Journal of Universal Computer Science, vol. 23, no. 3 (2017), 327-351
submitted: 15/9/16, accepted: 1/3/17, appeared: 28/3/17 © J.UCS

Framework for Affective News Analysis of Arabic News: 2014 Gaza Attacks Case Study

Mahmoud Al-Ayyoub
(Jordan University of Science and Technology, Irbid, Jordan
malayyoub@gmail.com)

Huda Al-Sarhan
(Jordan University of Science and Technology, Irbid, Jordan
huda.nimer@gmail.com)

Majd Al-So'ud
(Jordan University of Science and Technology, Irbid, Jordan
majdsoud5@gmail.com)

Mohammad Al-Smadi
(Jordan University of Science and Technology, Irbid, Jordan
smadi.mohammad@gmail.com)

Yaser Jararweh
(Jordan University of Science and Technology, Irbid, Jordan
yaser.amd@gmail.com)

Abstract: This paper aims at fostering the domain of Arabic affective news analysis through providing: (a) a benchmark annotated Arabic dataset of news for affective news analysis, (b) an aspect-based sentiment analysis (ABSA) approach for evaluating the sentimental affect of Arabic news posts on the reader, and (c) a baseline approach with a common evaluation framework to compare future research results with the baseline ones.

Key Words: Affective News; Emotional Affect; Aspect Based Sentiment Analysis; Natural Language Processing; Arabic Dataset.
Category: H.3, H.3.5, H.4.3, I.2.7, J.4

1 Introduction

With the explosive growth of social media networks, individuals and organizations are increasingly using the social media in order to improve their decision-making process. One example is concerned with the effect online contents have on the users and their reactions to them. Thus, there is a need to build automated techniques to evaluate news posts tone and sentimental affect on readers.

Researchers have started investigating the relation between the readers and the news articles in different topics. The readers are interested in detecting and tracking news stories [Allan et al., 1998], RSS feeds or news headlines [Wan and Yang, 2007] and classifying news [Sebastiani, 2002]. Most of the news categories listed on social networks, especially Twitter, are breaking news for cities, sports, brands, accidents, disasters and political events [Kwak et al., 2010, Farhi, 2009].

Peggy [Munezero et al., 2014, Thoits, 1989] defined affect as a "positive and negative evaluations of an object, behavior, or idea with intensity and activity dimensions." Affective news is an interesting domain for researchers where news posts are evaluated based on their sentimental affect on readers. It is a simpler problem than the emotional affect, which includes more involved types of emotions like happiness, sadness, angry, etc. Affective news analysis is of great importance in many domains. The tone of news posts could affect readers' attitudes towards: trading and stock market [Gentzkow and Shapiro, 2010], political parties [Eshbaugh-Soha, 2010, Farnsworth and Lichter, 2010], wars and conflicts [Cho et al., 2003], etc.

Local and international news are usually written by people with experience to attract readers and followers on social networking sites. A news post can be formulated about a certain event in a very emotional way by using some words and phrases that carry an affective meaning. The emotional affect of words/aspects depends significantly on text content, the context and the readers' convictions. For example, the term 'murder' has a negative emotional affect in general, but in the context of war; it might have a positive emotional affect if it is related to the killing of an enemy, which represents a victory.

This research aims to provide an online Arabic dataset for the purposes of research related to affective news analysis. The dataset consists of Facebook posts about the Gaza attacks of 2014 and their related comments. The dataset is annotated while keeping in mind the sentimental affect on the reader. Instead of simply annotating each post as merely having a positive, negative or neutral sentimental affect on the reader, we perform a more detailed and more challenging annotation by looking for each aspect contained in each post and the sentimental affect the reader might have with respect to this aspect. In a previous work of ours [AL-Smadi et al., 2015], we discussed the collection and annotation of the dataset; however, at that stage, the annotation was performed by a single human annotator with very limited experience in the non-trivial task of annotation for the purposes of aspect-based sentiment analysis (ABSA) of news posts. In this paper, we revisit the same dataset and perform a more thorough annotation. Specifically, two much more experienced human annotators are employed and their separate annotation efforts are compared and studied. To the best of our knowledge, studies on inter-annotator agreement are rare in the field of Arabic Natural Language Processing (NLP) despite their importance. Obviously, since the annotation of the dataset is modified, the baseline experiments reported in [AL-Smadi et al., 2015] have to be repeated in view of the new annotation.

Another significant advantage of this work compared to [AL-Smadi et al., 2015] is the expansion of the dataset to include the people's comments and replies to the news posts. Annotating these comments is a very challenging task for several reasons. The annotation of the comments takes into account the affect of the news posts on the commenters, who are pro-Palestinian Arabs. The comments are written in Modern Standard Arabic (MSA) as well as Dialectal Arabic (DA). MSA is the official version of Arabic used in formal communication, education, etc. It is the version used in the news post. On the other hand, DA is used in informal daily communications and it is the version common in the comments. Since DA has many "flavors" and does not obey standard grammar or vocabulary, it requires more complex pre-processing steps. Thus, expanding the dataset to include comments did not only involve significant annotation efforts, but it also required the tweaking and modification of existing tools to accommodate the dialects found in the expanded dataset. Finally, the dataset is evaluated using a baseline classifier as well as a lexicon-based classifier.

The rest of the paper is organized as follows. Section 2 presents some related work on affective news and how ABSA could be useful in this evaluation. The available dataset is prepared in readable format to support four ABSA tasks: aspect extraction, aspect polarity detection, aspect category identification and category polarity detection (see Section 3). Dataset collection and annotation process is discussed in Section 4. The baseline results to evaluate the prepared dataset are discussed in Section 5. Section 6 presents the lexicon-based algorithm and its evaluation results. Finally, we conclude this research and shed the light on future plans in Section 7.

2 Related Works

The rich field of NLP contains studies that focus on the viewpoint of the writer [Malandrakis et al., 2013]. It also contains other studies that are concerned with studying the emotions felt by the reader, and thus, analyzed text from the perspective of the reader, such as news headlines analysis [Strapparava and Mihalcea, 2007]. In such studies, news headlines (mainly written in English and contain a few words) are collected from trusted news website such Google news and CNN. Then, each headline is assigned to one of six predefined emotion labels based on the first impression of annotators' team. The considered emotions labels are: Anger, Sadness, Surprise, Disgust, Fear and Joy, which are widely accepted in the literature [Al-A'abed and Al-Ayyoub, 2016]. In addition, the authors of [Strapparava and Mihalcea, 2007] evaluated the same data based on three sentiment values (positive, negative and neutral). In [Papacharissi and de Fatima Oliveira, 2012], the authors offered an affective news theory for explaining a distinctive character of news tweets at a time of political crisis. The case study they considered is Egypt around the time of the resignation of President Hosni Mubarak. Some researchers studied affective news analysis using a Dictionary/Lexicon-based approach [Young and Soroka, 2012].

In contrast to the previous research, our idea adopts a novel approach of using contextual information of each user's reviews (i.e. news posts). Therefore, aspect based sentiment analysis (ABSA) is used to evaluate news posts affect on readers.

ABSA is a type of sentiment analysis (SA) applied at deeper level where the analyzed text may contain possibly different opinions about different topics (or different aspects of the same topic). The goal is to identify the aspects/features of the entities reviewed in a specific domain and discover the sentiment that reviewers express for each aspect [Liu, 2012, Liu and Zhang, 2012, Abdalkader, 2014, Pavlopoulos, 2014, Pontiki et al., 2014]. Examples of using ABSA for analyzing English reviews include movie [Thet et al., 2010], electronic products [Hu and Liu, 2004] and restaurants [Ganu et al., 2009, Brody and Elhadad, 2010].

The news articles in the proposed framework [Park et al., 2010] are labeled based on aspect-level news browsing. Aspect-level news browsing allows news followers and readers to read various news essays for any chosen event. Moreover, one news event could be evaluated from different biased perspectives. Readers can recognize these different viewpoints and create their own perspectives without any previous bias in the viewpoints.

Aspect-level classification can be applied to identify major aspects. Then, the articles can be categorized based on the similarity of the chosen aspects. Two approaches are developed to achieve aspect extraction and article classification tasks, namely: news structure-based extraction (NSE) as well as framing cycle-aware clustering (FCAC). NSE [Park et al., 2009] applies news writing rules to extract aspects automatically from each news article and it can be implemented to many articles regardless of the language and article topic. FCAC discovers the diversity of distinguished aspects of a news event.

Despite the fact that most of the available research aimed at analyzing affective news in English, few examples can be found for Arabic news. This can be due to the many difficulties faced by Arabic NLP researchers such as working on one of most Morphologically-Rich Languages (MRL) [Abdul-Mageed et al., 2011], having different dialects that are difficult to handle [Habash, 2010], having a very limited set of resources [Hmeidi et al., 2014], etc.

Another limitation with current works on Arabic NLP is the lack of support to ABSA. Existing datasets are mainly annotated with positive, negative and possibly neutral classes. Examples include the Opinion Corpus for Arabic (OCA) [Rushdi-Saleh et al., 2011a, Rushdi-Saleh et al., 2011b], which consists of movie reviews, the AWATIF dataset [Abdul-Mageed and Diab, 2012], the Yahoo!-Maktoob/Twitter datasets of [Abdulla et al., 2014] and the large Arabic language dataset of book reviews (LABR) [Aly and Atiya, 2013]. The only exception is the latter dataset which has been re-annotated for ABSA by [Al-Smadi et al., 2015].

3 ABSA Tasks

In general, the proposed news dataset covers the following six main research tasks:

3.1 Task 1: Aspect Term Extraction (T1)

Given a news post, the objective of this task is to extract all conceivable aspect terms related to Gaza domain regardless of their polarity. Annotated aspect terms examples are: (e.g., bombing / تفجير, military operations / العمليات العسكرية, Israeli occupation planes / طائرات الاحتلال الإسرائيلي). Note that the aspect terms should appear explicitly in the news post.

3.2 Task 2: Aspect Term Polarity estimation (T2)

Based on Task 1 (T1), the objective of this task is to determine the sentiment for each retrieved aspect terms and label it as neutral, positive, or negative. Instances with conflicting sentiments are not included in our study.

3.3 Task 3: Aspect Category Detection (T3)

In this task, predefined aspect categories are used to assign each news post to exactly one aspect category. Unlike aspect terms, the aspect category need not occur explicitly in the news post. They can be inferred using adjectives, sense words, or contextual meaning and not identified using aspect terms in the post.

3.4 Task 4: Aspect Category Polarity Estimation (T4)

This task is a subtask of the previous task (T3) and its objective is to determine the polarity (positive, neutral, or negative) for each identified aspect category.

3.5 Task 5: Comment Category Detection (T5)

In this task, the objective is to categorize comments into six predefined categories as followers: (تحليل / Analysis, دعاء / Supplication, معلومة / Information, استهزاء / Ridicule, استفسار / Query, رسالة اقتحامية / Spam). For the sake of this research, the spam category was removed in the cleaning/filtering stage. Each comment is assigned to one suitable category regardless of the contextual meaning of its designated news post.

Number	Description	Start	Finish
Event #1	The actual cause of the Gaza war in 2014: the kidnapping process of three young settlers on June 13. السبب الفعلي لحرب غزة 2014: عملية اختطاف الشبان المستوطنين الثلاثة يوم 13 يونيو/حزيران.	13-6-2014	30-6-2014
Event #2	The Kidnapping, torturing and burning of the child Mohammed Abu Khudair. خطف وتعذيب وحرق الطفل محمد أبو خضير.	1-7-2014	6-7-2014
Event #3	Operation "Protective Edge." عملية الجرف الصامد.	7-7-2014	8-7-2014
Event #4	"Bonian Marsous" Operation. عملية البنيان المرصوص.	8-7-2014	11-7-2014
Event #5	"Al-Asf Al-Ma'kool" Battle + Failed attempts to truce. معركة العصف المأكول + محاولات فاشلة للتهدئة.	11-7-2014	20-7-2014
Event #6	The kidnapping of the Israeli soldier Aaron Shaul. خطف وأسر الجندي الصهيوني شاؤول أرون.	20-7-2014	28-7-2014
Event #7	Children massacre in Shati refugee camp + continuation of failed attempts to truce. مجزرة أطفال مخيم الشاطئ + استمرار محاولات فاشلة للتهدئة.	28-7-2014	4-8-2014
Event #8	Declare victory + Results after accepting the terms of the truce. إعلان النصر + نتائج بعد قبول شروط التهدئة.	4-8-2014	18-9-2014

Figure 1: Basic eight events of attacks on Gaza 2014

3.6 Task 6: Comment Category Polarity Estimation (T6)

The objective of this task is to assign one sentiment value (positive, negative, or neutral) to each comment category. Note that we consider a pro-Palestinian point of view.

4 Dataset Collection and Annotation

4.1 Dataset Collection

The reviews that have been selected from social media are Arabic short posts representing Breaking News. News posts and their related comments are collected from well-known Arabic news networks such as Al Jazeera and Al Arabiya. An automatic tool named Netvizz v1.05 has been used to crawl news posts and comments out of Facebook pages. For this study, all the news posts and comments are taken from a Facebook page called "عاجل من غزة" / "Breaking news from Gaza".

The attacks on Gaza lasted for four months and during this interval, many significant events took place. We divide this interval into eight subintervals based on the most important events that took place and divide the posts accordingly. Figure 1 provides description of the main events we consider along with an explanation of the main reason for this division.

Since the number of collected posts is too large to be manually annotated by us, we randomly select 20% of the news posts on each event to undergo the

Table 1: The sizes of the annotated training and testing datasets.

Type	Training	Testing	Total
News Posts	1,811	454	2,265
Comments	10,902	2,726	13,628

manual annotation process. The resulting dataset consists of 2,265 news posts annotated with aspect terms (T1), aspect term polarity (T2), aspect category (T3), and aspect category polarity (T4). In addition, the dataset for the same news posts consists of 13,628 Arabic news comments annotated with comment category (T5), and comment category polarity (T6). The chosen posts and their related comments divided as training and testing data (80% to 20%). Table 1 provides the number of train and test dataset (size of each part).

4.2 Annotation Process

To annotate our dataset, the BRAT rapid annotation tool is used [Stenetorp et al., 2012].[1] BRAT is a web-based text annotation tool that can be configured to meet the needs of the annotation stage of our work. The tool properties have been adapted to provide six types of information: aspect terms extraction (T1), aspect term polarity estimation (T2), aspect category detection (T3), aspect category polarity estimation (T4), comment category detection (T5), and comment category polarity estimation (T6).

The annotation has been done by two members who are experienced with annotation for ABSA. The annotation is later validated by two senior researchers in Arabic SA. All annotators are native Arabic speakers. They have received training on annotation using BRAT. They are given guidelines for the required annotation as follows.

4.2.1 Aspect terms and polarities

During this stage, the annotator is asked to annotate all the explicit single/multiple terms (e.g., bombing / تفجير, military operations / العمليات العسكرية, Israeli occupation planes / طائرات الاحتلال الإسرائيلي) that are related to the basic target entity (i.e., Gaza news in our case). Taking into account that the selected terms are of the types noun or noun phrase, any words referring to feelings, such as adjectives and sense words, should not be annotated as aspect terms. The aspect terms are annotated as they appear in the original post even if they are misspelled. Then, the annotators are asked to give a polarity value (positive, neutral or negative) for each annotated aspect term. E.g., the post shown in Figure 2 is annotated

[1] `http://brat.nlplab.org/`

عباس ،،، تكلمنا مع الجانب الأمريكي وطلبنا أن يوقفوا العمليات العسكرية من جانب إسرائيل، ونحن نحاول ان نقنع حركة حماس بوقف العمليات، ولكن للاسف لم ننجح
Abbas ,,, we spoke with the American side and asked to stop the military operations from the Israeli side. We are trying to convince Hamas to halt the operations. Unfortunately, we did not succeed

Figure 2: Example of an Arabic post along with its English translation.

Table 2: Distribution of aspect terms over the polarity class

Domain	Positive	Negative	Neutral	Total
News Posts	4,165	4,805	685	9,655

as follows: {‘الجانب الأمريكي’ / ‘American side’: neutral, ‘العمليات العسكرية’ / ‘military operations’: positive, ‘اسرائيل’ / ‘Israel’: positive, ‘حركة حماس’ / ‘Hamas’: negative, ‘وقف العمليات’ / ‘stop the operations’: negative}. Table 2 summarizes the aspect terms distribution over the sentiment class in the annotated dataset.

4.2.2 Aspect categories and polarities

In this stage, the annotator should detect the appropriate category for each sentence and make sure that the pre-selected aspects are related to the chosen category. Four predefined aspect categories are considered (الخطط / Plans, النتائج / Results, التهدئة / Peace, الأطراف / Parties). Note that the most obvious category in the sentence is selected and the others are ignored. Then, each aspect category is assigned the most appropriate polarity (positive, negative, or neutral). E.g., the example of the previous paragraph is assigned the aspect category {التهدئة / Peace: negative}. Table 3 summarizes the aspect categories distribution over the sentiment classes.

Figure 3 depicts an example of annotated post in BRAT as is done by the annotators. Mainly, the two independent annotators annotated the complete dataset before any pre-processing step. The level of agreement between the two

Table 3: Aspect categories distribution over the polarity class

Category	Positive	Negative	Neutral	Total
Plans	317	453	43	813
Results	414	503	44	961
Peace	155	110	85	350
Parties	78	42	21	141
Total	964	1,108	193	2,265

Table 4: Conflict between the two human Annotators.

ABSA Tasks	Task 1	Task 2	Task 3	Task 4
Conflict	0.054	0.023	0.014	0.007

annotators is computed. Table 4 shows the conflict percentage between the two human annotated datasets according to the four tasks. Moreover, after clustering the differences, we find that the most difficulties are found in the extraction of the aspect terms (i.e. Task 1). It is the most critical task which has the highest conflict ratio. For example, in a certain news post, the first annotator chooses (Martyr / الشهيد) as an aspect term, while the second annotator chooses (Child Martyr / الشهيد الطفل) as an aspect term since the context highly focuses on the Child Martyr as a single aspect term. Differences in Task 2 (aspect polarity estimation) are clustered in the way that depends completely on the contextual meaning of the current post not as defined in the public lexicon/dictionary. For instance, (kill / القتل, prison / السجن, punishment / عقوبة) are classified as negative terms according to most lexicons, while, in the case of war, it may get a positive or negative emotional affect according to the post's context and which side the reader is supporting. The main reason of difference in Task 3 (aspect category detection) is due to the different views for both annotators about determining the appropriate categories for some posts either as (Parties / الأطراف) or (Results / النتائج). Finally, the main difference in task 4 is in identifying the polarity of the category (Peace / التهدئة), where, based on the text context, the news posts of the Peace category do not always have a positive emotional effect and could carry negative or neutral meaning. At the end, in order to get a more reliable and consistent dataset, a third party (two expert senior researchers) reviewed both the dataset and the differences between the two annotators in order to choose the best annotation according to the context of each post.

For testing the inter-annotator agreement, Cohen's kappa measure is used. It is one of the widely used measured for this purpose [Viera et al., 2005]. It depends on the proportion of observations in agreement ($Pr(a)$) as well as the proportion in agreement equivalent to chance ($Pr(e)$) [Viera et al., 2005]. In our case, it measures the agreement between the two annotators, and subtracts out the agreement due to chance. Table 5 shows Cohen's kappa measure for the first task (extraction of the aspect terms) by testing inter-annotator agreement between the two annotators, where (a) and (d) represent the times in which the two annotators agree and (b) and (c) represent the times in which they disagree.

The proportion of observations in agreement ($Pr(a)$) is computed using the following equation:

$$Pr(a) = \frac{a + d}{n} \tag{1}$$

Table 5: The Cohen's kappa of the first task (extraction of the aspect terms)

Inter-annotator agreement (Kappa measure)		Annotator 1 Results		
		Aspect Term	None Aspect	Total
Annotator 2 Results	Aspect Term	9,364 (a)	491 (b)	9,855 (m_1)
	Non-Aspect	428 (c)	28,098 (d)	28,526 (m_0)
	Total	9,792 (n_1)	28,589 (n_0)	38,381 (n)

Figure 3: Example of an annotated sentence in the BRAT tool

While the proportion in agreement due to chance $Pr(e)$ is computed using the following equation:

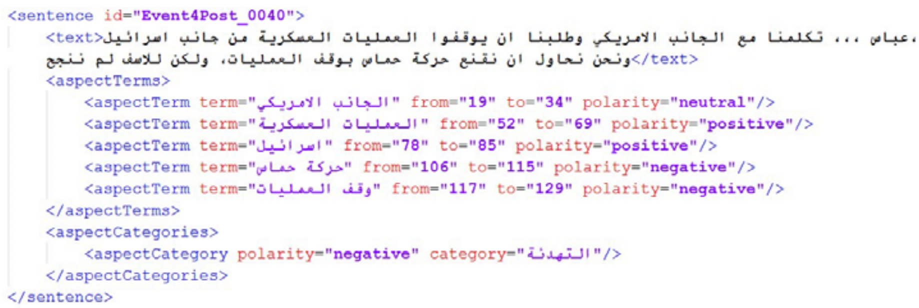
$$Pr(e) = \left[\frac{n_1}{n} \times \frac{m_1}{n}\right] + \left[\frac{n_0}{n} \times \frac{m_0}{n}\right] \tag{2}$$

Finally, Cohen's kappa measure is computed as follows.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3}$$

Table 6 shows the kappa values for: 1) two human annotated dataset, 2) the first human data set and the gold annotated dataset that has been reviewed by the two senior researchers, and 3) the second human and the gold according to the four tasks.

According to [Altman, 1990], a kappa value of 1 shows perfect agreement, while a kappa value of 0 shows agreement due to chance. Figure 5 shows the interpretation on Kappa as authors in [Viera et al., 2005] proposed. According



```
<sentence id="Event4Post_0040">
    <text>عباس ،،، تكلمنا مع الجانب الامريكي وطلبنا ان يوقفوا العمليات العسكرية من جانب اسرائيل، ونحن نحاول ان نقنع حركة حماس بوقف العمليات، ولكن للاسف لم ننجح</text>
    <aspectTerms>
        <aspectTerm term="الجانب الامريكي" from="19" to="34" polarity="neutral"/>
        <aspectTerm term="العمليات العسكرية" from="52" to="69" polarity="positive"/>
        <aspectTerm term="اسرائيل" from="78" to="85" polarity="positive"/>
        <aspectTerm term="حركة حماس" from="106" to="115" polarity="negative"/>
        <aspectTerm term="وقف العمليات" from="117" to="129" polarity="negative"/>
    </aspectTerms>
    <aspectCategories>
        <aspectCategory polarity="negative" category="التهدئة"/>
    </aspectCategories>
</sentence>
```

Figure 4: An XML snapshot that corresponds to the annotated sentence of Figure 3

Table 6: Cohens-kappa agreement between two human annotated dataset and compare it with gold annotated dataset according to four ABSA tasks.

	Two human annotators	Gold and first human annotator	Gold and second human annotator
Task #1	94%	95%	93%
Task #2	98%	97%	98%
Task #3	96%	95%	97%
Task #4	98%	98%	98%

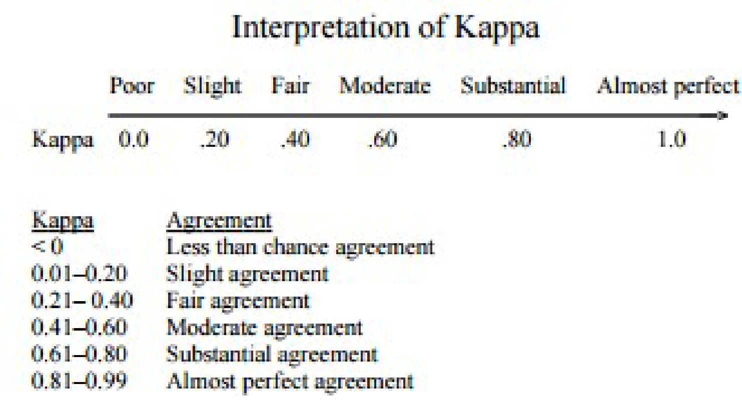


Figure 5: The interpretation of kappa according to [Viera et al., 2005].

to this interpretation, we have almost perfect agreements for the four tasks at hand as indicated by the standard error of the kappa values for the four tasks shown in Table 7.

Table 7: Standard error of Cohen's kappa for the four tasks

	Two human annotators	Gold and first human annotator	Gold and second human annotator
Task #1	0.2%	0.1%	0.2%
Task #2	0.1%	0.1%	0.1%
Task #3	0.3%	0.4%	0.3%
Task #4	0.3%	0.3%	0.3%

Table 8: News comments categories distribution over the polarity class

Category	Positive	Negative	Neutral	Total
Analysis	1,879	3,549	210	5,638
Supplication	2,646	2,245	108	4,999
Information	211	128	751	1,090
Ridicule	210	598	391	1,199
Query	29	125	548	702
Total	4,975	6,645	2,008	13,628

4.2.3 Comments Categories and Polarities

Each news posts that has been collected and manually annotated contains at least one comments. In this stage, the category and polarity are determined for each comment at the sentence level. We assume that each comment sentence holds only one category out of the six predefined categories (see section II). Then, each comment category is assigned the most appropriate polarity (positive, negative, or neutral).

We have noticed that most of the comments are either analysis or supplication classes and their feelings are clearly assigned as either positive or negative. In addition, most of the comments that belong to Information and Query classes have a neutral polarity. E.g., "اللهم الطف بـإخواننا في غزة / May God save our brothers in Gaza" is assigned the comment category {دعاء / Supplication: positive}, "اسالهم شو شعورهم نفسي / I would like to ask them about their feeling" is assigned the comment category {استفسار / Query: neutral} and " لاتاخذكم بالخائن شفقة / don't feel sympathetic with traitors" is assigned the comment category {تحليل / Analysis: positive}). Table 8 summarizes the categories of news comments distribution over the sentiment class.

4.3 Annotation Format

We now give more technical details about the annotation process. We follow the SemEval2014 Task 4 dataset schema [Pontiki et al., 2014, Al-Smadi et al., 2015]. Semantic Evaluation (SemEval) is a prominent annual international workshop in the NLP domain. The BRAT annotation files are mapped into the compliant XML file by using the selected schema supported in [Pontiki et al., 2014, Al-Smadi et al., 2015]. As depicted in Figure 4, each news post is annotated based on the three main XML tags: *text*, *aspectTerm* and *aspectCategory*. Text tag contains the original news post. AspectTerm tag contains four attributes: term (selected aspect), its polarity and the location of the aspect term in the post (from, to). AspectCategory contains two attributes: category type and the polarity of the selected category. Out of annotation phases we had 2,265 news posts annotated and formatted as XML format of [Pontiki et al., 2014, Al-Smadi et al., 2015]. The XML-based dataset is available publicly for noncommercial research.[2]

5 Baseline Evaluation

The prepared dataset can be used to evaluate the sentimental affect of Arabic news posts on the reader using ABSA technique. Our dataset is tested with the baseline classifiers of [Pontiki et al., 2014, Al-Smadi et al., 2015].

5.1 Baseline Description

T1: Aspect term extraction baseline: All the extracted aspect terms from the training dataset are collected to build a simple dictionary of aspect terms. Then, all the tokens in the test dataset that matches with the dictionary are tagged as aspect terms.

T2: Aspect term polarity baseline: The baseline checks if each returned aspect term t from the test sentence s has been encountered in the training dataset d. If yes, the baseline assigns to the aspect terms t in the test sentence the most frequent polarity in the training set (positive, negative or neutral). If aspect term t in test sentence has not seen in the training sentences, the baseline assigns t the most frequent aspect terms polarity tagged in the training sentences.

T3: Aspect category identification baseline: The baseline uses Dice coefficient similarity measure to calculate the distance between test sentences s and training sentences d based on the similar words and total number of words for t and d. The test sentence s is assigned the most frequent aspect category c presented in the training sentences.

T4: Aspect category polarity baseline: In test sentence s, each aspect category c has a polarity based on the most frequent polarity that the aspect category had in the most similar training set d. If aspect category c is not seen in the training sentences, then the most frequent polarity value in the whole training set is selected to aspect category c.

T5: Comment category detection baseline: It works with the same intuition like T3 but without having aspects. Whereas, the baseline algorithm compares every test sentence s with training sentences d and compute Dice coefficient similarity measure to retrieve the d most similar sentences in the training set.

[2] `https://github.com/malayyoub/ABSA-for-Affective-News-Analysis-`

Subsequently, the most frequent comment category found in d training set is assigned to test sentence s.

T6: Comment category polarity baseline: This task follows the same manner as task 4 in all steps except that it is applied to the comments rather than the posts.

5.2 Evaluation Measures

To measure the performance of ABSA tasks, the standard performance measures are applied. For the first, third and fifth task (T1, T3 and T5), the F1 measure is computed as follows.

$$F1 = \frac{2PR}{P + R} \tag{4}$$

where precision (P) and recall (R) are calculated as follows.

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

where TP is the set of true positives (relevant aspect terms for T1, aspect categories for T3 or comment categories for T5) which have been extracted from the test dataset. FP is the set of false positives (irrelevant aspect terms, aspect categories or comment categories) retrieved for the same test part. FN is the set of relevant aspects or categories not retrieved.

The simple accuracy measure is used to evaluate estimating aspect term polarity for T2, aspect category polarity for T4 and comment category polarity for T6. The accuracy is defined as the number of correctly retrieved polarities divided by the total number of aspect term polarity or aspect category polarity annotations.

$$A = \frac{\text{Number of Correctly retrieved polarities}}{\text{Total number of polarities}} \tag{7}$$

5.3 Baseline Results

The prepared dataset is evaluated by using the baseline model and the final results are presented in Table 9. We note that the results are promising and indicate a very good accuracy. Future research can compare the performance of their proposed approaches that using ABSA to study the affective news with the baseline results after applying the same common measures.

Table 9: Baseline Results

Baseline Tasks	P	R	F1	A
T1: Aspect Term Extraction	0.379	0.41	0.394	
T2: Aspect Term Polarity				0.659
T3: Aspect Category Identification	0.649	0.649	0.649	
T4: Aspect Category Polarity				0.74
T5: Comment Category Detection	0.667	0.667	0.667	
T6: Comment Category Polarity				0.64

6 Lexicon-Based Approach

We propose a lexicon-based approach to solve the first two tasks of ABSA, namely: aspect terms extraction (T1) and aspect term polarity estimation (T2). This approach overcomes baseline results in terms of evaluation measures.

6.1 T1: Aspect term extraction

The proposed approach to address T1 is summarized in Figure 6. It consists of the following steps: (A) manual aspect terms lexicon construction, (B) text pre-processing, (C) feature extraction.

6.1.1 Manual aspect terms lexicon

The lexicon is built by extracting all associated aspect terms from the training part of the human annotation dataset. Then, the aspects list is stored in an external file after removing the duplicate phrases and sorting them in alphabetical order. Note that the aspect terms that have been selected by the human annotators are only nouns and noun phrases and, thus, our aspects lexicon is created accordingly.

6.1.2 Text pre-processing

Several pre-processing steps are used. The first one is related to the fact that some sentences start with a news source, which could be a person, a news agency, etc. To determine the source boundaries, double commas (،،) or triple commas (،،،) have been added to text. E.g., in the example of Figure 3, the news source is Abbas / عباس. These parts are not deleted in the pre-processing step since the news source can be important for other research ideas, such as building a social network out of the news posts. However, for ABSA, the news source is not considered as an aspect term and is not used to determine polarities. Therefore, it has been deleted temporarily.

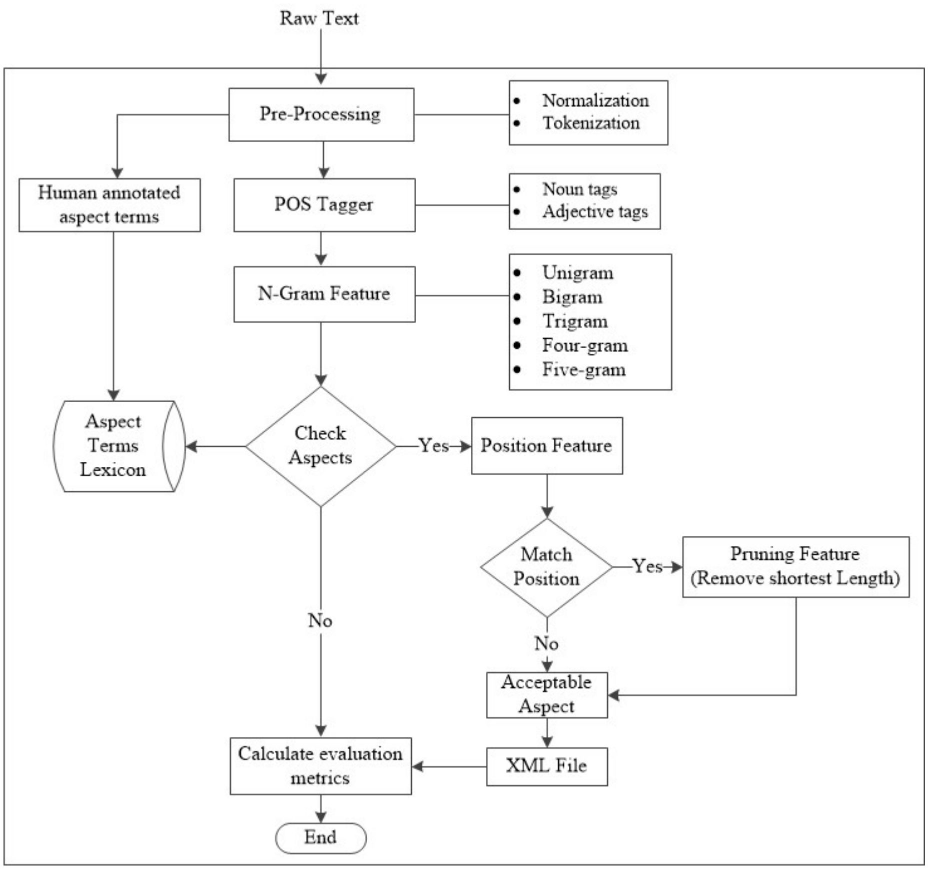


Figure 6: A diagram of the lexicon-based algorithm to solve T1 (Aspect Term Extraction).

Although candidate news posts are written in modern standard Arabic (MSA), a pre-processing and cleaning phase is required to prepare the dataset for further processing.

For the cleaning and filtering purpose, a Java-based library called AraNLP is used to cover the tasks of pre-processing [Althobaiti et al., 2014]. The pre-processing step can be summarized in the following points:

- Removal of non-Arabic words.
- Removal of hyperlinks and URLs in all posts.
- Removal of Arabic diacritics (Tashkeel) such as “ًٌٍَُِّْ”.
- Removal of punctuations and symbols such as “? ’ ! @ $ # |”.

(1)	**Original sentence**	تكلمنا مع الجانب الأمريكي وطلبنا أن يوقفوا العمليات العسكرية من جانب إسرائيل، ونحن نحاول ان نقنع حركة حماس بوقف العمليات، ولكن للأسف لم ننجح
(2)	**Tagged sentence (complete version)**	تكلمنا/VBD مع/NN الجانب/DTNN الامريكي/DTJJ و/CC طلبنا/VBD ان/IN يوقفوا/VBP العمليات/DTNNS العسكرية/DTJJ من/IN جانب/NN اسرائيل/PUNC ,/NNP و/CC نحن/PRP نحاول/VBP ان/IN نقنع/VBP حركة/NN حماس/NNP ب/IN و/CC قف/NN العمليات/PUNC ,/DTNNS و/CC لكن/VBP للأسف/NN ل/IN م/WP ننجح/VBP
(3)	**Tagged sentence (Keeping noun / adjective tags)**	الجانب/DTNN الامريكي/DTJJ العمليات/DTNNS العسكرية/DTJJ جانب/NN اسرائيل/NNP حركة/NN حماس/NNP وقف/NN العمليات/DTNNS للأسف/NN
(4)	**Sentence after removing tags**	الجانب الامريكي العمليات العسكرية جانب اسرائيل حركة حماس وقف العمليات للأسف

Figure 7: Level of filtering sentence by using the POS Tags feature.

– Normalization, which is used to remove "HAMZA / ء" from the "ALEF / ا" (i.e., the "آ أ إ" are all replaced with the abstract letter "ا").

– Tokenization and Word Segmentation: to segment news posts into their meaningful tokens.

The other pre-processing step we employ is segmentation and part-of-speech (POS) tagging. The Arabic language is rich of affixes like prepositions, conjunctions and pronouns. It is considered one of the highly morphological languages and there is a need to apply involved segmentation and POS tagging algorithms [Althobaiti et al., 2014]. In this way, each word of the post is tagged with the convenient category (nouns, verbs, adjectives, adverbs, pronouns, and so on). E.g., the sentence shown in Figure 3 can be tagged as shown in the Figure 7.

6.1.3 Feature extraction

One of the most critical tasks in NLP is selecting a set of effective features in order to improve the results. We consider several features as follows.

POS Tagging Feature: POS tagging for each words in each post is performed. Note that the aspect terms in the prepared lexicon are often nouns or noun phrases. Therefore, the POS tags are used to retrieve words from noun/adjective categories and discard other tag types (e.g., verbs, adverb, prepositions). Figure 7 presents an example of a news post and how the POS tagging and the filtration step based on it alters the post.

N-Gram Feature: The next step after executing the POS tagging feature is applying n-gram feature to extract all the possible words/phrases of each post.

An n-gram is a contiguous sequence of n words. In our algorithm, a minimum of n = 1 and a maximum of n = 5 are used. So, we create all possible sequences of 1 word, 2 words, . . ., 5 words. Then, all these sequences are compared with the prepared aspects lexicon. If there is a matching result, the words/multiple words are retrieved to check the position feature explained in the next sub-section. E.g., the example of Figure 7 has many n-grams, but the acceptable list with correct meanings which can be found in the prepared lexicon is: 'الجانب الأمريكي' / 'American side', 'العمليات' / 'operations', 'العمليات العسكرية' / 'military operations', 'اسرائيل' / 'Israel', 'حركة حماس' / 'Hamas', 'وقف العمليات' / 'stop the operations'. Actually, this feature is very important because, in many cases, the human annotators select multiple contiguous words as the aspect and we want to capture this.

Position Feature: The successfully matched phrases are marked and their positions are checked in the original text. StartIndex and EndIndex for each returned word/phrase are computed. It is an important step in the algorithm because the lexicon contains some aspect terms that might be parts of other phrases, e.g., {"العمليات / operations", "العمليات العسكرية / military operations"}, {"الطائرات / Planes", "الطائرات الحربية / Warplanes", "الطائرات الحربية الاسرائيلية / Israeli warplanes"}, {"حماس / Hamas", "حركة حماس / Hamas"}. In this example, three groups contain the retrieved aspect terms from the different posts and can be found in the prepared aspects lexicon. The first two groups have the same StartIndex and the last group has the same EndIndex. The essential question is which one of the retrieved aspects in each group is the correct aspect? Inspecting the StartIndex and EndIndex (position) of such words/phrases can help in answering this question accurately.

Pruning Feature: When executing pruning feature, the phrases with the same StartIndex or the same EndIndex (same position) are captured. Then, all the aspects that have shorter lengths are neglected. In contrast, the aspect that has the longest length are stored in the final XML file. E.g., for the example of the previous subsection, the correctly returned aspects from each group are {"العمليات العسكرية / military operations"}, {"الطائرات الحربية الاسرائيلية / Israeli warplanes"} and {"حركة حماس / Hamas"}.

6.2 T2: Aspect term polarity estimation

Our task is to provide a mechanism to automatically determine the sentiment values of the given aspect terms. Then, the accuracy measure is used for evaluation. The proposed lexicon-based approach is summarized in Figure 8.

To clarify the work in this task, two sentiment lexicons have been prepared:

1. A simple sentiment lexicon file is built from the training part of the human annotation XML file. This lexicon is limited and just related to the 2014

Gaza attacks posts. Our sentiment lexicon consists of six different values where each value has a specific meaning as follows:

(a) In all training dataset, the aspect terms that belong exclusively to one sentiment class (positive, negative or neutral) are given values (1, -1, 0) respectively. This case is considered as the highest priority. Examples include {"أبطال / Heroes: positive"} → (1), {"مجازر / Massacres: negative"} → (-1), {"أوضاع / Situations: neutral"} → (0).

(b) If an aspect term belongs to more than one sentiment class in the training data, the frequency of its appearance in each class is computed. The class that has the highest frequency is the class assigned to this aspect. If the highest frequency is for the positive class, then the label value is (11). While if the highest frequency is for negative class, then the aspect provides value (-11). Finally, if the highest frequency is for neutral class, then the aspect provides value (00)). This case is considered as the second highest priority. E.g., {"الطواقم الطبية / Medical staff: positive frequency"} → (11), {"أشلاء / Remains: negative frequency"} → (-11), {"اجتماع / Meeting: neutral frequency"} → (00).

(c) If the frequency is similar for all classes, the given aspect term is removed from the sentiment lexicon because we cannot give an accurate decision about its polarity. E.g., "المبادرة المصرية / Egyptian initiative" → Frequency value for each class (positive, negative, neutral) = (1, 1, 1).

2. An external open source sentimental lexicon (Positive & Negative Arabic sense words) is used and downloaded from the Data Science Lab website.[3] In addition, some general military terms that are collected from the public news articles have been added such as حرب / war, أسر / capture, اغتيال / assassination, قتلى / killed, جرحى / wounded, أسرى / captured, etc. The sense words are labeled by (+1) for the positive words and (-1) for the negative words. In this way, the external and general sentiment lexicon is ready to be integrated into our approach.

After finishing the preparation of the required lexicons, they are called by particular priority and based on specific steps as follows.

1. Each post consisting of noun and noun phrases is retrieved and split it into tokens. In addition, the human annotated aspect terms are read.

2. The sentiment lexicon that was created from the training dataset is opened. Then, each aspect term that belongs to a specific post is compared with terms in the lexicon. If found, then we return its polarity value, which is

[3] `https://sites.google.com/site/datascienceslab/projects/multilingualsentiment`

either (1, -1, 0) → (positive, negative, neutral) for the highest priority, or (11, -11, 00) → (positive frequency, negative frequency, neutral frequency) for the second highest priority.

3. If the aspect term does not exist in the sentiment lexicon, then we check the words (tokens) around each aspect term. In other words, from the given aspect term, we move one word/token backward and one word/token forward and check the tokens' polarities. If they are found in the sentiment lexicon, we return the token values and the cases we have are (1, -1, 0) or (11, -11, 00) → (positive Before/After, negative Before/After, neutral Before/After). This case is referred to as the third highest priority.

4. The external lexicon that contains positive and negative words is opened to check if the given aspect term is found or not. If yes, then the aspect is assigned with different label like (1, -1) → (neg, pos) to distinguish the priority. This case has the fourth highest priority.

5. The last check also involves moving one word/token backward and one word/token forward around each aspect in an attempt to find a polarity for the adjacent tokens. This case has the least priority which is the fifth and it is labeled as (0, 1) → (pos Before/After, neg Before/After).

6. In some cases, aspect terms have a blank polarity value because all previous cases did not find a match, either aspect term does not exist in two lexicons or the token before and after aspect term as well cannot be found in both lexicons. In this case, the algorithm of repairing empty polarities is applied and it has three cases.

   (a) Fill the empty polarity value of the current aspect term with the polarity of the previous aspect term in the same post.

   (b) If there is no previous value, then fill the empty polarity value with the polarity of the next (after) aspect term in the same post.

   (c) The worst case is when all aspects in some posts are still empty, despite all of our attempts. In this case, all the empty polarities will be assigned with the sentiment class that has the highest frequency in the whole training file.

The results of the lexicon-based approach for T1 and T2 are illustrated in Table 10. According to task 1 (T1), the lexicon-based method overcomes the baseline method by 11%, 14% and 12% for precision, recall and F1, respectively. Although the lexicon-based and baseline methods relied on using a simple lexicon to retrieve the terms from the training part of the database, the improvement

Table 10: Lexicon-Based Algorithm Results

Baseline Tasks	P	R	F1	A
T1: Aspect Term Extraction	0.485	0.546	0.514	
T2: Aspect Term Polarity				0.673

ratio for the lexicon-based method is due to applying extra features that are not implemented in the baseline model (such as POS, position and N-gram features).

For task 2 (T2), the improvement in the accuracy measure when it is compared to baseline result is low (around 1%). The reason for this might be due to the nature of the news sentences that carry a small proportion of feelings terms. In addition, there is a certain bias for the Palestinian side and this will affect the annotator decisions. For example, in general the murder is negative and located in a negative sense file, but in our case, it provides a positive polarity affect if it is about killing the enemy.

7 Conclusion

The goal of this research is to evaluate news affect on reader by providing a dataset of Arabic text in news domain. News posts and their comments have been collected during Israel-Gaza conflict of 2014 from well-known Arabic news networks such as Al-Jazeera and Al-Arabiya. They are annotated based on SemEval 2014 guidelines: Task4 for aspect-based sentiment analysis ABSA [Cho et al., 2003]. Moreover, the annotator evaluated each post based on the sentimental affect of news posts on Arab readers with bias to the Palestinian side. Thus, Arab readers' comments clearly reflected their opinion and feelings from the perspective of pro-Palestinian side.

The Arabic news dataset consists of 2265 news posts and 13628 news comments. Dataset has been prepared and annotated to provide important information for all tasks related to the ABSA. The six tasks that were covered in the annotation stage are: aspect terms extraction (T1), aspect term polarity identification (T2), aspect category selection (T3), aspect category polarity identification (T4), comment category detection (T5), and comment category polarity estimation (T6).

Thereafter, the final XML file is evaluated by executing baseline model for six ABSA tasks. The common measurement equations (precision, recall, F1 measure and accuracy) are applied to get the baseline results. Therefore, any proposed approaches concerned of studying the affective news by using ABSA technique will be compared with the results of the baseline model.

As future work, we are working on novel approaches of using aspect-based sentiment analysis (ABSA) in evaluating Arabic news posts affect on readers. In

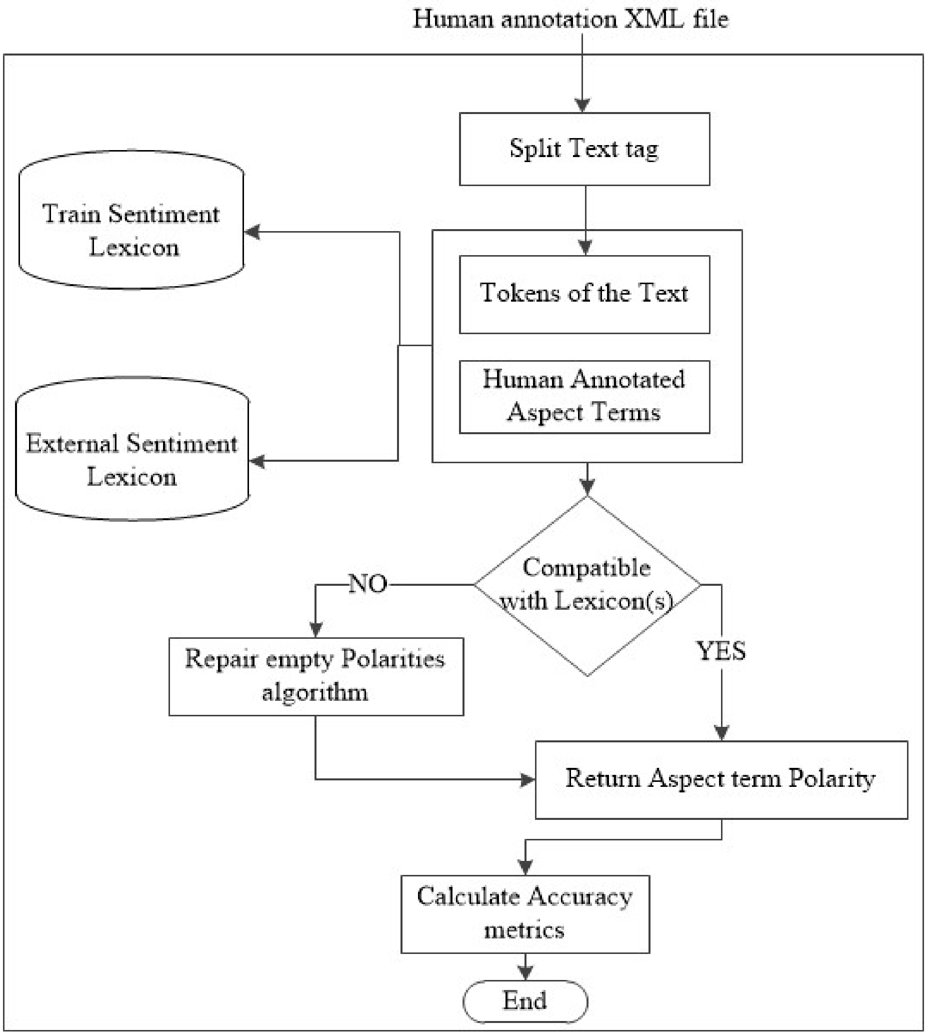


Figure 8: A diagram of the lexicon-based algorithm to solve T2 (Aspect Term Polarity Estimation).

addition, we aim to develop a new approach overcomes baseline results in task 3 and 4 of ABSA tasks, which are: aspect categories detection and aspect category polarity estimation. Finally, we plan to study the level of agreement in opinion and sentiment value between news posts affect and user's comments (replies).

Acknowledgments

This research is partially funded by Jordan University of Science and Technology, Research Grant Number: 20150164.

References

[Abdalkader, 2014] Abdalkader, M. (2014). Sentiment analysis of egyptian arabic in social media. Technical report, Haverford College. Department of Computer Science. B.A. Thesis.

[Abdul-Mageed and Diab, 2012] Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer.

[Abdul-Mageed et al., 2011] Abdul-Mageed, M., Diab, M. T., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 587–591. Association for Computational Linguistics.

[Abdulla et al., 2014] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N., and Al-rifai, S. (2014). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, 9(3):55–71.

[Al-A'abed and Al-Ayyoub, 2016] Al-A'abed, M. and Al-Ayyoub, M. (2016). A lexicon-based approach for emotion analysis of arabic social media content. In *The International Computer Sciences and Informatics Conference (ICSIC)*.

[AL-Smadi et al., 2015] AL-Smadi, M., Al-Ayyoub, M., Al-Sarhan, H., and Jararweh, Y. (2015). Using aspect-based sentiment analysis to evaluate arabic news affect on readers. In *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, pages 436–441. IEEE.

[Al-Smadi et al., 2015] Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M. (2015). Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *3rd International Conference on Future Internet of Things and Cloud (FiCloud), 2015*, pages 726–730. IEEE.

[Allan et al., 1998] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study final report. In *the DARPA Broadcast News Transcription and Understanding Workshop*.

[Althobaiti et al., 2014] Althobaiti, M., Kruschwitz, U., and Poesio, M. (2014). Aranlp: A java-based library for the processing of arabic text. In *the 9th Language Resources and Evaluation Conference (LREC 2014)*.

[Altman, 1990] Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

[Aly and Atiya, 2013] Aly, M. A. and Atiya, A. F. (2013). Labr: A large scale arabic book reviews dataset. In *ACL (2)*, pages 494–498.

[Brody and Elhadad, 2010] Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

[Cho et al., 2003] Cho, J., Boyle, M. P., Keum, H., Shevy, M. D., McLeod, D. M., Shah, D. V., and Pan, Z. (2003). Media, terrorism, and emotionality: Emotional differences in media content and public reactions to the september 11th terrorist attacks. *Journal of Broadcasting & Electronic Media*, 47(3):309–327.

[Eshbaugh-Soha, 2010] Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, 27(2):121–140.

[Farhi, 2009] Farhi, P. (2009). The twitter explosion. *American journalism review*, 31(3):26–32.

[Farnsworth and Lichter, 2010] Farnsworth, S. J. and Lichter, R. S. (2010). *The nightly news nightmare: Media coverage of US presidential elections, 1988-2008*. Rowman & Littlefield Publishers.

[Ganu et al., 2009] Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6. Citeseer.

[Gentzkow and Shapiro, 2010] Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

[Habash, 2010] Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

[Hmeidi et al., 2014] Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., and Mahyoub, N. A. (2014). Automatic arabic text categorization: A comprehensive comparative study. *Journal of Information Science*.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Liu and Zhang, 2012] Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

[Malandrakis et al., 2013] Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2013). Distributional semantic models for affective text analysis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(11):2379–2392.

[Munezero et al., 2014] Munezero, M. D., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

[Papacharissi and de Fatima Oliveira, 2012] Papacharissi, Z. and de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #egypt. *Journal of Communication*, 62(2):266–282.

[Park et al., 2009] Park, S., Kang, S., Chung, S., and Song, J. (2009). Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM.

[Park et al., 2010] Park, S., Lee, S., and Song, J. (2010). Aspect-level news browsing: understanding news events from multiple viewpoints. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 41–50. ACM.

[Pavlopoulos, 2014] Pavlopoulos, I. (2014). *Aspect based sentiment analysis*. PhD thesis, Athens University of Economics and Business.

[Pontiki et al., 2014] Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 27–35.

[Rushdi-Saleh et al., 2011a] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011a). Bilingual experiments with an arabic-english corpus for opinion mining. In *Recent Advances in Natural Language Processing*.

[Rushdi-Saleh et al., 2011b] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011b). Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

[Thet et al., 2010] Thet, T. T., Na, J.-C., and Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, page 0165551510388123.

[Thoits, 1989] Thoits, P. A. (1989). The sociology of emotions. *Annual review of sociology*, pages 317–342.

[Viera et al., 2005] Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363.

[Wan and Yang, 2007] Wan, X. and Yang, J. (2007). Collabsum: exploiting multiple document clustering for collaborative single document summarizations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 143–150. ACM.

[Young and Soroka, 2012] Young, L. and Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.