L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language

Hala Mulki[*§], **Hatem Haddad**[†§], **Chedi Bechikh Ali**[**] **and Halima Alshabani**[***]

[*]Department of Computer Engineering, Konya Technical University, Turkey
[†]RIADI Laboratory, National School of Computer Sciences, University of Manouba, Tunisia
[**]LISI Laboratory, INSAT, Carthage University, Tunisia
[***]Department of Computer Engineering, Kırıkkale University, Turkey
[§]iCompass Consulting, Tunisia
halamulki@selcuk.edu.tr,haddad.Hatem@gmail.com
chedi.bechikh@gmail.com,halima.alshabani@gmail.com

Abstract

Hate speech and abusive language have become a common phenomenon on Arabic social media. Automatic hate speech and abusive detection systems can facilitate the prohibition of toxic textual contents. The complexity, informality and ambiguity of the Arabic dialects hindered the provision of the needed resources for Arabic abusive/hate speech detection research. In this paper, we introduce the first publicly-available Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset with the objective to be a benchmark dataset for automatic detection of online Levantine toxic contents. We, further, 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 has been 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.

1 Introduction

With the freedom of expression privilege granted to social media users, it became easy to spread abusive/hate propaganda against individuals or groups. Beyond the psychological harm, such toxic online contents can lead to actual hate crimes (Matsuda, 2018). This provoked the need for automatic detection of hate speech (HS) and abusive contents shared across social media platforms.

In (Nockleby, 2000), hate speech (HS) is formally defined 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"*. HS detection can be conducted as a subtask of the abusive language detection (Waseem et al., 2017); yet, HS detection remains challenging since it requires to consider the correlation between the abusive language and the potential groups that are usually targeted by HS as per the definition of (Nockleby, 2000). Further challenges could be met when HS detection is investigated with complex, rich and ambiguous languages such as the Arabic language which combines different informal language variants known as dialects.

Compared to the increasing studies of abusive/HS detection in Indo-European languages, similar research for Arabic dialects is still very limited. This is due to the lack of the publicly-available resources needed for abusive/HS detection in Arabic social media texts. Building such resources involves several difficulties in terms of data collection and annotation, especially for underrepresented dialects such as Syrian, Lebanese, Palestinian and Jordanian dialects which are all combined within the Levantine dialect.

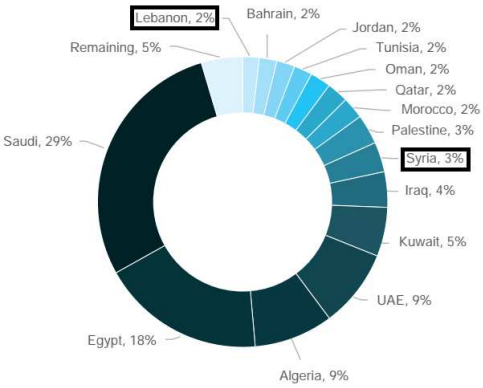

Figure 1: Twitter usage in the Arab region, 2017

Although Levantine is not among the top-ranking Arabic dialects used on Twitter (Salem) (see Figure 1), the volatile political/social atmosphere in Levantine-speaking countries, have been always associated with intensive debates; a considerable part of which took place on Twitter. With

Proceedings of the Third Workshop on Abusive Language Online, pages 111–118
Florence, Italy, August 1, 2019.

Study	Type	Platform	Size	Language
(Alakrota et al., 2018)	offensive	Youtube	16K	Egyptian, Iraqi and Libyan
(Mubarak et al., 2017)	obscene, offensive, and clean	Twitter	1.1K and 32K	MSA/DA
(Albadi et al., 2018)	religious hate, not hate	Twitter	6.6K	Arabic
(Al-Ajlan and Ykhlef, 2018)	bullying, nonbullying	Twitter	20K	Arabic

Table 1: Arabic Hate/Abusive Speech Presented Datasets

multiple opposite parties being involved in such debates, the relevant tweets tend to contain abusive and HS content. Thus, we believe that providing a Levantine abusive/HS dataset would support the research of automatic detection of abusive/HS in underrepresented Arabic dialects.

In this study, we introduce the first **L**evantine **H**ate **S**peech and **AB**usive (**L-HSAB**) Twitter dataset. Our dataset combines 5,846 tweets labeled as Hate, Abusive or Normal[1]. With the objective of providing a reliable, high quality benchmark dataset, and unlike the previous studies whose proposed Arabic corpora lack the needed annotation evaluation, we provide a comprehensive quantitative evaluation for L-HSAB. This was done through using agreement without chance correction and Inter-annotator agreement (IAA) reliability measures. In addition, our dataset was examined as a benchmark abusive/HS dataset by subjecting it to supervised machine learning experiments conducted by SVM and NB classifiers.

2 Dialectal Arabic Hate/Abusive Speech

As seeking to propose a new dialectal Arabic dataset for abusive and HS, we opted to review the Arabic abusive 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. According to (Al-Hassan and Al-Dossari, 2019), the toxic online content on social media can be classified into: Abusive, Obscene, Offensive, Violent, Adult content, Terrorism and Religious hate speech. Table 1 lists a summary of the proposed abusive/HS datasets while a detailed review of these datasets is provided below.

In (Alakrota et al., 2018), the authors investigated the offensive language detection in Youtube comments. A dataset of 16K Egyptian, Iraqi and Libyan comments was created. Three annotators from Egypt, Iraq and Libya were asked to annotate the comments as: offensive, inoffensive and neutral. The annotation evaluation measurements for the Egyptian and Lybian annotators were 71% and 69.8% for inter-annotator agreement and Kappa metric, respectively. With Support Vector Machines (SVM) algorithm applied for classification, the best achieved F-measure was 82%.

(Mubarak et al., 2017) proposed two datasets: a Twitter datasets of 1,100 dialectal tweets and a 32K inappropriate comments dataset collected from a popular Arabic news site. To support the offensive content detection, the authors relied on on common patterns used in offensive and rude communications to construct a list of obscene words and hashtags. Three Egyptian annotators annotated the data as obscene, offensive, and clean. With only obscene instances considered, the average inter-annotator agreement was 85% for the Twitter dataset and 87% for the comments dataset.

The religious HS detection was investigated in (Albadi et al., 2018) where a multi-dialectal Arabic dataset of 6.6K tweets was introduced. The annotation task was assigned to 234 different annotators; each of which was provided with an identification of the religious groups targeted by HS such as Muslims, Jews, Christians, Sunnis, Shia and so forth. Out of the resulting annotated corpus, three Arabic lexicons were constructed using chi-square, Pointwise Mutual Information (PMI) and Bi-Normal Separation (BNS) scoring methods. Each lexicon combined the terms commonly used in religious discussions accompanied with scores representing their polarity and strength. As an annotation evaluation, the authors indicated that 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 were targeted by the religious HS. The proposed corpus was further examined 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

[1] will be made publicly available on github.

word embedding was the best-performing model where it achieved an F-measure of 77%.

Another type of HS was tackled by (Al-Ajlan and Ykhlef, 2018) where the authors presented a Twitter dataset for bullying detection. A dataset of 20K multi-dialectal Arabic tweets was collected and annotated manually with bullying and non-bullying labels. In their study, neither inter-rater agreement measures nor classification performances were provided.

3 L-HSAB Dataset

L-HSAB can be described as a political dataset since the majority of tweets was collected from the timelines of politicians, social/political activists and TV anchors. In the following subsections, we provide a qualitative overview of the proposed dataset, while a detailed quantitative analysis is presented in Section 5.

3.1 Data Collection and Processing

The proposed dataset was constructed out of Levantine tweets harvested using Twitter API[2]. We collected the tweets based on multiple queries formulated from the potential entities that are usually targeted by abusive/hate speech such as "اللاجئين" (*refugees*), "البنات" (*females*), "العرب" (*Arabs*), "الدروز" (*Druze*), etc. In addition, some user timelines (verified or having more than 100k followers) which belong to certain politicians, social/political activists and TV anchors, were adopted as data resources, since their tweets and tweets' replies are rich of the abusive/hate content. Aiming to maximize the size of the abusive/HS tweets, relevant to hot debates and major events, we scraped tweets posted within the time period: March 2018- February 2019.

Initially, we retrieved 57,058 tweets; to cope with goal of the paper which is to provide a Levantine dataset, we manually reduced the non-Levantine tweets. In addition, we filtered out the non-Arabic, non-textual, promoted and duplicated instances. Thus, we ended up with 6,000 tweets, written in the Levantine dialect (Syrian and Lebanese).

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

[2] http://www.tweepy.org

3.2 Annotation Guidelines

The annotation task requires labeling the tweets of L-HSAB dataset as Hate, Abusive or Normal. Based on the definition of hate and abusive speech stated in the introduction, differentiating HS from abusive is quite difficult and is usually prone to personal biases; which, in turn, yields low inter-rater agreement scores (Waseem et al., 2017; Schmidt and Wiegand, 2017). 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 abusive 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 tweets are those instances which have no offensive, aggressive, insulting and profanity content.
- Abusive tweets are those instances which combine offensive, aggressive, insulting or profanity content.
- Hate tweets are those instances that: (a) contain an abusive language, (b) dedicate the offensive, insulting, aggressive speech towards a specific person or a 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 2 lists the relevant examples to each class.

Label	Example
Normal	أحلى شي دولة طفرانة بدّا تشلّح شعب مفلس The nicest thing is that a government in an abject poverty loots its own bankrupt people
Abusive	انت كعب صرمايتي القديمه اطهر من نيعك I consider the bottom of my old nasty shoes more clean than your own mouth
Hate	أصلن خلفة البنات بتجيب العار To have a girl kid brings disgrace

Table 2: Tweet examples of the annotation labels

3.3 Annotation Process

The annotation task was assigned to three annotators, one male and two females. All of them are Levantine native speakers and at a higher educational level (Postdoc/PhD).

Besides the previous annotation guidelines, and based on the domain and context of the proposed dataset, we had the annotators aware of the ethnic origin, religion, and the geographic region represented by each political party. Moreover, we provided them with the nicknames usually used to refer to certain political parties, minorities and ethnic/religion groups. For example, "تيار المستقبل" (*Future Movement Party*), which represents the Sunnis ethnic group, is usually called by its nickname "تيار المستهبل" (*Dumb Party*) in hate speech contexts. More examples are shown in Table 3. Having all the annotation rules setup, we asked the three annotators to label the 6,000 tweets as Normal, Abusive or Hate. For the whole dataset, we received a total of 18,000 judgments. By exploring the annotations, we faced three cases:

1. Unanimous agreement: the three annotators annotated a tweet with the same label. This was encountered in 4,222 tweets.
2. Majority agreement: two out of three annotators agreed on a label of a tweet. This was encountered in 1,624 tweets.
3. Conflicts: each annotator annotated a tweet differently. They were found in 154 tweets.

Nickname	Entity	Ethnic/Religion
تيار المستهبل	تيار المستقبل	السنة (*Sunnis*)
العونية	تيار الوطني الحر	الموارنة (*Maronites*)
عرب الشمال		السوريين (*Syrians*)
سكان الضاحية		الشيعة (*Shia*)
أهل الجبل		الدروز (*Druze*)

Table 3: A sample of the entities targeted by HS

Annotation Case	# Tweets
Unanimous agreement	4,222
Majority agreement (2 out of 3)	1,624
Conflicts	154

Table 4: Summary of annotation statistics

After excluding the tweets that have 3 different judgments, the final version of L-HSAB composed of 5,846 tweets. A summary of the annotation statistics is presented in Table 4.

4 Annotation Results

With the annotation process accomplished, we decided the final label of each tweet in the dataset considering the annotation cases in Section 3.3. For tweets 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. Consequently, we got 3,650 normal tweets, 1,728 abusive and 468 hate tweets. A detailed review of the statistics of L-HSAB final version is provided in Table 5 where Avg-S-L denotes the average length of tweets in the dataset, calculated based on the number of words in each tweet.

	Normal	Abusive	Hate
# Tweets	3,650	1,728	468
Avg-S-L	9	7	10
Word Count	31,598	11,938	4,380
Vocabulary	14,064	7,059	2,971
Ratio	62.43%	29.55 %	8.00 %

Table 5: Tweets distribution across 3 classes

Hate	Dist.	Abusive	Dist.
كلب (*dog*)	1%	هوا (*sh*t*)	1.58%
كلاب (*dogs*)	0.98%	كول (*swallow*)	1.52%
لبنان (*Lebanon*)	0.55%	كلب (*dog*)	0.97%
قطر (*Qatar*)	0.55%	حمار (*donkey*)	0.59%
سوري (*Syrian*)	0.39%	خراس (*chup*)	0.52%
العرب (*Arabs*)	0.39%	يلعن (*damn*)	0.45%
شعب (*people*)	0.37%	لبنان (*Lebanon*)	0.39%
ولاك (*jerk*)	0.37%	بشار (*Bashar*)	0.37%
حزب (*party*)	0.34%	واطي (*mean*)	0.36%
معراب (*a region*)	0.30%	صرماية (*shoe*)	0.30%

Table 6: Distribution of ten most frequent terms

As seeking to identify the words commonly used within hate and abusive speech contexts, we investigated the lexical distribution of the dataset words across both hate and abusive classes. Therefore, we subjected L-HSAB to further normalization, where we removed stopwords based on a manually built Levantine stopwords list. Later, we constructed a visualization map for the most frequent occurring words/terms under each of Hate

and Abusive categories (Figure 2 and Figure 3). The ten most frequent words and their frequencies in each class are reviewed in Table 6, where Dist. denotes the word's distribution in a specific class.

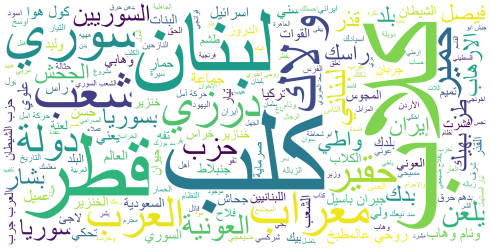

Figure 2: Most frequent terms in hate tweets

Figure 3: Most frequent terms in abusive tweets

As it can be seen from Table 6, Figure 2 and Figure 3, both Hate and abusive classes have many terms in common. 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 annotators while recognizing HS tweets. On the other hand, it justifies our annotation guidelines for hate tweets 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.

In order 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 equation 1:

$$PMI_c(w) = log(P_c(w)/P_c) \quad (1)$$

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

$$HtS(w) = PMI(w, hate) - PMI(w, normal) \quad (2)$$

$$AbS(w) = PMI(w, abusive) - PMI(w, normal) \quad (3)$$

Then, to decide whether the words under the hate/abusive classes are discriminating, their correlation with the Normal class should be identified as well (de Gibert et al., 2018). This is done by assigning a hate score (HtS) and an abusive score (AbS) for each of the most/least words under Hate and Aabusive classes. Both scores indicate the difference of the PMI value of a word w under a hate/abusive category and its PMI value with the Normal category. The formula to calculate HtS and AbS is given in equations 2 and 3.

Most hate	HtS	Least hate	HtS
كلاب (*dogs*)	5.85	الوزير (*minister*)	-2.30
ولاك (*jerk*)	4.86	حق (*right*)	-2.23
كلب (*dog*)	4.77	معالي (*highness*)	-2.07
معراب (*a region*)	3.96	شكرا (*thanks*)	-1.58
سوري (*Syrian*)	2.15	العهد (*promise*)	-1.09
العرب (*Arabs*)	1.39	وطن (*homeland*)	-1.00
حزب (*party*)	1.36	العربية (*Arabic*)	-1.00
شعب (*people*)	1.17	الإعلام (*media*)	-0.96
قطر (*Qatar*)	0.42	حكومة (*government*)	-0.31
لبنان (*Lebanon*)	-0.06	كبير (*big*)	-0.28

Table 7: HtS score for most/least hateful words

Table 7 and Table 8 list the HtS and AbS scores calculated for the 10 most and least words under hate/abusive category against the normal category.

Most abusive	AbS	Least abusive	AbS
حمار (*donkey*)	4.25	فخامة (*excellency*)	-1.37
يلعن (*damn*)	3.99	الجمهورية (*republic*)	-1.27
كول (*swallow*)	3.82	نعيم (*bless*)	-1.27
واطي (*mean*)	3.76	مبروك (*congrats*)	-1.20
كلب (*dog*)	3.65	خير (*good*)	-1.20
هوا (*sh*t*)	3.63	حليف (*ally*)	-0.86
خراس (*chup*)	3.43	قصة (*story*)	-0.46
صرماية (*shoe*)	2.89	طبيعي (*natural*)	-0.64
بشار (*Bashar*)	0.79	يحتاج (*need*)	-0.55
لبنان (*Lebanon*)	-1.49	المنطقة (*region*)	-0.50

Table 8: AbS score for most/least abusive words

It could be observed from Table 7 and Table 8 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 hate/abusive words are negative which emphasizes their appearance within Normal tweets more than hate/abusive ones. On the other hand,

given the specificity of the HS, used in our dataset, it is common to involve named entities such as location, person or a party name while disgracing, dehumanizing certain entities; this justifies why the country name "لبنان" (*Lebanon*) has a negative HtS and AbS scores as this word can be among the most hateful/abusive words, yet, it is naturally used in Normal tweets.

5 Annotation Evaluation

We conducted the annotation evaluation following the study of (Artstein and Poesio, 2008). Observed agreement A_0, All categories are equally likely (S) and Cohen's kappa as agreement without chance correction measures, were adopted for evaluation. For agreement with chance correction, we used Krippendorff's α.

5.1 Agreement Without Chance Correction

Observed agreement A_0 is the simplest measure of agreement between annotators. It is defined as the proportion of the agreed annotations out of the total number of annotations (Artstein and Poesio, 2008). For our annotations, A_0 is 81.5%; while Pairwise Percent Agreement Measure (PRAM) values between each pair of the three annotators are 78.43%, 87.24% and 78.77% (Table 9). However, observed agreement and Pairwise Percent Agreements are criticized for their inability to account for chance agreement (McHugh, 2012). Therefore, to take into account the chance agreement described in (Artstein and Poesio, 2008), 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 (3) categories. With S value deduced as high as 72.3%, it could be said that for an agreement constant observation, the coefficient S is not sensitive to the elements distribution across the categories.

Annotators	PRAM	Cohen's K
1 & 2	78.43%	0.599
1 & 3	87.24%	0.758
2 & 3	78.77%	0.594

Table 9: PRAM and pairwise Cohen's K results

Cohen's kappa (Cohen's K) (Cohen, 1960) 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 9, the agreement values between annotators 1 & 2 and 2 & 3 are moderate while the agreement between annotators 1 & 3 is substantial.

It is 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 as the categories were found highly unbalanced (Table 5). Here, we can observe that, despite the strong agreement on the prevailing category, the coefficients seem to be very sensitive to disagreements over the minority categories. Thus, to ensure 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 (Artstein and Poesio, 2008).

5.2 Inter-Annotator Agreement (IAA)

According to (Artstein and Poesio, 2008), weighted coefficients can give more importance to certain disagreements. 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, the agreement coefficient 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 these judgments. Based on Krippendorff's α value, 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 with an agreement below 0.67. For L-HSAB dataset, the obtained Krippendorff's α was 76.5% which indicates the agreement on the minority categories without considering the majority category.

5.3 Discussion

The agreement measures with/without chance correlation show a clear agreement about the categories Normal and Abusive. Indeed, our detailed study of the annotation results revealed that the three annotators identified abusive tweets in the same way while conflicts were encountered in tweets having an ironic content. On the other hand, more disagreement is observed when it comes to the Hate category and it is mainly related to the annotator's background knowledge, their personal taste and personal assumptions. In addition, the conflicts are not related to the annotator's gender; since, although annotator 1 & 3 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 which is 76.5%, we can conclude that L-HSAB is a reliable HS and abusive dataset.

6 Classification Performance

L-HSAB dataset was used for the abusive/HS detection task within two experiments:

1. Binary classification: tweets are classified into Abusive or Normal. This requires merging the Hate class with the Abusive one.

2. Multi-class classification: tweets are classified into Abusive, Hate or Normal.

We filtered out the Levantine stopwords, then split the dataset into a training and a test set as it is shown in Table 10, where Classes denotes the number of classification classes.

Classes	Training			Test		
	Abusive	Normal	Hate	Abusive	Normal	Hate
2	1,708	2,968	-	488	682	-
3	1,369	2,968	339	359	682	129
Total	4,676			1,170		

Table 10: Training and Test sets of L-HSAB

We employed two supervised classifiers: SVM (Chang and Lin, 2011) and NB from NLTK (Bird et al., 2009). Both classifiers were trained with several n-gram schemes: unigrams, unigrams+bigrams and unigrams+bigrams+trigrams. Term frequency (TF) weighting was employed to reduce the features size according to two predefined frequency thresholds: 2 and 3. Among several runs with various n-gram schemes and TF values, we selected the best results to be listed in Table 11, where the classification algorithm, Precision, Recall, F-measure and Accuracy are referred to as Alg., P., R., F1 and Acc., respectively.

Classes	Alg.	P.(%)	R.(%)	F1(%)	Acc.(%)
2	NB	**90.5**	**89.0**	**89.6**	**90.3**
	SVM	84.7	81.1	82.0	83.2
3	NB	**86.3**	**70.8**	**74.4**	**88.4**
	SVM	74.0	64.2	66.8	78.6

Table 11: Classification results over L-HSAB

As it can be observed in Table 11, NB classifier performed better than SVM in both classification experiments. This is due to the fact that NB from NLTK is implemented as a multinomial NB decision rule together with binary-valued features (Bird et al., 2009). This explains its effectiveness while dealing with our feature vectors that were formulated from binary values denoting the presence/absence of n-gram schemes.

7 Conclusion

In this paper, we introduced L-HSAB, the first publicly available Levantine dataset for HS and abusive Language. The proposed dataset was aimed to be a benchmark dataset for automatic detection of online Levantine toxic contents. To build L-HSAB, we crawled Twitter for tweets while 3 annotators manually labeled the tweets following a set of rules. The dataset combined 5,846 tweets with 3 categories: Normal, Abusive and Hate. High values were achieved in agreement without chance correction and inter-annotator agreement which indicates the reliability of annotations. However, the agreement between annotators remains an issue when it comes to identify HS. This is because HS annotation does not only rely on rules, but it is also related to the annotators' background knowledge, their personal tastes and assumptions. L-HSAB was subjected to machine learning-based classification experiments conducted using NB and SVM classifiers. The results indicated the outperformance of NB over SVM in both binary and multi-class classification experiments. A natural future step would involve building publicly-available HS and abusive datasets for other underrepresented Arabic dialects such as Tunisian and Gulf.

References

Monirah A. Al-Ajlan and Mourad Ykhlef. 2018. Optimized twitter cyberbullying detection based on deep learning. In *Proceedings of the 21st Saudi Computer Society National Computer Conference (NCC)*, pages 52–56.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science & Information Technology (CS & IT)*, 9(2):83–100.

Azalden Alakrota, Liam Murray, and Nikola S.Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1(20):37–46.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20.

Mari J Matsuda. 2018. Public response to racist speech: Considering the victim's story. In *Words that wound*, pages 17–51. Routledge.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

John T. Nockleby. 2000. *Hate Speech*, volume 1. Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al. New York: Macmillan, New York: Macmillan.

Fadi Salem. Social media and the internet of things towards data-driven policymaking in the arab world: potential, limits and concerns.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.