Journal of Universal Computer Science, vol. 22, no. 5 (2016), 630-649
submitted: 15/9/15, accepted: 29/4/16, appeared: 1/5/16

An Aspect-Based Sentiment Analysis Approach to Evaluating Arabic News Affect on Readers

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

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)

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

Abstract: Great challenges arise due to the rapid growth of online data. The widespread use of online social networks (OSN) have enabled the generation of massive amounts of raw data where users post their own material. One interesting example of user generated data is their political views and opinions. The ability to crawl OSN and automatically analyze their political content is of undeniable importance. However, this requires automated methods for posts' tone analysis, sentiment analysis, and emotional affect. The purpose of this paper is to evaluate Arabic news posts affect on readers using a novel approach of aspect-based sentiment analysis (ABSA). There are many tasks typically associated with ABSA such as the extraction and polarity identification of aspect terms and categories. The focus of this work is on the tasks related to aspect terms. A typical approach to address these tasks goes through several stages of text pre-processing, features extraction and classification. This paper follows this approach and makes use of widely used features and classifiers. The features considered include Part of Speech (POS) tagging, Named Entity Recognition (NER), and N-Grams. As for the considered classifiers, they are: Conditional Random Fields (CRF), Decision Tree (J48), Naive Bayes and K-Nearest Neighbor (IBk). A set of experiments are conducted to compare the considered classifiers against each other and against a baseline classifier that is very common for ABSA. The results show that the extracted features allow all of the four considered classifiers to significantly outperform the baseline classifier. They also show that J48 performs the best for the task of aspect terms extraction whereas CRF and Naive Bayes are slightly better in aspect terms polarity identification.

Key Words: Affective News; Aspect-Based Sentiment Analysis; Aspect Terms Extraction; Conditional Random Fields; J48; Naive Bayes; IBk

Category: H.3, H.3.5, H.4.3, I.2.7, J.4

1 Introduction

Affective news is the domain of research where a news posts sentiment or tone is evaluated according to their emotional affect on readers. In general, the affect can be positive, negative, or neutral. However, more detailed level of emotional affect can be considered such as happiness, sadness, surprise, frustration, etc. Plenty of studies are showing the affect media could play on citizens' behavior. For instance, the studies that evaluate the tone of news covering presidents' and political parties' views [Eshbaugh-Soha 2010, Farnsworth and Lichter 2010], economics [Gentzkow and Shapiro 2010], or Wars and conflicts between nations [Cho et al. 2003]. The tone of news or sentiment affects readers' behavior and attitudes. Therefore, scholars have started to investigate possible techniques of automated news posts analysis. However, available research focuses on document level sentiment of news and deploys approaches that merely focus on sense words or words emotional affect compiled into dictionaries or lexicons [Young and Soroka 2012].

Words emotional affect highly depends on news post context. For instance, although the term 'kill' has a negative emotional affect in general, in the context of Wars, it may have two contradictory emotions depending on which side you support in the war. If the news post is reporting that the army of your country has killed soldiers of the enemy, you may get a positive emotional affect as your side has made a victory. The same situation happens in news reporting on sport events such as Football matches; your team's win is an opponent's loss. Therefore, investigating approaches other than lexicon-based ones is needed.

Despite the availability of research targeting affective news problem in different languages such as English, few examples - if any - can be found for Arabic news. In the context of this study, Arabic refers to Modern Standard Arabic (MSA). MSA is derived from Classical Arabic (Quranic Arabic) and it is the official Arabic language used in media, education, literature, official documents, and most of new books throughout the Arab world. MSA is the native language of 420 million people and spans many regions of the Middle East and North Africa (MENA) in addition to parts of East Africa (Horn of Africa). It is one of the six official languages used in the United Nations [Abdulla et al. 2014, Habash 2010].

In this research, we present a novel research of evaluating Arabic news tone and their affect on readers. Moreover, we report on using aspect-based sentiment analysis (ABSA) as a novel technique to tackle the related research problems. News posts evaluated as part of this research were collected out of social media during the Israel-Gaza conflict in the summer of 2014. There are many tasks typically associated with ABSA such as the extraction and polarity identification of aspect terms and categories, however, the focus of this work is on the tasks related to aspect terms. A typical approach to address these tasks goes through several stages of text pre-processing, features extraction and classification. This paper follows this approach and makes use of widely used features and classifiers. The features considered include Part of Speech (POS) tagging, Named Entity Recognition (NER), and N-Grams. As for the considered classifiers, they are: Conditional Random Fields (CRF), Decision Tree (J48), Naive Bayes and K-Nearest Neighbor (IBk). The results show the superiority of the proposed approach compared with a well-known baseline classifier for ABSA.

The rest of this paper is organized as follows. Section 2 provides a literature review of affective news analysis and related work of ABSA. Section 3 explains the dataset collection and annotation. Section 4 discusses the research methodology. Section 5 presents the experiments and their settings while Section 6 discusses the results. Section 7 concludes the paper and sheds light onto future directions of this work.

2 Related Works

Our coverage of the existing literature is divided into two parts. One is dedicated to discuss the works related to the problem at hand, i.e., affective news analysis, while the other is dedicated to the works related to the approach followed in this work, i.e., Sentiment Analysis (SA).

2.1 Affective News Analysis

Much research effort has been made on news articles in various topics such as: topic detection and tracking on news stories [Allan et al. 1998], news summarization [Wan and Yang 2007] and news categorization [Sebastiani 2002]. They have essentially focused on helping people proficiently browse an almost unlimited amount of news articles and easily get the required data. Subsequently, these researches, for the most part, saw the connection and relation between readers and news articles.

In [Park et al. 2009], the authors presented NewsCube, an Internet news service designed for relieving and decreasing the impact of media bias. It gathers news articles reporting on an event from various news sources, arranges them and presents articles with diverse perspectives. The adequacy of the model was assessed and reported through different client studies. The outcomes demonstrated that the model helps readers thoroughly analyze diverse perspectives of news events and build up their own particular perspectives.

The framework in [Park et al. 2010] is based on aspect-level news browsing which provides readers with a labeled news articles from different perspectives. It encourages dynamic interactions whereas readers effectively find and look at different biased perspectives over a news event. Therefore, it adequately helps readers comprehend the news event from a plural of perspectives and create their own adjusted perspectives free from particular biased perspectives.

Aspect-level news browsing allows readers to read different news articles for a selected event. Thus, there are diverse viewpoints and need to apply aspect-level classification in order to identify prominent aspects. In this way, the articles are classified based on the similarity of these selected aspects. They focused on solving two tasks: aspect extraction and article classification. They developed two approaches namely: news structure-based extraction (NSE) and framing cycle-aware clustering (FCAC). NSE [Park et al. 2009] uses the news writing rules to abstract the aspects of each news article automatically. It can be applied to many articles regardless of the language and article topic. FCAC successfully uncovers the variety of distinctive aspects of a news event.

O'Connor et al. in paper [O'Connor et al. 2010] depended on link measures of general opinion obtained from polls with sentiment that can be measured through analysis of texts on Twitter [Malandrakis et al. 2013]. The results demonstrated that there is a high degree of compatibility between textual sentiment in tweets and the traditional polling results.

The goal of [Wu and Ren 2011] is to design a model that has the ability to study the sentimental influence in OSN in general and tweets in particular. Moreover, this model checks if a user has high or low sentimental influence on his/her surrounding neighbors such as user's friends and followers by calculating the *influencing probability*. Thereafter, the model calculates the *influenced probability* of the user's neighbors on him/her. The tweets are collected from user's accounts and the approach depend on using an unsupervised lexicon based method to count the number of positive and negative words in each tweet. If the number of positive words is greater than the negative ones, then the tweet is considered as a positive tweet, vice versa. The results demonstrated that the number of negative tweets is more than the number of positive tweets and, thus, concluded that the users prefer to change the current status when feeling bad. In addition, there is a close correlation between the influencing and influenced probabilities. Consequently, most of the users who have low influencing and influenced probabilities maintain balanced sentimental influence. On the other hand, people who have high influencing and influenced probabilities are going to be either positively or negatively influencing/influenced by other people.

Some of the existing applications are focusing on the feeling and emotion communicated by the writer (the viewpoint of the writer), instead of the emotion felt by the reader [Malandrakis et al. 2013]. Other applications may concentrate on the reader or media customer point of view to study their emotion recognition, for example, the analysis of news headlines [Strapparava and Mihalcea 2007].

News describing the regional and international events and can carry a high level of emotional expressive content. Some news posts contain words or phrases that convey affective meaning and are often written by experts to attract the readers' attention. The authors of [Strapparava and Mihalcea 2007] used an unsupervised setting which does not require training. The data are collected from news web sites (i.e., CNN and Google news) and concentrate on news headlines that have a few words. Then, a team of specialist annotators labeled each headline with an appropriate emotion based on their first intuition. There are six predefined emotion labels (such as Sadness, Anger, Disgust, Surprise, Fear, Joy) and three valence sentimental value for each title as positive, negative and neutral. They applied three systems (namely: SWAT, UA and UPAR7) to evaluate emotion labeling. Thereafter, they applied five systems (namely: UPAR7, SICS, CLaC, CLaC-NB, UA) to evaluate valence classification (positive, negative and neutral). The results show that the emotion annotation task is difficult to apply. In addition, there is disparity between the results acquired by the systems due to the low Pearson correlation measure for the inter-annotator agreement. Consequently, there is a great opportunity for improvement in future work.

In [Sun and Ng 2014], the authors built an emotional dictionary in both English and Chinese that is compatible with familiar OSN. The proposed method has the ability to detect the emotions and estimate the sentimental influence of posts. In this research, the authors are interested in differentiating between personal and public topics to increase the percentage of accuracy when measuring the sentimental influence. A personal topic is mainly produced by users who upload the posts via social network. The replies are clearly affected by the sentiments of user's post and are therefore less affected by outsourcing. In contrast, a public topic is published in the media before the users write on it such as news and current affairs. Therefore, the replies will be affected by outsourcing substantially. In general, the influence of the post can be judged based on the number of replies that commented on it. Furthermore, these replies should carry a sentimental value of either positive or negative.

The authors of paper [Barbagallo et al. 2012] studied the tourism domain and collected their tweets from Twitter. The experiments have shown that the negative tweets are more retweeted. On the other hand, they did not study the sentimental influence of posts and whether negative tweets got approval or disapproval of the readers. Thus, the research [Sun and Ng 2014] aimed to fill this gap and address how much the sentiment of influential posts (influencer) will affect the sentiment of readers (receptors). Thus, if the sentimental influence of the influencer and the receptor are equal, then they are in a "compliance" situation. Otherwise, they are in an "opposition" situation. The final results demonstrated that positive polarity in personal topics and the negative polarity of the influencers in public topics had the greatest compliance in the sentimental influence on their readers (receptors).

To evaluate the actual mood state of the user, a rating system called Profile Of Mood States (POMS) is used and many researchers checked its validity on different cases. Wyrwich and Yu [Wyrwich and Yu 2011] studied the dataset related to postmenopausal women and applied POMS system to evaluate their mood states. This system has the ability to measure seven emotion labels (active, friendship, angry, anxious, depressed, confused, and fatigue). In another work, Gibson [Gibson 1997] applied POMS system on samples of older and young adults and examined its reliability to assess their current mood. The same seven predefined emotion labels are used to compare between the two samples. The results proved the validity and reliability of POMS system and its ability to truly reflect the current mode status of a human. In addition, it is useful for extracting and detecting emotional words from the posts and build an emotional word dictionary.

2.2 Sentiment Analysis (SA)

Sentiment Analysis (SA) is concerned with extracting the sentiments conveyed in a piece of text. Existing works focus on the sentiment of an entire document (article, review, essay, etc.) or a single sentence. However, this is not very useful in all cases. For example, a review about a camera might contain conflicting sentiments in different parts of it. The reviewer might be praising the camera's focus and image quality and in the same sentence criticizing its power usage. Reviews conveying different sentiments about different aspects of the same product require an aspect-based SA (ABSA) approach to process them. Compared with traditional SA, ABSA poses several added challenges such as linking each part of the text with the aspect to which it refers, identifying the parts of text discussing the same aspect (e.g., one reviewer might be talking about battery life while the other is discussing power consumption; obviously, both are referring to the same aspect), dealing with comparative sentences, etc. Most of the existing works on ABSA focus on English with very rare works on other languages. Below, we discuss some of the works on ABSA in English.

One of the earliest papers on ABSA is [Hu and Liu 2004], which followed a frequency-based approach. The basic idea is that frequently mentioned nouns are more likely to be aspects. To compensate for the errors resulting from ignoring infrequent names, the authors suggested exploiting opinion words to find aspects. For this part, they followed a very simple approach of considering the nearest opinion word. This idea was used in many papers [Zhuang et al. 2006, Qiu et al. 2009]. In [Popescu and Etzioni 2007], the authors suggested using the part-of relationship to remove frequent noun phrases that are not aspects. Benefiting from statistics about the use of nouns in the English language, Scaffidi et al. [Scaffidi et al. 2007] improved the general approach of relying frequent nouns to extract aspects. In a more recent work [Long et al. 2010], Long et al. added the use of information distance, and dependent words (adjectives). Another interesting approach is the use of centering theory and leveraging the use of title words for the extraction of aspects in news comments

[Ma and Wan 2010]. Finally, SemEval, one of the prestigious events in NLP that is dedicated to the evaluation of computational semantic analysis systems, paid special attention to ABSA by dedicating one of its tasks in 2014 to ABSA [Pontiki et al. 2014]. The success of this task has led to repeating it in 2015 [Pontiki et al. 2015] and 2016 [Pontiki et al. 2016]. Such an event presents an opportunity to introduce rapid improvements in the considered problem and it is a testimony to the importance of the considered problem.

Based on the SemEval guidelines, few papers addressing the ABSA of Arabic text have recently appeared.For instance, the authors of [Al-Smadi et al. 2015] filtered the Large Arabic Book Reviews (LABR) dataset of [Aly and Atiya 2013] and re-annotated it for ABSA purposes creating the Human Annotated Arabic Dataset of Book Reviews (HAAD). Another ABSA dataset of Arabic reviews is the one prepared for the 2016 version of the SemEval ABSA task [Pontiki et al. 2016] for Arabic hotel reviews. Clearly, none of these datasets consider the domain of politics or affective news.

3 Dataset Collection and Annotation

3.1 Related Tasks

As this research adopts ABSA as the main approach to evaluate news posts sentiment and their affect on readers. The following ABSA related tasks from [Al-Smadi et al. 2015] are targeted.

T1: Aspect Terms Extraction. Given a news post, this task deals with extracting all possible aspect terms with respect to the news event. Examples of annotated aspect terms are: (bombing, ground invasion, Israeli Air Force). The extraction of aspects is done regardless to their polarity. For instance, conflict and neutral aspect terms should be extracted as well.

T2: Aspect Terms Polarity Identification. Depending on previous task (T1), this task focuses on assigning the extracted aspects to the polarity class (positive, negative, and neutral).

T3: Aspect Category Identification. Having a predefined aspect categories and a collection of review sentences (without any annotations), this task investigate the ability of assigning each review sentence to one or more aspect category. The difference between this task and T1 is that the aspect terms are more fine-grained and should appear in the review sentence, whereas the aspect category is coarser category of the sentence and do not appear in the review sentence. Moreover, the aspect categories are not identified using aspect terms in the sentence, but rather inferred using sense words, adjectives, or context of the sentence meaning.

T4: Aspect Category Polarity. Having that the aspect categories of the review sentences are given, this task investigates the possibilities of assigning a specific polarity (positive, negative, and neutral) to each aspect category.

This paper is related to research conducted on the first two tasks namely, T1 and T2 only. According to these tasks, this research aims at answering the following questions:

– What are the main entities in news posts covering the 2014 Israel-Gaza conflict and what are their sentimental affect on readers?

– What are the main text features that may influence the research of Affective news?

– How news affect can be evaluated using aspect based sentiment analysis.

3.2 Data Collection

Our study deals with short news posts on social media - i.e. Breaking News. News posts are collected from well-known Arabic news networks such as Al Jazeera[1] and Al Arabiya.[2] A tool named netvizz v1.05 has been used to crawl news posts out of a Facebook page related to this purpose. For the sake of our case study - the 2014 Israel-Gaza conflict - all the news posts were taken from a Facebook page ("عاجل من غزة" / "Breaking news from Gaza"). The Facebook page was collecting breaking news from Arabic news networks during the 2014 Israel-Gaza conflict. Over 10K posts were collected covering different events during the conflict (e.g., the kidnapping process of three young settlers on June 13, the start of Operation Protective Edge, etc.). In total, 2,265 Arabic news posts were annotated to support the research tasks.

3.3 Annotation Process

The dataset was annotated by humans to support the aforementioned tasks: aspect terms extraction (T1), and aspect terms polarity identification (T2). For the sake of annotation, the annotators used the BRAT web-based annotation tool [Stenetorp et al. 2012] which was configured to meet the needs of the annotation phase. Figure 1 depicts an example for an annotated post in BRAT as done by annotators.

The annotation process has been done by a group of three members, a graduate student and two senior researchers. All participants were native Arabic speakers. In the first phase, the graduate student annotated the news posts, and then some posts were randomly evaluated by the other two senior researchers.

[1] `http://www.aljazeera.net/`
[2] `http://www.alarabiya.net/`

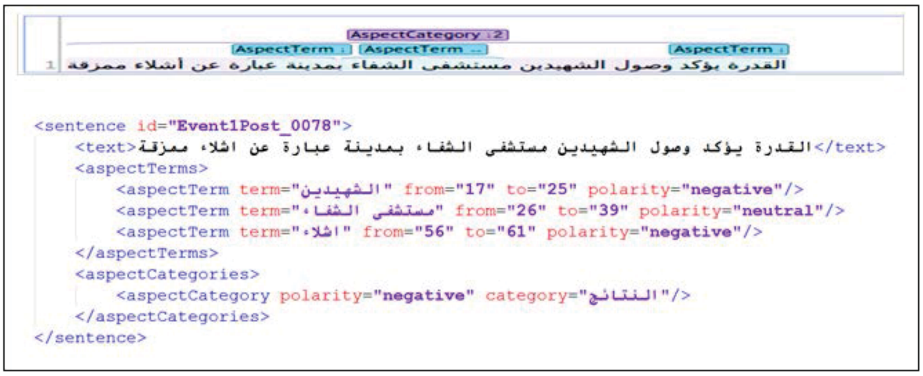

Figure 1: Example of Brat annotated news post and its XML representation

Dataset	Polarity			Total
	Positive	Negative	Neutral	
Overall Dataset	4,136	4,649	807	9,592

Table 1: Aspect Terms and Their Polarities

The conflict and incorrectly annotated posts were discussed with the graduate student and a second round of annotation was conducted. Out of annotation phases we had 2,265 Arabic news posts annotated and formatted as XML.

The annotators had a training session on annotation using BRAT and were given guidelines for the required annotation as follows: the annotators were asked to annotate all single/multiple-terms that have a positive or negative affect on readers (e.g., bombing, ground invasion, Israeli Air Force). The annotation process adopted a pro-Palestinians point of view. The aspect terms were annotated as they appeared in the original post even if they were misspelled. For each annotated aspect term, the annotators were asked to provide a polarity value (positive, negative, and neutral). In total 9,592 aspect terms were annotated, divided as follows (Positive Class = 4,136), (Negative Class = 4,649) and (Neutral Class = 807). Table 1 summarizes the aspect terms distribution over the sentiment class in the annotated dataset.

The SemEval-2014 Task 4 dataset schema [Pontiki et al. 2014] has been selected as a dataset possible format. The selected schema was used to configure the BRAT tool annotation and a mapping layer where used to map the BRAT annotated file to the SemEval-2014 Task 4 compliant XML file.

As depicted in Figure 1, each news post is annotated based on the following XML format.

– `<aspectTerm term=" " polarity=" " from="" to=""/>` XML element for each occurrence of an aspect term. In addition to the aspect term polarity, its location in the text is provided based on start and end index of text characters.

– `<aspectCategory category=" " polarity=" "/>` XML element for each occurrence of an aspect term category.

The XML-based dataset is publicly available upon request for non-commercial research.

4 Research Methodology

As the dataset was annotated and formatted based on the SemEval-2014 Task 4 guidelines, the task's baseline evaluation is used. It is explained as follows [Pontiki et al. 2014].

4.1 Baseline Evaluation

T1: Aspect terms extraction baseline: the baseline tags all the tokens in the test dataset if they are listed in the human annotations list of aspect terms from the training dataset.

T2: Aspect terms polarity identification baseline: for each aspect term t in the test sentence s, the baseline checks if t has been seen in the training sentences. If yes, the baseline retrieves the d most similar sentences of the training to s, and assigns to the aspect term t the most frequent polarity in the d sentences. The Dice coefficient similarity measure is used to compute the distance between sentences s and d. If not, if t has not been seen in the training set, t is assigned the most frequent aspect term polarity label in the training set.

4.2 Supervised Approach of Affective News Evaluation

Our approach uses a supervised learning approach to evaluate the news post affect on the readers. The approach consists of three phases, namely: (a) text pre-processing, (b) features extraction, and (c) news post affect classification.

4.2.1 Text Pre-processing

A pre-processing and cleaning phase is has been conducted to prepare the dataset for further processing. Since the candidate news posts are written in modern standard Arabic (MSA), which is easier to handle compared with Dialectal Arabic (DA), one of the standard Arabic text processing tools can be used. For this work, a Java-based toolkit called AraNLP was used to handle text pre-processing

[Althobaiti et al. 2014]. AraNLP collects widely used NLP tools for Arabic text mining and place them into a single package. Examples of available services/tools include: tokenization, segmentation, stemming, part-of-speech tagging (POS), punctuation and stop words removal.

Text pre-processing was used to ensure that noisy unwanted words are removed from the news posts. The pre-processing phase involved the following steps.

- Removal of non-Arabic words.
- Removal of hyperlinks and hash-tags in all posts.
- Removal of Arabic diacritics such as "ًٌٍَُِّْ".
- Removal of punctuations and symbols such as "? ' ! @ $ # |".
- Normalization, which is used to remove "HAMZA/ء" from the "ALEF/ا" (i.e., the "آ إ أ إ" are all replaced with the abstract version of the letter "ا").
- Tokenization & Word Segmentation in order to segment news posts into their meaningful tokens.

4.2.2 Features Extraction

The output of the processed posts was used to extract a set of features through which the classification process will be enhanced. Based on our literature review, Part of Speech (POS) tagging, named entity recognition (NER), and N-Grams features are widely used in the domain of ABSA [Pontiki et al. 2014, Pontiki et al. 2015, Pontiki et al. 2016]. Therefore, in our approach, these main features were extracted to support both tasks T1 and T2. Moreover, the aspect-term position in the news post represented by the count of the starting character to the ending character was used as an additional feature in our approach. AraNLP provides an implementation for Stanford POS tagger for Arabic and N-Grams. As for the NER feature, a web service based on the work of [Al-Rfou et al. 2015] was used. The tool supports the recognition of three major classes of entities (namely: Person, Location, and Organization), in addition to another class represented by the letter 'O' to represent any other entity class different of the three major ones.

Based on our data annotation for the task of aspect terms extraction, aspect terms can be a single term or multiple terms (phrase). Therefore, N-Grams were used as a feature pruning aid following the work of [Hu and Liu 2004]. A priority for the N-Grams has been given over Bi-grams and Uni-grams for the terms that appear in single and multiple forms. For instance, the Tri-gram aspect term "Israeli Air Force/سلاح الجو الاسرائيلي" is given a higher priority over the Bi-gram

TermEN/AR		From	To	ANER	POS	Human Labeling
Abbas	عباس	1	4	B-PER	NNP	Non-Aspect
we spoke	تكلمنا	6	11	O	VBD	Non-Aspect
With	مع	13	15	O	NN	Non-Aspect
Side	الجانب	17	23	O	DTNN	First-Aspect
the American	الامريكي	25	32	O	DTJJ	Second-Aspect
and asked	وطلبنا	34	39	O	VBD	Non-Aspect
To	ان	40	41	O	IN	Non-Aspect
Stop	يوقفوا	43	48	O	VBP	Non-Aspect
Operations	العمليات	50	57	O	DTNNS	First-Aspect
the military	العسكرية	59	66	O	DTJJ	Second-Aspect
From	من	68	69	O	IN	Non-Aspect
Side	جانب	71	74	O	NN	Non-Aspect
Israel	اسرائيل	76	82	B-LOC	NNP	Unique-Aspect
and we	ونحن	84	87	O	PRP	Non-Aspect
are trying	نحاول	89	93	O	VBP	Non-Aspect
To	ان	95	96	O	IN	Non-Aspect
Convince	نقنع	98	101	O	VBP	Non-Aspect
Hamas	حركة	103	107	B-ORG	NN	First-Aspect
	حماس	109	112	I-ORG	NNP	Second-Aspect
to halt	وقف	113	115	O	NN	First-Aspect
the operations	العمليات	117	124	O	DTNNS	Second-Aspect
But	ولكن	126	129	O	VBP	Non-Aspect
Unfortunately	للاسف	132	136	O	NN	Non-Aspect
do not	لم	138	139	O	NEG	Non-Aspect
Succeed	ننجح	141	144	O	VBP	Non-Aspect

Figure 2: The tokens of a news post sample along with their translations, computed features and human annotation

aspect term “Air Force/سلاح الجو” if they appear in a news post. Although both cases represent aspect terms that appear in the human annotated training file, this step is needed to avoid having a large number of “True Negatives” during the test phase.

Figure 2 shows a news post sample consisting of 25 Arabic tokens. The figures shows each Arabic token along with its rough English translation. It also shows the computed features for each token as well as the human annotation.

4.2.3 News Post Affect Classification

The extracted features were then used to train four different classifiers namely: Conditional Random Filed (CRF) classifier, Decision Tree (J48) classifier, Naive Bayes classifier and K-Nearest Neighbor (IBk) classifier. More details about these classifiers are given below.

- Apart from other classifiers, CRF considers the classification problem context and is able to predict tags or labels for a sequence of input data (such as a sentence of tokens) based on their conditional dependence. CRF builds on the assumption that tags/labels do not occur independently, but rather each tag/label has a conditional dependence on the tag/label of the previous word. CRF plays a prominent role in classification problems such as POS tagging and NER.

- J48 is a Java-based implementation of the C4.5 decisions tree algorithm [Quinlan 1993], which builds an enhanced decision tree classifier based on the concept of information entropy. The algorithm eliminates bad features during the training phases as it chooses the most appropriate feature on which to split a node based on the difference of the entropy value (information gain).

- The Naive Bayes classifier is a simple linear classifier based on Bayes rule. Despite relying on the unrealistic assumption that the features are mutually independent, the Naive Bayes classifiers performs well for many classification problems including text classification.

- Finally, IBk is an implementation of the K-Nearest Neighbor classifier. IBk is a lazy classifier that does not build a classification model. Instead, it relies on computing the distance (using some similarity measure such as the cosine measure) between the feature vector of the unseen example and the feature vectors of the training examples. The decision is then made based on the k nearest neighbors of the unseen example.

All of these classifiers are selected based on their appropriateness to the research domain and their prominent role in similar classification problems.

As part of task T1: Aspect terms extraction, the processed news post was provided as a sequence of tokens associated with their features, which were then used to train the classifiers. For the sake of N-Grams aspect-terms, a label is assigned to each token in the aspect phrase representing the token's order in the phrase. For instance, the "Israeli Air Force" is a Trigram aspect-term that is represented as a sequence of labels as follows: "Force: First-Aspect", "Air: Second-Aspect", and "Israeli: Third-Aspect". Each token can be labeled as a "Unique-Aspect" as well if it appears alone in the news post. The rest of the tokens (that are not annotated as aspect term) are labeled by the "Non-Aspect" label.

Measures	Method				
	Baseline	CRF	J48	Naive Bayes	IBk
Precision	0.3822	0.73095	0.8134	0.7978	0.7892
Recall	0.37081	0.78212	0.825	0.7772	0.7908
F_1 measure	0.3762	0.7554	0.817	0.7814	0.7898
ROC	–	–	0.9212	0.9318	0.881

Table 2: Experimentation Results of T1

In task T2: Aspect Term Polarity Identification, the task focuses on polarity identification of a given aspect term. The classifiers were trained with the aspect terms labels and their polarity value (i.e. positive, negative, and neutral).

5 Experimentation and Results

The annotated dataset was used in evaluating our proposed method with respect to the research defined tasks. The experiments used 5-fold cross validation technique to split the dataset into training and testing files. The training file was used to train the classifiers with the tasks associated features as discussed earlier. Nevertheless, the same dataset of training and testing files was used for the baseline evaluation approach as well. In order to evaluate the proposed method, the results out of the proposed approach were compared to the baseline evaluation results.

In order to evaluate the aspect terms extraction (T1), the F_1 measure, precision (P) and recall (R) were computed. On the other hand, the accuracy of the approach was used to evaluate the aspect term polarity identification (T2). The accuracy measure is defined as the number of correctly predicted polarities divided by the total number of aspect terms polarity annotations. In order to evaluate the effectiveness and performance of the classifiers, the Receiver Operating Characteristics (ROC) [Hanley and McNeil 1982] measure has been computed for both tasks. The baseline results for the two tasks are as follows: T1 (F_1 = 0.3762) and T2 (Accuracy = 0.6147)

5.1 T1: Aspect Terms Extraction

In order to train and test our method based on the selected classifiers, the SimpleTagger[3] (Mallet implementation of the CRF) and the Weka 2.7 implementation of J48, Naive Bayes and IBk were used. Experimentation results showed that all classifiers outperform the baseline one ($F_1 = 0.3762$). The CRF classifier achieved

[3] http://mallet.cs.umass.edu/sequences.php

Measures	Method				
	Baseline	CRF	J48	Naive Bayes	IBk
Accuracy	0.6147	0.865	0.8628	0.8647	0.8515
ROC	–	–	0.9632	0.9782	0.936

Table 3: Experimentation Results of T2

F_1 value of 0.7554 whereas the Naive Bayes and IBk had slight enhancements of 3% over CRF with F_1 values of 0.7814 and 0.7898, respectively. Finally, the J48 classifier gave the best performance with F_1 value of 0.817. Table 2 summarizes T1 experimentation results for the baseline and the proposed classifiers.

As presented in Table 2, the performance of the classifiers was ranging form very good ($ROC = 0.881$) for the IBK to excellent ($ROC = 0.9212$, $ROC = 0.9318$) for both the J48 and the Naive Bayes respectively. The missing ROC values for the baseline and the CRF classifiers are due to the limitations in the used packages to implement them. These packages lack the ability of classifiers' prediction values (probability) output printing. Thus, the '–' ROC values in Table 2.

5.2 T2: Aspect Terms Polarity Identification

As this task deals with evaluating the proposed approach's ability to predict the aspect term polarity given that the aspect term is known, the same training file of T1 in addition to the aspect term feature was used to train the classifiers. Experimentation results showed that the proposed approach outperforms the baseline evaluation (Accuracy = 0.6147). The accuracy of the CRF classifier was 0.865 whereas the accuracies of the J48 classifier and the Naive Bayes classifiers were almost the same (0.8628 and 0.8647). On the other hand, the accuracy of the IBk classifier is slightly lower (0.8515). Moreover, classifiers' performance in T2 is excellent as the ROC values are greater than 90%. Table 3 summarizes T2 results.

6 Discussion

According to the experimentation results, it can be noticed that classification techniques trained on text features namely, POS, NER, and N-grams outperforms baseline approaches based on terms frequency and Bag-of-Words (BoW) containing the human annotated aspect terms. However, in order to evaluate the impact of the selected features the experiments were repeated for both tasks several times and each time a specific feature was hidden. Surprisingly, hiding some features during experimentation of both tasks had a noticeable influence

Feature	Method			
	CRF	J48	Naive Bayes	IBk
Without POS	0.7019	0.812	0.7946	0.7906
Without NER	0.7540	0.815	0.7824	0.7912
Without Position	0.7758	0.816	0.805	0.8164
All Features	0.7554	0.817	0.7814	0.7898

Table 4: F_1 Results with Different Feature Settings in T1

Feature	Method			
	CRF	J48	Naive Bayes	IBk
Without POS	0.863	0.8629	0.8652	0.853
Without NER	0.864	0.8629	0.8657	0.8523
Without Position	0.879	0.8630	0.87	0.8639
All Features	0.865	0.8628	0.8647	0.8515

Table 5: Accuracy Results with Different Feature Settings in T2

only on the CRF classifier whereas the remaining classifiers were marginally affected. The most noteworthy observation is that the J48 results were almost the same with a very small negligible change in F_1 measure and accuracy values (less than 0.01). As for the case of most apparent influence, which was on CRF both tasks T1 and T2, hiding the POS feature had a drop in the F_1 measure value of almost 6% in T1 and almost 1% in T2, hiding the NER feature had a negligible loss in both tasks, whereas hiding the position feature increased the classifier F_1 measure value by almost 3% in T1 and almost 2% in T2.

The above findings highlight the importance of POS tags as a main feature for both tasks and gives an empirical evidence that the term position in the text represented as ‘from’ ‘to’ numbers acts as a bad feature and has a negative impact on the CRF classifier’s accuracy. This can be explained, as in the case of POS feature, by the case that consequent tokens which have dependency in terms of their syntax lead to have a conditional dependency in terms of their probability to appear consequently in the text. The position feature represented as abstract numbers lack such logic and, therefore, had a negative influence on the CRF classifier. Moreover, the CRF algorithm is not able to eliminate bad features during the training phase as in the case of the J48 (C4.5) one. J48 (C4.5) builds the decision tree based on data divergence where the data with the highest information gain (difference in entropy values) are used to make the split decision on each node. Therefore, J48 results were almost the same when features were hidden (See Tables 4 and 5).

Finally, based on our study of the dataset and the extracted features, It was noticed that the entities that were affecting the readers the most vary depending on the stage of the conflict. For example, during the early stages of the conflict, entities such as "خطف" / "kidnapping" "تعذيب" / "torture" were the most influential. On the other hand, during the peak of the conflict, entities such as "قصف" / "bombing" "شهيد" / "martyr" were more influential. Finally, towards the end of the conflict, entities such as "تهدئة" / "truce" "نصر" / "victory" became more influential.

7 Conclusions and Future Work

Over the past couple of decades, the world witnessed a significant increase in the amount of interactive contents on the Internet where websites are allowing users to post virtually whatever they want. Among the things users post are their thoughts and feelings about different things in life. News posts on social media platforms have a prominent influence on people's attitude and behavior. Mining through such data is of great importance to governments, political parties, corporations, etc.

This research evaluates news posts affect on readers. In order to meet research objectives, a novel approach of using ABSA is proposed. The proposed approach provides empirical results of the applicability of using ABSA in the affective news context.

A dataset of news post covering the 2014 Israel-Gaza conflict were collected and used to evaluate the proposed ABSA approach. Experimentation results show that features such as POS tags and NER types plays an important role in news affect evaluation. Moreover, the CRF, J48, Naive Bayes and IBk classifiers were evaluated and all of them have a great potential in ABSA in general and affective news research in particular.

Future directions of this work include performing a more comprehensive study of the proposed approach using additional features and classifiers and working on more detailed framework focusing on detailed levels of emotional affect such as sadness, happiness, etc. Moreover, the approach can be evaluated on different contexts of news such as sports, politics and technology in order to enrich the domain of news sentimental/affective analysis.

Acknowledgments

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

References

[Abdulla et al. 2014] Abdulla, N. A., Al-Ayyoub, M., and Al-Kabi, M. N.: ”An extended analytical study of arabic sentiments”; International Journal of Big Data Intelligence 1, 1(1-2):103–113, 2014.

[Al-Rfou et al. 2015] Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S.: ”Polyglotner: Massive multilingual named entity recognition”. In the 2015 SIAM International Conference on Data Mining.(2015).

[Al-Smadi et al. 2015] Al-Smadi, M., Qawasmeh, O., Talafha, B., and Quwaider, M.: ”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, 2015.

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

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

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

[Barbagallo et al. 2012] Barbagallo, D., Bruni, L., Francalanci, C., and Giacomazzi, P.: ”An empirical study on the relationship between twitter sentiment and influence in the tourism domain”. In Information and Communication Technologies in Tourism 2012, pages 506–516. Springer, (2012).

[Cho et al. 2003] Cho, J., Boyle, M. P., Keum, H., Shevy, M. D., McLeod, D. M., Shah, D. V., and Pan, Z.: ”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, (2003).

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

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

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

[Gibson 1997] Gibson, S. J.: ”The measurement of mood states in older adults”. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 52(4):167–174, (1997).

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

[Hanley and McNeil 1982] Hanley, J. A. and McNeil, B. J.: ”The meaning and use of the area under a receiver operating characteristic (roc) curve”. Radiology, 143(1):29–36, (1982).

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

[Long et al. 2010] Long, C., Zhang, J., and Zhut, X.: A review selection approach for accurate feature rating estimation. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 766–774. Association for Computational Linguistics, (2010).

[Ma and Wan 2010] Ma, T. and Wan, X.: ”Opinion target extraction in chinese news comments”. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 782–790. Association for Computational Linguistics, (2010).

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

[Obaidat et al. 2015] Obaidat, I., Mohawesh, R., Al-Ayyoub, M., AL-Smadi, M., and Jararweh, Y.: "Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches". In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on, pages 1–6. IEEE, (2015).

[O'Connor et al. 2010] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A.: "From tweets to polls: Linking text sentiment to public opinion time series". In the Fourth International AAAI Conference on Weblogs and Social Media, (2010).

[Park et al. 2009] Park, S., Kang, S., Chung, S., and Song, J.: "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,(2009).

[Park et al. 2010] Park, S., Lee, S., and Song, J.: "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, (2010).

[Pontiki et al. 2016] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., Clercq, O. D., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G.: "SemEval-2016 task 5: Aspect based sentiment analysis". In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California. Association for Computational Linguistics, (2016).

[Pontiki et al. 2015] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I.: "Semeval-2015 task 12: Aspect based sentiment analysis". In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, pages 486–495, (2015).

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

[Popescu and Etzioni 2007] Popescu, A.-M. and Etzioni, O.: "Extracting product features and opinions from reviews". In Natural language processing and text mining, pages 9–28, Springer, (2007).

[Qiu et al. 2009] Qiu, G., Liu, B., Bu, J., and Chen, C.: "Expanding domain sentiment lexicon through double propagation". In IJCAI, volume 9, pages 1199–1204, (2009).

[Quinlan 1993] Quinlan, J. R.: "C4.5: Programs for machine learning". (1993).

[Scaffidi et al. 2007] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C.: "Red opal: product-feature scoring from reviews". In Proceedings of the 8th ACM conference on Electronic commerce, pages 182–191. ACM, (2007).

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

[Stenetorp et al. 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J.: "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, pp. 102–107. Association for Computational Linguistics, (2012).

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

[Sun and Ng 2014] Sun, B. and Ng, V. T.: "Analyzing sentimental influence of posts on social networks". In Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on, pp. 546–551. IEEE, (2014).

[Wan and Yang 2007] Wan, X. and Yang, J.: ”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, (2007).

[Wu and Ren 2011] Wu, Y. and Ren, F.: ”Learning sentimental influence in twitter”. In Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pages 119–122. IEEE, (2011).

[Wyrwich and Yu 2011] Wyrwich, K. W. and Yu, H.: ”Validation of poms questionnaire in postmenopausal women”. Quality of Life Research, 20(7):1111–1121, (2011).

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

[Zhuang et al. 2006] Zhuang, L., Jing, F., and Zhu, X.-Y.: ”Movie review mining and summarization”. In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 43–50. ACM, (2006).