T-HSAB: A Tunisian Hate Speech and Abusive Dataset

Hatem Haddad[1,3](✉), Hala Mulki[2,3], and Asma Oueslati[1]

[1] RIADI Laboratory, National School of Computer Sciences,
Manouba University, Manouba, Tunisia
haddad.hatem@gmail.com, asmaoueslati67@gmail.com

[2] Department of Computer Engineering, Konya Technical University,
Konya, Turkey
hallamulki@gmail.com

[3] iCompass Consulting, Tunis, Tunisia

Abstract. Since the "Jasmine Revolution" at 2011, Tunisia has entered a new era of ultimate freedom of expression with a full access into social media. This has been associated with an unrestricted spread of toxic contents such as Abusive and Hate speech. Considering the psychological harm, let alone the potential hate crimes that might be caused by these toxic contents, automatic Abusive and Hate speech detection systems become a mandatory. This evokes the need for Tunisian benchmark datasets required to evaluate Abusive and Hate speech detection models. Being an underrepresented dialect, no previous Abusive or Hate speech datasets were provided for the Tunisian dialect. In this paper, we introduce the first publicly-available Tunisian Hate and Abusive speech (T-HSAB) dataset with the objective to be a benchmark dataset for automatic detection of online Tunisian toxic contents. We provide a detailed review of the data collection steps and how we design the annotation guidelines such that a reliable dataset annotation is guaranteed. This was later emphasized through the comprehensive evaluation of the annotations as the annotation agreement metrics of Cohen's Kappa (k) and Krippendorff's alpha (α) indicated the consistency of the annotations.

Keywords: Tunisian dialect · Abusive speech · Hate speech

1 Introduction

Tunisia is recognized as a high contact culture with dense social networks and strong social ties; where online social networks play a key role in facilitating social communications [1]. With the freedom of expression privilege granted after the Tunisian revolution, sensitive "taboo" topics such as the religion have become popular and widely discussed by Tunisians across social media platforms. However, on the down side, it became easy to spread abusive/hate propaganda against individuals or groups. Indeed, recent events like the legalization of gender equality in inheritance, the appointment of a Jewish as the Tourism Minister

K. Smaïli (Ed.): ICALP 2019, CCIS 1108, pp. 251–263, 2019.
https://doi.org/10.1007/978-3-030-32959-4_18

of Tunisia and the murder of a Sub-Saharan African student caused intensive debates between Tunisians, most of which took place on social media networks leading to a high emergence of abusive/hate speech. This evoked the need for tools to detect such online abusive/hate speech contents.

In the literature, there has been no clear distinction between Abusive speech (AS) and Hate speech (HS). [2] defined HS as *"any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic"*. Although HS can be conducted as a subtask of the abusive language detection [3], it remains challenging since it requires to consider the correlation between the abusive language and the potential groups that are usually targeted by HS. Further challenges could be met when HS detection is investigated with complex, rich and ambiguous languages/dialects such as the Arabic language and its relevant dialects.

Compared to the increasing studies of AS/HS detection in Indo-European languages, similar research for Arabic dialects is still very limited. This is mainly attributed to the lack of the needed publicly-available AS/HS resources. Building such resources involves several difficulties in terms of data collection and annotation especially for underrepresented Arabic dialects such as the Tunisian dialect.

Tunisian dialect, also known as "Tounsi" or "Derja", is different from Modern Standard Arabic; where the Tunisian dialect features Arabic vocabulary spiced with words and phrases from Amazigh, French, Turkish, Italian and other languages [4]. In this study, we introduce the first **T**unisian **H**ate **S**peech and **AB**usive (**T-HSAB**) dataset. The dataset combines 6,039 comments labeled as Abusive, Hate or Normal[1]. With the objective of building a reliable, high quality benchmark dataset, we provide a comprehensive qualitative evaluation of the annotation process of T-HSAB. To achieve this goal, agreement without chance correction and Inter-annotator agreement (IAA) reliability measures are employed. In addition, our dataset was examined as a benchmark AS/HS dataset through subjecting it to supervised machine learning experiments conducted by SVM and NB classifiers.

To the best of our knowledge, this is the first study on Tunisian Abusive and Hate speech. This could be deduced in the next section where we will present the state-of-the-art of the Arabic AS and HS detection.

2 Arabic Abusive/Hate Speech Detection

As seeking to propose a new dialectal Arabic dataset for AS and HS, we opted to review the Arabic AS and HS datasets proposed in the State-Of-The-Art focusing on their characteristics in terms of: source, the tackled toxic categories, size, annotation strategy, metrics, the used machine learning models, etc. Recently, [5] presented a preliminary study of the Arabic AS and HS detection domain,

[1] Will be made publicly available on github.

where they aimed to classify the online toxic content on social media into: Abusive, Obscene, Offensive, Violent, Adult content, Terrorism and Religious Hate Speech.

The first attempt to detect Arabic abusive language was performed by [6]. A dataset of 25 K Arabic tweets was manually annotated as abusive or not abusive. However, annotation evaluation measures were not provided.

To detect offensive speech in Youtube, the authors in [7] created a data set of 16 K Egyptian, Iraqi and Libyan comments. The comments were annotated as offensive, inoffensive and neutral by three annotators from Egypt, Iraq and Libya. The annotation evaluation measurements of the Egyptian and Libyan annotators were 71% and 69.8% for inter-annotator agreement and Kappa metric, respectively. The best achieved F-measure was 82% with Support Vector Machines (SVM) algorithm used for classification. Similarly, to detect offensive speech, [8] proposed two datasets:a dataset of 1,100 dialectal tweets and a 32 K inappropriate comments dataset collected from a popular Arabic news site. To support the detection of the offensive content, the authors relied on common patterns used in offensive and rude communications to construct a list of obscene words and hashtags. The tweets and comments were annotated as obscene, offensive, and cleaned by three annotators. With only obscene instances considered, the average inter-annotator agreement was 85% for the Twitter dataset and 87% for the comments dataset.

[9] focused on religious HS detection to identify religious groups targeted by HS such as Muslims, Jews, Christians, Sunnis, Shia and so forth. For this purpose, a multi-dialectal Arabic dataset of 6.6 K tweets was introduced and annotated by 234 different annotators. As a result, three Arabic lexicons were constructed. Each lexicon combined the terms commonly used in religious discussions accompanied with scores representing their polarity and strength. The inter-rater agreement regarding differentiating religious HS tweets from non-religious ones was 81% while this value decreased to 55% when it comes to specify which religious groups are targeted by the religious HS. The proposed dataset was evaluated as a reference dataset using three classification models: Lexicon-based, SVM and GRU-based RNN. The results revealed that the GRU-based RNN model with pre-trained word embedding was the best-performing model where it achieved an F-measure of 77%.

In order to detect bullying in social media texts, [10] presented a Twitter dataset of 20K multi-dialectal Arabic tweets annotated manually with bullying and non-bullying labels. In their study, neither inter-rater agreement measures nor classification performances were provided.

More recently, a Twitter dataset, called L-HSAB, about AS and HS was introduced in [11] as a benchmark dataset for automatic detection of online Levantine AS and HS contents. The dataset composed of 6K tweets, manually annotated as Normal, Abusive and Hate. The high obtained values of agreement without chance correction and inter-annotator agreement indicated the reliability of the dataset. The inter-rater agreement metric denoted by Krippendorff's alpha (α) was 76.5% and indicated the consistency of the annotations. Given that their study is the first attempt to measure the reliability and the consistency of a

Levantine Abusive/Hate speech dataset, we followed the same approach, in our paper. Therefore, to verify reliability and the consistency of our T-HSAB dataset, we adopted agreement without chance correction, inter-annotator agreement and inter-rater agreement metrics within the annotation evaluation task.

3 T-HSAB

T-HSAB can be described as a sociopolitical dataset since the comments are mainly related to politics, social causes, religion, women rights and immigration. In the following subsections, we provide the annotation guideline, the annotation process and the annotation quantitative/qualitative results.

3.1 Data Collection and Processing

The proposed dataset was constructed out of Tunisian comments harvested from different social media platforms. We collected the dataset comments based on multiple queries, each of which represents a potential entity that is usually attacked by abusive/hate speech. Among the used queries, we can mention: “اليهود” (*Jews*), “الأفارقة” (*Africans*) and “المساواة في الميراث” (*gender equality in inheritance*). To harvest the query-related comments, the collection process focused on the comments posted within the time period: October 2018-March 2019. Initially, we retrieved 12,990 comments; after filtering out the non-Arabic, non-textual, AD-containing and duplicated instances, we ended up with 6,075 comments, written in the Tunisian dialect.

In order to prepare the collected comments for annotation, they were normalized through eliminating platform-inherited symbols such as Rt, @ and #, Emoji icons, digits, in addition to non-Arabic characters found in URLs and user mentions.

3.2 Annotation Guidelines

The annotation task requires labeling the comments of T-HSAB dataset as Hate, Abusive or Normal. Based on the definition of Abusive and Hate speech stated in the introduction, differentiating HS from AS is quite difficult and is usually prone to personal biases; which, in turn, yields low inter-rater agreement scores [3]. However, since HS tends to attack specific groups of people, we believe that, defining the potential groups to be targeted by HS, within the scope of the domain, time period and the context of the collected dataset, can resolve the ambiguity between HS and AS resulting in better inter-rater agreement scores. Hence, we designed the annotation guidelines such that all the annotators would have the same perspective about HS. Our annotation instructions defined the 3 label categories as:

- Normal comments are those instances which have no offensive, aggressive, insulting and profanity content.
- Abusive comments are those instances which combine offensive, aggressive, insulting or profanity content.

- Hate comments are those instances that: (a) contain an abusive language, (b) dedicate the offensive, insulting, aggressive speech towards a person or a specific group of people and (c) demean or dehumanize that person or that group of people based on their descriptive identity (race, gender, religion, disability, skin color, belief).

Table 1 lists the relevant examples to each class.

Table 1. Comment examples of the annotation labels

Label	**Example**
Normal	انا ما فهمت شي الا تتحدثون ألعربيه I couldn't understand anything, don't you speak Arabic?
Abusive	انسانة ساقطة مريضة نفسيا What a despicable psychopath woman
Hate	اللاسف رجال تونس اصبحو نسوان Unfortunately, Tunisian men have become women

3.3 Annotation Process

The annotation task was assigned to three annotators, two males and one female. All of them are Tunisian native speakers and at a higher education level (Master/PhD).

Besides the previous annotation guidelines, and based on the domain and context of the proposed dataset, we provided the annotators with the nicknames usually used to refer to certain minorities and ethnic groups. For instance, within an insulting context, Sub-Saharan African ethnic groups are usually referred to using these nicknames: " عبيد" (*slaves*), "أوصيف" (*black*), "نيغرو" (*nig**a*) and "كحلوش" (*of a dark skin*).

Having all the annotation rules setup, we asked the three annotators to label the 6,075 comments as Normal, Abusive or Hate. For the whole dataset, we received a total of 18,225 judgments. When exploring these annotations, we faced three cases:

1. Unanimous agreement: the three annotators annotated a comment with the same label. This was encountered in 4,941 comments.
2. Majority agreement: two out of three annotators agreed on a label of a comment. This was encountered in 1,098 comments.
3. Conflicts: each rater annotated a comment differently. They were found in 36 comments.

After excluding the comments having 3 different judgments, the final released version of T-HSAB is composed of 6,039 comments. A summary of the annotation statistics is presented in Table 2.

Table 2. Summary of annotation statistics

Annotation case	#Comments
Unanimous agreement	4,941
Majority agreement (2 out of 3)	1,098
Conflicts	36

4 Annotation Results

Having all the annotations gathered in one data file, we decided the final label of each comment in the dataset according to the annotation cases in Sect. 3.3. For comments falling under the first annotation case, the final labels were directly deduced, while for those falling under the second annotation case, we selected the label that has been agreed upon by two annotators out of three. Thus, we got 3,834 Normal, 1,127 Abusive and 1,078 Hate comments. Bearing in mind that HS is naturally a limited phenomenon [3], we kept the data unbalanced in order to have a dataset that reflects the actual distribution of HS in an Arabic dataset. A detailed review of the statistics of T-HSAB final version is provided in Table 3, where Avg-S-L denotes the average length of comments in the dataset, calculated based on the number of words in each comment.

Table 3. Comments distribution across 3 classes.

	Normal	Abusive	Hate
# Comments	3,834	1,127	1,078
Avg-S-L	11	8	12
Word Count	43,254	9,320	13,219
Vocabulary	19,162	6,358	7,789
Ratio	63.49%	18.66%	17.85%

As seeking to identify the words commonly used within AS and HS contexts, we investigated the lexical distribution of the dataset words across both Abusive and Hate classes. Therefore, we subjected T-HSAB to further normalization, where we removed stopwords based on our own manually-built Tunisian stopwords list. Later, we constructed a visualization map for the most frequent occurring words/terms under the Hate category (Fig. 1. The ten most frequent words and their frequencies in each class are reviewed in Table 4, where Dist. denotes the word's distribution under a specific class.

As it can be seen from Table 4 and Fig. 1, both Abusive and Hate classes can have terms in common such as "تونس" (*Tunisia*). These terms are not only limited to the offensive/insulting words but also combine entity names representing ethnic groups. This on one hand, explains the difficulty faced by

Table 4. Distribution of ten most frequent terms

Hate	Dist.	Abusive	Dist.
تونس (*Tunisia*)	2.01%	تونس (*Tunisia*)	1.39%
شعب (*people*)	0.74%	ع**ة (*Fu*k you!*)	0.40%
الاسلام (*Islam*)	0.43%	مي*ون (*faggot*)	0.32%
اليهود (*Jews*)	0.39%	لعنة (*curse*)	0.31%
التونسي (*Tunisian*)	0.26%	كلب (*dog*)	0.30%
عورة (*private parts*)	0.24%	تفوه (*«t'fu»*[a])	0.26%
لعنة (*curse*)	0.23%	جبري (*a boor*)	0.24%
العرب (*Arabs*)	0.23%	عاه*ة (*bi*ch*)	0.17%
مرأة (*woman*)	0.23%	تافه (*silly*)	0.17%
كافر (*pagan*)	0.22%	كلاب (*dogs*)	0.15%

[a] transliteration of an angry act of spitting on someone

annotators while recognizing HS comments. On the other hand, it justifies our annotation guidelines for hate comments identification, where we stressed that the joint existence of abusive language and an entity cannot indicate a HS, unless the abusive language is targeting that entity.

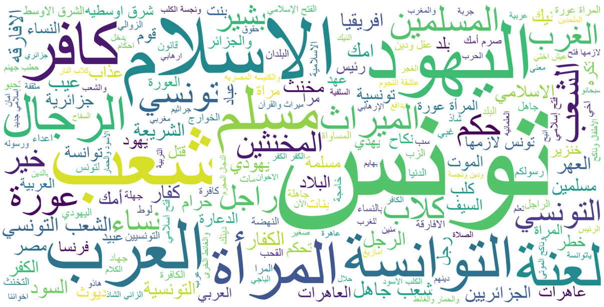

Fig. 1. Most frequent terms in hate comments.

To evaluate how distinctive are the vocabulary of our dataset with respect to each class category, we conducted word-class correlation calculations. First, we calculated the Pointwise Mutual Information (PMI) for each word towards its relevant category such that, for a word w and a class c, PMI is calculated as in Eq. 1.

$$PMI(w,c) = log(P_c(w)/P_c) \tag{1}$$

Where $P_c(w)$ denotes the appearance of the word w in the comments of the class c, while P_c refers to the number of comments of the class c.

$$HtS(w) = PMI(w,hate) - PMI(w,normal) \tag{2}$$

$$AbS(w) = PMI(w,abusive) - PMI(w,normal) \tag{3}$$

Then, to decide whether the words under the Abusive/Hate classes are discriminating, their correlation with the Normal class should be identified as well [12]. This is done by assigning a hate score (HtS) and an abusive score (AbS) for each of the most/least words under Hate and Abusive classes. Both scores indicate the difference of the PMI value of a word w under a Abusive/Hate category and its PMI value with the Normal category. The formula to calculate HtS and AbS is given in Eqs. 2 and 3.

Table 5. HtS score for most/least hateful words

Most Hate	**HtS**	**Least Hate**	**HtS**
كافر (*pagan*)	3.53	عجبني (*liked it*)	-1.13
اليهود (*Jews*)	2.08	الدولة (*the state*)	-1.04
لعنة (*curse*)	2.06	قانون (*law*)	-0.99
عورة (*private parts*)	1.51	الميراث (*inheritance*)	-0.38
شعب (*people*)	1.49	صادق (*honest*)	-0.35
الاسلام (*Islam*)	0.89	طفلة (*girl*)	-0.35
العرب (*Arabs*)	0.88	باهية (*nice*)	-0.35
تونس (*Tunisia*)	0.67	حاب (*I'd like*)	-0.35
التونسي (*Tunisian*)	0.36	عقبة (*obstacle*)	-0.35
المرأة (*woman*)	-0.08	احترام (*respect*)	-0.12

Table 6. AbS score for most/least abusive words

Most Abusive	**AbS**	**Least Abusive**	**AbS**
عاه*ة (*bi*ch*)	4.01	الرجل (*man*)	-3.43
مي*ون (*faggot*)	3.60	الدولة (*the state*)	-2.16
ع**ة (*Fu*k you!*)	3.36	نحبو (*we like*)	-1.16
كلب (*dog*)	2.78	فهمتش (*understand*)	-0.96
جبري (*a boor*)	2.54	سيناريو (*scenario*)	-0.70
تفوه (≪*t'fu*≫)	2.47	قلبك (*your heart*)	-0.70
لعنة (*curse*)	1.97	بالكلام (*by words*)	-0.70
تافه (*silly*)	1.53	الرئيس (*president*)	-0.70
كلاب (*dogs*)	1.32	الرجولية (*braveness*)	-0.37
تونس (*Tunisia*)	-0.07	مخو (*his mind*)	-0.15

It could be observed from Tables 5 and 6 that HtS and AbS scores for the most hateful and abusive words are positive indicating that they appear significantly under Hate and Abusive categories. In contrast, HtS and AbS scores for the least abusive/hate words are negative which emphasizes their appearance within Normal comments more than abusive/hate ones. On the other hand, given the specificity of the AS and HS used in Arabic, it is common to involve named entities such as locations, persons or organizations while disgracing, dehumanizing certain individuals or groups; this justifies why the country name "تونس" (*Tunisia*) has a small HtS and negative AbS scores as this word can be among the most abusive/hate words, yet, it is naturally used in normal contexts. Similarly, the word "المرأة" (*woman*) was found among the most hateful words but it had a negative hate score HtS. This is because this word is usually used when attacking women within hate contexts, whilst it can be mentioned in normal comments as well.

5 Annotation Evaluation

The annotation evaluation is based on the study of [14]. We used observed agreement A_0, all categories are equally likely S and Cohen's kappa as agreement without chance correction. For agreement with chance correction we used Krippendorff's α.

5.1 Agreement Without Chance Correction

Observed agreement A_0 is defined as the proportion of the agreed annotations out of the total number of annotations [14]. For our three annotators, the A_0 value was found of 81.82%. On the other hand, Pairwise Percent Agreement values between each pair of the three annotators are 97.963%, 83.11% and 82.563% (Table 7). Nevertheless, as observed agreement and Pairwise Percent Agreements are usually criticized for their inability to account for chance agreement [16]. Therefore, to take into account the chance agreement described by [14], we considered that all the categories are equally likely and computed the S coefficient which measures if the random annotations follow a uniform distribution in the different categories, in our case: three. With S having a high value of 72.73%, it could be said that, for an agreement constant observation, the coefficient S is not sensitive to the distribution of the elements in the categories.

Table 7. Pairwise Percent Agreement (PRAM) and pairwise Cohen's K results

Annotators	PRAM	Cohen's K
1 & 2	97.963%	0.961
1 & 3	83.11%	0.638
2 & 3	82.563%	0.624

Cohen's kappa (Cohen's K) [13] is another metric that also considers the chance agreement. It represents a correlation coefficient ranged from −1 to +1, where 0 refers to the amount of agreement that can be expected from random chance, while 1 represents the perfect agreement between the annotators. As it can be seen from Table 7, the agreement values between annotators 1 & 2 and 2 & 3 are moderate while the agreement between annotators 1 & 3 is substantial. It could be noted that, A_0, S and Cohen's K values obtained based on the annotations of our dataset, are high and show a little bias. Nevertheless, they put, on the same level, very heterogeneous categories: two minority but significant which are Abusive and Hate categories, and a non significant majority which is the Normal category.

Indeed, the categories were found unbalanced (Table 3). Here, we can observe that, despite the strong agreement on the prevailing category, the coefficients seem to be very sensitive to the disagreements over the minority categories. Thus, to make sure that the calculated coefficients for the three categories, reflect a significant agreement on the two minority categories: Abusive and Hate, we used a weighted coefficient (Inter-annotator agreement) which gives more importance to certain disagreements rather than treating all disagreements equally, as it is the case in A_0, S and Cohen's K [14].

5.2 Inter-Annotator Agreement (IAA)

According to [14], weighted coefficients make it possible to give more importance to certain disagreements. Thus, Inter-Annotator Agreement (IAA) measures can estimate the annotation reliability to a certain extent, on the assigned category. The kind of extent is determined by the method chosen to measure the agreement. For annotation reliability, Krippendorff's α has been used in the vast majority of the studies. Krippendorff's α is based on the assumption that expected agreement is calculated by looking at the overall distribution of judgments regardless of the annotator who produced those judgments. Based on Krippendorff's α, the annotation is considered: (a) Good: for any data annotation with an agreement in the interval [0.8, 1], (b) Tentative: for any data annotation with an agreement in the interval [0.67, 0.8] or (c) Discarded: for any data annotation where agreement is below 0.67. For T-HSAB dataset, the obtained Krippendorff's α was 75% which indicates the agreement on the minority categories without considering the majority category.

5.3 Discussion

The agreement measures with/without chance correlation have shown a clear agreement about the categories Normal and Abusive (Table 7). This was emphasized through our detailed study of the annotation results as the three annotators annotated abusive comments in the same way. Indeed, with the annotators following the annotation guidelines, they decided that a comment is abusive if it contains an abusive word of the Tunisian dialect. Hence, only few conflict comments (36 comments) where observed. These conflicts are mainly encountered in comments having no explicit abusive words where the annotator has to analyze the whole meaning to associate the comment with an abusive judgment.

On the other hand, more disagreement is observed when it comes to the Hate category (Table 7) and it is mainly related to the annotators' background knowledge, their personal taste and personal assumptions. The comments related to the religion topic, for instance, can be judged as Hate or not Hate according to the annotator believes. As it is seen from the examples in Table 8, where M and F denote Male and Female, respectively, the conflicts are occurred with

Table 8. Examples of annotators conflicts over the Hate category

Comment	**A.#1(M)**	**A.#2(F)**	**A.#3(M)**
حقيرة نحنو مسلمين يا بهلوانية (Hey scummy girl we are Muslims you clown)	Abusive	Abusive	Hate
قبح الله وجه كل من أيد هذا القانون الوسخ **المقزم لبلاد الإسلام بلاد الزيتونة والقروان** (God cursed all those who supported this nasty law which demeans our Muslim country, the country of Al-Zaytoonah and Kairouan)	Hate	Hate	Abusive

regard to the legalization of gender equality in inheritance. This indicates the sensitivity of this subject and reflects the divergence of opinions on social media towards such hot debates within the Tunisian community.

Another observation, is that the conflicts are not related to the annotator's gender. Indeed, despite that annotator 1 & 2 are from different genders, they achieved the highest Pairwise Percent Agreement) and pairwise Cohen's K results. Finally, based on the deduced value of Krippendorff's α, we can conclude that T-HSAB is a reliable dataset [14].

6 Classification Performance

T-HSAB dataset was used for the AS/HS detection within two experiments:

1. Binary classification: comments are classified into Abusive or Normal. This requires merging the Hate class instances with the Abusive ones.
2. Multi-class classification: comments are classified as Abusive, Hate or Normal.

For experiments setup, we first randomized the order of the comments in the dataset, then, we filtered out the Tunisian stopwords, then split the dataset into a training and a test set where 80% of the comments formed the training set. The comments distribution among the three categories in Training and Test sets is shown in Table 9, where Exp. denotes the experiment's number.

Table 9. Training and Test sets of T-HSAB

Exp.	Training			Test		
	Abusive	Normal	Hate	Abusive	Normal	Hate
1	1,770	3,061	–	434	774	–
2	889	3,061	881	237	774	197
Total	4,831			1,208		

We employed two supervised classifiers: SVM [15] and NB from NLTK [17]. Both classifiers are trained with several n-gram schemes: unigrams (uni), unigrams+bigrams (uni+bi) and unigrams+bigrams+trigrams (uni+bi+tri).

Term frequency (TF) weighting was employed to reduce the features size. Among several runs with various n-gram schemes and TF values, we selected the best results to be listed in Table 10, where the Precision, Recall, F-measure and Accuracy are referred to as P., R., F1 and Acc., respectively.

Table 10. Classification results over T-HSAB

Classes	Algorithm	Features	P.(%)	R.(%)	F1(%)	Acc.(%)
2	NB	uni+bi(TF≥2)	**93.5**	**91.5**	**92.3**	**92.9**
	SVM	uni	76.4	73.8	74.7	77.7
3	NB	uni+bi(TF≥2)	**89.5**	**79.8**	**83.6**	**87.9**
	SVM	uni	66.5	59.9	62.2	73.9

As it can be observed in Table 10, NB classifier performed, remarkably, better than SVM for both binary and multi-class classification experiments. This could be attributed to the fact that NB variant from NLTK is implemented as a multinomial NB decision rule together with binary-valued features [17]. This explains its effectiveness in dealing with our feature vectors that are formulated from binary values indicating the presence/absence of n-gram schemes.

7 Conclusion

This paper introduced a Tunisian dataset for Abusive speech and Hate speech known as T-HSAB. T-HSAB is the first public Tunisian dataset with the objective to be a benchmark dataset for automatic detection of online Tunisian toxic contents. To build our dataset, Tunisian comments were harvested from different social media platforms and 3 annotators conducted the manual annotation following an annotation guideline. The final version of the dataset combined 6,039 comments. While achieving high values of agreement without chance correction and inter-annotator agreement indicated the reliability of T-HSAB dataset, the agreement between annotators is still an issue when it comes to identify HS. This is attributed to the fact that, HS annotation does not only rely on rules but also is related to the annotators' background knowledge, their personal tastes and assumptions. The machine learning-based classification experiments conducted with NB and SVM classifiers to classify the AS/HS content in T-HSAB dataset, indicated the outperformance of NB over SVM for both binary and multi-class classification of AS and HS comments. T-HSAB was made publicly available to intensify the progress in this research field. In addition, a lexicon of Tunisian abusive words and a lexicon of hate words will be built based on the annotated comments. Both lexicons will be made publicly available. A natural future step would involve building further publicly-available Abusive speech and Hate Speech datasets for Algerian dialect and Moroccan dialect as Algeria and Morocco have an identical linguistic situation and share socio-historical similarities with Tunisia [18].

References

1. Skandrani, H., Triki, A.: Trust in supply chains, meanings, determinants and demonstrations: a qualitative study in an emerging market context. Qual. Res. Int. J. **14**(4), 391–409 (2011)
2. Nockleby, J.T.: Hate speech. In: Levy, L.W., Karst, K.L., et al. (eds.) Encyclopedia of the American Constitution, 2nd edn. Macmillan, Detroit (2000)
3. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: a typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online, pp. 78–84. Association for Computational Linguistics, Vancouver (2017)
4. Stevens, P.B.: Ambivalence, modernisation and language attitudes: French and Arabic in Tunisia. J. Multiling. Multicult. Dev. **4**(2–3), 101–114 (1983)
5. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in social networks: a survey on multilingual corpus. J. Comput. Sci. Inf. Technol. (CS IT) **9**(2), 83–100 (2019)
6. Abozinadah, E., Mbaziira, A., Jones, J.: Detection of abusive accounts with Arabic Tweets. Int. J. Knowl. Eng. **1**(2), 113–119 (2015)
7. Alakrota, A., Murray, L., Nikolov, N.: Dataset construction for the detection of anti-social behaviour in online communication in Arabic. J. Procedia Comput. Sci. **142**, 174–181 (2018)
8. Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on Arabic social media. In: Proceedings of the First Workshop on Abusive Language Online, pp. 52–56. Association for Computational Linguistics, Vancouver (2017)
9. Albadi, N., Kurdi, M., Mishra, S.: Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 69–76. IEEE, Barcelona (2018)
10. Al-Ajlan, M., Ykhlef, M.: Optimized Twitter cyberbullying detection based on deep learning. In: Proceedings of the 21st Saudi Computer Society National Computer Conference (NCC), pp. 52–56. IEEE, Riyadh (2018)
11. Mulki, H., Haddad, H., Bechikh Ali, C., Alshabani, H.: L-HSAB: a Levantine Twitter corpus for hate speech and abusive language. In: Proceedings of the Third Workshop on Abusive Language Online. Association for Computational Linguistics, Florence (2019)
12. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online, pp. 11–20. Association for Computational Linguistics, Brussels (2018)
13. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Comput. Linguist. **34**(4), 555–591 (2008)
14. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Psychol. Bull. **70**(4), 213–220 (1968)
15. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011)
16. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochem. Med. **22**(3), 276–282 (2012)
17. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text With the Natural Language Toolkit. O'Reilly Media Inc., Sebastopol (2009)
18. Harrat, S., Meftouh, K., Smaili, K.: Maghrebi Arabic dialect processing: an overview. J. Int. Sci. Gen. Appl. **1**, (2018)