Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

SAMAR: Subjectivity and sentiment analysis for Arabic social media☆

Muhammad Abdul-Mageed[a,b,*], Mona Diab[c], Sandra Kübler[a]

[a] Department of Linguistics, Indiana University, 1021 E 3rd. St., Bloomington, IN 47405, USA
[b] School of Library and Information Science, 1320 East 10th Street, Bloomington, IN 47405, USA
[c] Department of Computer Science, School of Engineering & Applied Science, The George Washington University, Washington, DC, USA

Received 16 August 2012; received in revised form 16 December 2012; accepted 13 March 2013

Abstract

SAMAR is a system for subjectivity and sentiment analysis (SSA) for Arabic social media genres. Arabic is a morphologically rich language, which presents significant complexities for standard approaches to building SSA systems designed for the English language. Apart from the difficulties presented by the social media genres processing, the Arabic language inherently has a high number of variable word forms leading to data sparsity. In this context, we address the following 4 pertinent issues: how to best represent lexical information; whether standard features used for English are useful for Arabic; how to handle Arabic dialects; and, whether genre specific features have a measurable impact on performance. Our results show that using either lemma or lexeme information is helpful, as well as using the two part of speech tagsets (RTS and ERTS). However, the results show that we need individualized solutions for each genre and task, but that lemmatization and the ERTS POS tagset are present in a majority of the settings.

Keywords: Subjectivity and sentiment analysis; Morphologically rich language; Arabic; Social media data

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,

☆ This paper has been recommended for acceptance by R.K. Moore.
* Corresponding author at: Department of Linguistics, Indiana University, 1021 E 3rd. St., Bloomington, IN 47405, USA. Tel.: +1 812 855 6456.
E-mail address: mabdulma@indiana.edu (M. Abdul-Mageed).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.03.001

only few attempts have been made to build SSA systems for *morphologically-rich languages*, i.e., languages in which a significant amount of information concerning syntactic units and relations is expressed at the word-level (Tsarfaty et al., 2010), such as Finnish or Arabic, cf. (Abbasi et al., 2008; Abdul-Mageed et al., 2011a; Mihalcea et al., 2007). Thus, we aim at partially bridging these two gaps in research by presenting an SSA system for Arabic social media genres as Arabic is one of the most morphologically 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 for four different genres: Synchronous chat, Twitter, Web discussion fora, and Wikipedia Talk Pages. These genres vary considerably in terms of their functions and the language variety employed. While the chat genre is mostly in dialectal Arabic, the other genres are mixed between Modern Standard Arabic (MSA) and dialectal Arabic to varying degrees.

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? To date most robust SSA systems have been developed for English, which has relatively little morphological variation. In such systems most of the features are highly lexicalized, hence a direct application of these methods would not be quite as successful for Arabic since a lemma in Arabic may be associated with hundreds if not thousands of variant surface forms. Accordingly, 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 characterize our problem in two spaces: the lexical space comparing simple lexeme tokenization with full lemmatization (lexemes vs. lemmas); abstracting away from the lexical form to the part of speech class, we investigate using two different POS tag sets for Arabic that encode a significant amount of morphological information.
- **RQ2:** Can standard features be effective for SSA when handling social media despite the inherently short texts typically used in these genres? In this prong of the research we investigate the impact of using two standard features frequently employed in SSA studies (Wiebe et al., 2004; Turney, 2002) on social media data that employes dialectal Arabic usage and the text inherently varying in length (i.e., the text being very short, 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 certain 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. Moreover, it should be noted that dialectal Arabic, to date, does not have a standardized orthography, therefore low frequency content words will be pervasive in social media genres since most of these genres employ dialectal Arabic. Second, we test whether a polarity lexicon that was extracted from a standard Modern Standard Arabic (MSA) newswire domain is useful for processing SSA for social media data.
- **RQ3:** How do we handle dialects in an SSA system for Arabic? For Arabic, there are significant differences between dialects on all levels of linguistic representation: morphology, lexical, phonology, syntax, semantics, and pragmatics. This difference is even more pronounced between the dialects and MSA. However, existing robust Arabic NLP tools such as tokenizers, Part of speech (POS) taggers, and syntactic parsers are exclusively trained on and for MSA newswire genres. Therefore we would like to measure the impact on SSA performance of explicitly modeling for dialectal usage.
- **RQ4:** Which features specific to social media can we leverage? We are interested in investigating the impact of using information that is typically present in social media (meta) data such as gender, author and document id information on SSA performance.

The remainder of the paper is organized as follows: In Section 2, we give an overview of the linguistic characteristics of Arabic that are important for our work; Section 3 describes the social media corpora and the polarity lexicon used in the experiments; In Section 4, we review related work; Section 5 describes the SSA system, SAMAR, used for the current research, as well as the features used in the experiments; Section 6 describes the experiments and discusses the results; In Section 7, we give an overview of the best settings for the different corpora, followed by a conclusion in Section 8.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

2. Arabic facts

Arabic is a Semitic language known for its rich morphology. For our purposes, we define a word to be a space delimited token in naturally occurring written Arabic text. A typical word in Semitic languages packs more information than a typical word in a language such as English. A word in Arabic exhibits several morphological aspects: derivation, inflection, and agglutination.

Morphemes. New words can be derived from existing words or morphemes. A morpheme is the basic, minimal linguistic unit that bears meaning, regardless of whether it can stand alone as a word. For example, the inflectional plural marker suffix *-s*, the lexical prefix *anti-*, or the word *table* are all morphemes in English; the word *tables* is comprised of two morphemes: *table* +*-s*. There are four ways to combine morphemes, where typically one morpheme (called the *base*) bears the core meaning of the word: before the base (prefix or proclitic; e.g., *Al+*[1] 'the'), after the base (suffix or enclitic; e.g., *+At* feminine plural form), or both before and after (circumfix; e.g., *ta+...+uwA* present tense second person masculine plural, or *ma+...+$* negation in some Arabic dialects).

2.1. Derivational morphology

Semitic languages are largely templatic. Most derived words are comprised of a root and a pattern/template. However the semantics of the word is not always predictable due to some arbitrary idiosyncratic features due to semantic drift in its meaning, with usage over time.The derivational mechanism typically takes two morphemes and creates a new word with a part of speech possibly different from that of any of the participating morphemes. For example, *speak* (verb) +*-er* → *speaker* (noun) in English. Derivational morphology in Arabic, like most Semitic languages, is made up of roots and patterns.

Roots. The *root* is an ordered tuple of consonants, a.k.a. *radicals*, most often a triplet (but could be a pair or quadruple). For example, the three radicals ك ت ب *k t b* are a root related to 'writing' in Arabic. Roots are not necessarily monosemic, i.e., a root could have multiple meanings. Typically a root is unpronounceable. It is meant to be an abstract concept around which words cluster. The vowels are added by merging with a pattern/template. Roots and patterns are inherent parts of Semitic languages and unique to them.

Pattern/wazn/template. The *template/pattern* is a fixed sequence of vowels and interleaved place-holders for the root's consonants. For example *1a2a3* is a perfective verb, where the digits correspond to the root radicals. A pattern may also contain additional consonants. Like the root, the pattern/template is not necessarily monosemic. Roots and patterns are each morphemes in their own right. The root carries a basic meaning of an action or an object (in our example above, ك ت ب *k t b* are a root related to writing), while the template reflects a modifying meaning over the root's basic meaning, as in conveying a verb/action or a noun/object, mutual or reflexive action, etc. For example كَاتَب *kAtab* 'corresponded', كَتَّب *kat ab* 'dictated', and كِتَاب *kitAb* 'book (noun)' – each having a different pattern, merged with the same root ك ت ب *k t b*.

2.2. Inflectional morphology

Arabic inflects for various features: number, gender, grammatical aspect, syntactic case, mood, definiteness, person, voice, and tense. The set of words sharing meaning but varying over these features is called a *lexeme* class. It is typically represented as a citation form by one of the set members – called the *lemma*. The lemma is used for linguistic purposes, e.g., as the citation/head form in a dictionary, or as a feature in parsing or semantic processing. The lemma form is typically the masculine singular active perfective verb form for verbs and the masculine singular for nouns. Lemmas are typically fully diacritized where the vowels and gemination forms are present. The lexeme can inflect with certain combinations of these features depending on the part of speech (noun, verb, adjective, etc.) Table 1 illustrates some examples of words. It should be noted that lemmas are also inflected words.

In many cases the meaning of the resulting word is compositional, i.e., it can be constructed from the meaning of the participating morphemes.

[1] We use the Buckwalter transliteration scheme for rendering romanized Arabic script (Buckwalter, 2002) http://www.qamus.org.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 1
Arabic root and template combination for the root ك ت ب *k t b*, 'write'

Root	Template	Lemma	Lexeme	Gloss	Inflection
k t b	1a2a3	katab	katabat	'she wrote'	verb, past.3rd.sing.fem
k t b	1A2a3	kAtab	takAtabuw	'they corresponded'	verb, past.2nd.pl.masc
k t b	ma1o2a3	makotab	makotab	'office, desk'	noun, masc.sing
k t b	ma1A2i3 / ma1o2a3	makAtib/makotab	makAtib	'offices, desks'	noun, masc.pl
k t b	ma1o2a3/ap	makotab/makotabap	makotabap	'library'	noun, fem.sing
k t b	ma1o2a3/ap	makotab/makotabap	makotabAt	'libraries'	noun, fem.pl

In the example in Table 1, we note two types of plurals: sound مَكْتَبَات *makotabAt* and broken مَكَاتِب *makAtib*. The sound plurals are typically predictable, hence their underlying lemma is the same as the singular form of the word. Conversely the broken plurals are unpredictable, different from the general (sound) affixational case, hence their underlying lemma is often represented as the same as the broken form.

2.3. Agglutinative morphology

Lexeme members (including lemmas) could further agglutinate to certain closed class words (*clitics*), forming complex surface words in naturally occurring written Arabic texts (where in English, these would typically be space-delimited tokens). These agglutinative clitics may be particles (such as negation particles), prepositions, and grammatical markers of various sorts, such as aspectual and future marker morphemes, conjunctions, and pronouns, some of which occur as enclitics (suffixes) and others as proclitics (prefixes). It is worth noting that agglutinative morphology is more complex in dialects of Arabic than they are in MSA, thereby rendering the tokenization of Arabic dialects much more complex as a process. For example, the Arabic word وبِحَسَنَتِهِم *wabi. Hasanathim*, 'and by their virtue (sing.)', comprises the morphemes illustrated in Table 2.

In Table 2, the stem is simply the base word without handling of morphotactics: namely in this case the conversion of the ت *t* word finally to a ة *p* once separating the clitic هم *him* from the stem. The lexeme is by definition inflectional. In this case the lemma and the lexeme are the same form. Agglutinative morphology is different from derivational and inflectional morphology in that the resulting word retains a complex meaning, as in the example above: conjunction + preposition + the stem + possessive pronoun (equivalent to an English conjunction of a prepositional phrase), and may reach a complexity equivalent to a whole sentence in English. The tokenization process typically handles clitic segmentation and morphotactic normalization.

3. Data genres, corpora & resources

To our knowledge, to date, no gold-labeled social media SSA data exist for Arabic. For this reason, 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 of 'Egypt's room' in Maktoob chat http://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. The language variety in this collection is mostly dialectal Arabic (DA).

Table 2
Arabic surface word wbHsnthm fully diacritized and analyzed: وبِحَسَنَتِهِم *wabiHasanathim*.

Morpheme	Type	Morphological class	POS	Gloss
wa	Proclitic	Agglutinative	Conjunction	'and'
bi	Proclitic	Agglutinative	Preposition	'by'
Hasanat	Stem	–	Noun	'virtue'
Hasanap	Lexeme	Inflectional	Noun	'virtue'
Hasanap	Lemma	Derivational	Noun	'virtue'
H s n	Root	Derivational	–	'good'
him	Enclitic	Agglutinative	Masc plural possessive pronoun	'their'

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 3
Types of annotation labels (features) manually assigned to the data.

Data set	SUBJ	GEN	LV	UID	DID
DAR	√	√			
MONT	√	√			√
TGRD	√	√	√	√	
THR	√				√

Table 4
Annotation statistics of the different genres.

Data set	# Instances	# Tokens	# Types	# OBJ	# S-POS	# S-NEG	# S-MIXED
DAR	2798	11,810	3133	328	1647	726	97
MONT	3097	82,545	20,003	576	1101	1027	393
TGRD	3015	63,383	16,894	1428	483	759	345
TGRD-MSA	1466	31,771	9802	960	226	186	94
TGRD-DIA	1549	31,940	10,398	468	257	573	251
THR	3008	49,425	10,489	1206	652	1014	136

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

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

MONTADA (MONT): ('forum') comprises 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 discussion topics covered in the forums are hand selected to be in the domains of social issues, religion, and/or politics. The sentences are automatically filtered to exclude non-MSA threads. It is worth noting that over 20% of the data used in this corpus is classical Arabic since some of it was drawn from literary criticism writings.

Each of the data sets is labeled at the sentence level by two college-educated native speakers of Arabic. For each sentence, the annotators assigned one of 4 possible labels: (1) objective (OBJ), (2) subjective-positive (S-POS), (3) subjective-negative (S-NEG), and (4) subjective-mixed (S-MIXED). Following Wiebe et al. (1999), if the primary goal of a sentence is judged to be the objective reporting of information, it is labeled OBJ. Otherwise, a sentence is a candidate for one of the three SUBJ classes. We also label the data with a number of other *metadata* tags.[2] 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 dialect) used, tagged at the level of each unit of analysis (i.e., sentence, tweet). Annotators are instructed to label a tweet as MSA if it mainly employs MSA words and adheres syntactically to MSA rules, otherwise it should be labeled DA.

Table 3 shows the types of annotations available for each data set. Data statistics, such as the distribution of classes, are provided in Table 4, and inter-annotator agreement in terms of Kappa (K) is shown in Table 5. For evaluation purposes, we chose the labels assigned by one of the annotators for each data set.

Polarity lexicon: We manually created a lexicon of 3982 adjectives labeled with one of the following tags {*positive*, *negative*, *neutral*}, (c.f. Abdul-Mageed et al., 2011a). We focus on adjectives since they are primary sentiment bearers. However, the adjectives pertain to the newswire domain: they were extracted from the first four parts of the Penn Arabic Treebank (Maamouri et al., 2004). Consequently, they are out of domain for all the corpora studied here.

4. Related work

There is a vast amount of literature on subjectivity and sentiment analysis, an excellent overview of this field can be found in Liu's overview (Liu, 2012). In this section, we will focus on research in areas that are close to the research

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

Table 5
Inter-annotator agreement statistics on the individual corpora.

Data set	Kappa (*K*)
DAR	0.89
MONT	0.88
TGRD	0.85
TGRD-MSA	0.85
TGRD-DIA	0.82
THR	0.85

questions that we address. We will start by giving an overview of approaches to SSA for English (Section 4.1), then we will concentrate on work on languages other than English (Section 4.2) with a specific focus on work on Arabic (Section 4.3).

4.1. SSA for English

Opinion detection has mostly been performed on the document level (cf. e.g., Bruce and Wiebe, 1999; Wiebe et al., 1999; Wiebe, 2000). Another focus that can be discerned is in terms of genre, on movie and product reviews (cf. e.g., Dave et al., 2003; Hu and Liu, 2004; Turney, 2002). However, there are a number of sentence- and phrase-level classifiers (e.g., Wiebe et al., 1999; Morinaga et al., 2002; Yi et al., 2003; Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004). Yu and Hatzivassiloglou (2003) propose a three-stage approach that performs subjectivity analysis first on the document level, then on the sentence level. In a final step, they classify the sentences into positive, negative, or neutral opinions.

If sentiment analysis is performed on the sentence level, it is generally with regard to a target concept (Morinaga et al., 2002; Yi et al., 2003; Kim and Hovy, 2004), i.e., the system has as its goal to identify sentiment towards a concept such as “cellular phone” or one of its attributes such as “size”.

Standard approaches to subjectivity include rule-based approaches (Morinaga et al., 2002), supervised classifiers such as Naive Bayes (Yu and Hatzivassiloglou, 2003) or statistical approaches using linguistic features as well as meta-data (Dave et al., 2003). Sentiment is often determined based on lexical resources such as wordnet (Dave et al., 2003; Yi et al., 2003; Kim and Hovy, 2004), sentiment lexicons (Yi et al., 2003), or bootstrapped word lists based on seeds (Kim and Hovy, 2004). Since sentiment is often determined for given target concept, a subtask of extracting opinion targets has been investigated separately using support vector machines (Kessler and Nicolov, 2009) or conditional random fields (Jakob and Gurevych, 2010).

More recently, the focus on movie and product reviews is becoming less prominent, with work on blogs (Vechtomova, 2010) and twitter data (Davidov et al., 2010; Speriosu et al., 2011) becoming available. Since tweets are rather short, often additional information such as twitter hashtags and smileys (Davidov et al., 2010) or from label propagation (Speriosu et al., 2011) are used. More recent developments with regard to machine learning techniques use graph-based methods, which allow a more global view of the problem as well as joint inference with sentence cohesion (Pang and Lee, 2004), agreement between speakers (Thomas et al., 2006; Bansal et al., 2008), or discourse relations (Somasundaran et al., 2009).

4.2. Multilingual SSA

Work on SSA for languages other than English is still in its infancy. The most important reason for this is the lack of resources, annotated data collections on the one hand and sentiment lexicons on the other. Thus, work in this area has concentrated to a considerable degree on approaches that leverage multilingual resources such as machine translation. We focus here on four seminal works that use machine translation in different ways to obtain SSA systems for Chinese, Japanese, and Romanian as well as one approach using parallel corpora rather than MT resources. Another approach is the creation of a multilingual corpus.

Kanayama et al. (2004) develop a high-precision sentiment analysis system at low development cost by making use of an existing transfer-based machine translation (MT) system between Japanese and English. The original MT system

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

consists of three parts: (1) a source language syntactic parser, (2) a bilingual transfer component, which converts the syntactic tree structures from Japanese to an English tree structure, and (3) a target language generator that lexicalizes the trees in the target language via a bilingual lexicon. In order to use the approach for SSA, Kanayama et al. replace the translation patterns in the transfer component by sentiment patters and the bilingual lexicons by a sentiment polarity lexicon. Thus, they generate sentiment units rather than another language.

Yao et al. (2006) propose an automatic translation method for determining the sentiment orientation of individual Chinese words. They use machine translation to translate sentiment lexicons rather than text. Overall, they use 10 bilingual lexicons to ensure wide coverage. Chinese words are translated into English. Yao et al. then extract all the English translations of a word and use them to form a vector that is later used for classifying the sentiment of the Chinese word. This lexical information is then used in a supervised classification approach.

Mihalcea et al. (2007) present an approach to automatically generate a subjectivity lexicon as well as a subjectivity-annotated corpus for Romanian from similar resources available for English. They create a target-language subjectivity lexicon by translating an existing source language lexicon and then build a classifier that relies on the resulting lexicon. The translation is based on two English-Romanian dictionaries. This new subjectivity lexicon is then used in a rule-based classifier to annotate Romanian texts. To our knowledge, this is the first approach to SSA for a morphologically rich language. However, as a consequence of translating a dictionary, Mihalcea et al. circumnavigate the problem of multiple morphological forms since all words are in their lemma form. As a consequence, the corpus had to be lemmatized, too.

Wan (2008) extends the machine translation approach to a method in which Chinese reviews are first translated into English using MT services, then the sentiment polarity of English reviews are identified by directly leveraging English resources. Consequently, sentiment is detected both on the English and the Chinese text, using Chinese and English lexicons that include (1) polarity terms, (2) negation terms, and (3) intensification terms. Finally, an ensemble method is used to combine the results from the English and the Chinese analysis.

Lu et al. (2011) also perform bilingual sentiment analysis for the language pair English–Chinese. However, rather than using machine translation, their approach is based on a parallel corpus, with some manual annotations available for both languages. The approach uses EM to jointly learn a classifier for each language, treating the sentiment labels in the unannotated text as latent variables. Lu et al. show that their approach can also deal with automatically translated texts rather than a parallel corpus.

All the methods described in this section depend heavily on the availability of a machine translation system that translates between the target language and English or a parallel corpus of the two languages. However, the best results are still reached when SSA specific resources are available in both languages (cf. the approach by Wan (2008)). Additionally, none of these approaches deal with morphological complexity.

Another approach is presented by Steinberger et al. (2011), who annotate a multilingual corpus for sentiment annotation. While the paper focuses on the annotation of the corpus, the authors also present an evaluation of their system, which is based on sentiment dictionaries for the individual languages. These dictionaries can be automatically extracted from the parallel corpus (Steinberger et al., 2012), but their approach to SSA does not seem to exploit the parallel nature of the corpus. For these reasons, we will look at Arabic-specific approaches to SSA next.

4.3. SSA for Arabic

To our knowledge, only two studies have been performed on Arabic. Abbasi et al. (2008) perform a sentiment analysis of English and Arabic Web forums, with an overarching goal of identifying hostility in computer-mediated communication. They make use of not only syntactic but also stylistic features. Syntactic features include word, *n*-grams, POS tag *n*-grams, and word roots. Stylistic features include character *n*-grams among other types of information. Thus, Abbasi et al. deal with the morphological richness of Arabic by indirect means in the form a *n*-grams as well as by reducing words to their roots.

Abbasi et al. use an entropy-weighted genetic algorithm (EWGA) as a feature selection technique on (1) an English benchmark movie review database taken from the IMDb movie review archive (http://www.imdb.com) and (2) a testbed of messages from two major extremist forums (a U.S. American one in English and a Middle Eastern one in Arabic). Their EWGA uses information gain as a heuristic to weight the various sentiment attributes. Abbasi et al. find that stylistic features on their own were outperformed by syntactic features (i.e., word *n*-grams, punctuation, word roots), but when triangulated with syntactic features, stylistic features helped gain higher classification accuracy of approximately 5%. A number of stylistic features were found to be specifically helpful, including the total number of characters, use

of digits and emphasizing symbols, and vocabulary richness. The root extraction is handled by a clustering algorithm (Abbasi and Chen, 2005), which compares words against a list of roots. The number of roots is manually set to 50. From the publication, it is unclear how well this algorithm performs on identifying roots, especially since only the most frequent roots are recognized. More related to the current research is our previous work (Abdul-Mageed et al., 2011a) on building an SSA system that exploits newswire data from the Penn Arabic Treebank (Maamouri et al., 2004). 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. This system reaches an *F*-score of 71.54 for subjectivity classification and 95.52 for sentiment detection. Our current work is an extension of this previous work. Major differences are that now, we use automatically predicted morphological features and work on data belonging to more genres and dialect varieties, hence addressing a more challenging task.

5. SAMAR

5.1. System

SAMAR is a supervised machine learning system for Arabic SSA. For classification, we use SVM^{light} (Joachims, 2008). In our experiments, we investigate various kernels and associated parameter settings and found that linear kernels yield the best performance. We perform all experiments with *presence* vectors, i.e., 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:

5.2. Morphological features

Word form. 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 processing of MSA, trained on Penn Arabic Treebank (Maamouri et al., 2004) data, which consists of newswire text. We experiment with two different configuration settings to extract base forms of words: (1) *Lexeme* (LEX), where the surface words are tokenized and the morphotactics at clitic boundaries are handled; (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). Table 6 illustrates the difference between the various word base forms. Note that the first example is plural while the second is singular: 'and by their virtues' vs. 'and by their virtue', which affects the form of the word if we are extracting the lexeme not the lemma form.

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 article definiteness (excluding construct state definiteness) information for nominals and person information for verbs and pronouns. Diab (2007a,b) shows that using ERTS improves performance for higher processing tasks such as Base Phrase Chunking for MSA.

Table 6
Example illustrating difference between lemma and Lexeme word forms.

	Surface form	Clitics	Lemma	Lexeme	Gloss
UTF8	وبحسَناتهم	و+ ب+ ... +هم	حسنة	حسنَات	'and by their virtues'
BW	*wbHsnAtHm*	*w+b+... +hm*	*Hsnp*	*HsnAt*	
UTF8	وبحستهم	و+ ب+ ... +هم	حسنة	حسنة	'and by their virtue'
BW	*wbHsntHm*	*w+b+... +hm*	*Hsnp*	*Hsnp*	

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

5.3. Standard features

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

Unique (*Q*). Following Wiebe et al. (2004), we apply a UNIQUE feature: We replace low frequency words with the token "UNIQUE". Our tuning experiments show that choosing a threshold of 3 results in the optimal parameter setting for the token frequency.

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) and Abdul-Mageed et al. (2011a) 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.

5.4. 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.

5.5. 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., 2011) as opposed to lay users. We hence employ two features from the set {*PERSON*, *ORGANIZATION*} to the 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.

6. Experiments, results & discussion

6.1. Data set up

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

6.2. Impact of morphology on SSA

For RQ1, we investigate the treatment of morphological richness in the context of Arabic SSA. We run two experimental conditions: 1. A comparison of lexemes (LEX) to lemmas (LEM) (cf. Section 5); 2. Adding POS tag features to LEX and LEM respectively. We experiment with two tagsets: RTS and ERTS. We report the results of these two sets of experiments for both subjectivity and sentiment classification. It is worth noting that LEX preserves some more morphological information on the lexical word, especially for grammatical case for sound plurals, when compared with the ERTS tag set which does not explicitly model grammatical case, construct state definiteness, indefiniteness markers, nor verbal mood information. Accordingly, LEM + RTS has the least amount of explicit morphological information. LEX + ERTS has redundant morphological features on both the words and in the tagset features (number, voice,

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 7
SSA results for LEX and LEM separately and then combined with different POS tagsets RTS and ERTS. Numbers in bold show the highest improvement per data set and task over the baseline.

Data	Condition	Subjectivity			Sentiment		
		Acc	*F*-O	*F*-S	Acc	*F*-P	*F*-N
DARDASHA	Baseline	84.30	0.00	91.48	64.74	78.60	0.00
	LEX	84.30	0.00	91.48	67.71	77.04	45.61
	LEX + RTS	84.30	0.00	91.48	66.15	76.36	40.37
	LEX + ERTS	84.30	0.00	91.48	68.23	77.82	44.04
	LEM	84.58	0.00	91.65	**70.16**	**78.65**	**50.43**
	LEM + RTS	84.30	0.00	91.48	67.19	77.09	42.20
	LEM + ERTS	**84.65**	0.00	**91.69**	68.75	77.78	47.37
TAGHREED	Baseline	52.40	0.00	68.76	56.45	0.00	72.16
	LEX	**71.38**	**69.62**	72.95	**65.32**	**49.41**	**73.62**
	LEX + RTS	67.87	61.42	72.47	62.90	43.90	72.29
	LEX + ERTS	67.21	60.63	71.91	62.90	42.50	72.62
	LEM	70.16	63.75	**74.65**	62.10	41.98	71.86
	LEM + RTS	68.85	62.15	73.54	62.90	46.51	71.60
	LEM + ERTS	68.85	62.15	73.54	64.52	46.34	73.49
TAHRIR	Baseline	59.09	0.00	74.29	75.00	0.00	85.71
	LEX	61.04	38.14	71.56	60.47	37.04	71.19
	LEX + RTS	59.42	34.55	70.59	59.30	33.96	70.59
	LEX + ERTS	61.36	38.34	71.87	59.30	36.36	70.09
	LEM	**61.69**	**38.54**	72.17	63.37	**38.83**	73.86
	LEM + RTS	61.36	37.70	72.00	58.14	28.00	70.49
	LEM + ERTS	58.44	34.69	69.52	59.30	35.19	70.34
MONTADA	Baseline	80.32	0.00	89.08	86.76	92.91	0.00
	LEX	**83.17**	0.00	**90.81**	74.55	83.63	42.86
	LEX + RTS	83.17	0.00	90.81	69.09	79.27	39.29
	LEX + ERTS	83.17	0.00	90.81	70.00	80.36	36.54
	LEM	**83.17**	0.00	**90.81**	72.27	81.68	**42.99**
	LEM + RTS	83.17	0.00	90.81	70.00	80.24	37.74
	LEM + ERTS	83.17	0.00	90.81	68.64	79.15	36.70

tense, aspect, gender, determiner definiteness), lexically we find grammatical case, and verbal mood present for plural words, and potentially we might find construct state definiteness, and indefiniteness explicitly marked. LEM + RTS and LEX + ERTS present two opposite ends of the spectrum with the following order in terms of morphological richness of feature representation: LEM + RTS, followed by LEM + ERTS, followed by LEX + RTS, then LEX + ERTS. Table 7 presents all the results of these experiments.

6.2.1. Results per data set

DARDASHA. This data set is mostly dialectal Arabic, hence we expect that our MSA preprocessing tools for tokenization and POS tagging would perform sub optimally on this data set.

For *subjectivity classification*, using lemmatization (LEM) yields a small improvement over the baseline while lexemes do not provide any gain. Adding the ERTS with its partial morphological information adds another small gain to LEM however no impact on LEX. In the LEM condition, ERTS outperforms RTS with a marginal improvement. We note that none of the conditions yields a performance above 0% for the objective classification. This is due to the fact that the data set is heavily skewed towards the subjective class, which is reflected in the high scores in the baseline condition; most examples are classified as subjective, and only a small part of the data is actually manually classified as objective. Overall for subjectivity classification, in this data set which is highly dialectal and skewed, we observe only minimal gains by preprocessing for morphology or even adding explicit POS tag information as features.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

For *sentiment classification*, we observe a similar trend where LEM outperforms LEX, which in turn outperforms Base, indicating that the least amount of morphological information packed onto the lexical items, the better the performance. LEM is more of an abstraction over LEX which seems to be favored in this condition. By adding POS tag information as features to both conditions, we note an improvement from adding ERTS to LEX yet a deterioration in performance for LEX + RTS. Contrarily, adding POS information to the LEM condition yields worse results than LEM features alone, but still ERTS marginally outperforms RTS. The results suggest that for this data set, POS tag information is adding noise which is due to the fact that the POS tagger is designed for MSA not for DA.

TAGHREED. The TAGHREED data set is partially MSA and partially dialectal Arabic, so we expect AMIRA, our POS tagger, to be able to correctly handle the MSA portion better than the DA portion. However, it should be noted that there are genre variations for the MSA between the AMIRA system (newswire) vs. the social media genre which results in sub-par performance by AMIRA.

For *subjectivity classification*, LEX outperforms LEM in accuracy as well as the *F*-score for objectivity, and they outperform the baseline overall. Thus, LEX seems to prefer classifying sentences as objective (69.62% vs. 63.75%), which leads to an overall higher accuracy while LEM prefers subjectivity, resulting in a higher *F*-score (74.65% vs. 72.95%). We note a relative degradation in performance to both LEX and LEM by adding POS features to the basic LEX and LEM conditions respectively. We also note that in the LEM condition there is no difference between RTS and ERTS, yet in the LEX condition RTS outperforms ERTS.

For *sentiment classification*, LEX yields the best performance across all metrics outperforming both LEM and the baseline. Adding POS features to the LEX conditions results in a deterioration in performance. Again RTS outperforms ERTS in the LEX condition. In the LEM condition we note that adding POS information results in performance gains with ERTS features outperforming the RTS features. Overall the results show that LEX is the best performing condition for this data set.

TAHRIR. The TAHRIR data set is mostly MSA but the genre is very different from MSA newswire as this is wikipedia talk pages and discussion groups. Again, we would expect AMIRA to suffer from the difference in genre.

For *subjectivity classification*, LEM yields the best performance in terms of accuracy and the *F*-score for objectivity. For subjectivity, the baseline yields the highest *F*-score. This suggests that adding morphological information allows the classifier to decide for the objective class more often, thus increasing overall performance, but decreasing the *F*-score for subjectivity. Adding POS information does not help. But it is interesting that adding ERTS to LEX yields higher results than adding RTS features. For LEM, however, RTS gives higher results.

For *sentiment classification*, none of the conditions outperform the baseline in terms of accuracy, but LEM improves over the baseline in terms of the *F*-score for positive sentiment. Adding POS information also has a detrimental effect on all scores.

MONTADA. The MONTADA data set is a combination of MSA and classical Arabic which is in some manner even more of a challenge since the DA data would have lexical items that are shared with MSA even if they are faux amis but the classical Arabic data typically has completely unknown lexical items for the AMIRA tool.

For *subjectivity classification*, we note an improvement over the baseline for accuracy and the *F*-score for subjectivity when we analyze the word forms. However, there is no difference between LEX and LEM, and adding POS information does not change the results in any way. Again, we note that the heavy skewing in the data set leads to an *F*-score of 0% for the objectivity *F*-score.

For *sentiment classification*, the baseline shows a preference for positive sentiment, the majority class. The only improvement over the baseline is in the LEM condition for the *F*-score for negative sentiment, i.e., in the LEM condition, the classifier has a higher tendency to classify sentences as negative.

6.2.2. LEX vs. LEM

Given the results, we observe no clear trends for the impact for morphological preprocessing alone on performance.

Our results suggest that the DARDASHA and MONTADA data sets are easier to classify than TAGHREED and TAHRIR. Note that both DARDASHA and MONTADA are highly skewed, resulting in extremely high baselines for subjectivity. For these latter data sets, neither classifier improves over the baseline. With regard to subjectivity, LEM results in slightly higher accuracy and *F*-score for subjectivity for DARDASHA and in higher accuracy and *F*-score for objectivity for TAHRIR, while LEX performs better for TAGHREED. For MONTADA, LEM and LEX yield the same scores. For sentiment, DARDASHA shows a clear preference for LEM while TAGHREED prefers LEX. The other two data sets do not improve over the baseline in accuracy.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

For subjectivity, differences between LEM and LEX scores are small (<1.5% accuracy). For sentiment and for all data sets, the differences in accuracy between LEX and LEM are higher than their counterpart experimental conditions differences for subjectivity (>2.4% for sentiment compared to <1.5% for subjectivity).

6.2.3. Adding POS tags

For subjectivity, the results show that adding POS information improves accuracy only for one data set, DARDASHA. A comparison between the performance of RTS vs. ERTS across the different data sets is inconclusive.

The sentiment task shows a similar trend: 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, comparing RTS with ERTS for sentiment classification shows that in many cases, ERTS outperforms RTS, thus indicating that the additional morphological features are helpful being explicitly represented in the tagset. One possible explanation may be that variations of some of the morphological features (e.g., existence of a gender, person, adjective feature) correlate more frequently with positive or negative sentiment.

6.3. Standard features for social media data

RQ2 addresses the question whether standard features can be used successfully for classifying social media text which are characterized by the usage of informal language and by varying text length. We add the standard features, polarity lexicon (PL) and UNIQUE (Q), to the two tokenization schemes and the POS tag sets as illustrated in Table 8. For subjectivity classification, we see improvement in accuracy over the baseline and the results in Table 7 for TAGHREED and MONTADA. For TAGHREED, the highest accuracy is reached by the combination of LEM, ERTS, the polarity lexicon, and the unique feature; for MONTADA, the highest accuracy is achieved with LEX, RTS, and the two standard features. However for DARDASHA and TAHRIR there is no improvement over the baseline for either of these data sets for any of the conditions. The performance of all the conditions for DARDASHA are very close with LEX based conditions slightly outperforming LEM based conditions, and in both cases adding RTS is better than ERTS in the presence of standard features. For TAHRIR we observe a similar trend with the LEX condition outperforming the LEM condition. However comparing LEM + RTS + PL + Q3 to LEX+RTS + PL + Q3 we note that the former condition yields better accuracy.

For sentiment classification, we observe an improvement in accuracy over the baseline for DARDASHA and TAGHREED, but for the latter, the LEX setting without standard features gives the highest results. For MONTADA, adding the standard features to LEX improves the *F*-score for negative sentiment, however decreasing the other two metrics. It must be noted that we are employing here a polarity lexicon that is tailored to and extracted from MSA newswire, hence there is room for improvement with a polarity lexicon that has wider coverage.

6.4. 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 TAGHREED 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 TAGHREED data set into two subsets, one containing only MSA, the other one containing only dialectal Arabic. 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 9, and the sentiment results are shown in Table 10. For both tasks, the results show considerable differences between MSA and dialectal Arabic: For TGRD-MSA, the results are higher in terms of subjectivity and lower for sentiment than for TGRD-DIA, which is a direct consequence of the difference in distribution of sentiment between the two subcorpora. TGRD-DIA is mostly subjective while TGRD-MSA is more balanced. With regard to sentiment, TGRD-DIA 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 approximately 13% over the baseline in the LEM condition. For TGRD-DIA, the improvement is modest, <0.5%. We assume that this is partly due to the higher skew towards subjectivity in TGRD-DIA. Moreover, our preprocessing tools yield better performance on MSA data. For

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 8
SSA results with standard features polarity lexicon (PL) and UNIQUE feature (Q3). Numbers in bold indicate improvements over the results in Table 7.

Data	Condition	Subjectivity			Sentiment		
		Acc	*F*-O	*F*-S	Acc	*F*-P	*F*-N
DARDASHA	Baseline	84.30	0.00	91.48	64.74	78.60	0.00
	LEX + PL + Q3	83.42	0.00	90.96	68.55	78.07	44.44
	LEX + RTS + PL + Q3	83.33	0.00	90.91	69.03	78.18	46.67
	LEX + ERTS + PL + Q3	82.98	0.00	90.70	69.03	78.18	46.67
	LEM + PL + Q3	83.25	0.00	90.86	67.86	76.72	48.08
	LEM + RTS + PL + Q3	83.25	0.00	90.86	69.88	78.45	50.00
	LEM + ERTS + PL + Q3	82.91	0.00	90.66	**70.30**	78.41	**52.43**
TAGHREED	Baseline	52.40	0.00	68.76	56.45	0.00	72.16
	LEX + PL + Q3	70.00	64.00	74.29	64.52	48.84	72.84
	LEX + RTS + PL + Q3	70.00	62.50	75.00	62.90	43.90	72.29
	LEX + ERTS + PL + Q3	71.00	63.60	75.90	62.10	41.98	71.86
	LEM + PL + Q3	71.10	65.34	75.21	62.90	42.50	72.62
	LEM + RTS + PL + Q3	71.33	65.32	75.57	62.90	43.90	72.29
	LEM + ERTS + PL + Q3	**73.00**	66.67	**77.31**	62.10	41.98	71.86
TAHRIR	Baseline	59.09	0.00	74.29	75.00	0.00	85.71
	LEX + PL + Q3	58.82	28.41	71.10	62.35	36.00	73.33
	LEX + RTS + PL + Q3	57.24	29.35	69.34	58.82	31.37	70.59
	LEX + ERTS + PL + Q3	58.69	29.21	70.83	61.18	37.74	71.79
	LEM + PL + Q3	57.47	25.99	70.16	63.95	39.22	74.38
	LEM + RTS + PL + Q3	58.36	29.05	70.53	59.30	31.37	71.07
	LEM + ERTS + PL + Q3	56.21	23.86	69.27	61.05	37.38	71.73
MONTADA	Baseline	80.32	0.00	89.08	86.76	92.91	0.00
	LEX + PL + Q3	83.92	**3.85**	91.23	79.45	87.25	**47.06**
	LEX + RTS + PL + Q3	**84.36**	0.00	**91.52**	72.02	81.79	39.60
	LEX + ERTS + PL + Q3	84.09	0.00	91.36	77.06	85.38	46.81
	LEM + PL + Q3	83.60	0.00	91.07	78.08	86.36	44.19
	LEM + RTS + PL + Q3	84.14	0.00	91.39	75.34	84.30	42.55
	LEM + ERTS + PL + Q3	83.87	0.00	91.23	72.60	82.56	36.17

sentiment classification on TGRD-MSA, LEX and LEM equally improve over the baseline. On TGRD-DIA, LEM leads to higher results.

The results for subjectivity on the MSA and dialect 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-DIA data sets. This is not surprising since the TGRD-DIA is not very different from the newswire data on which AMIRA

Table 9
Dialect-specific subjectivity experiments.

Condition	TAGHREED			TGRD-MSA			TGRD-DIA		
	Acc	*F*-O	*F*-S	Acc	*F*-O	*F*-S	Acc	*F*-O	*F*-S
Baseline	52.40	0.00	68.76	62.91	77.24	0.00	66.67	0.00	80.00
LEX	71.38	69.62	72.95	70.20	77.83	54.55	66.67	3.57	79.85
LEX + RTS	67.87	61.42	72.47	72.85	80.38	55.91	66.67	3.57	79.85
LEX + ERTS	67.21	60.63	71.91	74.83	81.73	59.57	66.67	3.57	79.85
LEM	70.16	63.75	74.65	76.16	82.35	63.27	67.28	3.64	80.30
LEM + RTS	68.85	62.15	73.54	74.83	81.73	59.57	66.05	0.00	79.55
LEM + ERTS	68.85	62.15	73.54	75.50	81.77	62.63	66.05	0.00	79.55

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 10
Dialect-specific sentiment experiments.

Condition	TAGHREED			TGRD-MSA			TGRD-DIA		
	Acc	*F*-P	*F*-N	Acc	*F*-P	*F*-N	Acc	*F*-P	*F*-N
Baseline	56.45	0.00	72.16	52.00	68.42	0.00	65.22	0.00	78.95
LEX	65.32	49.41	73.62	54.00	58.18	48.89	66.30	20.51	78.62
LEX + RTS	62.90	43.90	72.29	54.00	58.18	48.89	67.39	21.05	79.45
LEX + ERTS	62.90	42.50	72.62	50.00	52.83	46.81	66.30	20.51	78.62
LEM	62.10	41.98	71.86	54.00	58.18	48.89	69.57	30.00	80.56
LEM + RTS	62.90	46.51	71.60	44.00	48.15	44.00	66.30	16.22	78.91
LEM + ERTS	64.52	46.34	73.49	44.00	46.15	41.67	66.30	20.51	78.62

was trained: Twitter users discuss current event topics typically found in newswire. While MSA tweets are expected to come from news headlines (Abdul-Mageed et al., 2011), and headlines usually are not loci of explicitly subjective content and thereby lack typical sentiment cues and hence are difficult to classify and in essence harder to preprocess since the genre is different from regular newswire even if MSA. There is also a considerable lexical overlap between MSA and dialectal Arabic. Furthermore, dialectal data have 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-DIA.

In a second experiment, we used the original TAGHREED corpus but added the language variety (LV) (i.e., MSA and dialect) features. For both subjectivity and sentiment, the best results are acquired using the LEM + ERTS + LV settings. However, for subjectivity, we observe a drop in accuracy from 73.00% (LEM + ERTS + PL + Q3) to 69.21%. For sentiment, we also observe a performance drop in accuracy, from 65.32% (LEX) to 63.71%. This means that explicitly modeling the language variety as a singleton feature does not provide enough information for successfully conquering the differences between those varieties.

6.5. Leveraging genre specific features

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

Table 11 shows the results of adding the genre features to the subjectivity classifier. For this task, no data sets profit from these features.

Table 12 shows the results of adding the genre features to the sentiment classifier. Here, all the data sets profit from the new features in some way. For TAHRIR and subjectivity, adding document ID (DID), results in a moderate

Table 11
Subjectivity results with genre features.

Data	Condition	Acc	*F*-O	*F*-S
DARDASHA	LEM + ERTS + GEN	84.23	0.00	91.44
TAGHREED	LEX + ERTS + PL + Q3 + GEN	72.37	64.41	77.42
TAHRIR	LEM + RTS + PL + Q3 + DID	**62.99**	**42.42**	72.73
MONTADA	LEM + ERTS + PL + Q3 + GEN	84.36	0.00	91.52

Table 12
Sentiment results with genre features. Numbers in bold show improvement over Table 8.

Data	Condition	Acc	*F*-P	*F*-N
DARDASHA	LEX + RTS + PL + Q3 + GEN	69.70	**78.99**	45.65
TAGHREED	LEX + PL + Q3 + GEN + UID + LV	**66.67**	**55.32**	73.42
TAHRIR	LEX + PL + Q3 + GEN + DID	66.28	36.96	76.98
MONTADA	LEX + PL + Q3 + GEN + DID	81.82	88.89	**50.00**

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Table 13
Subjectivity results for the balanced DARDASHA corpus.

Condition	SUBJ		
	Acc	F-O	F-S
LEX	43.46	28.98	55.41
LEX + RTS	39.6	28.14	47.69
LEX + ERTS	42.01	28.68	55.03
LEX + PL + Q3	52.85	29.72	67.68
LEX + RTS + PL + Q3	49.64	31.95	67.2
LEX + ERTS + PL + Q3	52.13	32.76	66.94
LEM	42.85	27.17	54.24
LEM + RTS	39.53	27.18	47.69
LEM + ERTS	41.01	27.98	54.05
LEM + PL + Q3	49.84	28.43	61.36
LEM + RTS + PL + Q3	46.36	28.63	56.56
LEM + ERTS + PL + Q3	47.82	28.99	58.4

improvement in accuracy and a stronger one (approx. 4 percent points) in the *F*-score for objectivity. For sentiment classification, all data sets but TAHRIR profit in one of the two *F*-scores, the positive one for DARDASHA and TAGHREED, and the negative one for MONTADA. The gender feature was part of all best performing combinations, except for the subjectivity classification on TAHRIR; document ID also helped for those corpora where they were annotated (TAHRIR and MONTADA). The user ID (UID, which is only present in TAGHREED), was useful for sentiment classification.

6.5.1. Balanced data sets

We note the poor performance on Objective classification for both DARDASHA and MONTADA data sets. We carry out a set of experiments by applying cross validation to the data.

DARDASHA. We divide the DARDASHA data into seven training folds. Each fold had a 50% objective and 50% subjective data points. The size of the folds was 558 (1/7th of the original training data set for the DARDASHA corpus). We evaluate each training model against the same test data for DARDASHA used above which comprises 263 data points. We report the results of averaging across the different folds per the various conditions in Table 13.

In general we note that the objective class reaches a positive performance with the balanced data sets. The overall results are much lower than training with the entire data set but the relative trends are similar to our observations with the entire data set. LEX outperforms LEM, morphological POS tagging information does not help performance, however ERTS improves over RTS. This is expected since DARDASHA is mostly dialectal Arabic, hence the AMIRA performance is expected to be better on LEX than on LEM. Adding PL and Q3 significantly improves performance.

MONTADA. We divide the MONTADA data into four training folds. Each fold had a 50% objective and 50% subjective data points. The size of the folds was 501 (1/4th of the original training data set for the MONTADA corpus). We evaluate each training model against the same test data for MONTADA used above which comprises 467 data points. We report the results of averaging across the different folds per the various conditions in Table 14.

In general we note that the objective class reaches a positive performance with the balanced data sets. The overall results are much lower than training with the entire data set but the relative trends are similar to our observations with the entire data set. LEX outperforms LEM but with marginal gains, however adding POS tag information to the LEM improves performance. The improvement of LEX over LEM with slight performance gains especially compared tot he DAR corpus may be attributed to the fact that this data set, although is in MSA, comprises a significant classical Arabic component has a significant negative impact on AMIRA performance. Similar to the DAR corpus as well as overall trends, adding ERTS improves over adding RTS. Adding PL and Q3 significantly improves performance.

7. Overall performance

Table 15 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

Table 14
Subjectivity results for the balanced MONTADA corpus.

Condition	SUBJ		
	Acc	F-O	F-S
LEX	38.97	34.09	47.43
LEX + RTS	38.57	34.35	44.25
LEX + ERTS	38.41	34.09	44.9
LEX + PL + Q3	52.97	36.85	69.71
LEX + RTS + PL + Q3	52.44	35.88	68.78
LEX + ERTS + PL + Q3	52.35	35.59	66.67
LEM	38.65	34.26	43.2
LEM + RTS	40.55	34.8	41.44
LEM + ERTS	42.01	35.11	41.79
LEM + PL + Q3	51.77	36.33	68.93
LEM + RTS + PL + Q3	53.24	36.55	72.04
LEM + ERTS + PL + Q3	51.08	35.59	59.95

Table 15
Overall best SAMAR performance. Numbers in bold show improvement over the baseline.

Data	Condition	SUBJ			Condition	SENTI		
		Acc	*F*-O	*F*-S		Acc	*F-P*	*F-N*
DAR	LEM + ERTS	**84.65**	0.00	91.69	LEM + ERTS + PL + Q3	**70.30**	78.41	52.43
TGRD	LEM + ERTS + PL + Q3	**73.00**	66.67	77.31	LEX + PL + Q3 + GEN + UID + LV	**66.67**	55.32	73.42
THR	LEM + RTS + PL + Q3 + DID	**62.99**	42.42	72.73	LEX + PL + Q3 + GEN + DID	66.28	36.96	76.98
MONT	LEM + ERTS + PL + Q3	**84.36**	0.00	91.52	LEX + PL + Q3 + GEN + DID	81.82	88.89	50.00

improves over the baseline on all the data sets, except MONTADA. The results also show that all optimal feature settings for subjectivity, except with the TAHRIR data set, include the ERTS POS tags while the results in Section 6.2 showed that adding POS information without additional features, while helping in most cases with subjectivity, does not help with sentiment classification.

8. Conclusion

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 characteristics of Arabic can be handled in the context of SSA. Coming back to our research questions, we will review here what we have learned in our experiments: RQ1 was concerned with the morphological richness of Arabic and how this has to be addressed in an SSA system. We saw that adding morphological information either in form of lexemes or lemmas, as well as adding POS tags, in general has a positive effect on SSA. However, there are no overall valid decisions which combination of those types of information is the best for the individual corpora and for the tasks (subjectivity vs. sentiment analysis). For subjectivity analysis, lemma information plus the extended POS tagset generally work best while for sentiment analysis, lexemes are more appropriate.

RQ2 was concerned with standard features given the short texts. Again, the results are inconclusive, showing that for some combinations, the polarity and the unique words have a favorable effect. RQ3 addressed the problem of dialects in Arabic. The results show that the linguistic preprocessing does suffer in the presence of dialectal data. This means that this component must be adapted to handle dialects in order to reach good results for SSA. RQ4 was concerned with leveraging features specific to social media. The results show that document ID is a useful feature if it is available.

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. We are also considering the identification of irony in the texts.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

References

Abbasi, A., Chen, H., 2005. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems 20 (5), 67–75.

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Transactions on Information Systems 26, 1–34.

Abdul-Mageed, M., Diab, M., 2012a. AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In: Proceedings of LREC, Istanbul, Turkey.

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

Abdul-Mageed, M., Albogmi, H., Gerrio, A., Hamed, E., Aldibasi, O., 2011. 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.

Abdul-Mageed, M., Diab, M., Korayem, M., 2011a. Subjectivity and sentiment analysis of modern standard Arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, pp. 587–591.

Abdul-Mageed, M., Korayem, M., YoussefAgha, A., 2011c. "Yes we can?": subjectivity annotation and tagging for the health domain. In: Proceedings of RANLP2011, Hissar, Bulgaria.

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

Bansal, M., Cardie, C., Lee, L., 2008. The power of negative thinking: exploiting label disagreement in the min-cut classification framework. In: Coling 2008: Companion volume: Posters, Manchester, UK, pp. 15–18.

Brown, P., Levinson, S., 1987. Politeness: Some Universals in Language Usage. Cambridge University Press, Cambridge, UK.

Bruce, R., Wiebe, J., 1999. Recognizing subjectivity. A case study of manual tagging. Natural Language Engineering 5 (2), 187–205.

Buckwalter, T., 2002. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania. LDC Catalog No.: LDC2004L02, Tech. Rep., ISBN 1-58563-324-0, 2004.

Dave, K., Lawrence, S., Pennock, D., 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, ACM, Budapest, Hungary, pp. 519–528.

Davidov, D., Tsur, O., Rappoport, A., 2010. Enhanced sentiment learning using twitter hashtags and smileys. In: Coling 2010: Posters, Beijing, China, pp. 241–249.

Diab, M., Jurafsky, D., Hacioglu, K., 2007. Automatic processing of modern standard Arabic text. In: Soudi, A., van den Bosch, A., Neumann, G. (Eds.), Arabic Computational Morphology. Springer, Dordrecht, pp. 159–179.

Diab, M., Habash, N., Rambow, O., Altantawy, M., Benajiba, Y., 2010. COLABA: Arabic dialect annotation and processing. In: LREC Workshop on Semitic Language Processing, Valetta, Malta, pp. 66–74.

Diab, M., 2007a. Towards an optimal POS tag set for modern standard Arabic processing. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.

Diab, M., 2007b. 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, Prague, Czech Republic, pp. 89–96.

Diab, M., 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, Cairo, Egypt, pp. 285–288.

Habash, N., Rambow, O., Roth, R., 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, Cairo, Egypt, pp. 102–109.

Herring, S., 1996. Bringing familiar baggage to the new frontier: gender differences in computer-mediated communication. In: Selzer, J. (Ed.), Conversations. Allyn & Bacon, Boston, pp. 1069–1082.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, pp. 168–177.

Jakob, N., Gurevych, I., 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, pp. 1035–1045.

Joachims, T., 2008. SVMLight: Support Vector Machine. http://svmlight.joachims.org/, Cornell University.

Kanayama, H., Nasukawa, T., Watanabe, H., 2004. Deeper sentiment analysis using machine translation technology. In: Proceedings of Coling 2004, Geneva, Switzerland, pp. 494–500.

Kessler, J., Nicolov, N., 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In: Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, CA.

Kim, S.-M., Hovy, E., 2004. Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, pp. 1367–1373.

Liu, B., 2012. Sentiment Analysis and Opinion Mining Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Lu, B., Tan, C., Cardie, C., Tsou, B.K., 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, pp. 320–330.

Maamouri, M., Bies, A., Buckwalter, T., Mekki, W., 2004. The Penn Arabic Treebank: building a large-scale annotated Arabic corpus. In: NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, pp. 102–109.

Mihalcea, R., Banea, C., Wiebe, J., 2007. Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 976–983.

Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T., 2002. Mining product reputations on the web. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 341–349.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001

Pang, B., Lee, L., 2004. A sentimental education: sentimental analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics 27, pp. 1–278.

Somasundaran, S., Namata, G., Getoor, L., Wiebe, J., 2009. Opinion graphs for polarity and discourse classification. In: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4), Suntec, Singapore, pp. 66–74.

Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J., 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First Workshop on Unsupervised Learning in NLP, Edinburgh, Scotland, pp. 53–63.

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R., van der Goot, E., 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, pp. 770–775.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., Zavarella, V., 2012. Creating sentiment dictionaries via triangulation. Decision Support Systems 53, 689–694.

Thomas, M., Pang, B., Lee, L., 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, pp. 327–335.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., Tounsi, L., 2010. Statistical Parsing of Morphologically Rich Languages (SPMRL) what, how and whither. In: Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, Los Angeles, CA, pp. 1–12.

Turney, P., 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, pp. 417–424.

Vechtomova, O., 2010. Facet-based opinion retrieval from blogs. Information Processing and Management 46 (1), 71–88.

Wan, X., 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, pp. 553–561.

Wiebe, J., Bruce, R., O'Hara, T., 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), University of Maryland, pp. 246–253.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M., 2004. Learning subjective language. Computational Linguistics 30, 227–308.

Wiebe, J., 2000. Learning subjective adjectives from corpora. In: Proc.17th National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, pp. 735–741.

Yao, J., Wu, G., Liu, J., Zheng, Y., 2006. Using bilingual lexicon to judge sentiment orientation of chinese words. In: The Sixth IEEE International Conference on Computer and Information Technology, 2006. CIT'06, p. 38.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W., 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, Melbourne, FL, pp. 427–434.

Yu, H., Hatzivassiloglou, V., 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, pp. 129–136.

Please cite this article in press as: Abdul-Mageed, M., et al., SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.03.001