

BiSAL- A Bilingual Sentiment Analysis Lexicon to Analyze Dark Web Forums for Cyber Security

Khalid Al-Rowaily^a, Muhammad Abulaish, SMIEEE^{b,*}, Nur Al-Hasan Haldar^c, Majed Al-Rubaian^a

^aCollege of Computer and Information Sciences, King Saud University, Riyadh, KSA

^bDepartment of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India

^cCenter of Excellence in Information Assurance, King Saud University, Riyadh, KSA

Abstract

In this paper, we present the development of a Bilingual Sentiment Analysis Lexicon (BiSAL) for cyber security domain, which consists of a Sentiment Lexicon for ENGLISH (SentiLEN) and a Sentiment Lexicon for ARABIC (SentiLAR) that can be used to develop opinion mining and sentiment analysis systems for bilingual textual data from dark web forums. For SentiLEN, a list of 279 sentiment bearing English words related to cyber threats, radicalism, and conflicts are identified and a unifying process is devised to unify their sentiment scores obtained from four different sentiment data sets. Whereas, for SentiLAR, sentiment bearing Arabic words are identified from a collection of 2000 message posts from Alokab Web forum, which contains radical contents. The SentiLAR provides a list of 1019 sentiment bearing Arabic words related to cyber threats, radicalism, and conflicts along with their morphological variants and sentiment polarity. For polarity determination, a semi-automated analysis process by three Arabic language experts is performed and their ratings are aggregated using some aggregate functions. A Web interface is developed to access both the lexicons (SentiLEN and SentiLAR) of BiSAL data set online, and a beta version of the same is available at <http://www.abulaish.com/bisal>.

Keywords: Sentiment analysis lexicon, Sentiment lexicon for English, Sentiment lexicon for Arabic, Cyber security, Dark Web forum

1. Introduction

Sentiment analysis aims to identify subjective information components from textual data and determine their sentiment polarity. Though a number of sentiment analysis lexicons for English language texts for rating the polarity of sentiment bearing words exist in literature, a major bottleneck is created in the process of sentiment analysis due to the variability and complexity of multilingual documents and the heterogeneous nature of the existing sentiment lexicons. On the other hand, no sentiment analysis lexicon exists for Arabic language texts due to which sentiment analysis over Arabic language texts seems almost an infeasible task. The primitive task in the process of sentiment analysis involves determining the polarity of opinion bearing words in text documents. A bottleneck is created in this process due to the presence of multiple heterogeneous classifiers that produce a varying output using different polarity scales, creating issues while developing an opinion mining and sentiment analysis system [3]. For example, *AFINN* [13] uses polarity scale ranging from -5 to +5, *General Inquirer* [15] uses a polarity vector, and *SentiWordNet* [10] uses -1 to +1 as a polarity scale. In lexical-based sentiment analysis approach, a collection of related words with known sentiment scores and polarities are used for sentiment polarity determination. However, domain-specific sentiment data sets are more

effective than general-purpose sentiment data set for analyzing sentiment of a sentiment-bearing document [14]. For example, the word *shooting* has negative sentiment intensity in crime domain, whereas it bears neutral or positive sentiment in movie and sports domains. Similarly, some non-sentiment terms like *bluetooth* make positive sense for mobile communication domain.

Though, a general-purpose lexicon can be built semi-automatically from general-purpose corpus using heuristics with a set of known polarity sentiment words, there exists some limitations with general lexical methods that they can assign incorrect sentiments on the basis of sentiment terms only [16]. For example, the term *border* may have neutral association in normal usage, but it carries negative sentiment in terms of homeland security. BiSAL (Bilingual Sentiment Analysis Lexicon), which consists of SentiLEN (Sentiment Lexicon for English) and SentiLAR (Sentiment Lexicon for Arabic), is developed as a bilingual sentiment analysis lexicon through a semi-automatic process wherein each sentiment bearing word is assigned a unified sentiment score in the range of [-1, +1]. A positive score assigned to a word determines that the words subjectivity is positive, whereas a negative-scored word confirms that the word is associated with some sort of negative sentiment. The numeric value assigned to a word determines its sentiment strength. For Arabic language having no sentiment analysis lexicon, we have analyzed a corpus of message posts extracted from a dark web form, where users use cyber

*To whom correspondence should be addressed.

E-mail: mAbulaish@jmi.ac.in, Telefax: +91-11-26980014

threats, radicalism, and conflicts bearing words to share their views and experiences [18]. A list of 1019 sentiment bearing Arabic seed words along with their morphological variants and polarity scores is compiled. Whereas, for English language, we have identified a list of 279 cyber crime, border security, and terrorism related words from the department of homeland security website [2] and Ansar1 forum [11], and considered existing sentiment analysis lexicons to determine their polarity scores. We have also proposed a sentiment unifying process to unify sentiment polarity scores from four different sentiment lexical data sets. Besides sentiment analysis over Dark Web forum data, the proposed sentiment lexicons can be used to identify and cluster cyber threats, radicalism, and conflicts related digital forensic traces embedded within textual data, as attempted in [4, 6, 7].

The rest of the paper is organized as follows. Section 2 presents the procedural details of the proposed sentiment lexicon development process. Section 3 presents the experimental results for both SentiLEN and SentiLAR data sets. Finally, section 4 concludes the paper and provides future directions of work.

2. Proposed Method

In this section, we present the procedural details of the development of sentiment lexicons, including the process of seed words identification and sentiment scores determination.

2.1. Seed Words Identification

The words used in BiSAL are collected from two sources - topic-wise word list provided by the Department of Homeland Security (DHS) and a Dark Web portal. The word list provided by the DHS contains the words that are frequently associated with homeland security, terrorism, cyber security, etc. Dark Web portal is a large collection of Dark Web forums that are used to encourage violence activities and distribute dispute-related information related to cyber threats, radicalism, and conflicts in grouped manner. Dark Web portal contains 29 jihadist forums [9], among which 17 and 7 forums are in Arabic and English language, respectively, and rest of the forums are in French, German, and Russian languages [8].

A set of frequently used 370 words found in Ansar1 forum is publicly available in [11], out of which 73 are Arabic words and remaining 297 are English words. These words represent some degree of violence with multiple occurrence in different message posts. After grouping the words based on their common stems, a total number of 208 unique English seed words are identified from Ansar1 data set. The difference in number of unique seeds (208) and number of English words (297) is due to grouping of words having common seeds. For example, *kill*, *killing*, and *kills* that are available in Ansar1 data set have the common seed *kill*. Moreover, there are three words in Ansar1 data set which are abbreviations like CIA, IED, etc. and do not carry proper meaning. Hence, a total number of 205 seed words are finally chosen from Ansar1 data set for SentiLEN. Table 1 presents a partial list of English words and their frequency from Ansar1 data set.

Table 1: A partial list of English words representing cyber crimes and their frequency from Ansar1 data set

Words	Frequency
Weapons	918
Terrorists	738
Violence	1089
Brothers	1692
Militants	1680
Suicide	1148
Bomb	1317
Killing	1589
Fighting	1785
Death	1214
Injured	1008

In addition, other English words are collected from the topic-wise word list provided by the Department of Homeland Security (DHS). This list is used by NOC (National Operations Center) for the purpose of media monitoring. In this list, the words are divided into various categories depending on topic characterization. Since the goal behind the development of BiSAL is to provide a seed set of sentiment bearing Arabic and English words for cyber security domain, only six DHS topics related to security or violence (i.e., domestic security, HAZMAT and nuclear, terrorism, infrastructure study, south west border violence, and cyber security) are considered, whereas other three topics related to “health+H1N1”, “weather/disaster/emergency”, and “DHS & other agencies” are discarded. Category-wise details along with exemplar words are provided in Table 2. From first six categories, a list of 265 words is generated, which is reduced to 110 on the basis of the common stems. After filtering the stems not directly related to violence or conflict, finally 99 stems from this list are considered for SentiLEN. It may be noted that while analyzing the terms for their violence or conflict related sentiments, the n-grams ($n > 1$) are decomposed into 1-grams. Thus, 205 words are taken from Ansar1 forum and 99 words are taken from DHS. However, on analysis we found that there are 25 common stems in both of the lists, which resulted in total 279 words for SentiLEN. A summary of the relevant English words extraction process discussed above is given in Table 3.

For SentiLAR, 2000 message posts of Alokab forum, which is one of the Arabic forums available as a part of dark web portal [1] are processed using Natural Language Processing (NLP) techniques and top-ranked words are considered as the seed words. A partial list of Arabic seed words along with their frequency count is given in Table 4.

2.2. Morphological Variants Identification

An important task related to the generation of BiSAL is to compile the morphological variants of both English and Arabic seed words, which could be useful for the bilingual text processing systems in pattern matching or query expansion. Morphology is the study of internal structure of words [5]. The smallest part of a word, which has grammatical function or independent meaning, is termed as *morpheme*. For example, *attacks*, *attacked*, *attackers*, and *attacking* can be studied in terms of morphemes like *attack* followed by “s”, “ed”, “ers”, and “ing”,

Table 2: Sample DHS categories and exemplar words related to cyber crimes

Category	No. of words	Exemplar words
Domestic security	53	assassination, attack, drill, exercise, recovery, shooting, evacuation, deaths, explosion, gangs, security, threat, bomb
HAZMAT and nuclear	34	hazmat, nuclear, toxic, plume, radiation, radioactive, chemical, biological, epidemic, hazardous, incident, infection
Terrorism	55	terrorism, terror, attack, target, jihad, extremism, radicals, plot, nationalist, fundamentalism
Infrastructure study	35	collapse, subway, power
South west border violence	63	violence, gang, drug, border, gunfight, trafficking, kidnap, bust
Cyber security	25	botnet, virus, trojan, hacker, worm, scammer
Total	265	

Table 3: Statistics of English words used in SentiLEN

Data source	Total no. of words	No. of English words	No. of stems generated from Eng. words	No. of stems used in SentiLEN	No./Name of stems not used in SentiLEN	No./Name of common stems across both data sources	Total no. of stems used in SentiLEN
Ansar1	370	297	208	205	3 (CIA, Hamas, IED)	25 (attack, body, bomb, death, extreme, illegal, improve, incident, jihad, kidnap, militia, nation, nuclear, plot, police, power, radical, response, shoot, swat, target, terror, threat, violent, weapon)	(205+99-25) = 279
Department of Homeland Security (DHS)	265	265	110	99	11 (gas, hamas, leak, mitigate, pirate, recruit, airport, cancel, delay, cartel, decapitate)		

Table 4: A partial list of cyber crime related Arabic seed words and their frequency count identified from the message posts of Alokab forum

Words	Frequency
بقوة	1113
والتطرف	171
الوحشي	435
إرهابية	445
وعذاب	208
الحقد	445
الفاحش	376
الفاجر	264
اغتصاب	423
الإجرامية	172

respectively. These parts cannot be further divided into any other meaningful units. Here, the morpheme *attack* is called *free* morpheme as it can stand alone as a word, whereas other morphemes are called *bound* morpheme as they can only occur in combination as parts of a word and consequently they must be attached as parts of a word.

In addition to *free* or *bound* categories of morphemes, they can also be classified as *inflectional* or *derivational* morphemes. Derivational morphology involves prefixing as well as suffixing to transform a new but morphologically related word, which is often a different class. For example, the suffix “*ation*” converts the verb *normalize* into a new noun form *normalization*, whereas the suffix “*ize*” converts the noun *crystal* into *crystalize* which denotes its verb form. Similarly, *subgroup*, *inactivate*, *deactivate* are the examples of some other category of *derivational* morphological variants that are generated through prefixing the root words. Some derivational prefixes like “*non*”, “*un*” converts a word into negation like *nonscientific*, *unable*, etc. It should be noted that all prefixes used in English are *derivational* prefixes, whereas suffixes can be *derivational* as well as *inflectional*.

Table 5: A sample list of root words and their morphological variants

Root word	Morphological variants	Affix
Act	Acts	-s
	Action	-ion
	Actions	-ions
	Acting	-ing
Bomb	Bombs	-s
	Bombers	-ers
	Bombing	-ing
	Bombings	-ings
Terror	Terrorism	-ism
	Terrorist	-ist
	Terrorists	-ists
Rich	Enrich	en-

On the other hand, *inflectional* morphology generally distinguishes a change in the root form of a word, keeping its syntactic class unchanged. Rather, it indicates grammatical properties in terms of comparison degree, tense, and quantities. For example, the root verb *prove* has inflectional variations like *proves*, *proved*, *proving*, and *proven* that also belong to verb category. A sample list of morphological variants identified while processing BiSAL is shown in Table 5. For morphological variants of English words, various databases are searched and all relevant morphemes for each word are compiled and stored in SentiLEN, whereas the morphological variants of Arabic words are identified by three different Arabic language experts independently and stored in SentiLAR.

2.3. Sentiment Polarity Determination

In this section, we present the proposed technique to determine the polarity of sentiment bearing words identified for BiSAL. For English words, we have used four publicly available sentiment corpora AFFIN [13], SentiWordNet [10], General Inquirer [15], and SentiStrength [17] that assign polarity score to sentiment bearing words in different range. Since there

is no such sentiment corpus for Arabic words, we have applied semi-automated analysis by Arabic language experts to assign polarity scores in a pure scientific manner. Further details about polarity determination for English and Arabic words are provided in the following sub-sub-sections

2.3.1. Sentiment Polarity Determination of English Words

As discussed earlier, SentiLEN (Sentiment Lexicon for English) consists of a list of 279 unique words related to the various categories of cyber crime. For polarity determination, these words are searched in four different sentiment lexicons - AFFIN [13], SentiWordNet [10], General Inquirer [15], and SentiStrength [17], and their scores are unified using an aggregate function. AFFIN data set [13] consists of total 2477 words and each word is assigned a polarity score between -5 to +5 based on its sentiment intensity. In case a word has no match in AFFIN data set, polarity score is assigned as 0.

Second sentiment data set used in our experiment is the SentiWordNet [10], which is a lexical resource for sentiment analysis based on WordNet [12]. It assigns each synset of WordNet with three sentiment scores *positive*, *negative*, and *objective*. Each word is assigned a numeric number between 0 and 1 with its polarity (positive or negative). Like AFFIN, English words are searched in SentiWordNet dataset and the corresponding polarity score is assigned to them when there is a hit, otherwise 0 is considered as the polarity score. On analysis, we found that some words have multiple meanings based on the context and each corresponding meaning has different sentiment scores. In such cases, the words describing conflict activities are gathered together and their average sentiment score is assigned to the word. The web interface of SentiWordNet [10] is used to compare multiple meanings (if exist) of a particular word.

Third sentiment data set used in our experiment is General Inquirer [15], which does not have any numeric score for English words. However, based on the words' sentiment intensity, General Inquirer data set categorizes them as *positive*, *negative*, *strong*, *hostile*, etc. It is observed that no word is annotated as positive if it is hostile in nature, whereas a hostile word is noted as bearing maximum negative sentiment. If a word has a match in General Inquirer data set, its respective annotations are discretized in the range of [-4, +2] using the rules given in Table 6. A word that does not satisfy any combination listed in Table 6 is assigned polarity score as 0. Like SentiWordNet, a word in General Inquirer data set may appear in multiple forms having different meanings. In this case, only those word-forms having some crime or conflict sense are considered. For example, the word *shoot* has three different entries in General Inquirer data set, out of which two entries are related to violence and terror and the third one doesn't have any violence sense as it is used as *shoot up*, which means "to grow or rise rapidly". Moreover, a particular word may have various meanings and all meanings have some violence sense. In this case, the maximum numeric score calculated using Table 6 is considered. For example, the word *arm* is available in General Inquirer data set which has three different forms (arm#1, arm#2, and arm#3). The first form (arm#1) is basically used to describe body part which has no sense related to terror or violence. However, the

Table 6: Discretization of General Inquirer sentiment annotations

General Inquirer annotation				Score
Positive	Negative	Strong	Hostile	
-	Yes	Yes	Yes	-4
-	-	Yes	Yes	-4
-	-	-	Yes	-3
-	Yes	-	Yes	-3
-	Yes	Yes	-	-2
-	Yes	-	-	-1
Yes	-	-	-	+1
Yes	-	Yes	-	+2

Table 7: Exemplar SentiLEN words and their polarity scores from different sentiment data sets

Word	AFFIN Score [-4,+2]	SentiWordNet Score [-0.88,+0.63]	GI Score [-4,+2]	SentiStrength Score [-4,+3]
Attack	-1	-0.25	-4	-3
Harm	-2	-0.42	-4	-3
Bomb	-1	-0.19	-4	-2
Secret	0	+0.13	-1	-1

other two forms (arm#2 and arm#3) refer some violence sense, but the latter one (arm#3) describes the violence intensity more than arm#2. Therefore, the score of arm#3, which is defined as negative, strong, and hostile is considered as the final score for *arm*.

Fourth sentiment data set used in our experiment is SentiStrength [17] in which words are available in stemmed forms with their sentiment scores in the range of [-5, +5]. Therefore, before searching a word in this data set, the word is stemmed using Porter stemmer. In case of a hit, the word is assigned a sentiment score in the range of [-5, +5], whereas 0 is assigned in case of a miss.

In this way, each word of SentiLEN is assigned four independent polarity scores based on their match in the sentiment representing data sets mentioned above. Table 7 shows the polarity scores of some exemplar words. After searching the sentiment data sets for all 279 words of SentiLEN, it is found that the scores obtained from AFFIN, SentiWordNet, General Inquirer, and SentiStrength data sets lie in the range of [-4, +2], [-0.875, +0.625], [-4, +2], and [-4, +3], respectively.

Therefore, the next task is to map the four different scores of each word to a unified scale and normalize them in the range [-1, +1]. To this end, min-max normalization, given in equation 1, is applied to map each score in the range of [-1, +1]. In Equation 1, $P_n(w_i, d_j)$ represents the new polarity value of word w_i in data set d_j , $P_o(w_i, d_j)$ represents the original polarity value of w_i in d_j , $\min(d_j)$ and $\max(d_j)$ are the lowest and highest scores of a word in data set d_j , respectively. After normalization, the mean score of each word w_i , $\delta(w_i)$, is calculated using equation 2, where m is the number of data sets. The normalized and mean scores of the exemplar words of Table 7 are shown in Table 8.

$$P_n(w_i, d_j) = \frac{P_o(w_i, d_j) - \min(d_j)}{\max(d_j) - \min(d_j)} \times (\text{newMax} - \text{newMin}) + \text{newMean} \quad (1)$$

Table 8: Normalized and mean scores of the exemplar words of Table 7

Word(w_i)	AFFIN(d_1)		SentiWordNet(d_2)		GeneralInquirer(d_3)		SentiStrength(d_4)		$\delta(w_i)$
	$P_0(w_i, d_1)$	$P_n(w_i, d_1)$	$P_0(w_i, d_2)$	$P_n(w_i, d_2)$	$P_0(w_i, d_3)$	$P_n(w_i, d_3)$	$P_0(w_i, d_4)$	$P_n(w_i, d_4)$	
	[-4,+2]	[-1,+1]	[-0.88,+0.63]	[-1,+1]	[-4,+2]	[-1,+1]	[-4,+3]	[-1,+1]	
Attack	-1	0	-0.25	-0.17	-4	-1.0	-3	-0.71	-0.47
Harm	-2	-0.33	-0.42	-0.39	-4	-1.0	-3	-0.71	-0.61
Bomb	-1	0	-0.19	-0.08	-4	-1.0	-2	-0.43	-0.38
Secret	0	+0.33	+0.13	+0.33	-1	0	-1	-0.14	+0.13

Table 9: Final sentiment scores of the exemplar words considered in Tables 7 and Table 8

Word(w_i)	$P_n(w_i, d_1)$	$P_n(w_i, d_2)$	$P_n(w_i, d_3)$	$P_n(w_i, d_4)$	$\delta(w_i)$	$\Delta(w_i)$	$\rho(w_i)$	$\eta(w_i)$
Attack	0	-0.17	-1.0	-0.71	-0.47	-1	-1.47	-0.90
Harm	-0.33	-0.39	-1.0	-0.71	-0.61	-1	-1.61	-0.98
Bomb	0	-0.08	-1.0	-0.43	-0.38	-1	-1.38	-0.85
Secret	+0.33	+0.33	0	-0.14	+0.13	-0.33	-0.20	-0.17

$$\delta(w_i) = \frac{\sum_{j=1}^m P_n(w_i, d_j)}{m} \quad (2)$$

$$\Delta(w_i) = \frac{p - n}{p + n} \quad (3)$$

$$\rho(w_i) = \delta(w_i) + \Delta(w_i) \quad (4)$$

$$\eta(w_i) = \frac{\rho(w_i) - \min[\rho(w)]}{\max[\rho(w)] - \min[\rho(w)]} \times (newMax - newMin) + newMean \quad (5)$$

In order to take into consideration the majority voting for determining the positive or negative sentiment of a word w_i , an additive factor $\Delta(w_i)$ is introduced and defined using Equation 3, where p and n represent the number of positive and negative scores of the word w_i , respectively. The additive factor $\Delta(w_i)$ is added to the mean score $\delta(w_i)$ to get an intermediate polarity score $\rho(w_i)$, which is finally normalized using min-max normalization (Equation 5) to get the final polarity score $\eta(w_i)$ of the word w_i . Table 9 shows the final sentiment scores of the exemplar words considered in Tables 7 and 8.

2.3.2. Sentiment Polarity Determination of Arabic Words

As stated earlier, unlike English language, there is no existing sentiment lexicon for Arabic language. Therefore, we have adopted a semi-automated analysis to create Arabic sentiment lexicon based on three domain experts' views taken independently. To start with, a collection of 2000 message posts from Alokab dark web forum [1] is processed using various NLP techniques. To facilitate domain experts for annotation, a GUI-based annotation application is developed in Java to parse documents and count the frequency of each word, which is not a member of the stop-words list identified for Arabic language. A total number of 7061 distinct words are compiled from the 2000 message posts. The GUI presents the words in decreasing order of their frequency count and provides a list of all matching sentences while clicking on a word. This facilitates the experts

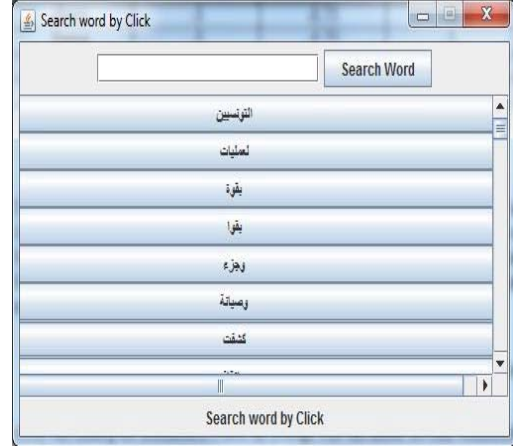


Figure 1: A snapshot of the GUI of the annotation application showing exemplar words extracted from the message posts of Alokab forum

to know the context of a word in which it is used by the users of Alokab forum. Figure 1 presents a list of exemplar Arabic words extracted from the message posts, and Figure 2 shows the list of sentences from the message posts retrieved in response of clicking the first word in Figure 1.

The annotation application was given to three Arabic language experts to annotate them as *positive*, *negative*, *strong*, and/or *hostile*. Scores in the range of [0, 1] for *positivity* and *negativity* are assigned to each word by the experts. In case a word is always used as positive its positive polarity is considered as 1 and the negative polarity as 0. Similarly, if a word is always used in negative sense its negative polarity is set to 1 and positive polarity as 0. For other words that are used in both positive as well as negative sense, positive and negative polarity scores are assigned in the range of 0 and 1 in such a way that their sum is 1. Depending on the degree of positivity of a word, it is also marked as *strong*, and similarly depending on the degree of negativity of a word it is marked as *strong* and/or *hostile*. A sample list of Arabic words and their polarity scores assigned by three different domain experts is shown in Tables 10, 11, and 12. Those words having total score as 0 by every expert are filter out from the list considering them as non-sentiment bearing words. As a result, a total number of 1019 words are retained as sentiment representing words.

For experts scores aggregation, average of individual polarity categories is calculated, as shown in Table 13. For a given word w_i the larger of the average positive score and average negative score is considered as initial polarity score $\delta(w_i)$ and the corre-

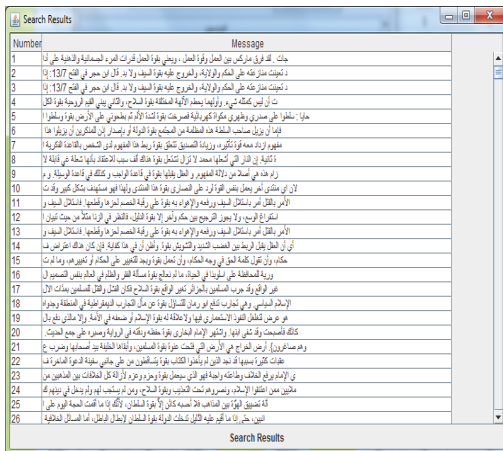


Figure 2: A list of sentences extracted from the message posts of Alokab forum containing the first word of Figure 1

Table 12: Polarity scores assigned to Arabic words by Expert-3

Word	Positive	Negative	Strong	Hostile
بقوة	0.0	0.0	0	0
رائدة	1.0	0.0	1	0
وإنقاذ	0.0	0.0	0	0
والتطرف	0.0	1.0	0	1
استطاعت	0.0	0.0	0	0
ابعد	0.4	0.6	0	0
الظالمون	0.0	1.0	0	1
للظالمين	0.0	1.0	0	1
وضيق	0.0	0.0	0	0
متفقون	0.0	0.0	0	0

Table 13: Average polarity scores of all three domain experts

Word	Positive	Negative	Strong	Hostile
بقوة	0.30	0.03	0	0
رائدة	0.63	0.03	1	0
وإنقاذ	0.30	0.37	1	0
والتطرف	0.00	1.00	1	1
استطاعت	0.16	0.16	0	0
ابعد	0.13	0.20	0	0
الظالمون	0.00	1.00	1	1
للظالمين	0.03	0.96	1	1
وضيق	0.00	0.66	1	1
متفقون	0.30	0.30	0	0

Table 10: Polarity scores assigned to Arabic words by Expert-1

Word	Positive	Negative	Strong	Hostile
بقوة	0.9	0.1	0	0
رائدة	0.0	0.0	0	0
وإنقاذ	0.0	1.0	1	0
والتطرف	0.0	1.0	0	1
استطاعت	0.0	0.0	0	0
ابعد	0.0	0.0	0	0
الظالمون	0.0	1.0	1	0
للظالمين	0.0	1.0	1	0
وحشية	0.0	1.0	0	1
متفقون	0.0	0.0	0	0

Table 11: Polarity scores assigned to Arabic words by Expert-2

Word	Positive	Negative	Strong	Hostile
بقوة	0.0	0.0	0	0
رائدة	0.9	0.1	0	0
وإنقاذ	0.9	0.1	0	0
والتطرف	0.0	1.0	1	1
استطاعت	0.5	0.5	0	0
ابعد	0.0	0.0	0	0
الظالمون	0.0	1.0	1	1
للظالمين	0.1	0.9	1	1
وضيق	0.0	1.0	1	1
متفقون	0.9	0.1	0	0

sponding polarity (*positive* or *negative*) is treated as the words polarity. A word is considered as *positive* if its positive score is higher than the negative score and that a word is considered as *negative* if its negative score is higher than the positive score, otherwise the word is treated as *neutral* (polarity score 0). If total *hostile* score for a word by all three experts is greater than or equal to 1, the word is treated as *hostile* and the hostile score is assigned as 1. Similar approach is applied to determine whether a word is *strong* or not. Like English words, for each word w_i an additive factor $\Delta(w_i)$ is calculated and added to the initial polarity score $\delta(w_i)$ of the word w_i to obtain the next level polarity score $\rho(w_i)$. Finally, the $\rho(w_i)$ values are normalized in the range of [-1, +1] using min-max normalization defined in Equation 5 to obtain the final sentiment score $\eta(w_i)$ of the word w_i . Table 14 presents the final sentiment scores of some exemplar Arabic words.

3. Results

In this section, we present the result obtained through the Bilingual Sentiment Analysis Lexicon (BiSAL) development

Table 14: Sentiment scores of some Arabic sentiment-bearing words

Word (w_i)	Avg. Score (+ve)	Avg. Score (-ve)	$\delta(w_i)$	$\Delta(w_i)$	$\rho(w_i)$	$\eta(w_i)$
بقوة	0.30	0.03	+0.30	+1	+1.30	+0.65
رائدة	0.63	0.03	+0.63	+1	+1.63	+0.82
وإنقاذ	0.30	0.37	-0.37	0	-0.37	-0.18
والتطرف	0.00	1.00	-1.00	-1	-2.00	-1.00

process discussed in the previous sections. A partial list containing 20 entries from SentiLEN (Sentiment Lexicon for English) data set is given in Table 15. However, the complete SentiLEN data set including all 279 root words, their morphological variants, and sentiment scores can be downloaded using the URL: <http://www.abulaish.com/SentiLEN.pdf>.

Similarly, a partial list containing 20 entries from SentiLAR (Sentiment Lexicon for Arabic) data set is given in Table 16. However, the complete SentiLAR data set including all 1019 root words, their morphological variants, and sentiment scores can be downloaded using the URL: <http://www.abulaish.com/SentiLAR.pdf>. A Web interface is developed to access both the lexicons (SentiLEN and SentiLAR) of BiSAL data set online using a graphical interface, and a beta version of the same is available at <http://knp.com.sa/swap/>.

4. Conclusion and Future Work

In this paper, we have presented a bilingual lexical resource (BiSAL) for sentiment analysis over English and Arabic texts related to cyber threats, radicalism and conflicts. The BiSAL consists of two separate lexical resources, namely SentiLEN and SentiLAR. SentiLEN contains a list of 279 sentiment representing English words related to cyber threats, radicalism, and conflicts, along with their morphological variants and sentiment polarity, which is a unification of polarities from four different existing lexical resources. On the other hand, SentiLAR contains a list of 1019 sentiment representing Arabic words along with their morphological variants and sentiment polarity. Applying the process adopted for the development of SentiLAR to identify additional English words from a corpus of English texts from dark web forum is one of our future works to enhance the SentiLEN data set.

5. Acknowledgment

This work was supported by NSTIP strategic technologies program number *11-INF1594-02* in the Kingdom of Saudi Arabia. The authors would like to express their thanks to Abdullah Al-Sahli who contributed as an expert to analyze Arabic language texts for sentiment annotation and polarity determination process.

References

- [1] Alokab dark web forum ([online, last accessed: 10-Nov-2014]). URL <http://www.alokab.com>
- [2] Analyst's desktop binder 2011 ([online, last accessed: 10-Nov-2014]). URL <http://epic.org/foia/epic-v-dhs-media-monitoring/Analyst-Desktop-Binder-REDACTED.pdf>
- [3] M. Annett, G. Kondrak, A comparison of sentiment analysis techniques: Polarizing movie blogs, in: *Advances in Artificial Intelligence*, Springer, 2008, pp. 25–35.
- [4] T. Anwar, M. Abulaish, A social graph based text mining framework for chat log investigation, *Digital Investigation* 11 (4) (2014) 349–362.
- [5] M. Aronoff, *Morphology by itself: Stems and inflectional classes*, No. 22, MIT Press, 1994.
- [6] N. L. Beebe, L. Liu, Clustering digital forensic string search output, *Digital Investigation* 11 (4) (2014) 314–322.
- [7] N. L. Beebe, L. Liu, Ranking algorithms for digital forensic string search hits, *Digital Investigation* 11 (2014) S124–S132.
- [8] H. Chen, *Dark web: Exploring and data mining the dark side of the web*, vol. 30, Springer Science & Business Media, 2011.
- [9] H. Chen, D. Denning, N. Roberts, C. Larson, X. Yu, C.-N. Huang, et al., The dark web forum portal: From multi-lingual to video, in: *IEEE International Conference on Intelligence and Security Informatics (ISI)*, IEEE, 2011, pp. 7–14.
- [10] A. Esuli, F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in: *Proceedings of LREC*, vol. 6, Citeseer, 2006, pp. 417–422.
- [11] S. Kramer, Anomaly detection in extremist web forums using a dynamical systems approach, in: *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ACM, 2010, p. 8.
- [12] G. A. Miller, Wordnet: a lexical database for english, *Communications of the ACM* 38 (11) (1995) 39–41.
- [13] F. Å. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903. URL http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6010/zip/imm6010.zip
- [14] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, *Journal of Informetrics* 3 (2) (2009) 143–157.
- [15] P. J. Stone, *Thematic text analysis: New agendas for analyzing text content*, Text Analysis for the Social Sciences. Mahwah, NJ: Lawrence Erlbaum (1997) 33–54.
- [16] M. Thelwall, K. Buckley, Topic-based sentiment analysis for the social web: The role of mood and issue-related words, *Journal of the American Society for Information Science and Technology* 64 (8) (2013) 1608–1617.
- [17] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, *Journal of the American Society for Information Science and Technology* 63 (1) (2012) 163–173.
- [18] C. C. Yang, X. Tang, B. M. Thuraisingham, An analysis of user influence ranking algorithms on dark web forums, in: *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ACM, 2010, p. 10.

Table 15: A partial list containing 20 entries from SentiLEN data set

Words	Morphological variations	Sentiment polarity
ambush	ambush, ambushed, ambushes, ambushing, ambushade	-0.86
arm	arm, armed, arms, arming	-0.54
assassin	assassination, assassinate, assassinating, assassinated, assassin, assassinator, assassinator	-0.95
assault	assault, assaulting, assaulted, assaulter, assaultive, assaulters	-0.84
attack	attack, attacked, attacking, attacks, attacker, attackers	-0.90
belief	belief, beliefs, believe, believer, believing, believed, believes, believable, believably, believing	+0.68
blast	blast, blasts, blasting, blaster, blasted	-0.61
blow	blow, blows, blew, blown, blower, blowing, blowy	-0.75
blood	blood, bloody, blooded, bleeding, bloods	+0.08
body	body, bodies, bodily, bodied, bodying	+0.08
bomb	bomb, bombs, bombing, bombings, bomber, bombers, bombed	-0.85
burn	burn, burning, burns, burned, burner, burnable	-0.69
business	business, businesses, businessman	+0.73
bust	bust, busts, buster, busty, busting, busted	-0.53
camp	camp, camps, camper, camping, campy, camply, camped	-0.52
capture	capture, captured, captures, capturing, capturer, capturers	-0.35
casualty	casualty, casualties	-0.85
change	change, changes, changing, changed	+0.68
checkpoint	checkpoint, checkpoints	+0.08
chief	chief, chiefs	+0.73
support	support, supports, supporter, supporters, supporting, supported, supportive	+0.95

Table 16: A partial list containing 20 entries from SentiLAR data set

Words	Morphological variants	Sentiment polarity
اتفق	تتفق الاتفاقية، متفقة، متفق، اتفق، يتفق، اتفق	+0.67
احترام	احترامي، محترم، يحترم، محترمة، الاحترام	+0.65
احيي	يحيي، احيي، محيي، محيبة، تحيي، احياء، الاحياء	+0.67
ادى	سيؤدي، أد، مؤد، مؤدية، تؤدي، مؤدى، سيؤدي	+0.67
استغل	يستغل، استغل، مستغل، مستغلة، تستغل، استغلال، لاستغلال	+0.67
استقلت	يستقل، استقل، مستقل، مستقلة، تستقل، استقلال، مستقل	-0.08
استمر	يستمر، استمر، مستمر، مستمرة، تستمر، استمرار، مستمر	-0.08
استوطن	يستوطن، استوطن، مستوطن، مستوطنة، تستوطن، استيطان، الاستيطاني	+0.67
الانم	لانم، لاما، لومت، لام، يلام، لامة، ملامة، يلوم، لوما، لؤماء، ملامان، لؤمان، اللؤم، الأم، الأما، استلام، ملام، الملام، بلامهما	-0.83
الاستراتيجية	الإستراتيجية	+0.67
الأمل	الأمال، آمال، أملته، أمله، يأمله، أملاً، تأملاً	+0.67
التصدية	الصدى، صدي، يصدى، يصدد، الصدى، يصدون، تصديت، صداه، يتصدد، صديداً، تصده، صدده	-0.67
الجبين	الجبناء، الجبينان، أجبن، أجبنه، جبن، جبان	-0.67
الجرم	جريمة، جرمه، جرائم، أجرام، جروم، جريم، مجرم، اجترمه، جرم، المجرمين، أجمت، جرم، يجرمك، جرمك، اجرم	-1.00
الجمال	أجمل، جميل، جمال، المجمال، جمالاً، أجملت، جمالات، جملة، أجملت، إجمال	+0.83
الرحمة	الرحيم، الرحمة، رحمة، رحماً، رحمت، رحمتنا، الرحمن، الأرحام، الرحم، رحمن، رحيماً	+0.67
السطوة	سطوة، السطوات، سطا، سطا، يسطوان، يسطو	-0.67
الشرير	الشر، بشر، شرور، شراً، شرارة، أشر، أشراء، الشرار، أشرار، شربرين	-1.00
الصراف	الصراف	+0.67
الغالية	الغالية	+0.67