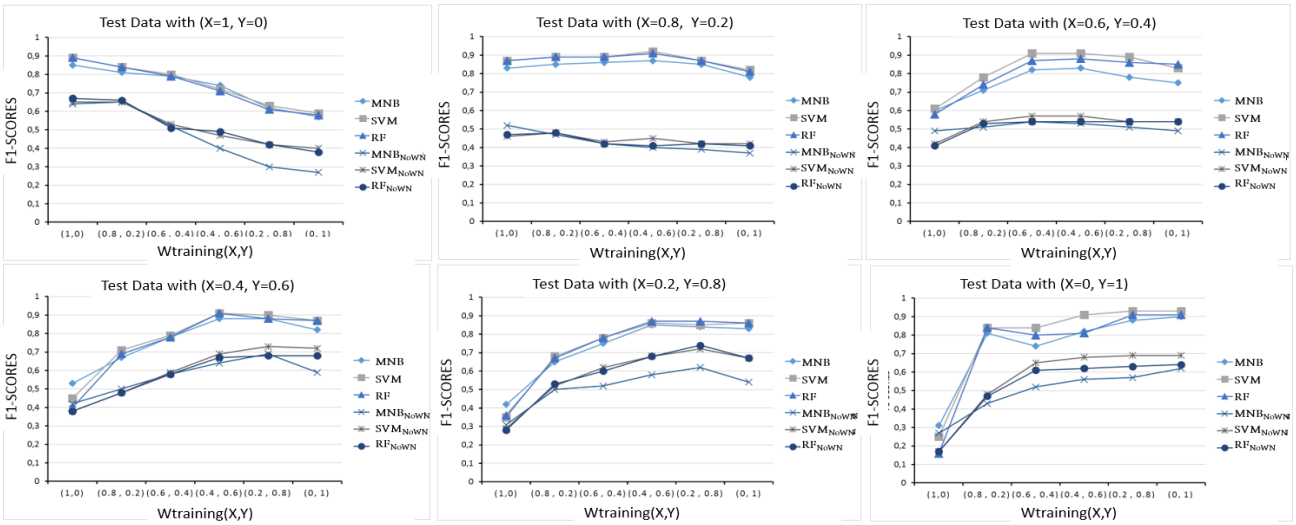

Fig. 6. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on CRIME CORPUS 2 for IAI Phase.

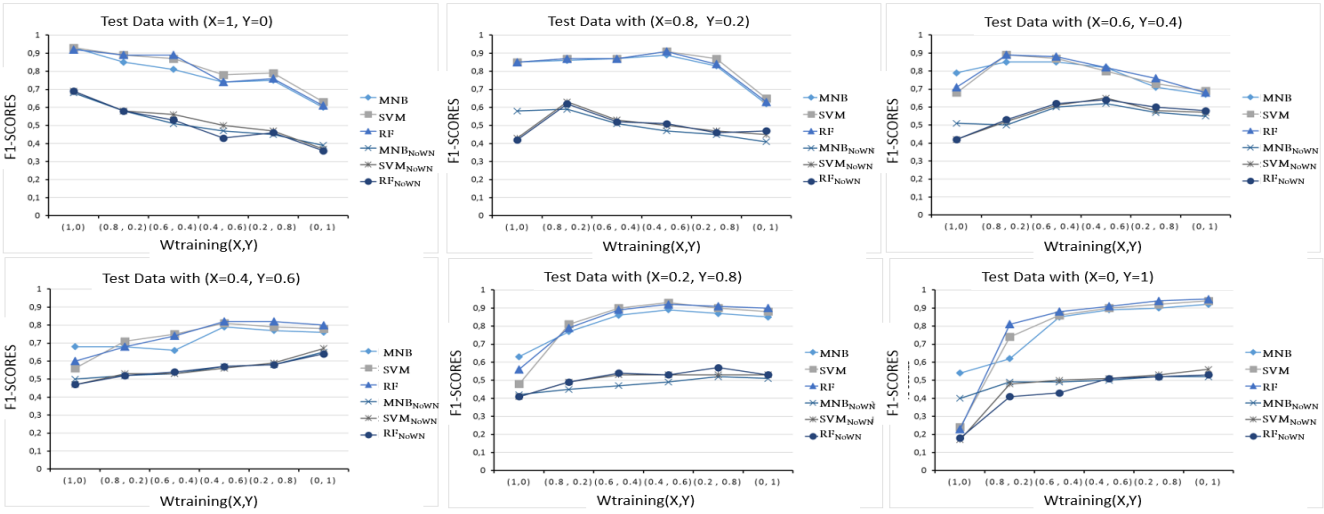

Fig. 7. F1-Performances of MNB, SVM and RF using Different W-Training Data Splits, with and without the best WN Subsets on HATE CRIME CORPUS for IAI Phase.

2) Crime implicit aspect identification IAI: Fig. 5, 6 and 7 show the performances of the three classifiers obtained from experiments pertaining to the two points already introduced in experimental protocols for this phase. For each testing data, X-axis denotes the different weights (x, y) assigned to adjectives and verbs and Y-axis indicates F-1 performances. The number of adjectives and verbs of the three crime corpora is shown in table 3.

TABLE III. NUMBER OF ADJECTIVES AND VERBS IMPLYING IMPLICIT ASPECT FOR EACH CRIME DATASET

	Crime dataset 1	***Crime dataset 2***	***Hate crime dataset***
Number of sentences	357	641	648
Number of adjectives	406	841	773
Number of verbs	446	872	729

As shown from Fig. 5, 6 and 7, the use of different weights (x, y) assigned to adjectives and verbs leads to variant F1-performances for MNB, SVM and RF.

The case of training and test data involves adjectives only (test data with (x=1, y=0) and W-training with (1,0)) leads to the best performances for all classifiers. One unanticipated finding is that when considering only verbs for training (Wtraining with (0,1)), all classifiers are able to achieve considerable F1- performances that exceed 60% for implicit aspect identification implied by adjectives.

In contrast, adjectives do not support verb identification (test data with (x=0, y=1)). In this test data case and for the three datasets, classifiers achieve their worst performances when verbs are completely absent in training data (W-training with (1 ,0)). For the same test data, the best F-1 scores are attained when considering only verbs for training (Wtraining with (0,1)).

Using adjectives in training data supports IAI involving more adjectives than verbs (test data with (x=1, y=0) and (x=0.8, y=0.2)). Implicit aspect identification including verbs is known to be more challenging than adjective. Using only adjectives in training to predict implicit verbs does not support classifiers identifying the implicit aspect for crime datasets. However, using verbs for implicit aspect identification is more beneficial for classifiers. This can be explained by the fact that, verbs used to imply a crime aspect are more descriptive and useful than adjectives. In other words, for each crime aspect there are a number of verbs specifically used to imply this aspect, for example '*to kill'*, '*to kidnap'* and *'to steal'*, each verb is used to imply a single type of crime which is *'Homicide'*, *'Kidnapping'* and *'Robbery'* respectively. However, as often happens, one adjective can be used to imply different crime aspects such as *'blooded'*, *'atrocious'*, '*hostile'*, *'agonizing'*, *'cruel'* that can be used not only for *'homicide'* but for *'violent crime'* and *'kidnapping'* as well. As a result, adjective extracted terms can represent more than one aspect. However, when considering more verbs for training than adjectives, the WN extracted terms are more descriptive and contain more reliable terms that better represent the implicit aspect which supports adjective and verb identification.

For training and test data using a combination of adjectives and verbs, and for the same reasons explained above, the highest performance is achieved in general when considering training with more verbs than adjectives. Overall, for the three datasets, the best performing W-training is $(0.4, 0.6)$.

On the other hand, for each testing data of the three crime corpora, MNB_{NoWn}, SVM_{NoWn} and RF_{NoWn} have the same behavior than MNB, SVM and RF, but the performances reached, without WN extracted terms for verbs and adjectives, are consistently lower.

Considering more verbs than adjectives in training data supports implicit aspect identification for adjectives and verbs. While using more adjectives for learning conducts to better classifiers performances for test data involving only adjectives. However, the observed decrease in F1-performances can be attributed to the lack of WN extracted terms. Without WN, classifiers are not able to enlarge training vocabulary. This makes it extremely hard to identify adjectives and verbs appearing only twice in datasets. Even worse, it's completely impossible to identify terms appearing only once either in training or test set. The Absence of WN terms of adjectives and verbs severely penalizes performances of all classifiers for crime IAI. Hence, considering a weighted training data based on verbs and their WN extracted terms not only is required and undeniable but also improves the performance of the considered classifiers for implicit aspect identification for crimes.

Finally, Fig. 8 presents the extracted implicit aspects of the three considered crime datasets using the proposed framework.

V. Conclusion

We presented a hybrid approach for training data improvement of MNB, SVM and RF classifiers to address Aspect Based Sentiment Analysis for Crime datasets. We conduct an empirical and analytical study at the level of:

1) The crime implicit aspect sentence Detection IASD phase, where experiments are conducted according to three points: (1) the use of TF-IDF versus TF-ICF for document representation (2) The integration of WordNet Synonym relations of adjectives and verbs in document representation model and (3) The integration of the best WN subsets of adjectives and verbs in document representation.

2) The crime implicit aspect identification IAI phase, where experiments are carried out according to two points: (1) The use of adjectives and verbs for training data enhancement and (2) The Absence of WN terms of adjectives and verbs in document representation.

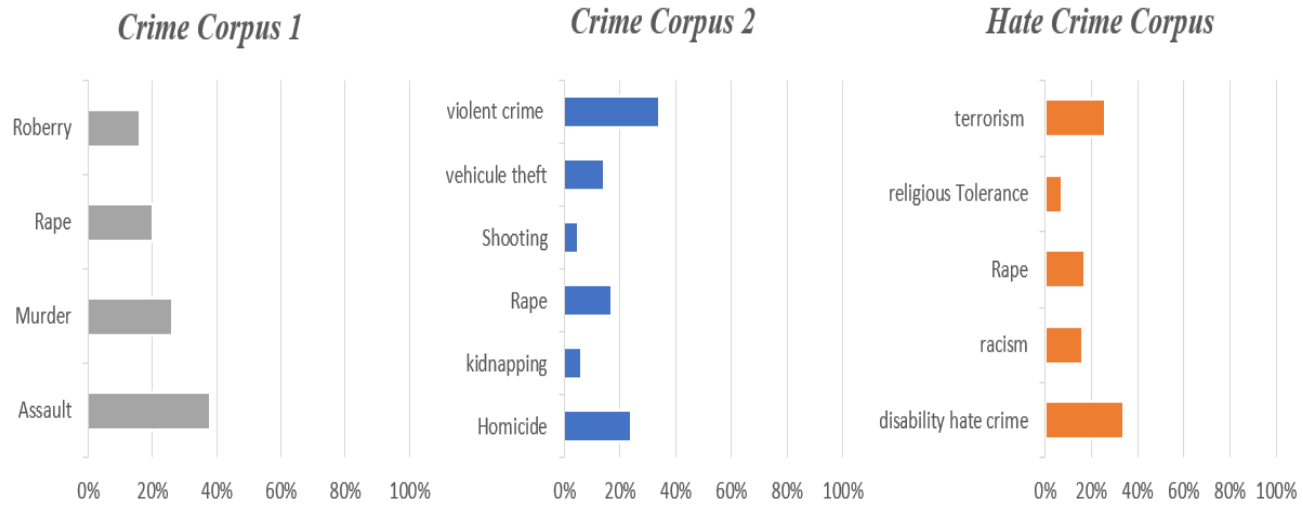

Fig. 8. Percentage of Implicit Aspects of Crime for the Three Crime Datasets.

The major findings of the work include:

- For the three imbalanced crime datasets, using TF-ICF is shown to help the three classifiers achieve better performances for IASD than using TF-IDF. This is true with and without the integration of WordNet terms.
- Using the synonyms relations for adjectives and verbs are shown to support better classifiers for IASD phase.
- Using an appropriately selected WN semantic relations for adjectives and verbs (Best WN subsets) improves training data for crime IASD and IAI and thus helps classifiers performing better for these two phases.
- Comparing to adjectives, verbs and their WN extracted terms are empirically proven to be as the key element for training data enhancement that allows classifiers to be more performant for crime implicit aspect identification.

Further work will investigate those findings to deal with the problem of the identification of crimes committed by the same individual or same group which became an important and challenging task of crime prevention systems.

Another interesting future perspective is applying the proposed approach for crime detection from variant resources of data such as weather data which significantly influence crime rates and criminal behavior.

REFERENCES

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[2] Doaa Mohey El-Din, "Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis" International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016.

[3] B. Liu, Sentiment analysis: mining opinions, sentiments, and emotions. New York, NY: Cambridge University Press, 2015.

[4] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," 2015, pp. 63–68.

[5] Jermy Prichard, Paul Watters, Tony KRONE, Caroline Spiranovic, and Helen Cockburn, "Social Media Sentiment Analysis: A New Empirical Tool for Assessing Public Opinion on Crime?," pp. 217–236, 2015.

[6] Nisal Waduge, "Machine Learning Approaches For Detect Crime Patterns - Data Gathering and Analysing Techniques," 2017.

[7] P. Burnap et al., "Detecting tension in online communities with computational Twitter analysis," Technol. Forecast. Soc. Change, vol. 95, pp. 96–108, Jun. 2015.

[8] N. Zainuddin, A. Selamat, and R. Ibrahim, "Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach," in Intelligent Information and Database Systems, vol. 9621, N. T. Nguyen, B. Trawiński, H. Fujita, and T.-P. Hong, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 151–160.

[9] Hissah AL-Saif and Hmood Al-Dossari, "Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 9(10), 2018.

[10] B. Keith, E. Fuentes, and C. Meneses, "A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews," N. S., p. 10, 2017.

[11] K. Schouten, P. O. Box, and D. Rotterdam, "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data," IEEE Trans. Cybern., p. 13, 2017.

[12] V. S. Jagtap and K. Pawar, "Sentence-Level Analysis of Sentiment Classification," Natl. Conf. Emerg. Trends Eng. Technol. Archit., p. 6, 2013.

[13] K. Gull, S. Padhye, and D. S. Jain, "A Comparative Analysis of Lexical/NLP Method with WEKA's Bayes Classifier," Int. J. Recent Innov. Trends Comput. Commun., vol. 5, no. 2, p. 7, 2017.

[14] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[15] A. Alghunaim, M. Mohtarami, S. Cyphers, and J. Glass, "A Vector Space Approach for Aspect Based Sentiment Analysis," 2015, pp. 116–122.

[16] G. Abdulsattar A. Jabbar Alkubaisi, S. Sakira Kamaruddin, and H. Husni, "Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers," Int. J. Eng. Technol., vol. 7, no. 2.14, p. 57, Apr. 2018.

[17] V. N. Patodkar and S. I.R, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," IJARCCE, vol. 5, no. 12, pp. 320–322, Dec. 2016.

[18] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter," p. 17, 2017.

[19] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and As. Perera, "Opinion mining and sentiment analysis on a Twitter data stream," 2012, pp. 182–188.

[20] M. Ishtiaq, "Sentiment Analysis of Twitter Data Using Sentiment Influencers," vol. 6, no. 1, p. 9, 2015.

[21] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," p. 6, 2012.

[22] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," Expert Syst. Appl., vol. 36, no. 1, pp. 690–701, Jan. 2009.

[23] El Hannach, H. and Benkhalifa , M., "Using Synonym and Definition WordNet Semantic relations for implicit aspect identification in Sentiment Analysis," Pap. Present. 1st Int. Conf. Netw. Inf. Syst. Secur. NISS 2018 Conf. Tangier Morocco., p. 8, 2018.

[24] Junseok Song, Kyung Tae Kim, Byungjun Lee, Sangyoung Kim, and Hee Yong Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," KSII Trans. Internet Inf. Syst., vol. 11, no. 6, Jun. 2017.

[25] R. Sergienko, M. Shan, and A. Schmitt, "A Comparative Study of Text Preprocessing Techniques for Natural Language Call Routing," in Dialogues with Social Robots, vol. 427, K. Jokinen and G. Wilcock, Eds. Singapore: Springer Singapore, 2017, pp. 23–37.

[26] G. Loosli, S. Canu, and L. Bottou, "Training Invariant Support Vector Machines using Selective Sampling," p. 26, 2005.

[27] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design," Neural Comput., vol. 13, no. 3, pp. 637–649, Mar. 2001.

[28] B. Xu, X. Guo, Y. Ye, and J. Cheng, "An Improved Random Forest Classifier for Text Categorization," J. Comput., vol. 7, no. 12, Dec. 2012.

[29] LEO BREIMAN, "Random Forests," Machine Learning, The Netherlands, pp. 45, 5–32, 2001.

[30] C. J. van RIJSBERGEN, "Information Retrieval," 2nd ed. Butterworth-Heinemann, 1979.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, p. 10, Nov. 2009.