SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media

Muhammad Abdul-Mageed, Sandra Kübler
Indiana University
Bloomington, IN, USA
{mabdulma,skuebler}@indiana.edu

Mona Diab
Columbia University
New York, NY, USA
mdiab@ccls.columbia.edu

Abstract

In this work, we present SAMAR, a system for Subjectivity and Sentiment Analysis (SSA) for Arabic social media genres. We investigate: how to best represent lexical information; whether standard features are useful; how to treat Arabic dialects; and, whether genre specific features have a measurable impact on performance. Our results suggest that we need individualized solutions for each domain and task, but that lemmatization is a feature in all the best approaches.

1 Introduction

In natural language, *subjectivity* refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982) and, as such, it incorporates sentiment. The process of subjectivity classification refers to the task of classifying texts as either *objective* (e.g., *The new iPhone was released.*) or *subjective*. Subjective text can further be classified with *sentiment* or *polarity*. For sentiment classification, the task consists of identifying whether a subjective text is *positive* (e.g., *The Syrians continue to inspire the world with their courage!*), *negative* (e.g., *The bloodbaths in Syria are horrifying!*), *neutral* (e.g., *Obama may sign the bill.*), or, sometimes, *mixed* (e.g., *The iPad is cool, but way too expensive*).

In this work, we address two main issues in Subjectivity and Sentiment Analysis (SSA): First, SSA has mainly been conducted on a small number of genres such as newspaper text, customer reports, and blogs. This excludes, for example, social media genres (such as Wikipedia Talk Pages). Second, despite increased interest in the area of SSA, only few attempts have been made to build SSA systems for *morphologically-rich languages* (Abbasi et al., 2008; Abdul-Mageed et al., 2011b), i.e. languages in which a significant amount of information concerning syntactic units and relations is expressed at the word-level, such as Finnish or Arabic. We thus aim at partially bridging these two gaps in research by developing an SSA system for Arabic, a morphologically highly complex languages (Diab et al., 2007; Habash et al., 2009). We present SAMAR, a sentence-level SSA system for Arabic social media texts. We explore the SSA task on four different genres: chat, Twitter, Web forums, and Wikipedia Talk Pages. These genres vary considerably in terms of their functions and the language variety employed. While the chat genre is overridingly in dialectal Arabic (DA), the other genres are mixed between Modern Standard Arabic (MSA) and DA in varying degrees. In addition to working on multiple genres, SAMAR handles Arabic that goes beyond MSA.

1.1 Research Questions

In the current work, we focus on investigating four main research questions:

- **RQ1:** How can morphological richness be treated in the context of Arabic SSA?
- **RQ2:** Can standard features be used for SSA for social media despite the inherently short texts typically used in these genres?
- **RQ3:** How do we treat dialects?

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 19–28,
Jeju, Republic of Korea, 12 July 2012.

- **RQ4:** Which features specific to social media can we leverage?

RQ1 is concerned with the fact that SSA has mainly been conducted for English, which has little morphological variation. Since the features used in machine learning experiments for SSA are highly lexicalized, a direct application of these methods is not possible for a language such as Arabic, in which one lemma can be associated with thousands of surface forms. For this reason, we need to investigate how to avoid data sparseness resulting from using lexical features without losing information that is important for SSA. More specifically, we concentrate on two questions: Since we need to reduce word forms to base forms to combat data sparseness, is it more useful to use tokenization or lemmatization? And given that the part-of-speech (POS) tagset for Arabic contains a fair amount of morphological information, how much of this information is useful for SSA? More specifically, we investigate two different reduced tagsets, the RTS and the ERTS. For more detailed information see section 4.

RQ2 addresses the impact of using two standard features, frequently employed in SSA studies (Wiebe et al., 2004; Turney, 2002), on social media data, which exhibit DA usage and text length variations, e.g. in twitter data. First, we investigate the utility of applying a UNIQUE feature (Wiebe et al., 2004) where low frequency words below a threshold are replaced with the token "UNIQUE". Given that our data includes very short posts (e.g., twitter data has a limit of only 140 characters per tweet), it is questionable whether the UNIQUE feature will be useful or whether it replaces too many content words. Second, we test whether a polarity lexicon extracted in a standard domain using Modern Standard Arabic (MSA) transfers to social media data. Third, given the inherent lack of a standardized orthography for DA, the problem of replacing content words is expected to be increased since many DA content words would be spelled in different ways.

RQ3 is concerned with the fact that for Arabic, there are significant differences between dialects. However, existing NLP tools such as tokenizers and POS taggers are exclusively trained on and for MSA. We thus investigate whether using an explicit feature that identifies the dialect of the text improves SSA performance.

RQ4 is concerned with attempting to improve SSA performance, which suffers from the problems described above, by leveraging information that is typical for social media genres, such as author or gender information.

The rest of the paper is organized as follows: In Section 2, we review related work. Section 3 describes the social media corpora and the polarity lexicon used in the experiments, Section 4 describes SAMAR, the SSA system and the features used in the experiments. Section 5 describes the experiments and discusses the results. In Section 6, we give an overview of the best settings for the different corpora, followed by a conclusion in Section 7.

2 Related Work

The bulk of SSA work has focused on movie and product reviews (Dave et al., 2003; Hu and Liu, 2004; Turney, 2002). A number of sentence- and phrase-level classifiers have been built: For example, whereas Yi et al. (2003) present a system that detects sentiment toward a given subject, Kim and Hovy's (2004) system detects sentiment towards a specific, predefined topic. Our work is similar to Yu and Hatzivassiloglou (2003) and Wiebe et al. (1999) in that we use lexical and POS features.

Only few studies have been performed on Arabic. Abbasi et al. (2008) use a genetic algorithm for both English and Arabic Web forums sentiment detection on the document level. They exploit both syntactic and stylistic features, but do not use morphological features. Their system is not directly comparable to ours due to the difference in data sets.More related to our work is our previous effort (2011b) in which we built an SSA system that exploits newswire data. We report a slight system improvement using the gold-labeled morphological features and a significant improvement when we use features based on a polarity lexicon from the news domain. In that work, our system performs at 71.54% F for subjectivity classification and 95.52% F for sentiment detection. This current work is an extension on our previous work however it differs in that we use automatically predicted morphological features and work on data belonging to more genres and DA varieties, hence addressing a more challenging task.

3 Data Sets and Annotation

To our knowledge, no gold-labeled social media SSA data exist. Thereby, we create annotated data comprising a variety of data sets:

DARDASHA (DAR): (Arabic for "chat") comprises the first 2798 chat turns collected from a randomly selected chat session from "Egypt's room" in Maktoob chat `chat.mymaktoob.com`. Maktoob is a popular Arabic portal. DAR is an Egyptian Arabic subset of a larger chat corpus that was harvested between December 2008 and February 2010.

TAGREED (TGRD): ("tweeting") is a corpus of 3015 Arabic tweets collected during May 2010. TRGD has a mixture of MSA and DA. The MSA part (TRGD-MSA) has 1466 tweets, and the dialectal part (TRGD-DA) has 1549 tweets.

TAHRIR (THR): ("editing") is a corpus of 3008 sentences sampled from a larger pool of 30 MSA Wikipedia Talk Pages that we harvested.

MONTADA (MONT): ("forum") comprises of 3097 Web forum sentences collected from a larger pool of threaded conversations pertaining to different varieties of Arabic, including both MSA and DA, from the COLABA data set (Diab et al., 2010). The discussions covered in the forums pertain to social issues, religion or politics. The sentences were automatically filtered to exclude non-MSA threads.

Each of the data sets was labeled at the sentence level by two college-educated native speakers of Arabic. For each sentence, the annotators assigned one of 3 possible labels: (1) objective (OBJ), (2) subjective-positive (S-POS), (3) subjective-negative (S-NEG), and (3) subjective-mixed (S-MIXED). Following (Wiebe et al., 1999), if the primary goal of a sentence is judged as the objective reporting of information, it was labeled as OBJ. Otherwise, a sentence was a candidate for one of the three SUBJ classes. We also labeled the data with a number of other *metadata*[1] tags. Metadata labels included the user gender (GEN), the user identity (UID) (e.g. the user could be a *person* or an *organization*), and the source document ID (DID). We also mark the language variety (LV) (i.e., MSA or DA) used, tagged at the level of each unit of analysis (i.e., sentence, tweet, etc.). Annotators were instructed to label a tweet as MSA if it mainly employs MSA words and adheres syntactically to MSA rules, otherwise it is treated as dialectal. Table 1 shows the annotations for each data set. Data statistics, distribution of classes, and inter-annotator agreement in terms of Kappa (K) are provided in Table 2.

Data set	SUBJ	GEN	LV	UID	DID
DAR	✓	✓			
MONT	✓	✓			✓
TRGD	✓	✓	✓	✓	
THR	✓				✓

Table 1: Types of annotation labels (features) manually assigned to the data.

Polarity Lexicon: We manually created a lexicon of 3982 adjectives labeled with one of the following tags {*positive, negative, neutral*}, as is reported in our previous work (2011b). We focus on adjectives since they are primary sentiment bearers. The adjectives pertain to the newswire domain, and were extracted from the first four parts of the Penn Arabic Treebank (Maamouri et al., 2004).

4 SAMAR

4.1 Automatic Classification

SAMAR is a machine learning system for Arabic SSA. For classification, we use SVM$^{\text{light}}$ (Joachims, 2008). In our experiments, we found that linear kernels yield the best performance. We perform all experiments with *presence* vectors: In each sentence vector, the value of each dimension is binary, regardless of how many times a feature occurs.

In the current study, we adopt a *two-stage* classification approach. In the first stage (i.e., *Subjectivity*), we build a binary classifier to separate objective from subjective cases. For the second stage (i.e., *Sentiment*) we apply binary classification that distinguishes S-POS from S-NEG cases. We disregard the neutral and mixed classes for this study. SAMAR uses different feature sets, each of which is designed to address an individual research question:

4.2 Morphological Features

Word forms: In order to minimize data sparseness as a result of the morphological richness of Arabic, we tokenize the text automatically. We use AMIRA (Diab, 2009), a suite for automatic

[1]We use the term 'metadata' as an approximation, as some features are more related to social interaction phenomena.

Data set	# instances	# types	# tokens	# OBJ	# S-POS	# S-NEG	# S-MIXED	Kappa (*K*)
DAR	2,798	11,810	3,133	328	1647	726	97	0.89
MONT	3,097	82,545	20,003	576	1,101	1,027	393	0.88
TRGD	3,015	63,383	16,894	1,428	483	759	345	0.85
TRGD-MSA	1,466	31,771	9,802	960	226	186	94	0.85
TRGD-DIA	1,549	31,940	10,398	468	257	573	251	0.82
THR	3,008	49,425	10,489	1,206	652	1,014	136	0.85

Table 2: Data and inter-annotator agreement statistics.

processing of MSA, trained on Penn Arabic Treebank (Maamouri et al., 2004) data, which consists of newswire text. We experiment with two different configurations to extract base forms of words: (1) *Token* (TOK), where the stems are left as is with no further processing of the morpho-tactics that result from the segmentation of clitics; (2) *Lemma* (LEM), where the words are reduced to their lemma forms, (citation forms): for verbs, this is the 3rd person masculine singular perfective form and for nouns, this corresponds to the singular default form (typically masculine). For example, the word وبحسناتهم (*wbHsnAtHm*) is tokenized as و + ب + حسنات + هم (*w+b+HsnAt+Hm*) (note that in TOK, AMIRA does not split off the pluralizing suffix ات (*At*) from the stem حسن (*Hsn*)), while in the lemmatization step by AMIRA, the lemma rendered is حسنه (*Hsnp*). Thus, SAMAR uses the form of the word as *Hsnp* in the LEM setting, and *HsnAt* in the TOK setting.

POS tagging: Since we use only the base forms of words, the question arises whether we lose meaningful morphological information and consequently whether we could represent this information in the POS tags instead. Thus, we use two sets of POS features that are specific to Arabic: the reduced tag set (RTS) and the extended reduced tag set (ERTS) (Diab, 2009). The RTS is composed of 42 tags and reflects only number for nouns and some tense information for verbs whereas the ERTS comprises 115 tags and enriches the RTS with gender, number, and definiteness information. Diab (2007b; 2007a) shows that using the ERTS improves results for higher processing tasks such as base phrase chunking of Arabic.

4.3 Standard Features

This group includes two features that have been employed in various SSA studies.

Unique: Following Wiebe et al. (2004), we apply a UNIQUE (Q) feature: We replace low frequency words with the token "UNIQUE". Experiments showed that setting the frequency threshold to 3 yields the best results.

Polarity Lexicon (PL): The lexicon (cf. section 3) is used in two different forms for the two tasks: For subjectivity classification, we follow Bruce and Wiebe (1999; 2011b) and add a binary *has_adjective* feature indicating whether or not any of the adjectives in the sentence is part of our manually created polarity lexicon. For sentiment classification, we apply two features, *has_POS_adjective* and *has_NEG_adjective*. These binary features indicate whether a POS or NEG adjective from the lexicon occurs in a sentence.

4.4 Dialectal Arabic Features

Dialect: We apply the two gold language variety features, {*MSA, DA*}, on the Twitter data set to represent whether the tweet is in MSA or in a dialect.

4.5 Genre Specific Features

Gender: Inspired by gender variation research exploiting social media data (e.g., (Herring, 1996)), we apply three *gender* (GEN) features corresponding to the set {*MALE, FEMALE, UNKNOWN*}. Abdul-Mageed and Diab (2012a) suggest that there is a relationship between politeness strategies and sentiment expression. And gender variation research in social media shows that expression of linguistic politeness (Brown and Levinson, 1987) differs based on the gender of the user.

User ID: The *user ID* (UID) labels are inspired by research on Arabic Twitter showing that a considerable share of tweets is produced by organizations such as news agencies (Abdul-Mageed et al., 2011a) as opposed to lay users. We hence employ two features from the set {PERSON, ORGANIZATION} to

classification of the Twitter data set. The assumption is that tweets by persons will have a higher correlation with expression of sentiment.

Document ID: Projecting a *document ID* (DID) feature to the paragraph level was shown to improve subjectivity classification on data from the health policy domain (Abdul-Mageed et al., 2011c). Hence, by employing DID at the instance level, we are investigating the utility of this feature for social media as well as at a finer level of analysis, i.e., the sentence level.

5 Empirical Evaluation

For each data set, we divide the data into 80% training (TRAIN), 10% for development (DEV), and 10% for testing (TEST). The classifier was optimized on the DEV set; all results that we report below are on TEST. In each case, our baseline is the majority class in the training set. We report accuracy as well as the F scores for the individual classes (objective vs. subjective and positive vs. negative).

5.1 Impact of Morphology on SSA

We run two experimental conditions: 1. A comparison of TOK to LEM (cf. sec. 4.2); 2. A combination of RTS and ERTS with TOK and LEM.

TOK vs. LEM: Table 3 shows the results for the morphological preprocessing conditions. The baseline, Base, is the majority class in the training data. For all data sets, Subjective is the majority class. For subjectivity classification we see varying performance. DAR: TOK outperforms LEM for all metrics, yet performance is below Base. TGRD: LEM preprocessing yields better accuracy results than Base. LEM is consistently better than TOK for all metrics. THR: We see the opposite performance compared to the TGRD data set where TOK outperforms LEM and also outperforming Base. Finally for MONT: the performance of LEM and TOK are exactly the same yielding the same results as in Base.

For sentiment classification, the majority class is positive for DAR and MONT and negative for TGRD and THR. We note that there are no obvious trends between TOK and LEM. DAR: we observe better performance of LEM over Base and TOK. TGRD: Both preprocessing schemes outperform Base on all metrics with TOK outperforming LEM across the board. THR: LEM outperforms TOK for all metrics of sentiment, yet they are below Base performance. MONT: TOK outperforms LEM in terms of accuracy, and positive sentiment, yet LEM slightly outperforms TOK for negative sentiment classification. Both TOK and LEM are beat by Base in terms of accuracy and positive classification. Given the observed results, we observe no clear trends for the impact for morphological preprocessing alone on performance.

		SUBJ			SENTI		
Data	**Cond.**	**Acc**	***F*-O**	***F*-S**	**Acc**	***F*-P**	***F*-N**
DAR	Base	*84.75*	0.00	*91.24*	*63.02*	*77.32*	0.00
	TOK	**83.90**	0.00	**91.24**	67.71	77.04	45.61
	LEM	83.76	0.00	91.16	**70.16**	**78.65**	**50.43**
TRGD	Base	*61.59*	0.00	*76.23*	*56.45*	0.00	*72.16*
	TOK	69.54	64.06	73.56	**65.32**	**49.41**	**73.62**
	LEM	**71.19**	**64.78**	**75.63**	62.10	41.98	71.86
THR	Base	*52.92*	0.00	*69.21*	*75.00*	0.00	*85.71*
	TOK	**58.44**	**28.09**	**70.78**	60.47	37.04	71.19
	LEM	57.79	26.97	70.32	**63.37**	**38.83**	**73.86**
MONT	Base	*83.44*	0.00	*90.97*	*86.82*	*92.94*	0.00
	TOK	83.44	0.00	90.97	**74.55**	**83.63**	42.86
	LEM	83.44	0.00	90.97	72.27	81.68	**42.99**

Table 3: SSA results with preprocessing TOK and LEM.

Adding POS tags: Table 4 shows the results of adding POS tags based on the two tagsets RTS and ERTS. Subjectivity classification: The results show that adding POS information improves accuracy and F score for all the data sets except MONT which is still at Base performance. RTS outperforms ERTS with TOK, and the opposite with LEM where ERTS outperforms RTS, however, overall TOK+RTS yields the highest performance of 91.49% F score on subjectivity classification for the DAR dataset. For the TGRD and THR data sets, we note that TOK+ERTS is equal to or outperforms the other conditions on subjectivity classification. For MONT there is no difference between experimental conditions and no impact for adding the POS tag information. In the sentiment classification task:

The sentiment task shows a different trend: here, the highest performing systems do not use POS tags. This is attributed to the variation in genre between the training data on which AMIRA is trained (MSA newswire) and the data sets we are experimenting with in this work. However in relative compari-

Data	Cond.	SUBJ Acc	SUBJ *F*-O	SUBJ *F*-S	SENTI Acc	SENTI *F*-P	SENTI *F*-N
DAR	Base	*84.75*		*91.24*	*63.02*	*77.32*	
	TOK+RTS	**84.32**	0.00	**91.49**	66.15	76.36	40.37
	TOK+ERTS	83.90	0.00	91.24	67.19	77.09	42.20
	LEM+RTS	83.47	0.00	90.99	67.71	77.21	44.64
	LEM+ERTS	83.47	0.00	90.99	**68.75**	**77.94**	**46.43**
TGRD	Base	*61.59*		*76.23*	*56.45*		*72.16*
	TOK+RTS	70.20	64.57	74.29	62.90	43.90	72.29
	TOK+ERTS	71.19	65.06	**75.49**	62.90	42.50	72.62
	LEM+RTS	70.20	64.57	74.29	62.90	46.51	71.60
	LEM+ERTS	**72.19**	**76.54**	71.19	**65.32**	**48.19**	**73.94**
THR	Base	*52.92*		*69.21*	*75.00*		*85.71*
	TOK+RTS	57.47	28.42	69.75	59.30	33.96	70.59
	TOK+ERTS	**59.42**	**28.57**	**71.66**	59.88	**38.94**	70.13
	LEM+RTS	**59.42**	**28.57**	**71.66**	59.88	33.01	**71.37**
	LEM+ERTS	58.77	25.73	71.46	**60.47**	37.04	71.19
MONT	Base	*83.44*		*90.97*	*86.82*	*92.94*	
	TOK+RTS	83.44	0.00	90.97	69.09	79.27	39.29
	TOK+ERTS	83.44	0.00	90.97	**71.82**	**81.55**	**40.38**
	LEM+RTS	83.44	0.00	90.97	70.00	80.36	36.54
	LEM+ERTS	83.44	0.00	90.97	69.55	79.64	39.64

Table 4: SSA results with different morphological preprocessing and POS features.

son between RTS and ERTS for sentiment shows that in a majority of the cases, ERTS outperforms RTS, thus indicating that the additional morphological features are helpful. One possible explanation may be that variations of some of the morphological features (e.g., existence of a gender, person, adjective feature) may correlate more frequently with positive or negative sentiment.

5.2 Standard Features for Social Media Data

RQ2 concerns the question whether standard features can be used successfully for classifying social media text characterized by the usage of dialect and by differing text lengths. We add the standard features, polarity (PL) and UNIQUE (Q), to the two tokenization schemes and the POS tag sets. We report only the best performing conditions here.

Table 5 shows the best performing settings per corpus from the previous section as well as the best performing setting given the new features. The results show that apart from THR and TGRD for sentiment, all corpora gain in accuracy for both subjectivity and sentiment. In the case of subjectivity, while considerable improvements are gained for both DAR (11.51% accuracy) and THR (32.90% accuracy), only slight improvements ($< 1\%$ accuracy) are reached for both TGRD and MONT. For sentiment classification, the improvements in accuracy are less than the case of subjectivity: 1.84% for DAR and 6.81% for MONT. The deterioration on THR is surprising and may be a result of the nature of sentiment as expressed in the THR data set: Wikipedia has a 'Neutral Point of View' policy based on which users are required to focus their contributions not on other users but content, and as such sentiment is expressed in nuanced indirect ways in THR. While the subjectivity results show that it is feasible to use the combination of the UNIQUE feature and the polarity lexicon features successfully, even for shorter texts, such as in the twitter data (TGRD), this conclusion does not always hold for sentiment classification. However, we assume that the use of the polarity lexicon would result in higher gains if the lexicon were adapted to the new domains.

5.3 SSA Given Arabic Dialects

RQ3 investigates how much the results of SSA are affected by the presence or absence of dialectal Arabic in the data. For this question, we focus on the TGRD data set because it contains a non-negligible amount (i.e., 48.62%) of tweets in dialect.

First, we investigate how our results change when we split the TGRD data set into two subsets, one containing only MSA, the other one containing only DA. We extract the 80-10-10% data split, then train and test the classifier exclusively on either MSA or dialect data. The subjectivity results for this experiment are shown in Table 6, and the sentiment re-

	SUBJ				SENTI			
Data	**Best condition**	**Acc**	***F*-O**	***F*-S**	**Best condition**	**Acc**	***F*-P**	***F*-N**
DAR	TOK+RTS	84.32	0.00	91.49	LEM+ERTS	68.75	77.94	46.43
	TOK+ERTS+PL+Q3	**95.83**	0.00	97.87	LEM+ERTS+PL+Q3	**70.59**	79.51	47.92
TGRD	LEM+ERTS	72.19	76.54	71.19	LEM+ERTS	65.32	73.94	48.19
	LEM+ERTS+PL	**72.52**	65.84	77.01	LEM+ERTS+PL	65.32	73.94	48.19
THR	L./T.+ERTS	59.42	28.57	71.66	LEM+ERTS	63.37	38.83	73.86
	TOK+ERTS +PL+Q3	**83.33**	0.00	90.91	LEM+RTS+PL+Q3	61.05	34.95	72.20
MONT	LEM+ERTS	83.44	0.00	90.97	TOK	74.55	83.63	42.86
	LEM+RTS+PL+Q3	**84.19**	3.92	91.39	TOK+PL+Q3	81.36	88.64	48.10

Table 5: SSA results with standard features. Number in bold signify improvements over the best results in section 5.1.

	TGRD			TGRD-MSA			TGRD-DA		
Cond.	**Acc**	***F*-O**	***F*-S**	**Acc**	***F*-O**	***F*-S**	**Acc**	***F*-O**	***F*-S**
Base	61.59	0.00	76.23	51.68	68.14	0.00	78.40	0.00	87.89
TOK	69.54	64.06	73.56	**61.74**	70.16	46.73	78.40	5.41	87.80
LEM	71.19	64.78	75.63	**65.10**	72.04	53.57	**79.01**	15.00	88.03

Table 6: Dialect-specific subjectivity experiments.

sults are shown in Table 7. For both tasks, the results show considerable differences between MSA and DA: For TGRD-MSA, the results are lower than for TGRD-DA, which is a direct consequence of the difference in distribution of subjectivity between the two subcorpora. TGRD-DA is mostly subjective while TGRD-MSA is more balanced. With regard to sentiment, TGRD-DA consists of mostly negative tweets while TGRD-MSA again is more balanced. These results suggest that knowing whether a tweet is in dialect would help classification.

For subjectivity, we can see that TGRD-MSA improves by 13.5% over the baseline while for TGRD-DA, the improvement is more moderate, $< 3\%$. We assume that this is partly due to the higher skew in TGRD-DA, moreover, it is known that our preprocessing tools yield better performance on MSA data leading to better tokenization and lemmatization.

For sentiment classification on TGRD-MSA, neither tokenization nor lemmatization improve over the baseline. This is somewhat surprising since we expect AMIRA to work well on this data set and thus to lead to better classification results. However, a considerable extent of the MSA tweets are expected to come from news headlines (Abdul-Mageed et al., 2011a), and headlines usually are not loci of explicitly subjective content and hence are difficult to classify and in essence harder to preprocess since the genre is different from regular newswire even if MSA. For the TGRD-DA data set, both lemmatization and tokenization improve over the baseline.

The results for both subjectivity and sentiment on the MSA and DA sets suggest that processing errors by AMIRA trained exclusively on MSA newswire data) result in deteriorated performance. However we do not observe such trends on the TGRD-DA data sets. This is not surprising since the TGRD-DA is not very different from the newswire data on which AMIRA was trained: Twitter users discuss current events topics also discussed in newswire. There is also a considerable lexical overlap between MSA and DA. Furthermore, dialectal data may be loci for more sentiment cues like emoticons, certain punctuation marks (e.g. exclamation marks), etc. Such clues are usually absent (or less frequent) in MSA data and hence the better sentiment classification on TGRD-DA.

We also experimented with adding POS tags and standard features. These did not have any positive effect on the results with one exception, which is shown in Table 8: For sentiment, adding the RTS tagset has a positive effect on the two data sets.

In a second experiment, we used the original TGRD corpus but added the language variety (LV) (i.e., MSA and DA) features. For both subjectivity and sentiment, the best results are acquired using the LEM+PL+LV settings. However, for subjectivity, we observe a drop in accuracy from 72.52% (LEM+ERTS+PL) to 69.54%. For sentiment, we also observe a performance drop in accuracy, from 65.32% (LEM+ERTS+PL) to 64.52%. This means that knowing the language variety does not provide

Cond.	TGRD			TGRD-MSA			TGRD-DA		
	Acc	***F*-P**	***F*-N**	**Acc**	***F*-P**	***F*-N**	**Acc**	***F*-P**	***F*-N**
Base	56.45	0.00	72.16	53.49	69.70	0.00	67.47	0.00	80.58
TOK	65.32	49.41	73.62	53.49	56.52	50.00	**68.67**	23.53	80.30
LEM	62.10	41.98	71.86	48.84	52.17	45.00	**73.49**	38.89	83.08
TOK+RTS	70.20	64.57	74.29	55.81	61.22	48.65	**71.08**	29.41	81.82

Table 7: Dialect-specific sentiment experiments.

Data	SUBJ				SENTI			
	Condition	**Acc**	***F*-O**	***F*-S**	**Condition**	**Acc**	***F*-P**	***F*-N**
DAR	TOK+ERTS+PL+Q3	**95.83**	0.00	97.87	LEM+PL+GEN	**71.28**	79.86	50.00
TGRD	LEM+ERTS+PL	**72.52**	65.84	77.01	TOK+ERTS+PL+GEN+LV+UID	**65.87**	49.41	74.25
THR	TOK+ERTS+PL+Q3	**83.33**	0.00	90.91	TOK+PL+GEN+UID	**67.44**	39.13	77.78
MONT	LEM+RTS+PL+Q3	**84.19**	3.92	91.39	TOK+PL+Q3	81.36	88.64	48.10

Table 8: Overall best SAMAR performance. Numbers in bold show improvement over the baseline.

Data	Condition	Acc	*F*-O	*F*-S
DAR	TOK+ERTS+PL+GEN	84.30	0.00	91.48
TGRD	LEM+RTS+PL+UID	71.85	65.31	76.32
THR	LEM+RTS+PL+GEN+UID	66.67	0.00	80.00
MONT	LEM+RTS+PL+DID	83.17	0.00	90.81

Table 9: Subjectivity results with genre features.

Data	Condition	Acc	*F*-P	*F*-N
DAR	LEM+PL+GEN	**71.28**	79.86	50.00
TGRD	TOK+ERTS+PL+GEN+LV +UID	**65.87**	49.41	74.25
THR	TOK+PL+GEN+UID	**67.44**	39.13	77.78
MONT	LEM+PL+DID	76.82	47.42	85.13

Table 10: Sentiment results with genre features. Numbers in bold show improvement over table 5.

enough information for successfully conquering the differences between those varieties.

5.4 Leveraging Genre Specific Features

RQ4 investigates the question whether we can leverage features typical for social media for classification. We apply all GENRE features exhaustively.We report the best performance on each data set.

Table 9 shows the results of adding the genre features to the subjectivity classifier. For this task, no data sets profit from these features.

Table 10 shows the results of adding the genre features to the sentiment classifier. Here, all the data sets, with the exception of MONT, profit from the new features. In the case of DAR, adding gender information improves classification by 1.73% in accuracy. For TGRD, the combination of the gender (GN), language variety (LV), and user ID slightly (0.52%) improves classification over previous best settings. For THR, adding the gender and user ID information improves classification by 4.07%.

Our results thus show the utility of the gender, LV, and user ID features for sentiment classification. The results for both subjectivity and sentiment show that the document ID feature is not a useful feature.

6 Overall Performance

Table 8 provides the best results reached by SAMAR. For subjectivity classification, SAMAR improves on all data sets when the POS features are combined with the standard features. For sentiment classification, SAMAR also improves over the baseline on all the data sets, except MONT. The results also show that all optimal feature settings for subjectivity, except with the MONT data set, include the ERTS POS tags while the results in Section 5.1 showed that adding POS information without additional features, while helping in most cases with subjectivity, does not help with sentiment classification.

7 Conclusion and Future Work

In this paper, we presented SAMAR, an SSA system for Arabic social media. We explained the rich feature set SAMAR exploits and showed how complex morphology characteristic of Arabic can be handled in the context of SSA. For the future, we plan to carry out a detailed error analysis of SAMAR in an attempt to improve its performance, use a recently-developed wider coverage polarity lexicon (Abdul-Mageed and Diab, 2012b) together with another DA lexicon that we are currently developing.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems*, 26:1–34.

Muhammad Abdul-Mageed and Mona Diab. 2012a. *AWATIF*: A multi-genre corpus for Modern Standard Arabic subjectivity and sentiment analysis. In *Proceedings of LREC*, Istanbul, Turkey.

Muhammad Abdul-Mageed and Mona Diab. 2012b. Toward building a large-scale Arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet Conference*, Matsue, Japan.

Muhammad Abdul-Mageed, Hamdan Albogmi, Abdulrahman Gerrio, Emhamed Hamed, and Omar Aldibasi. 2011a. Tweeting in Arabic: What, how and whither. Presented at the 12th Annual Conference of the Association of Internet Researchers (Internet Research 12.0, Performance and Participation), Seattle, WA.

Muhammad Abdul-Mageed, Mona Diab, and Mohamed Korayem. 2011b. Subjectivity and sentiment analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, OR.

Muhammad Abdul-Mageed, Mohamed Korayem, and Ahmed YoussefAgha. 2011c. "Yes we can?": Subjectivity annotation and tagging for the health domain. In *Proceedings of RANLP2011*, Hissar, Bulgaria.

Ann Banfield. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge, Boston.

Penelope Brown and Stephen Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Rebecca Bruce and Janyce Wiebe. 1999. Recognizing subjectivity. A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

Kushal Dave, Steve Lawrence, and David Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528, Budapest, Hungary. ACM.

Mona Diab, Dan Jurafsky, and Kadri Hacioglu. 2007. Automatic processing of Modern Standard Arabic text. In Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors, *Arabic Computational Morphology*. Springer.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassin Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74, Valetta, Malta.

Mona Diab. 2007a. Improved Arabic base phrase chunking with a new enriched POS tag set. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 89–96, Prague, Czech Republic.

Mona Diab. 2007b. Towards an optimal POS tag set for Modern Standard Arabic processing. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Mona Diab. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 285–288, Cairo, Egypt.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Susan Herring. 1996. Bringing familiar baggage to the new frontier: Gender differences in computer-mediated communication. In J. Selzer, editor, *Conversations*. Allyn & Bacon.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA.

Thorsten Joachims. 2008. Svmlight: Support vector machine. http://svmlight.joachims.org/, Cornell University, 2008.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva, Switzerland.

Mohamed Maamouri, Anne Bies, Tim Buckwalter, and W. Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA.

Janyce Wiebe, Rebecca Bruce, and Tim O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th*

Annual Meeting of the Association for Computational Linguistics (ACL-99), pages 246–253, University of Maryland.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30:227–308.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434, Melbourne, FL.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sapporo, Japan.