Sentiment Analysis for Micro-blogging Platforms in Arabic

Eshrag Refaee[1,2(✉)]

[1] Department of Computer Sciences and Information Systems, Jazan University, Jazan 45 142, Kingdom of Saudi Arabia
erefaie@jazanu.edu.sa
[2] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH11 4AS, UK

Abstract. Most previous Sentiment Analysis (SA) work has focused on English with considerable success. In this work, we focus on studying SA in Arabic, as a less-resourced language. SA in Arabic has been previously addressed in the literature, but has targeted text genres of more formal/edited domains (e.g. news-wire) and domains containing longer text instances, i.e. with more contextual information (e.g. reviews). That is, less work has focused on SA in Arabic for a noisy and short-length text genre, like micro-blogs. In addition, the time-changing nature of streaming data (e.g. the Twitter stream) has not been considered in previous work, as SA systems were mainly developed and evaluated on small test-sets that are sub-sets of the original data-set used for training.

This work reports on a wide set of investigations for SA in Arabic tweets, systematically comparing two existing approaches that have been shown to be successful in English. Unlike previous work, we benchmark the trained models against an independent test-set of >3.5k instances collected at different points in time to account for topic-shifts issues in the Twitter stream. Despite the challenging noisy medium of Twitter and the mixed use of Dialectal and Standard forms of Arabic, we show that our SA systems are able to attain performance scores on Arabic tweets that are comparable to the state-of-the-art SA systems for English tweets.

Keywords: Sentiment analysis · Machine learning · Arabic NLP · Twitter

1 Introduction

Over the past decade, there has been a growing interest in collecting, processing and analysing user-generated text from social media. As a sub-task of Affective Computing, Sentiment Analysis (SA) provides the means to mine the web automatically and summarise vast amounts of user-generated text into the sentiments they convey. The growth of research in automatic analysis of people's attitudes and sentiments has coincided with the increasing popularity of social

G. Meiselwitz (Ed.): SCSM 2017, Part II, LNCS 10283, pp. 275–294, 2017.
DOI: 10.1007/978-3-319-58562-8_22

media [22]. This is where the research area of SA plays a major role in capturing and analysing the subjective content from text produced by the general public on social media.

SA research on micro-blogging platforms (e.g. Twitter) is not only motivated by the vast amount of freely available data to crawl [30], but also by their popularity. The selection of Twitter and other sources of big data is motivated by the growing interest in studying content of social networks due to their influence both at social and individual levels [11]. In this context, research has pointed out the significance of Twitter in particular as a valuable resource with regard to the recent unstable political and social circumstances in the Middle East [37].

This work reports on a wide set of investigations for SA in Arabic tweets, systematically comparing two existing approaches that have been shown successful in English. Specifically, we report experiments evaluating fully-supervised-based (SL) and distant-supervision-based (DS) approaches for SA. The investigations cover training SA models on manually-labelled (i.e. in SL methods) and automatically-labelled (i.e. in DS methods) data-sets. Unlike previous work, we benchmark the trained models against an independent test-set of >3.5k instances collected at different points in time to account for topic-shifts issues in the Twitter stream. Despite the challenging noisy medium of Twitter and the mixed use of Dialectal and Standard forms of Arabic, we show that our SA systems are able to attain performance scores on Arabic tweets that are comparable to the state-of-the-art SA systems for English tweets.

The work also investigates the role of a wide set of features, including syntactic, semantic, morphological, language-style and Twitter-specific features. We introduce a set of affective-cues/social-signals features that capture information about the presence of contextual cues (e.g. prayers, laughter, etc.) to correlate them with the sentiment conveyed in an instance. Our investigations reveal a generally positive impact for utilising these features for SA in Arabic. Specifically, we show that a rich set of morphological features, which has not been previously used, extracted using a publicly-available morphological analyser for Arabic can significantly improve the performance of SA classifiers. We also demonstrate the usefulness of language-independent features (e.g. Twitter-specific) for SA. Our feature-sets outperform results reported in previous work on a previously built data-set.

2 Background

The growth of research in automatic analysis of people' s attitudes and sentiments has coincided with the increasing popularity of social media [22]. The ability to classify sentiments is important to understand attitudes, opinions, evaluations and emotions communicated among users across the world about current issues - answering the question of *'what is going on'*.

SA has been an active research area recently with a major focus on English, as a well-resourced languages. The most prominent effort for SA on English tweets has been made by a series of well-known international competition, namely SemEval. Between 2013 and 2016, four editions of this competition

have been successfully launched [25,33,34]. SemEval includes a number of sub-tasks, e.g. determining overall polarity. Our work is closely related to sub-task B, which aims to classify a given tweet instance into positive, negative or neutral (from its author's perspective). For this task, a benchmark data-set of nearly 10k tweets is created and manually annotated for positive, negative and neutral. The test-sets used were collected at different points in time than that of the training data, allowing for different topics to be covered in training and test data [33]. Results reported in this task ranged between 0.248–0.648 F-score on English tweets. It is interesting to mention that SemEval-2017 will include Arabic for the first time in the task of determining the overall polarity of a tweet.

As for SA in Arabic, less effort has been reflected in the literature. A major cause for this is the limited availability of SA-related resources, including annotated data-sets and subjectivity lexica. The limited availability of such resources can be partially attributed to the complexity of Arabic, as a morphologically-rich language. In addition, Arabic has two major language varieties: Modern Standard Arabic (MSA) and Dialectal Arabic (DA), which differs significantly [17]. The formal variety of the language, namely MSA, has been the subject of considerable efforts in developing NLP tools spanning various aspects. In contrast, NLP research on DA has only recently flourished to cope with the increasing prevalence of DAs on the web.

Most previous SA work on Arabic has targeted longer and more formal text instances like newswire, reviews and forums with accuracy rates of up to 95% [1,3]. Few recent attempts have addressed the problem of SA in social media platforms like Twitter (Table 1). However, studies on SA of Arabic tweets suffer from a number of shortcomings. For instance, some studies have only targeted a particular dialect, as in [20]. Others have considered only word-based n-gram features, e.g. [5] or use small sizes of data-sets (up to 3k tweets). In this work, we further expand previous work for SA on Arabic tweets by investigating the impact of: (1) expanded and more variant feature-sets, and (2) experimenting on larger and multi-dialectal training data. In addition, we test our models on

Table 1. Prominent previous work on SA for Arabic.

Paper	Data (size)	ML scheme	Results
Abdul-Mageed et al. [4]	Newswire (2.8k sentences)	SVM	95.52% acc.
Farra et al. [15]	Reviews (44 instances)	SVM	89.3% acc.
Abdul-Mageed et al. [3]	Tweets (3k instances)	SVM (held-out)	65.87% acc. and 61.83% F-score
Abbasi et al. [1]	Forums (1k instances)	SVM (CV)	93.60% acc.
Mourad and Darwish [23]	Tweets (<2k instances)	SVM and NB (CV)	71.9% acc. and 70.35% F-score
Duwairi et al. [12]	Tweets (1k Jordanian and MSA)	NB (CV)	76.78% acc.
Nabil et al. [24]	Tweets (10k Egyptian)	SVM (held-out)	69.10% acc. and 62.60% F-score

an independent test-set, collected at different points in time to explore the performance of our models for a dynamic medium like Twitter. In contrast, Mourad and Darwish [23] and Duwairi et al. [12] only use Cross-Validation (CV) to evaluate their classifiers, while Abdul-Mageed et al. [3] and Nabil et al. [24] use a held-out test-set, which is a sub-set of the original data set used for training. This can be less effective for real-world applications wherein the task is to use trained models for classifying a sample of Twitter feeds over a period of time.

3 Data Collection and Annotation

The Twitter API[1] allows Twitter data to be retrieved by external developers using some search criteria (i.e. keywords, user-names, locations, etc.). Following previous work [16], we search the Twitter API with a pre-prepared list of queries (see Table 2). For instance, in SemEval-2015 developers collected tweets that express sentiment about popular topics. Note that for training a classifier, query terms are replaced by place-holders to avoid bias.

Accessing Twitter API is rate limited (180 queries in a 15 min period), and so we set a delay/waiting time between requests of 2–3 min, as suggested by Go et al. [16]. Similar to the work of Purver and Battersby [30] and to avoid bias (i.e. weekends or active users), we collect data at random times of the day and on different days of week. In addition, we calculate the distribution of the number of tweets from individual users (using the unique IDs of authors). The recorded rate we observe in our data-sets is between 1.76 to 2.59 tweets per user showing no skew towards a group of users. To restrict the retrieved tweets to Arabic only, we set the language parameter of the API to *lang:ar*.

For training, we collected two data-sets: the gold-standard manually-annotated data-set and the distant-supervised automatically-labelled data-set.

3.1 Gold-Standard Manually Annotated Data-Set (GS)

This data-set contains a set of 9k tweets randomly sampled out of 57k tweets collected between January 2013 and February 2014. The 9k tweets were manually annotated by two native speakers of Arabic, using the guidelines displayed

Table 2. Examples of query-terms used for collecting the Arabic Twitter Corpus.

Products/brands	iPhone, channel
Social and religious Issues	Divorce, education, early/child marriage, Sheia
Public figures	Obama, Mandilla, Khamenei, Erdogan
Sport	Chelsea, Al-Ahli FC
Internet and technology	YouTube, Instagram, Google
Controversial topics	Isis

[1] https://dev.twitter.com/.

Table 3. Sentiment labelling criteria for Arabic Twitter Corpus

Label	Definition	Example	English
positive	Clear positive indicator	كم انت عظيم يَا بشَار الاسد	*How great you are, Bashar Al-Asad.*
negative	Clear negative indicator	حنَا للاسف نستخدم ايفون	*Unfortunately, we use the iPhone.*
neutral	• Simple factual statements / news	حَالة وفَاة جديده بَاتش ٧ان٩ بَالصين	*A new reported death case with H7N9 in China.*
	• Questions with no emotions indicated	بكم سعر الآيفون ٥ حَالياً؟	*What is the price of the iPhone 5 these days?*
mixed	Mixed positive and negative indicators (i.e. difficult to decide on the strongest)	فوضِيّ الاخوَان المسلمين التي تريد تدمير حريَاتنَا	*We love democracy, but hate the mess that Muslim Brotherhood is making to destroy our freedom*
uncertain	Undeterminable indicators/neither positive or negative/ lack subjective cues	احيَانَا فهمنَا الأمور بطريقه خطَا يكون هو الصح	*Sometimes, the wrong understanding of things leads to the right thing.*
skip	Redundant or advertising tweets	-	-

in Table 3. Table 5 shows the sentiment distribution of the resultant GS data-set. In order to measure the reliability of the annotations, we conducted an inter-annotator agreement study on the annotated tweets. We use Cohen's Kappa metric [9], which measures the degree of agreement among the assigned labels, correcting for agreement by chance. The resulting weighted Kappa reached $\kappa = 0.786$, which indicates reliable annotations.

What Happens with the Examples Where Both Annotators Disagree? A third annotator is employed to decide the selection of the final annotation, if the 3rd annotator disagrees with both annotators, the tweet will be assigned *uncertain* label. Data instances in this category are also excluded from the data-set [7].[2] This procedure is important for the quality of the gold-standard data-set. As provision of annotated data is a goal of this work, the GS data-set has already been made freely available to the research community via an ELRA repository and at the time of writing this work, the data-set has been accessed more than 162 times and downloaded more than 110 times [31].[3]

3.2 Distant-Supervised (DS) Automatically-Annotated Data-Set

Two DS-based data-sets were created using two popular conventional markers of Twitter, i.e. emoticons and hashtags, to collect and automatically label Twitter instances as positive or negative.

[2] The 9k tweets in this data-set represent the final number after all tweets labelled as uncertain were excluded. A total of 3,106 tweets were excluded from the Gold-Standard data-sets.

[3] Further information about how to access/download the corpus can be found at: goo.gl/qNLIZ2.

Table 4. Emoticons and hashtags used to automatically label the DS-based training data-sets.

Emoticon	Sentiment label	Hashtag	Sentiment label
:) , :-) , :)), (: , (-: , ((:	positive	(*happy*, سعَاده) (*joy*, بهجه) (hope, أمل)	positive
:(, :-(, :((, :((,): ,)):)-:	negative	(*sad*, حـزن) (*bane*, مصيبه) (despair,يَأس)	negative

Table 5. Sentiment label distribution of the training data-sets: gold standard manually annotated and distant supervision data-sets.

Data-set	Neutral	Polar[a]	Positive	Negative	Mixed	Total
Gold standard (GS)	4,854	4,327	1,346	2,408	573	9,181
Emoticon-based (Emo)	55,076	1,118,356	660,393	457,963	-	1,173,432
Hashtag-based (Hash)	55,076	130,160	59,990	70,170	-	185,236

[a]Polar = positive + negative + mixed

Following [8,16,30], we use a set of emoticons with pre-defined polarity and sentiment-bearing-hashtags (Table 4) to automatically label DS training sets. Conventional markers are merely used to assign the sentiment labels and removed from tweets to avoid any bias, following Go et al. [16].

The number of tweets collected varied in accordance with the popularity of conventional markers (i.e. emoticons and hashtags) that we used to query Twitter. That is, although Emo and Hash data-sets were collected over the same period of time, the total number of tweets retrieved using emoticons is 1,511,621 tweets, while the number of tweets collected using hashtag queries is 926,640 tweets. A similar behaviour was also observed by Purver and Battersby [30] on English tweets. Furthermore, we observe that removing duplicated instances from the emoticon-based and hashtag-based data-sets reveals a very high rate of noisy/repeated tweets in the hashtag-based data-set, resulting in reducing the hashtag-based data-set from 926,640 to 130,160 instances (see Table 5). To illustrate, the discarded content represents 85.9% of the originally collected hashtag-based data-set, as compared to 24.1% of the emoticon-based data-set. A closer look at a random sample of the Hash data-set reveals an extensive use of popular hashtags, e.g. happy, to post advertising content to a wider audience.

Test Data-Set. In order to compare SA systems trained on different training sets, we use an independent test-set to evaluate their performance. That is, considering the evolving nature of the Twitter stream [13], we built a test-set that is a collection of random samples retrieved over different periods of time (Table 6). In addition, the size of the data-set (as shown in the Table 6) is comparable to

that created and used in SemEval on English tweets (sizes for Twitter test-sets are 4,435 tweets in 2013 and 2,473 tweets in 2014). Previous studies on Arabic tweets, in contrast, have considered test-sets that are subsets of the original data-set (e.g. [3]) or used cross-validation (e.g. [23]). Both settings are problematic for Twitter due to its evolving nature and topic-shift issues that are likely to influence the predictive ability of a trained model over different points in time.

The test-set is manually annotated by two native speakers of Arabic, following the criteria presented in Table 3 (p. 5). The inter-annotator score for the test-set is at $\kappa = 0.69$. Our test-set is designed to provide a common ground to build and evaluate SA systems, as it (1) is built with a coverage that spans an extended period of time (see Table 6); (2) contains less bias to active users (observed distribution of the number of tweets from individual users is 1.16 tweet per user); (3) is annotated with a rich set of morphological, semantic, and stylistic features; and more importantly, (4) is publicly available.[4]

The class distribution in the test-set indicates the negative class as the majority class. This is in line with our previous manual annotations of the gold-standard training data. Following SemEval [33,34], the instances were randomly selected for manual annotation, which is likely to obtain a representative sample of the Twitter stream [8].

Table 6. Sentiment label distribution of the test data-set.

Data-set	Collection time	Neutral	Polar	Positive	Negative	Total
Test-sample1	Spring 2013	324	377	69	308	701
Test-sample2	Autumn 2013	480	621	285	336	1,101
Test-sample3	Winter 2014	333	518	169	349	851
Test-sample4	Summer 2014	218	667	208	459	885
Total	-	1,355	2,183	731	1,452	**3,538**

3.3 Data Pre-processing

We adapt pre-processing techniques to tackle informality and alleviate the noise typically encountered in social media. We use pre-processing techniques that have been previously employed and shown to be useful for improving performances of SA systems [16,23,32,34]. In particular, the extracted data is cleaned up in a computationally-motivated (i.e. reducing feature space) pre-processing step by:

– **Normalising conventional symbols of Twitter:** this involves detecting entities like: #hash-tags, @user-names, re-tweet (RT), and URLs; and replacing them with place-holders.

[4] http://www.macs.hw.ac.uk/~eaar1/Eshrag%20Refaee/myResearch1.html.

- **Normalising exchangeable Arabic letters**: mapping letters with various forms (i.e. *alef* and *yaa*) to their representative character.
- **Eliminating non-Arabic characters.**
- **Removing punctuation and normalising digits.**
- **Removing stop words:** this involves eliminating some frequent word tokens that are less likely to have a role in class prediction (e.g. prepositions).
- **Reducing emphasised words/expressive lengthening**: this involves normalising word-lengthening effects. In particular, a word that has a letter repeated subsequently more than two times will be reduced to two (e.g. *sadddd* is reduced to *sadd*).

Other text pre-processing steps involve:

Text Segmentation: This step is performed to separate tokens based on spaces and punctuation marks. For this, we use the publicly available tokeniser called TOKAN integrated into MADAMIRA [28].

Text Stemming: This is one step further in text pre-processing that aims to alleviate the high dimensionality of the text data by using reduced forms of words (e.g. stems). Abdul-Mageed et al. [3] argue about the importance of employing such a technique, in particular, when dealing with a morphologically rich and highly derivative language like Arabic, as the problem of high dimensionality becomes more pronounced. In this context, Abdul-Mageed [2] highlights the significance of this text pre-processing step and argues that SA on Arabic can be problematic without using the compressed forms of words, as it will result in the sentiment classifiers being exposed to a large number of previously unseen features (words), although they might be present in training and testing but in different forms. For instance, the words:

وبتَالقهَا *and+with+her+brilliance*, وبتَالقه *and+with+his+brilliance*, بتَالقه *with+his+brilliance* and بتَالقهَا *with+her+brilliance* can be reduced to the stem بتَالق meaning *brilliantly/brightly.*

In sum, stemming has shown to be beneficial for SA on Arabic newswire, reviews and social media posts [4–6].

4 Features Extraction

This section presents a number of feature-sets that we extract and employ to examine their utility for SA on Arabic tweets (Table 7). The categorisation and design of feature-sets is inspired by the work of Abbasi et al. [1].

Word-Token-Based Features: This set involves word-stem unigrams and bigrams, as they were found to perform better than other combinations of n-grams in our preliminary experiment.

Table 7. Summary of feature-sets used.

Feature-set	Features	Feature type
Syntactic	Word-stem n-grams	String
Morphological	Aspect	String
	Gender	String
	Mood	String
	Number	String
	Person	String
	POS:word	String
	State	String
	Voice	String
	Diacritics	String
	Has-morph-analysis	Binary
Semantic	Has-positive-lex	Binary
	Positive-lex-count	Numerical
	Has-negative-lex.	Binary
	Negative-lex-count	Numerical
	Has-neutral-lex.	Binary
	Neutral-lex-count	Numerical
	Has-negator	Binary
Affective-cues	Has-consent	Binary
	Has-dazzle	Binary
	Has-laughs	Binary
	Has-regret	Binary
	Has-prayer	Binary
	Has-sigh	Binary
Language-style	Tweet-length (char)	Numerical
	Word-length (char)	Numerical
	Word-offset (char)	Numerical
	Has-exclamation-mark	Binary
	Exclamation-mark-count	Numerical
	Has-question-mark	Binary
	Question-mark-count	Numerical
	Has-dots	Binary
	Dots-count	Numerical
	Has-lengthening	Binary
	Has-positive-emoticon	Binary
	Has-negative-emoticon	Binary
	MSA-or-DA	Binary
	Degree of dialectness	Numerical
Twitter-specific	is-Favourite	Binary
	Favourite-count	Numerical
	is-Retweet	Binary
	Retweet-count	Numerical
	Has-hashtag	Binary
	Has-URL	Binary
	Has-user-name	Binary

Morphological Features: The use of this feature-set is motivated by the rich morphology of Arabic, thus aiming to exploit this aspect by extracting a rich set of morphological features. For that, we employ a state-of-the-art morphological analyser for Arabic, namely MADAMIRA [28]. MADAMIRA on a gold annotated blind test

data by Pasha et al. [28] has achieved an accuracy of up to 95.9% for POS tagging and 84.1% for word-level morphological analysis on MSA.

Semantic Features: This feature-set includes a number of binary and numeric features that check the presence and number of occurrences of sentiment-bearing words in each given tweet (Table 7). To extract this feature-set, we utilise a combined sentiment lexicon. Our merged sentiment lexicon exhibits a reasonable degree of coverage/variation as ArabSenti and the Arabic translation of MPQA represent more formal language (both are in Standard Arabic), while our in-house Twitter-based lexicon[5] includes informal and dialectal entries, contributing words like:

طز *go to hell* and بلطجي *bully.*

Affective-Cues/Social-Signals: This feature-set comprises six binary features, indicating whether a tweet has any of these social signals: consent, dazzle, laughs, regret, prayer, and sigh. To obtain these features, we use six manually created dictionaries.[6] To avoid bias, the extracted dictionaries are based on an independent data-set that does not overlap with any of our data-sets. The use of this feature-set is motivated by the idea of finding a set of simple features that can correlate to users' culture and, at the same time, can be used as a means of conveying sentiments. For instance, Ptaszynski et al. [29] employ a manually collected lexicon of emotive expression, i.e. culturally-specific Japanese emotional expressions, and note that these features are useful for SA on Japanese blogs.

Twitter-Specific Features: This set utilises seven features characterising the way Twitter is being used (Table 7). Twitter can be used in various ways: for information sharing (via inclusion of URLs and hashtags) and/or for social networking (via inclusion of user-mentions and re-tweets), as such uses vary across languages [18]. For instance, Hong et al. [18] investigated behaviour differences among users of different languages and observed that communities like Korean and Indonesian tend to exhibit more for social networking, whereas English and German users tend to use Twitter more for information sharing. We are not aware of a similar study on Arabic. Thus, we explored one of our own data-sets comprising 1.2M Arabic tweets and observed a higher tendency for social networking (e.g. up to 36.80% of tweets included user-mentions), while only an average of 16.64% of tweets included hashtags/URLs, i.e. less use of tweets for information sharing.

Language Style Features: This set involves a number of features that characterise the language typically used in social media, including:

(A) **Stylistic features:** This set of features is also referred to as language independent. It captures information about the informal language used in social media and may convey sentiment. That is, stylistic features aim to unveil

[5] The lexicon is freely available at: goo.gl/qNLIZ2.
[6] The lists are freely available at: goo.gl/qNLIZ2.

latent patterns that can improve classification performance of sentiments [1]. This set comprises features checking for stylistic variation, i.e. presence of: emoticons, expressive lengthening (e.g. *sadddd*).

(B) **MSA-or-DA feature:** This is a binary feature to investigate the usefulness of employing an explicit feature that identifies the language variety of a tweet instance (MSA or DA). To automatically extract this feature, we use AIDA [14]. In addition to identifying the language variety of a tweet as MSA or DA, AIDA can provide a numerical value between [0,1] reflecting the degree of dialectness for the corresponding tweet, which we also exploit as a feature.
MSA-or-DA feature can be particularly useful for investigations on Arabic tweets to assess the impact of DA presence on the overall performance of SA. The use of this feature is also motivated by the fact that MSA is often referred to as "*the language of the mind*" while the DAs as "*the language of the heart*".[7]

5 Experimental Setup

In this section, we present the experimental setup we utilised in our empirical investigations.

Machine Learning Scheme: In this work, we use Support Vector Machines (SVMs) [21] as a machine learning scheme that is found to be particularly successful for text classification problems, including SA [7,27,33,34]. Since there are several implementations available for SVM, we follow guidelines by Hsu et al. [19] who show that LIBLINEAR is more efficient in tackling document/text classification problems – wherein both the number of instances and features are large – than LIBSVM, in terms of the time required to obtain a model with a comparable accuracy and memory consumption. Therefore, we use LIBLINEAR for all experiments reported in this work.[8]

Classification Levels: We experiment with two-level binary classification problem formulation (Fig. 1). The choice is based on the results of our preliminary experiments that showed steady better performance for two-level binary classification over single-level three-way classification.

Baselines: We compare our results against several baselines, including a majority baseline (B-Mjr) and a stem n-gram baseline (B-stem).

[7] For instance, we find that Dialectal tweets represent 34.12% of the negative tweets, 37.39% of the positive tweets, and only 13.52% of neutral tweets in the GS data-sets, suggesting subjective instances to be more dialectal as compared to neutral ones. In addition, Cotterell and Callison-Burch [10] reported 40% of their Arabic Twitter data-set comprising >40k tweets were manually annotated as highly dialectal.

[8] https://www.csie.ntu.edu.tw/~cjlin/liblinear/.

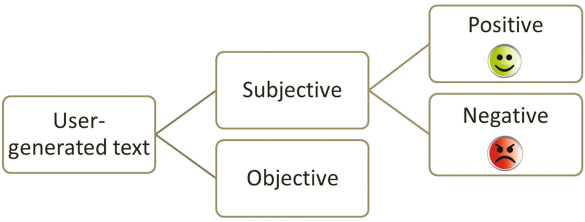

Fig. 1. Levels of sentiment classification.

Evaluation Metrics: The results are reported using two popular metrics: weighted F-score and accuracy.

Evaluations Methods: We use two methods for evaluating the performance of the trained models, namely cross-validation (CV) and independent test-set. CV relies on a fixed number of data proportions, i.e. folds. We also use our independent test-set account for the time-changing nature of the Twitter stream, following SemEval [25,33,34], that has not previously been considered for Arabic.

Statistical Tests: We employed two popular metrics, i.e. t-test and Chi-squared (χ^2) to provide evidence that variation among different classifiers is not caused by chance.

6 Experimental Results

This section displays the results of our empirical investigations.

6.1 Impact of Feature-Sets

First, we investigated the utility of a wide set of features (see Table 7) that has not previously been employed for SA on Arabic tweets. To assess the usefulness of the features, we conducted experiments on the only data-set available at that time. For that, we use the M&D data-set developed by Mourad and Darwish [23] (see class distribution in Fig. 2). The authors used SVM and experimented with CV setting. Table 8 displays the results of utilising our feature-sets on M&D data-set following similar experimental settings.

Subjectivity Classification (Polar vs. Neutral): The best performance is achieved with the morphological features at 66.25% accuracy. This is a 2.65% accuracy improvement compared to the top score originally reported by Mourad and Darwish [23] at 63.6% on this data-set. The addition of the morphological features has significantly improved performance over the stem n-grams baseline. Our morphological feature-set includes POS with 35 tags, as opposed to only five POS tags used by Mourad and Darwish [23]. We therefore concluded that a rich set of morphological features (e.g. gender, voice, aspect, among others) with an extended POS set is beneficial for Arabic SA.

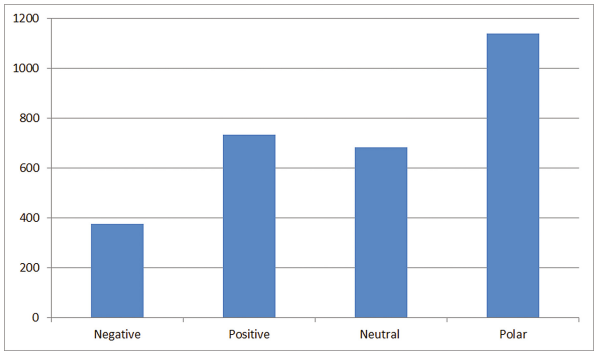

Fig. 2. Class distribution in M&D data-set.

Sentiment Classification (Positive vs. Negative): The average accuracy score is at 81.32%, which is 9.42% improvement as compared to 71.9% accuracy reported by Mourad and Darwish [23] on this task. The best performance is attained by the semantic features at 82.70% accuracy. For extracting the semantic features, Mourad and Darwish [23] used ArabSenti and a translated version of MPQA, which is similar to our work. However, they did not report on manually correcting/filtering the auto-translated entries of the MPQA in order to maintain its quality. We used a translated and manually filtered version of MPQA that

Table 8. Binary classification on M&D data-set: polar vs. neutral; positive vs. negative.

M&D data-set						
	Polar vs. neutral			Positive vs. negative		
	F	Acc.	SD	F	Acc.	SD
Majority baseline (B-mjr)	0.519	65.57	0.17	0.526	66.07	0.4
Stem n-grams[a]	0.620	65.13	2.81	0.818	82.05	2.64
Stem n-grams + Morph[a]	0.643	**66.25***	2.54	0.811	81.18	3.99
Stem n-grams + Semantic[a]	0.620	65.17	2.85	**0.827**	**82.70***	3.56
Stem n-grams + Affec-cues	0.624	65.27	2.87	0.816	81.85	2.93
Stem n-grams + Lang-style[a]	0.623	63.12*	3.51	0.776	77.61*	4.01
Stem n-grams + Twt-specific[a]	0.622	65.28	2.78	0.822	82.38	2.92
Comb. of all feat	**0.65**	66.14*	2.76	0.808	80.78	3.74
Average	0.628	65.19	2.88	0.812	81.32	3.54

Underline denotes a statistically-significant difference vs. majority baseline ($p < 0.05$).
*Denotes a statistically-significant difference vs. stem n-grams baseline ($p < 0.05$).
[a]Denotes that the feature-set or a subset of it has been used by Mourad and Darwish [23].

comprises 2.6k entries out of 8k in the original English MPQA. In addition, they automatically expanded the sentiment lexicon, which is likely to introduce more noise than benefit [36]. In our work, we utilised a new dialectal sentiment lexicon to adapt to the use of DAs in social media.

Use of M&D Data-Set in Other Studies: A recent study by Salameh et al. [35] on M&D data-set (positive vs. negative) with CV and an SVM classifier reported their best score at 74.62%. This still does not compete with our results on this data-set, with an average accuracy score of 81.32%. The performance variation can be attributed to the different feature-sets used. Salameh et al. [35] employed word-lemma n-grams and semantic features (leveraging manually and auto-generated sentiment lexica), while our system employs word-stem n-grams along with a wide set of semantic (manually created lexica) and a rich set of morphological features, among others.

In sum, our new, extended feature-sets have shown to outperform previous work on M&D data-set for both tasks: subjectivity (polar vs. neutral) and sentiment (positive vs. negative) classification.

6.2 Impact of Evaluation Method: CV vs. Independent Test-Set

To assess the impact of the time-changing nature of streaming data (e.g. Twitter stream) on the evaluation method employed, this section outlines experiments that compare the performance of classifiers when evaluated (1) using the standard CV and (2) using our independent test-set that was collected at different points in time than that of the training data (see Table 6). For that, we use a subset of 2.2k tweets of our manually annotated gold-standard (GS) data-set (see Table 5). Results of this set of experiments are displayed in Table 9.

For subjectivity classification (polar vs. neutral), we can observe a significant performance drop of 31.23% accuracy on average between CV and the results on independent test-set. This indicates that, despite the promising results with CV at an average accuracy of 95.49%, the classifiers do not generalise well to unseen topics.

For sentiment classification (positive vs. negative), again, testing on the independent test-set has resulted in an average performance drop of 17.03% in accuracy across all feature-sets compared to CV.

Conclusion: Unlike previous work, we re-evaluate our trained models on an independent, larger and more diverse test-set. We show that, despite very promising CV results, our models do not generalise well to data-sets collected at a later point in time, causing performance drops. The performance drop is likely to be caused by time-dependent topic-shifts issues in the Twitter stream and the prominent role of word n-gram features in our models [26,36].

Table 9. Binary classification on subset of 2.2k tweets of GS data.

	10 fold CV			Ind. test-set	
	F	Acc.	SD	F	Acc.
Polar vs. neutral					
Majority baseline (B-mjr)	0.578	70.08	0.1	0.471	61.70
Stem n-grams	0.905	91.01	2.24	0.557	**65.26**
Comb. of all feat	**0.998**	**99.93***	0.23	**0.594**	63.14*
Average	0.952	95.49	1.86	0.577	64.26
Positive vs. negative					
Majority baseline (B-mjr)	0.335	50.16	0.25	0.531	66.51
Stem n-grams	0.736	74.1	3.71	0.586	58.59
Comb. of all feat	**0.908**	**90.77***	2.41	**0.702**	**69.68***
Average	0.80	80.24	3.29	0.635	63.21

Underline denotes a statistically-significant difference vs. majority baseline ($p < 0.05$).
*Denotes a statistically-significant difference vs. stem n-grams baseline ($p < 0.05$).

Since Twitter experiences topic-shifts over time, the vocabulary, especially the content words, are likely to change as well [13]. Investigating this hypothesis, we find that the word frequency distribution differs amongst the training/test data-sets: the overall overlap of unique tokens is only 12.21%. Next, we will address this issue by using a larger gold-standard training data-set and by using semi-supervised approaches to automatically obtain larger training data.

6.3 Impact of Size of Training Data

To assess the impact of increasing the size of the training data in reducing the performance gap encountered when using the independent test-set for evaluating our classifiers, we experimented with different sizes of the GS data. Table 10 shows the results of three sets of experiments, each with different training data size. Results show that the average performance gap has been reduced from 24.13% with 2.3k instances to 7.63% with 6.8k instances, reaching only 4.9% with 9k instances.[9] This indicates a utility for expanding the training set on the classifiers' ability to attain better scores. Next, we examine the possibilities of further expanding training by exploiting existing clues (e.g. emoticons) to *automatically* obtain sentiment labels.

[9] The performance gap here is the average across subjectivity and sentiment classification.

Table 10. Summary of GS results on various sizes of training data.

Data-set (size)	Average acc.	Performance gap (CV - ind. test-set)
GS (2.3k)	63.74	24.13
GS (6.8k)	72.96	07.63
GS (9k)	74.10	04.93

6.4 Impact of Annotation Method: Manually vs. Automatically

To assess the utility of employing automatic means for obtaining larger annotated data as opposed to standard manually-based ones, we follow previous work by Go et al. [16] in using distant supervision (DS) approaches. DS approaches have been successfully used for SA in English (e.g. [8,16]. However, we are not aware of existing studies with investigation of DS for SA in Arabic.[10] As such, we collected and automatically labelled emoticon-based and hashtag-based DS data-sets (see Table 5). Table 11 shows that the best average accuracy performance is attained when combining emoticon- and hashtag-based data-sets (with 1.2M instances) at 62.22%. However, it is interesting to note that this score is still below the average accuracy score attained by the manually-annotated GS data-set (9k instances) at 75.91% (Fig. 3).

Table 11. Binary classification positive vs. negative on the emoticon and hashtag-based data-sets.

Positive vs. negative						
	Emo		Hash		Emo+Hash	
	F	Acc.	F	Acc.	F	Acc.
Majority baseline (B-mjr)	0.531	66.51	0.531	66.51	0.531	66.51
Stem n-grams	0.537	52.77	**0.674**	69.22	**0.621**	62.81
Comb. of all feat	0.531	64.41*	0.258	36.97*	0.565	62.53
Average	0.544	56.23	0.531	56.53	0.60	62.22

Underline denotes a statistically-significant difference vs. majority baseline ($p < 0.05$).
*Denotes a statistically-significant difference vs. stem n-grams baseline ($p < 0.05$).

As for the stem n-grams baseline, it can be seen that hashtag-based data-set (Hash) outperforms the emoticon-based data-set (Emo). This is interesting, considering that Emo is about 8.6 times larger than Hash. To clarify this, we

[10] Since the vast majority of previous work has used DS only with binary sentiment classification positive vs. negative (e.g. [8,16]) and due to the controversy in the existing means for automatic collection of neutral instances [23], we report the results in this section for the binary sentiment classification.

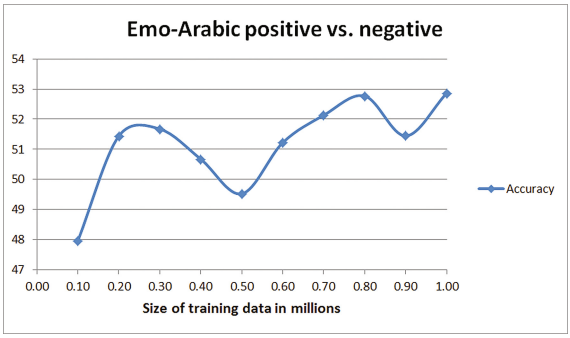

Fig. 3. Learning curve on a 1M Arabic emoticon-based data-set.

conducted an error analysis on a random sample of Emo data-set, in which we manually annotated a set of 303 tweets. We found that only in 34.32% of cases does the manual label match the automatically assigned label, i.e. using emoticons. Whereas, in 36.63% of the cases manual labels and automatically assigned labels do not match. This raises questions about the quality of automatically assigned labels using emoticons. A closer look at the sample reveals cases wherein emoticon-based labels do not match the emotion conveyed in the accompanying text, either due to sarcasm, as in example 1, or because of mistakenly interchanged parenthesis as a result of the right-to-left typing nature of Arabic, as presumably is the case in example 2.

1): جميل يَا اهلي
great job Ahli :(*- referring to a famous football team.*
2 البقَاء لّله :(اللّهم ارحمهم
Condolences :) May Allah shower their souls with mercy

7 SAAT: A System for Sentiment Analysis in Arabic Tweets

In this section we present SAAT, a java-based system we developed to automatically classify sentiments conveyed in Arabic tweets, utilising our best trained classifiers. The system will receive a query from systems users about entities, e.g. *'Trump'*. The query will then be sent to retrieve tweets containing the query text from the Twitter live stream. The retrieved tweets will pass through the pre-processing steps described earlier. Next, the trained subjectivity classifier will decide if the tweet is neutral or subjective. Finally, subjective tweets will be classified by the sentiment model as positive or negative (see Table 12).[11]

[11] Codes are available at: goo.gl/qNLIZ2.

Table 12. Examples of tweets about *'Trump'* auto-labelled via SAAT.

1 لو سَاندرز عمل معجزه وقَابل ترَامب، فَالَامر محسوم لهيلَاري ***If Sanders gets to the final with Trump, then things will be working really well for Hillary.***	positive
2 انونيموس تعلن حربَا شَامله علي دونَالد ترَامب ***Anonymous hackers declare total war on Donald Trump.***	negative
3 هل دونَالد ترَامب جورج وَاليس الجديد؟ ***Is Donald Trump the George Wallace of this time?***	neutral

8 Conclusion

In sum, DS-based approach using emoticons for SA in Arabic seem to be less useful as compared to English. The results indicate a tendency of a hashtag-based DS approach to be less noisy, attaining an accuracy score close to that achieved by manually annotated gold-standard (GS) data. As such, hashtag-based DS approach has the potential to obtain sentiment labels automatically and at the same time maintain quality levels close to GS, with the difference in performance as a trade off for the laborious effort required to obtain GS labels.

References

1. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in muliple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. (TOIS) **26**, 1–34 (2008)
2. Abdul-Mageed, M.: Subjectivity and sentiment analysis of Arabic as a morophologically-rich language. Ph.D. thesis, The School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA (2015)
3. Abdul-Mageed, M., Diab, M., Kübler, S.: SAMAR: subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. **28**(1), 20–37 (2014)
4. Abdul-Mageed, M., Diab, M.T., Korayem, M.: 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, HLT 2011, Stroudsburg, PA, USA, vol. 2, pp. 587–591. Association for Computational Linguistics (2011)
5. Ahmed, S., Pasquier, M., Qadah, G.: Key issues in conducting sentiment analysis on Arabic social media text. In: IIT, pp. 72–77. IEEE (2013)
6. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A.: Subjectivity and sentiment analysis of Arabic: trends and challenges. In: IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 148–155. IEEE (2014)
7. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 127–135. Association for Computational Linguistics (2008)

8. Bifet, A., Frank, E.: Sentiment knowledge discovery in Twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16184-1_1
9. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
10. Cotterell, R., Callison-Burch, C.: A multi-dialect, multi-genre corpus of informal written Arabic. In: LREC 2014, Reykjavik, Iceland. ELRA, May 2014
11. Dodds, P.S., Clark, E.M., Desu, S., Frank, M.R., Reagan, A.J., Williams, J.R., Mitchell, L., Harris, K.D., Kloumann, I.M., Bagrow, J.P., Megerdoomian, K., McMahon, M.T., Tivnan, B.F., Danforth, C.M.: Human language reveals a universal positivity bias. Proc. Natl. Acad. Sci. **112**(8), 2389–2394 (2015)
12. Duwairi, R., Marji, R., Sha'ban, N., Rushaidat, S.: Sentiment analysis in Arabic tweets. In: ICICS, pp. 1–6, April 2014
13. Eisenstein, J.: What to do about bad language on the internet. In: Proceedings of NAACL-HLT, pp. 359–369 (2013)
14. Elfardy, H., Al-Badrashiny, M., Diab, M.: AIDA: identifying code switching in informal Arabic text. In: EMNLP 2014, p. 94 (2014)
15. Farra, N., Challita, E., Assi, R.A., Hajj, H.: Sentence-level and document-level sentiment mining for Arabic texts. In: IEEE ICDMW 2010, pp. 1114–1119. IEEE (2010)
16. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pp. 1–12 (2009)
17. Habash, N.: Introduction to Arabic natural language processing. Synth. Lect. Hum. Lang. Technol. **3**, 1–189 (2010). Morgan & Claypool Publishers
18. Hong, L., Convertino, G., Chi, E.H.: Language matters in twitter: a large scale study. In: ICWSM (2011)
19. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification. National Taiwan University, Taipei, Taiwan (2003)
20. Ibrahim, H., Abdou, S., Gheith, M.: MIKA: a tagged corpus for modern standard Arabic and colloquial sentiment analysis. In: IEEE ReTIS, pp. 353–358, July 2015
21. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). doi:10.1007/BFb0026683
22. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**, 1–167 (2012). Morgan & Claypool Publishers
23. Mourad, A., Darwish, K.: Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In: WASSA 2013, p. 55 (2013)
24. Nabil, M., Aly, M., Atiya, A.: ASTD: Arabic sentiment tweets dataset. In: Proceedings of EMNLP 2015, Lisbon, Portugal, pp. 2515–2519. ACL, September 2015
25. Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T.: Semeval-2013 task 2: sentiment analysis in Twitter. In: *SEM, Atlanta, Georgia, USA, pp. 312–320. ACL, June 2013
26. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of EMNLP, pp. 79–86. ACL, 2002
28. Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A.E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.: MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: Proceedings of LREC 2014, Reykjavik, Iceland. ELRA, May 2014

29. Ptaszynski, M., Rzepka, R., Araki, K., Momouchi, Y.: Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. Comput. Speech Lang. **28**(1), 38–55 (2014)
30. Purver, M., Battersby, S.: Experimenting with distant supervision for emotion classification. In: Proceedings of EACL, Avignon, France, pp. 482–491. ACL, April 2012
31. Refaee, E., Rieser, V.: An Arabic Twitter Corpus for subjectivity and sentiment analysis. In: LREC 2014 (2014)
32. Refaee, E.A.: Sentiment analysis for micro-blogging platforms in Arabic. Ph.D. thesis, The School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK (2016)
33. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V.: SemEval-2015 task 10: sentiment analysis in Twitter. In: Proceedings of SemEval 201), pp. 451–463, Denver, Colorado. ACL, June 2015
34. Rosenthal, S., Ritter, A., Nakov, P., Stoyanov, V.: SemEval-2014 task 9: sentiment analysis in Twitter. In: SemEval, pp. 73–80, Dublin, Ireland. ACL, August 2014
35. Salameh, M., Mohammad, S., Kiritchenko, S.: Sentiment after translation: a case-study on Arabic social media posts. In: NAACL, Denver, Colorado, pp. 767–777. ACL, May–June 2015
36. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
37. Zaidan, O.F., Callison-Burch, C.: Arabic dialect identification. Comput. Linguist. **40**(1), 171–202 (2014)