Modeling Arabic Subjectivity and Sentiment in Lexical Space

Muhammad Abdul-Mageed
Natural Language Processing Lab
The University of British Columbia
muhammad.mageed@ubc.ca

Abstract

In spite of the vast amount of work on subjectivity and sentiment analysis (SSA), it is yet not very clear how lexical information can best be modeled in a morphologically-richness language. To bridge this gap, we report successful models targeting lexical input in Arabic, a language of very complex morphology. Namely, we measure the impact of both gold and automatic segmentation on the task and build effective models achieveing significantly higher than our baselines. In addition, we perform in-depth (error) analyses of the behavior of the models and provide detailed explanations of subjectivity and sentiment expression in Arabic against the morphological richness background in which the work is situated.

1 Introduction

In natural language, the concept of *subjectivity* refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and hence involves *sentiment*. The process of subjectivity classification refers to the task of classifying a unit of analysis (e.g., a sentence) as either *objective* (e.g., *The Prime Minister delivered a speech.*) or *subjective*. Subjective text is further labeled with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether a subjective text is *positive* (e.g., *The new models are ground-breaking!*), *negative* (e.g., *The war in Syria is horrifying!*), *neutral* (e.g., *The new brand might be released next month.*), or, sometimes, *mixed* (e.g., *I really like this camera, but it is very expensive.*).

Compared to English, Arabic *subjectivity and sentiment analysis* (SSA) (e.g., Abbasi, 2007; Abbasi, Chen, & Salem, 2008; Abdul-Mageed, Diab, & Korayem, 2011; Abdul-Mageed, Kübler & Diab, 2012; Abdul-Mageed, Diab, and Kübler, 2013; Abdul-Mageed, Diab, & Kübler, 2014; Al-Kabi, Abdulla, & Al-Ayyoub, 2013; Al-Ayyoub, Essa, Alsmadi, 2015; Al Shboul, Al-Ayyoub, & Jararweh, 2015; Faqeeh, Abdulla, Al-Ayyoub, Jararweh, & Quwaider, 2014), is still in a nascent stage. A language of complex morphology (e.g., Tsarfaty et al., 2010), Arabic raises questions different from those raised by a language of simple morphology like English. The focus of the current paper is one related question: Given the morphological richness of Arabic, how it is that SSA can be modeled within lexical space. Intuitively, a language with rich morphology could suffer from data sparsity in an SSA supervised machine learning setting. In other words, there will be too many word forms a classifier would need to consider that may have not been seen in training data. A simple solution would be to use a preprocessing tool that tokenoizes (or segments) the surface input. To what extent such a process is useful is not entirely clear. This is partly the case since in previous work different pre-processing tools have been used for Arabic SSA, but also because of lack of comparisons on benchmark datasets. As such, Arabic sentiment analysis is no exception to Arabic NLP in general as to the lack of standardization Habash (2010) discusses. To alleviate this problem, in this work we exploit Arabic treebank data and utilize treebank-style segmentation for modeling subjectivity and sentiment within lexical space. Given the popularity of the Penn Arabic treebank (ATB) (Maamouri, Bies, Buckwalter, & Mekki, 2004) and the availability of treebank data for training pre-processing tools like segmenters and POS taggers, it is hoped this endeavor is a step in the right direction.

By simply segmenting running lexical input using standard ATB style and providing it to a linear SVM classifier, we show how it is possible to acquire significant SSA gains. The focus of the current paper is thus exclusively on modeling subjectivity and sentiment within the lexical space. In other words, we investigate the utility of using only features based on input text words, rather than on, e.g., part-of-speech tags. To this end, we raise the research question "How can morphological richness in Arabic be handled in the context of SSA?" Clearly, 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 (i.e., they depend heavily on use of word *N*-grams), hence a direct application of these methods would not be quite as successful for Arabic since a content *segment*[1] (e.g., a segment of a noun) in Arabic may be associated with hundreds if not thousands of variant surface word forms. The utility of using segments rather than full surface word forms for SSA, however, is still unknown and hence a subsidiary research question that can be formulated as follows arises: "What is the effect of *segmentation* on Arabic SSA?"

Breaking down surface forms into their component segments is known as *segmentation*. Segmentation is possible when morphological boundaries within a word are identified. In the ATB, a segment can be stem, an inflectional affix, or a clitic. For example, the surface word *wbHsnAthm* 'and by their virtues' is segmented as *w+b+Hsn+At+hm* with the prefixal clitics (*w* and *b*, Eng. 'and' and 'by'), the stem *Hsn*, the inflection morpheme *At*, and the suffixal pronominal morpheme *hm*. Reducing a word to its component segments is expected to help SSA since this measure reduces the number of observed forms, which means reducing sparsity that results from seeing too many forms in a test set that have not been observed in the train set. Reducing sparsity, when it is possible, helps improve classification. We note that while modeling in lexical space can be done in different ways, the goal of the current paper is to focus on the role of segmentation and upack its utility and effect in the context of SSA.

The contributions of this paper are as follows:

1. We present a new human-labeled ATB dataset for SSA.
2. We measure the impact of tree-bank style segmentation on SSA.
3. We present detailed (error) analyses of the behaviors of the lexical models and present detailed linguistic explanations of subjectivity and sentiment expression in Arabic with morphological complexity in mind.

The rest of this paper is organized as follows: In Section 2, we provide a literature review. In Section 3, we describe our datasets and methods. Section 4 is where we present and discuss the results, and we conclude in Section 5.

2 Literature Review

[1] The term *segment* and the related process of *segmentation* are defined below.

Arabic as a Morphologically-Rich Language: *Arabic* is a term used to refer to a collection of languages, the formal (mostly written) modern variety of which is known as *Modern Standard Arabic (MSA)* (Badawi, 1973). MSA is the language of modern culture in the Arab world and as such, is used in educational settings and many media outlets. *Classical Arabic (CA)* is another variety of Arabic. CA is the language of the Qur'an (the Holy Book of Islam) and is currently a primarily written variety, with some use in spoken contexts (e.g., in religious sermons). There are few structural and syntactic differences between MSA and CA (see, e.g., Bateson, 1967; Ryding, 2005); the main difference between the two varieties is lexical and morphological. Lexical differences can be accounted for by modernization and the introduction of modern technologies. Moreover, dialectal Arabic is having an impact on the syntax of MSA. Rambow et al. (2005) and Chiang, Diab, Habash, Rambow, and Shareef (2006) found that Subject-Verb-Object (typically associated with the dialects) and Verb-Subject-Object constructions (typically observed in CA/MSA) have equal distribution in syntactically annotated newswire corpora.

In addition to MSA and CA, there are multiple dialects of Arabic (e.g., Egyptian, Levantine, Morrocan). Dialects are usually used for everyday informal communication and are not taught in school. For a long time, Arabic dialects have remained mainly spoken, not written. With the proliferation of electronic media, however, Arabic dialects are finding their way into written form in various computer-mediated communication (CMC) genres (e.g., blogs, social network sites, email). Arabic dialects can be classified in various ways. For example, according to Palva (2006), a synchronic approach where salient linguistic features of each dialect or group of dialects are measured and selected can be adopted, or, alternatively, a sociological, anthropological, and historical approach where the division between Bedouin and sedentary dialects is taken into consideration can be the criterion. Classifications of Arabic dialects remain, by far, arbitrary and primarily based on geographical divisions (see also Habash, 2010; Versteegh, 2001). Arabic dialects differ in various ways—phonological, morphological, lexical, and syntactic (see, e.g., Bassiouny, 2009; Holes, 2004; Palva, 1982)—from both CA and MSA. For instance, EGY is characterized by the use of the discontinuous morphemes *ma+… +ʃ*[2] (which is realized in Buckwalter transliteration as *mA…$*) for negation, as in the example below:

(1)	ENGLISH:	*He did not play.*
	MSA TRANSLITERATION:	*lam yalEab*
	EGY TRANSLITERATION:	*malEib$*

In Arabic, the *stem*[3] of a word can combine with various other units to form many word forms. For example, the Arabic stem *jy~d* can have many related word forms, including *jy~d* 'good + masc. sing.,' *jy~dp* 'good + fem. sing.,' *Aljy~d* 'the good

[2] We use the notation of the International Phonetic Alphabet (IPA) here. For more information about IPA, the reader is referred to the IPA website, accessible at http://www.langsci.ucl.ac.uk/ipa/.

[3] A *stem* is the base word segmented without handling the rules governing the combination of morphemes to form words. For example, the stem of the word *Hsnp* 'virtue + fem. sing' is *Hsn* 'virtue-related.'

+ masc. sing.,' *Aljy~dp* 'the good + fem. sing.,' *bAljy~d* 'by the good + masc. sing,' *bAljy~dp* 'by the good + fem. sing,' *wAljy~d* 'and the good + masc. sing.,' *wAljy~dp* 'and the good + fem. sing,' etc. This morphological richness of Arabic interacts in various ways with SSA. For example, an SSA classifier that has access to the information that a certain stem is positive will not be able to associate all the surface word forms related to that stem with positiveness. The solution to this problem is to identify the morphological boundaries within a word through the use of a segmenter as we explain later we do with ASMA (Abdul-Mageed, Diab, & Kübler, 2013). In the rest of this subsection, we give an overview of the morphological richness of Arabic as a motivation for the methods we employ to model SSA in lexical space.

Arabic morphology can be approached both from the perspective of *form* as well as *function* (see e.g., Farghaly & Shaalan, 2009; Habash, 2010).[4] Form-based morphology is focused on the ways units making up a word interact and relate to the word's overall form, whereas function-based morphology is about the functions of the units building up a word and how they affect its overall syntactic and semantic behavior. Habash (2010) divides form-based morphology into *concatenative* and *templatic* and function-based morphology into *derivational, inflectional,* and *cliticizational*. Most relevant to the current work is form-based morphology as well as cliticization functional morphology.

Concatenative morphology is concerned with how *stems, affixes,* and *clitics* form words. Arabic has prefixes (e.g., *y*+ 'third person singular of imperfective verbs'), suffixes (e.g., *+wn* 'nominative definite masculine sound plural'), circumfixes (e.g., *t+…+wn* 'second person masculine plural of imperfective verbs'), and clitics. *Clitics* in Arabic can attach before the stem (i.e., *proclitics*, e.g., the conjunction *f*+ 'then') or after the stem (i.e., *enclitics*, e.g., the third person female singular possessive pronoun *+ha* 'hers/its'). In Arabic, multiple stems, affixes, and clitics can occur in a word.

As Habash (2010, p. 45) points out, there are three types of *templatic morphemes* in Arabic, all of which are equally needed in creating a word templatic stem: *root, pattern,* and *vocalism*. The root is a sequence of 2, 3, 4, or 5 consonants (called *radicals*). Words derived from the same root usually have some meaning overlap. For example, the words *zrE* 'he planted,' *zArE* 'farmer,' and *mzrwE* 'is planted' are all derived from the root morpheme *z-r-E* 'planting-related.' A pattern (also known as *wazn* 'measure') is an abstract template in which roots and vocalisms are inserted. In the Arabic grammatical tradition, a basic citation form composed of the three radicals *f*, *E*, and *l* is used to refer to the radicals of a word root. These three radicals thus form the basic root *fEl* 'doing-related' with which vocalisms are combined to form patterns. The vocalism morpheme specifies the short vowels to be used with a pattern. For example, a word with a trilateral root like *Darab* 'hit+ perfective tense'[5] is composed of the root *D-r-b*, which has the same *wazn* (i.e., *form* or *measure*) as the citation form *f-E-l*, with the addition of the vowel *a* both after the first and the second radical. The word *Darab* is thus described as having the pattern *faEal* and a syllable structure *CVCVC* (with each C standing for one of the radicals and each V standing for one of the vowels). In Western treatments of Arabic grammar, the patterns are denoted via use of Roman numerals, and

[4] Habash (2010) indicates that this classification is influenced by Smrž's (2007).

[5] Perfective aspect indicates a completed action.

so the pattern *faEal* is usually referred to using the Roman numeral 'I,' since it has the minimal length required for a pattern.

As its name indicates, *cliticization morphology* is built around the concept of a clitic. As stated earlier, a *clitic* is a morpheme that behaves syntactically like a word, but depends phonologically on a neighboring word. An example of a clitic in Arabic is the definite article *Al-* 'the.' Unlike inflectional features, clitics are optional. While a word like *yaktubhu* 'write+2nd masc. sing.+it' must have a number of inflectional features (e.g., the gender and person), it does not have to have the *enclitic* (i.e., the clitic occurring after the based word) *-hu* 'he/it.' Clitics in Arabic are of two types, *proclitics* (i.e., clitics that attach before a base word) and *enclitics* (i.e., clitics that attach after a base word). For example, to repeat an example mentioned earlier, the word *wbhsnAtHm* 'and by their virtues' has two proclitics and an enclitic and is composed of the morphemes in Table 1:

Table 1. Clitics and morphemes of an example Arabic word

	Proclitic	**Proclitic**	**Stem**	**Affix**	**Enclitic**
TRANSLITERATION:	*w*	*b*	*Hsn*	*At*	*hm*
GLOSS	and	by	virtues	s	their
TRANSLATION	And by their virtues				

Performing Arabic SSA on unprocessed text (i.e., using a *bag-of-words* approach) will be problematic, because the system will have to classify texts with many word forms that it has not observed in the training data, even though the lexeme around which these words are made may have been observed in the training data. For example, word forms like *Aljmyl* 'the beautiful,' *bAljmyl* 'by the beautiful,' *wAljmyl* 'and the beautiful,' *wlljmyl* 'and to the beautiful,' *jmylp* 'beautiful+fem.,' *jmylAn* 'beautiful+masc.+dual,' *jmylAt* 'beautiful+fem.+pl.' are all built around the lexeme *jmyl* 'beautiful.' Habash, Rambow, and Roth (2009) maintain that a word from the Penn Arabic Treebank (PATB; Maamouri et al., 2004) has about 12 morphological analyses. Given the highly inflected nature of an Arabic word (i.e., the high number of word forms built around the same lexeme), performing SSA on unprocessed Arabic texts will result in *data scarcity* (i.e., a need to classify texts with many previously unobserved word forms).

As stated earlier, solution to this problem of high inflection is to use NLP tools to split a word into its component morphemes. The identification of morphological boundaries can be at various levels and can involve one of the following four processes:

- **Stemming:** Stemming is the process of splitting off clitics of a word without handling morphotactics, interactions on the boundaries of the morphemes. For example, the word *wbHsnthm* is stemmed as *w+b+HsnAt+hm* 'and by their virtue,' where the proclitics *w* 'and' and *b* 'with,' the enclitic *hm* 'their' are split off the stem *HsnAt*. However, it should be noted that the word *HsnAt* 'virtues' is in isolation a plural noun and that the actual underlying form of this word in the above context should be *Hsnp,* indicating the nominal reading in MSA, given that it is preceded by the proclitic preposition *b.*
- **Segmentation:** While in stemming clitics are split off a word, segmentation is the process of splitting off clitics and affixes of a word. For example, the word *wbHsnthm* is segmented as *w+b+Hsn+At+hm* 'and by their virtue,' where the

proclitics *w* ‘and’ and *b* ‘with,’ the enclitic *hm* ‘their,’ and the inflectional suffix *At* ‘plural’ are split off the token *HsnAt*. Segmentation is a generic term. Different NLP applications can require more or less segmentation.

- **Tokenization:** The term tokenization is used interchangeably with segmentation. However, in general, the result of tokenization should be morphemes that would exist in a dictionary for open class words and clear renderings of clitics. For example, the word *mktbth* ‘his library’ is segmented into *mktb+t+h* (Eng. ‘library’ + ‘fem. sing.’ + ‘his’). This is different from stemming in that the final *taa’ maftouha t* (which results from morphotactics) in the stem *mktbt* is split off from the possessive pronoun *h*. However, *mktbt* is not a word that one would find in an MSA dictionary. Its underlying form is *mktbp,* but due to morphotactics, the stem is transformed into *mktbt* (i.e., the *p* is transformed into a *t*).
- **Lemmatization:** Lemmatization is the process of splitting off clitics and affixes of a word while handling morphotactics. In other words, due to the fact that Arabic orthographic rules cause some parts of words to be deleted or modified after tokenization or stemming takes place, a *lemmatizer* (Diab et al., 2007; Habash et al., 2009) is used to restore these parts and hence acquire a word’s lemma. To illustrate, a final *taa’ marbuta p* is lost from the stem *Hsn* after performing tokenization on the word *wbHsnthm* above. A lemmatizer would restore the *taa marbuta p* to the stem *Hsn* to form the lemma *Hsnp*+fem. sing. ‘virtue.’

Example 3 below illustrates the four different automatic processing tasks of stemming, segmentation, tokenization, and lemmatization on the word *HsnAthm* ‘their virtues.’

(2)

Process	**Resulting units**	**Gloss**
Stemming	*HsnAt+hm*	‘virtues+their’
Segmentation/Tokenization	*Hsn+At+hm*	‘virtue+plural+their’
Lemmatization	*Hsnp+At+hm*	‘virtue+plural+their’

Accordingly, the question arises as to which NLP tools should be used to segment Arabic words as enabling technologies within the context of SSA systems. If a *stem* would be needed, then a *segmenter* would be sufficient as an enabling technology. If a lemma is the unit that will be used, then a lemmatizer will be needed. Since lemmatization typically occurs post tokenization (e.g., Diab et al., 2007), performing lemmatization would assume the existence of a tokenizer. Which is the better unit for SSA, a stem or a lemma, is a question that has been addressed to a certain extent in the literature. For example, in Abdul-Mageed et al. (2011), we report classification gains on both subjectivity and sentiment classification using human-labeled stems based on data from the PATB. In Abdul-Mageed et al. (2011), we find that stems (as acquired from treebank, gold-labeled data) outperform both surface word forms and lemmas on the two tasks. However, we did not report results on machine-predicted data in Abdul-Mageed et al. (2011). It remains an open question how machine-predicted stems would perform as compared to surface word forms and machine-predicted lemmas, for example. It also remains an empirical question whether keeping only the lemma or stem and removing the remaining morphemes of a word would hurt SSA. To illustrate, the English word ‘impolite’ is an adjective that is composed of two morphemes, the prefix *im-* and the morpheme *polite*. Obviously, the word ‘impolite’ has a negative prior polarity, and stemming it to keep only the stem ‘polite’ could hurt classification, as the prior polarity of the word shifts as a result of the stemming process. This would be the case especially if the problem is not modeled with a sequence labeler that takes context into account.

In addition, the highly inflected nature of Arabic results in the existence of a number of possible *N*-grams that can be mapped to a canonical *N*-gram. For example, in a product review domain focused on digital cameras, the positive English bigram ‘small screen’ can be mapped to the bigrams *$A$p Sgyrp* (lemma form) ‘small screen’ and possible surface forms: *$A$p Sgyrp* ‘small screen + fem.,’ *bAl$A$p AlSgyrp* ‘with the small screen + fem.,’ *kAl$A$p AlSgyrp* ‘like the small screen + fem.,’ *ll$A$p AlSgyrp* ‘to the small screen + fem.,’ *wAl$A$p AlSgyrp* ‘and the small screen + fem.,’ among others. Again, a classifier that has access to the information that the bigram ‘small screen’ is positive will not be able to map the various related *N*-grams without using an enabling technology that identifies the boundaries of the adjective *Sgyr* ‘small’ and the noun *$A$p* ‘screen.’ It will be important to have a mechanism that can relate the surface forms to the basic lemma underlying form. Otherwise, the morphological inflections render the space quite sparse.

The number of *N*-grams that can be mapped to canonical *N*-grams is also increased due to the relatively free word order of Arabic.[6] For example, the English bigram ‘student studies’ can be mapped to two bigrams in Arabic in example (4) below, due to the freer word order of Arabic. The example shows how the bigram ‘student studies’ can be mapped to *y*Akr AlTAlb* ‘studies + the student’ and *AlTAlb y*Akr* ‘the student studies.’[7]

(3)	(a)	**Word:**	*y*Akr*	*AlTAlb*	*Al*ky*	*fy*	*Almktbp*
		Gloss:	studies	the student	the intelligent	in	the library
		Translation:	‘The intelligent student studies in the library.’				
	(b)	**Word:**	*AlTAlb*	*Al*ky*	*y*Akr*	*fy*	*Almktbp*
		Gloss:	the student	the intelligent	studies	in	the library
		Translation:	‘The intelligent student studies in the library.’				
	(c)	**Word:**	*fy*	*Almktbp*	*y*Akr*	*AlTAlb*	*Al*ky*
		Gloss:	in	the library	studies	the student	the intelligent
		Translation:	‘In the library, the intelligent student studies.’				

[6] A more elaborate discussion of Arabic word order is beyond the scope of the current paper.

[7] Whereas in version (a) of example 6 the verb *y*Akr* ‘studies’ occurs first, in version (b) it is the subject *AlTAlb* ‘the student’ that occurs first. Version (c) is begun by the prepositional phrase *fy Almktbp* ‘in the library.’ Thus, while Arabic has both VSO (where the verb occurs first) and SVO (where the subject occurs first), English is an SVO language.

To account for the highly inflected nature of Arabic, along with its relatively free word order, NLP tools are needed. These tools can run the gamut from *stemming* to *clitic tokenization* to *parsing*, depending on the specific goals of the system. For example, a system can be built with tokenization being the only enabling technology, in which case POS tagging and parsing will not be necessary. Various tools have been developed for processing MSA. These include tools for (1) stemming only (Darwish, 2002; Lee, Papineni, Roukos, Emam, & Hassan, 2003), (2) tokenization, lemmatization, and POS tagging (Habash & Rambow, 2005), (3) segmentation and morphosyntactic disambiguation (Abdul-Mageed et al., 2013), (4) tokenization, lemmatization, and POS tagging, and base phrase chunking (e.g., Diab et al., 2004, 2007), and (5) parsing (e.g., Marton et al., 2010). While some of these tools can be used as enabling technologies in the proposed SSA system, we choose to use our system ASMA (Abdul-Mageed et al., 2013b) for segmentation. In addition to its state of the art performance on both segmentation and morphosyntactic disambiguation, ASMA is faster than many of the other available tools.

Subjectivity and Sentiment Analysis: There is a vast literature on especially sentiment detection, with researchers building systems attacking the problem at various levels of analysis. As such, SSA has been performed on the *term* (e.g., Esuli & Sebastiani, 2007; Kim & Hovy, 2007)*, phrase* (e.g., Wilson, Wiebe, & Hoffmann, 2005), *sentence levels* (Grefenstette, Qu, Evans, & Shanahan, 2006; Ikeda, Takamura, Ratinov, & Okumura, 2008; Yu & Hatzivassiloglou, 2003), and *document/product-review level* (e.g., Dalal and Zaveri, 2014; Dave et al., 2003; Hu & Liu, 2004; Kim & Hovy, 2004; Morinaga, Yamanishi, Tateishi, & Fukushima, 2002; Mullen & Collier, 2004; Pang et al., 2002; Terveen, Hill, Amento, McDonald, & Creter, 1997; Tsur et al., 2010)

Particularly relevant to the current paper is research targeting use of lexical information in SSA systems. The practice known as the *bag-of-words* approach, where only the words in a data point are used as features, is usually followed in building SSA systems. In this method, no further linguistic information (e.g., how words are arranged to form larger units like sentences) is provided to a classifier. The bag-of-words approach is usually not used alone in an SSA system, but as a baseline that researchers try to beat using richer feature sets. The way a *bag-of-words* approach is adopted depends on how text is represented. In the literature (e.g., Pang, Lee & Vaithyanathan, 2002), two approaches to representing text have been followed. The first approach is similar to the long-standing IR (Manning et al., 2008; Manning & Schütze, 1999) tradition of representing terms based on their *frequencies* and normalizing by document length. In IR this practice is referred to as TF*IDF (term frequency * inverse document frequency) and is calculated as follows:

$$TF * IDF_{ki} = f_{ki} \log(\frac{N_d}{d_k})$$

Where f_{ki} is the frequency of term *k* in document *I,* N_d is the number of documents in collection, and d_k is the number of documents in which term *k* occurs at least once. TF*IDF is thus a measure used to evaluate how important a term is to a document in a collection. This importance increases as the frequency of the term increases in a document, but it is still offset by the frequency of the term in the collection. However, Pang, Lee, and Vaithyanathan (2002) report better performance using another approach based on term *presence* (i.e., a term that occurs one or more times is given a value of 1, otherwise its value will be 0). The finding that in SSA term presence performs better than term frequency shows that the SSA task is different from topic-based text categorization.

In the SSA literature, which primarily concerns English, there has not been much focus on the effect of processing input text on the SSA task. Rather, many researchers either use input surface word forms (i.e., lexemes) (e.g., Pang et al., 2002; Yu & Hatzivassiloglou, 2003) or stemmed or tokenized text as input to a classifier (e.g., Dave et al., 2003; Hu & Liu, 2004). In many cases, however, the results acquired using stemmed or tokenized text are the baseline and are not compared to surface word forms. Only a few researchers, e.g., Dave et al. (2003), report results comparing processed and unprocessed input text. Dave et al. (2003) report gains using stems compared to surface word forms.

While processing input text may not be a very important step for SSA work on a morphologically simple language like English, it is important to consider text processing (e.g., tokenization, segmentation, stemming) for richer languages like Arabic and Hebrew. In our work on Arabic (Abdul-Mageed, Diab, & Korayem, 2011) we find human-marked segments to be useful classifier input as compared to surface word forms. However, we only exploit gold-processed data from the ATB (Maamouri et al., 2004), and hence it is not clear how a system will perform on machine-processed text. Similar to Abdul-Mageed, Diab, and Korayem's (2011) work on data where morphological boundaries are marked, Jang and Shin (2010) report building a system on Korean language texts where a morphological analyzer is used to identify morpheme boundaries. Jang and Shin (2010) make use of the dependence relations between morpheme sequences to detect the scope of opinionated terms and report classification gains with segmented text input.

Unique and Low-Frequency Words: *Hapax legomena* (words that occur only once in a corpus), as well as other low-frequency words (i.e., words with a frequency of 5 or less), have been shown to be useful for SSA (e.g., Abdul-Mageed, Diab, & Korayem, 2011; Wiebe et al., 2004). This finding can be accounted for by the observation that people can be creative (and hence use rare words) when expressing opinions. This is especially evident in some forms of CMC where users employ non-standard typography. For example, a dialectal Arabic word with repeated letters like *mrrrrrrrh* 'very' was observed to occur with low frequency in subjective texts in a recent study that focused on how Arabic is used on the microblogging site Twitter (Abdul-Mageed & Albogmi, 2011). Using corpus statistics, Wiebe et al. (2004) found that the use of a feature indicating the presence of unique and other low-frequency words results in increased precision, and that the difference between the proportion of unique words in subjective and objective documents is statistically significant ($p<0.001$). These findings suggest that using unique and low-frequency words is useful for document-level SSA.

Prior Polarity Lexica: A *lexical field* is the set of lexical items that covers a specific concept (Lyons, 1977; 1995). For example, the field of anger terms may contain units such as 'rage,' 'fume,' 'seethe,' 'boil over,' and 'look daggers.' Expressions are also said to have a *semantic orientation* (SO) (also called *prior polarity*). The SO of an expression indicates the direction the expression deviates from the norm for its lexical field: It is an evaluative characteristic of the meaning of the word that restricts its usage to an appropriate usage (Hatzivassiloglou & McKeown, 1997; Hatzivassiloglou & Wiebe, 2000). For example, in a field describing a lodging place, whereas the words ‘full’ and ‘empty’ may have a neutral SO, the words ‘fascinating’ and ‘admirable’ have a positive SO, and the words ‘shocking’ and ‘detestable’ have a negative SO.

The SO of a word or a phrase refers to its association with subjective language (e.g., positive or negative language) even outside context. For example, whereas the word ‘good’ has a positive SO, the word ‘bad’ has a negative SO. This use of lists of words and/or phrases for SSA is usually referred to as a *lexicon-based* approach, although the more general term *knowledge-based* approach is sometimes also used to refer to the same method (Liu, 2012). Researchers compile word lists and either manually or automatically label them with SO to acquire a *polarity lexicon*.

Two main approaches are usually followed when using a polarity lexicon. In the first, simple matching is used to check the presence or lack thereof of an entry in the lexicon in a given data point. For example, Hatzivassiloglou and Wiebe (2000) and Wiebe (2000) classify a sentence as subjective if it contains an occurrence of a polarized adjective. Another approach to using an opinion lexicon is to calculate an opinion score, based on the frequency of occurrence of entries from a polarity lexicon in a data point (e.g., Kim & Hovy, 2005; Taboada et al., 2011).

In developing polarity lexica, researchers label each entry (e.g., a word) with a polarity value (i.e., positive vs. negative). For example, the lexicon extracted by Hatzivassiloglou and Wiebe (2000) includes positive adjectives like ‘able,’ ‘above-average,’ ‘abundant,’ ‘acceptable,’ ‘accessible,’ and ‘accommodative,’ and negative adjectives like ‘abandoned,’ ‘abnormal,’ ‘abrupt,’ ‘absurd,’ ‘abusive,’ and ‘abysmal.’ In rule-based approaches, one way to identify the subjectivity or sentiment of a unit of analysis is to count the frequencies of the entries in the polarity lexicon in that unit. In machine learning approaches, one way to make use of a polarity lexicon is to apply a binary feature as to the existence or lack thereof of a word or phrase from the polarity lexicon. Lacking a polarity lexicon, a general-purpose lexicon with, e.g., synonyms and antonyms of entries can be used to expand an initial seed set that can later be used for enhancing classification. A significant amout of work in the literature has focused on learning polarity lexica. For example, a number of studies (e.g., Dave et al., 2003; Hu & Liu, 2004; Kamps & Marx, 2002; Kim & Hovy, 2004; Mullen & Collier, 2004; Yu & Hatzivassiloglou, 2003) have focused on exploiting the popular lexical resource WordNet (Miller, 1995) for expanding an initial polarized seed set. SentiWordNet (SWN) is a lexical resource developed by Esuli and Sebastiani (2005, 2006) and later by Baccianella, Esuli, and Sebastiani (2010) for SSA. SWN entries have also been exploited in a machine translation context to build polarity lexica for languages other than English (e.g., Abdul-Mageed & Diab, 2012a; 2014; Denecke, 2008; Perez-Rosas, Banea, & Mihalcea, 2012).

3 Datasets and Methods

For this work, we human-label a dataset from the first three parts of the ATB. More specifically, the data comprise the first 70 documents from ATB1V4.1, the first 50 documents from ATB2V3.1, and the first 58 documents from ATB3V3.2. The data belong to the newswire genre and were manually labeled by Linguistic Data Consortium (LDC)[8] for part-of-speech (POS), morphology, gloss, and syntactic treebank annotation. As is shown in Table 2, the three parts of the data were released by LDC at various times and comprise different sizes.[9]

Table 2. Release time, stories, and token statistics of LDC-DATA

Data set	Release year	Source	# Stories	# Token before cliticization	# Token after cliticization
ATB1V4.1	2010	Agence France Presse	734	154,386	167,280
ATB2V3.1	2011	Ummah Press	501	144,199	169,319
ATB3V3.2	2010	AnNahar News Agency	599	339,722	401,122

A single annotator, with a Ph.D. in linguistics and a native Arabic fluency, labeled the data after being provided written guidelines by the author and after several sessions of discussions and double-labeling of about 5% of the data (n=250 sentences) with inter-annotator agreement reaching > 95% after adjudication. Guidelines used for this study are similar to (Abdul-Mageed & Diab, 2011a; 2011b; 2012b), and will be published in an independent work as part of a wider array of datasets that will be released to the community.

Table 3. Class distribution of the different LDC-DATA genres

Data set	OBJ	S-POS	S-NEG	S-MIXED	# Instances
ATB1V4.1	582 (58.67%)	183 (18.45%)	188 (18.95%)	39 (3.93%)	992
ATB2V3.1	623 (62.05%)	151 (15.04%)	227 (22.61%)	3 (0.30%)	1,004
ATB3V3.2	1,472 (62.54%)	462 (19.63%)	414 (17.59%)	6 (0.25%)	2,354
ALL	2,677 (61.54%)	796 (18.30%)	829 (12.37%)	48 (1.10%)	4,350

Table 3 shows the statistics pertaining to the distribution of SSA tags assigned by the annotators to the different treebank parts. As Table 3 shows, all three Treebank parts have a more or less similar split between the OBJ and SUBJ classes: The distribution of OBJ cases ranges from 58.67% in ATB1V4.1 to 62.54% in ATB3V3.2. Since in the experiments exploiting data, the three parts will be concatenated, the last row in Table 3 provides the class distribution for the three parts combined. Across the three parts of the

[8] More information about the LDC is available at: http://www.ldc.upenn.edu.

[9] More information about each of the three parts comprising LDC-DATA is available online via the LDC site. More information on ATB1V4.1 is available at: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T13, and more on ATB3V3.2 is provided at: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T09, and further information about ATB3V2 is accessible at: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T08.

treebank (ALL), the OBJ class comprises 61.54% of the data. As such, there is a bias toward the OBJ class in the newswire genre. In addition, in ALL, the S-POS (%=18.30) and S-NEG (%=12.37) classes are not very unbalanced.

Procedure:

For each treebank part of the data, the part is divided into 80% training, 10% for development, and 10% for testing. The training parts from each Treebank are added up to build TRAIN, the development parts are added up to build DEV, and the test parts are combined to build TEST. For all the experiments related to each research question, results are reported both on DEV and TEST. Importantly, only the DEV set is used for tuning the classifier performance based on results acquired on DEV and related error analyses on predictions on DEV. The TEST set is used as a fully blind set and is never investigated for improving performance, as is the standard practice. The baseline we compare against is the majority class in the TRAIN with the surface form input text.

We follow a two-stage classification procedure: Stage one is *subjectivity classification* where objective (OBJ) and subjective (SUBJ) cases are teased apart, and stage two is *sentiment classification* where the subjective-positive (S-POS) and subjective-negative (S-NEG) cases are distinguished. Input to subjectivity classification is gold-labeled sentences with the respective text processing settings (e.g., surface word forms, segmented word), as is explained below. For sentiment classification, gold-labeled annotations are also used. As such, sentences with the gold tag S-POS are used as the positive class, and those gold-tagged as S-NEG are used as the negative class.

For each classification stage, results are reported in terms of overall accuracy (Acc) and average F_1-score. Average F_1-score is the average of the two F_1-scores of the two classes involved in each classification stage. In addition, within each classification stage, precision, recall, and F_1-score are reported for the individual classes. More focus on describing the results is laid on accuracy, and average F_1-score is only provided for completeness sake. Precision, recall, and F_1-score of the individual classes are discussed only in cases where there is an observable difference between the performances of the two classes with regard to these metrics. Each subsidiary research question is answered against two settings pertaining to pre-processing of input text: *Gold* and *machine-predicted*. For the gold setting, human-annotated segmentation and morphosyntactic disambiguation from as labeled by LDC are exploited. For the machine-predicted setting, ASMA (Abdul-Mageed, Diab, & Kübler, 2013b) segmentation and morphosyntactic tagging is used.

Given that SVMs have been used successfully for SSA (e.g., Mullen & Collier, 2004; Pang et al., 2002; Pang & Lee, 2004; Wiebe et al., 2004; Wilson et al., 2009), SVMs were chosen for the experimental work of this study. With all experiments, the default SVMLight[10] linear kernel is used, as we found it to perform best on DEV. The linear kernel usually performs best on text classification (e.g., Abdul-Mageed et. al, 2011; Ng et. al, 2006).

[10] SVMLight is an implementation of SVMs in the programming language C by Thorsten Joachims. More information about SVMLight is available at: http://svmlight.joachims.org.

4 Results and Discussion

To answer the research question “What is the effect of segmentation on Arabic SSA?” we now present the results of each of the two stages of classification. Within each stage, both GOLD and ASMA-predicted results are spelled out and discussed. For both subjectivity and sentiment classification, results of gold segmented (GOLD-SEGS) and ASMA-segmented (ASMA-SEGS) experiments are presented and discussed in the following subsections.

4.1 Subjectivity Classification

Table 4 below provides results for both GOLD-SEGS and ASMA-SEGS experiments with subjectivity classification for both the DEV and TEST data sets. Results acquired on GOLD are discussed first, followed by those on ASMA-processed data.

Table 4. Subjectivity classification results with GOLD-SEGS and ASMA-SEGS

				OBJ			**SUBJ**		
		Acc	**Avg-F**	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
DEV	**BASE**	62.07	57.40	78.75	29.86	43.30	58.31	92.41	71.50
	GOLD-SEGS	**68.28**	**65.31**	**87.63**	**40.28**	**55.19**	**62.72**	**94.64**	**75.44**
	ASMA-SEGS	65.98	63.02	81.19	38.86	52.56	61.38	91.52	73.48
TEST	**BASE**	60.58	60.55	90.91	44.22	59.50	46.41	91.61	61.61
	GOLD-SEGS	65.03	65.00	91.02	51.70	65.94	49.65	90.32	64.07
	ASMA-SEGS	**66.59**	**66.57**	**93.37**	**52.72**	**67.39**	**50.88**	**92.90**	**65.75**

4.1.1 GOLD

As Table 4 shows, on both the DEV and TEST sets, GOLD-SEGS improves subjectivity classification over the surface word forms input baseline (BASE). The improvement on the DEV set is 6.21% accuracy and 7.91% average F_1 and on TEST is 4.45% accuracy and 4.45% average F_1-measure. Likewise, for both the OBJ class and the SUBJ class, using GOLD-SEGS improves classification over the baseline across all the evaluation metrics on both DEV and TEST. On TEST data, GOLD-SEGS improves over BASE with 6.44% F_1 for OBJ class classification and 2.46% F_1 for SUBJ class classification.

The subjectivity classification improvement using GOLD-SEGS is quite predictable, as the gold segmentation is as accurate as it can get based on human effort. Gold segmentation helps reduce the data sparsity that using the surface word forms (SURF) can cause, hence the classification improvement. Table 5 below illustrates the scale of data sparsity in the data considered for the current research question for both GOLD-SEGS and ASMA-SEGS settings.[11] Data sparsity is expressed in terms of out of

[11] The term ‘type’ in Table 5 refers to the unique occurrence of each space delimited string in each setting. A type is usually contrasted with a ‘token,’ a term that refers to the different variants of a type. In BASE, a type is a surface word form, whereas in GOLD-SEGS and ASMA-SEGS, it refers to a segment as split from a surface word by humans and ASMA, respectively. In addition, the percentage (%) of out of vocabulary (OOV) units in a dataset *D* from the set *{DEV, TEST}* is calculated as the proportion of types in *D* that are not seen in TRAIN.

vocabulary (OOV) types. OOV is simply the number of word or segment types seen in TEST that do not exist in TRAIN.

Table 5. Token statistics and OOV for GOLD-SEGS and ASMA-SEGS data splits

	TRAIN	**DEV**		**TEST**	
	# of types	**# of types**	**% of OOV**	**# of types**	**% of OOV**
BASE	13,201	3,028	44.25%	3,268	46.05%
GOLD-SEGS	6,254	2,006	22.88%	2,307	32.12%
ASMA-SEGS	7,053	2,159	26.40%	2,425	31.79%

It is clear from Table 5 that segmentation, whether GOLD or ASMA-predicted, helps reduce sparsity significantly on both DEV and TEST. For the case of GOLD-SEGS, OOV is reduced from 44.25% to 22.88% on DEV and from 46.05% to 32.12% on TEST. The higher sparsity percentage in the case of TEST as compared to the percentage for DEV also accounts for the better results for DEV (as compared to TEST). Since classification here is based on lexical features exclusively, more data sparsity straightforwardly means a harder classification problem. TEST is thus a harder data set than DEV.

To illustrate how segmentation reduces data sparsity and hence improves classification, consider the surface form <*yjAbyp* (Eng. 'positivity'). The words <*yjAbyyn* (Eng. 'positive+masculine plural genitive')*,* <*yjAbywn* (Eng. 'positive+masculine plural nominative'), and <*yjAbythm* (Eng. 'their positivity') all share the same segment <*yjAby* (Eng. 'positive') and should be segmented as follows:

- <yjAby/ADJ +p/NSUFF_FEM_SG
- <yjAby/ADJ +yn/NSUFF_MASC_PL_GEN
- <yjAby/ADJ +wn/NSUFF_MASC_PL_NOM
- <yjAby/ADJ +hm/POSS_PRON_3MP

The classifier is not able to associate these words until they are segmented, otherwise each of them will be considered independently as a unique word. Thus, with the surface word forms setting, if any of the words <*yjAbyyn,* <*yjAbywn,* or <*yjAbythm* occurs in the test data but not in the training data, it will not be possible for the classifier to identify that word and associate it with the occurrence of the word <*yjAbyp* in the training data without segmentation.

To show how data sparsity is a significant factor in classification, an error analysis of the first 10% (more precisely, 10.34%, n= 45 out of 435) of predicted DEV set instances was performed. It was found that 26.66% of this subset of instances (n=12) includes words carrying prior polarity, whereas the remaining 74.44% of the instances (n=33) do not include words with prior polarity. Since words with prior polarity constitute more important features for the subjectivity classifier, the existence of these words in a given sentence affects the classification task. Examples of the words carrying prior polarity in the sample are *fAzt* (Eng. 'she won'), which carries positive prior polarity, and *dmrwA* (Eng. 'they destroyed'), which carries negative polarity. It is thus found that whereas the GOLD-SEGS TRAIN set does include segments preserving the prior polarity of all these 12 cases, the BASE TRAIN set includes only four of word forms that carry the same polarity as these 12 cases, which is a 66.66% drop.

4.1.2 ASMA

Table 4 also shows that ASMA-SEGS improves subjectivity classification, compared to BASE, on both DEV and TEST over BASE. The improvement on the DEV set is 3.91% accuracy and 5.62% average $F_{1,}$ and on TEST, the improvement is 6.01% accuracy and 6.02% average F_1 over BASE. Similarly, for both the OBJ class and the SUBJ class, ASMA-SEGS improves classification over BASE across all the evaluation metrics (with the exception only of recall (Rec) on the DEV set of the SUBJ class) on both DEV and TEST. On TEST data, ASMA-SEGS improves 7.89% F_1 for OBJ class classification and 4.14% F_1 for SUBJ class classification over BASE.

Similar to the case with GOLD-SEGS, using ASMA-SEGS results in a significant data sparsity percentage drop (from 44.25% OOV to 26.40% OOV on DEV and from 46.05% OOV to 31.79% OOV on TEST). This easing of the data sparsity problem accounts for the classification gain with ASMA-SEGS compared to BASE, which is predictable.

Comparing ASMA-SEGS to GOLD-SEGS shows that whereas ASMA-SEGS is outperformed by GOLD-SEGS on the DEV set (with 2.30% accuracy and 2.29% average F_1-measure), it achieves better results on TEST (where its performance is better with 1.56% accuracy and 1.57% average F_1-measure). This finding that ASMA-SEGS performs better than GOLD-SEGS is surprising, since gold segmentation is of higher quality than machine segmentation (as in the case of ASMA-SEGS), and this should translate into better subjectivity classification. This expectation does not seem to hold true with regard to the current situation. The case is that, compared to human segmentation as with GOLD-SEGS, ASMA processing with the TEST set results in further reducing data sparsity (see Table 5), which causes the slight classification improvement. The small margin of error characteristic of ASMA causes word segmentation at points where no segmentation is due in ways that do not seem to change segment polarity significantly enough to hurt classification compared to GOLD-SEGS. ASMA segmentation increases the number of overall types in a given data set. This is evident in the ASMA-SEGS setting with TRAIN (where the number of types increases from 6,254 to 7,053), DEV (where the number of types increases from 2,006 to 2,159), and TEST (where the number of types goes up from 2,307 to 2,425). This increase of overall types, especially in the TRAIN set, reduces the data sparsity on TEST (where the rate drops from 32.12% OOV on GOLD-SEGS to 31.79% OOV), and that accounts for improved performance on the TEST set as compared to performance on GOLD-SEGS. The discrepancy of ASMA-SEGS performance on DEV and TEST as compared to GOLD-SEGS's performance on these two settings is due to a better ASMA performance on TEST as compared to its performance on DEV. TEST seems to be easier for ASMA to segment than DEV, which is the cause of this performance difference across the two data sets.

To further back the argument above, an error analysis of the cases correctly classified with ASMA-SEGS but not with GOLD-SEGS on DEV was performed. These are five cases in total, and it was found that ASMA made 12 segmentation errors, none of which resulted in changing any of the segments' prior polarity. For example, the neutral word *AlvlAvyn* (Eng. 'the thirty'), which should be segmented into *Al+vlAv+yn* (Eng. 'the+three+3rd person masc. plural'), was wrongly segmented into *Al+vlA+v+yn,* whose segments are still neutral. Another example is the word *wAlsqwT* (Eng. 'and the downfall'), which carries a subjective negative polarity and should be segmented into *w+Al+sqwT* (Eng. 'and+the+downfall') but is wrongly segmented into *w+AlsqwT.* Even with this incorrect segmentation, the second segment still carries the subjective negative polarity. Thus, it is clear that ASMA-SEGS, even though these data do include segmentation errors, do not seriously hurt subjectivity classification (with 2.30% accuracy and 2.29% average F_1 loss as on DEV) and can even in some cases benefit classification (with 1.56% accuracy and 1.57% average F_1-measure), as on TEST.

Theoretically, however, incorrect segmentation can result in changing the polarity of a word. The following are some examples:

1. The OBJ word *AlmrHlp* (Eng. 'the stage') should be rightly segmented into *Al+mrHl+p* (Eng. 'the+stage-related+fem. sing.') in which case the segments would hold the OBJ polarity, but can be wrongly segmented into *Alm+rHl+p* (Eng. 'pain+he left+fem. sing.') and hence result in segments involving 'pain' which carries an S-NEG polarity. The same word *AlmrHlp* can also be wrongly segmented into *Al+mr+Hl+p* (Eng. 'the+sour+reached+fem. sing'), which has the possibly S-NEG segment *mr* (Eng. 'sour').
2. A similar case where an OBJ word would carry a different polarity upon wrong segmentation is *bAlTbE* (Eng. 'of course'). A correct segmentation of this word is *b+Al+TbE* ('with+the+trait') and is OBJ, but a wrong segmentation can be *bAl+TbE* ('he urinated+trait') and hence can be viewed as S-NEG in certain contexts (e.g., imagine someone saying the utterance 'he urinated' during dinner time).

Segmentation errors involving changes of segments' polarity do not seem, however, to be frequent enough to cause a considerable subjectivity classification loss. If ASMA-SEGS results on both DEV and TEST are averaged, it can be concluded that ASMA segmentation causes 0.74% accuracy and 0.72% average F_1 subjectivity classification loss as compared to GOLD-SEGS.

4.2 Sentiment Classification

Table 6 below provides results for both gold-segmented and ASMA-segmented (ASMA-SEGS) experiments with sentiment classification.

Table 6. Sentiment classification results with GOLD-SEGS and ASMA-SEGS

				S-POS			**S-NEG**		
		Acc	**Avg- F**	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
DEV	**BASE**	57.89	57.71	65.33	56.32	60.49	50.65	60.00	54.93
	GOLD-SEGS	**65.13**	**64.45**	**69.77**	**68.97**	**69.36**	**59.09**	**60.00**	**59.54**
	ASMA-SEGS	61.84	61.52	68.35	62.07	65.06	54.79	61.54	57.97
TEST	**BASE**	64.86	64.85	64.91	66.07	65.49	64.81	63.64	64.22
	GOLD-SEGS	67.57	67.56	67.86	67.86	67.86	67.27	67.27	67.27
	ASMA-SEGS	**69.37**	**69.35**	**71.15**	**66.07**	**68.52**	**67.8**	**72.73**	**70.18**

4.2.1 GOLD

As Table 6 shows, GOLD-SEGS improves sentiment classification on both DEV and TEST. On the DEV set, GOLD-SEGS outperforms BASE with 7.24% accuracy and 6.74% average F_1 and on the TEST set it beats BASE with 2.71% accuracy and 2.71% average F_1-measure. GOLD-SEGS also outperforms BASE for both the subjective positive (S-POS) and subjective negative (S-NEG) classes classification across all evaluation metrics on both DEV and TEST. On TEST, GOLD-SEGS improves S-POS classification with 2.37% F_1 and S-NEG classification with 3.05% F_1 as compared to performance on BASE.

The finding that GOLD-SEGS improves sentiment classification is due to the reduction of the OOV rate and hence reducing data sparsity. This is possible as a result of segmenting the data. Table 7 below shows type statistics and data sparsity percentages on DEV and TEST for both GOLD-SEGS and ASMA-SEGS. As Table 7 shows, on the DEV set, the OOV reduction rate is 27.50% and on the TEST set it is 24.18%.

Table 7. Token statistics and OOV percentages for GOLD-SEGS and ASMA-SEGS

	TRAIN	**DEV**		**TEST**	
	# of types	**# of types**	**% of OOV**	**# of types**	**% of OOV**
BASE	7,328	1,303	51.57%	1,061	44.49%
GOLD-SEGS	3,922	993	24.07%	842	20.31%
ASMA-SEGS	4,355	1,067	28.40%	888	24.32%

In spite of the performance improvement acquired with GOLD-SEGS, an error analysis on DEV shows that BASE correctly detects 15 cases where GOLD-SEGS errs. The errors GOLD-SEGS makes on these 15 instances can be categorized into two main types, as follows:

1. Errors caused by a change of the polarity of acquired segments. This accounts for 20% (n=3) of the errors where the segmented word ends up with either a) a different polarity than the original surface word or b) an ambiguous polarity. These three cases happen to be words where splitting off the suffixival singular feminine marker *taa' marbuta* causes the change of polarity in the words *rAHp* (Eng. 'rest'), *mbAlgp* (Eng. 'exaggeration'), and *mEArDp* (Eng. 'opposition'). The word *rAHp* carries a positive polarity, but segmenting it results in the segments *rAH+p where* the first segment *rAH* can mean 'he went' and hence carry no polarity (i.e., be neutral). The word *mbAlgp* carries a negative polarity and splitting of the *p* leaves the first segment *mbAlg* with an ambiguous polarity as it can mean 'sums (of money).' The last word *mEArDp* usually would carry a positive polarity, especially in democratic countries, but segmenting off the *p* results in *mEArD* (Eng. 'exhibition')*, which* doesn't necessarily carry a positive polarity.
2. Errors in sentences where sentiment is not expressed explicitly and/or exclusively at the lexical level, but in addition/rather at higher syntactic levels. In such instances, sentences include both words with positive polarity and others with negative polarity and the tag is based on the overall meaning communicated by the sentence as a whole. Segmentation of words in these sentences interacts with a number of linguistic phenomena in ways that biases the classifier toward

one of the two classes and ends up resulting in wrong predictions. One or more of these phenomena can be involved with sentiment expression in a given instance. Two such phenomena are relevant in the following discussion as they occur in the data:

- **Multiword expressions:** Constructions whose meanings span beyond a single word (usually with a certain frequency threshold) are usually referred to as multiword expressions (MWEs) and interact with SSA classification. An example of these expressions in English is a collocation like 'chain smoker' or an idiom like 'a storm in a teacup.' Ideally, each MWE should be identified and represented to the classifier as a single unit. Although with BASE this does not happen, the surface word forms combined to form a MWE are still used as independent features, which might help classification. Upon segmentation, such word forms are further fragmented which worsens the case as to the possibility of benefiting from such expressions. An example MWE that occurs with some degree of frequency in political discourse in Arabic texts, *tbdd Alt$A&m* (Eng. 'removes pessimism'), is observed in the considered instances. The expression is segmented into *t+bdd+Al+t$A&m.* Another expression that is also observed in the data investigated is *mn AlEbv >n* (Eng. 'It is absurd that') and is segmented into *mn+Al+Ebv+>n.*
- **Negation:** Negation can shift the polarity of words, which is theoretically what can also happen with segments. However, some segments can have ambiguous senses including ones whose polarity does not necessarily change in the context of negation. For example, the construction *lA yjwz* (Eng. 'it is not permissible') is segmented into *lA+y+jwz* where the last segment can mean 'walnut.' Since the word *jwz* 'walnut' can even have positive polarity, as compared to the orthographically identical root *jwz* 'permission-related,' which is neutral, the different senses of the word obfuscate the effect of negation on shifting the polarity.

In addition to the discussion based on the error analysis above, it is worth illustrating further how gold segmentation can cause a change of a resulting segment's prior polarity. The polarity of a word can change upon even gold segmentation, including the following:

1. From an ambiguous polarity in the surface form to a) OBJ or b) S-NEG upon segmentation. For example, the surface form *bgyrh* (Eng. 'without him' or 'with jealousy') should be rightly segmented into either the OBJ *b+gyr+h* (Eng. 'with+other than+him') or the S-NEG *b+gyrh* (Eng. 'with+jealousy').
2. From an ambiguous polarity in the surface form to a) OBJ or b) S-POS upon segmentation. For example, the surface form *lOnfh* (Eng. 'for/to his nose' or 'for a dignity') should be rightly segmented into the OBJ *l+Onf+h* (Eng. 'for/to+nose+his'), the S-POS *l+Onfh* (Eng. 'for+dignity').
3. From an S-POS polarity in the surface form to a) OBJ or b) S-POS upon segmentation. For example, the S-POS surface form *HsnAthm* (Eng. 'their virtues') should be rightly segmented into *Hsn+At+hm* (Eng. 'beauty+plural suffix+their') and hence result in involving the ambiguous segment *Hsn* which can act as a noun ('beauty') but also is a popular proper noun in Arabic. Given

that Arabic named entities are not capitalized like the practice in English and some other languages, this can result in heightened ambiguity.

It can be concluded from the discussion above that although segmentation helps reduce sparsity and hence potentially improve classification, it also interacts with ambiguity in intricate ways. As is explained above, segmentation can resolve some cases of ambiguous polarity but can also introduce ambiguity in other cases. This finding suggests that *word sense disambiguation* (where each sense of a given word is distinguished from other senses) may be needed as a means of resolving ambiguities caused by segmentation.

4.2.2 ASMA

Compared to BASE, ASMA-SEGS improves classification on both the DEV and TEST sets. On DEV, 3.95% accuracy and 3.81% average F_1 gains are achieved using ASMA-SEGS. On TEST, ASMA-SEGS improves sentiment classification over BASE with 4.51% accuracy and 4.50% average F_1-measure. ASMA-SEGS also improves both S-POS and S-NEG classification on both DEV and TEST across all evaluation metrics, with a single exception in the case of recall (Rec) on the TEST set for S-POS classification where it achieves identical scores with BASE (i.e., 66.07% Rec). On TEST, ASMA-SEGS improves S-POS class classification with 3.03% F_1 and the S-NEG class classification 5.96% F_1 against BASE.

There is a discrepancy between ASMA-SEGS performance as compared to GOLD-SEGS on DEV and TEST. While ASMA-SEGS is outperformed on DEV with 3.29% accuracy and 2.93% average F_1-measure, it improves over GOLD-SEGS with 1.80% accuracy and 1.79% average F_1-measure. Similarly, as to S-POS class classification and S-NEG class classification, ASMA-SEGS is outperformed across most metrics on DEV and improves over GOLD-SEGS on TEST across all evaluation metrics. On TEST, ASMA-SEGS improves S-POS class detection with 0.66% F_1 and also S-NEG class detection with 2.91% F_1-measure.

The fact that GOLD-SEGS outperforms ASMA-SEGS on DEV is due to the fact that segmentation with ASMA causes the change of the polarity of some segments from the polarity of their original word forms. An error analysis of correctly classified cases (14 in total) in the DEV set with the GOLD-SEGS that were misclassified when using the ASMA-SEGS setting is illustrative in this regard. Wrong segmentation results in changing the polarity of component segments of two words in two cases (%=14.28), as follows:

- The word *lAlqyAm* (Eng. ‘for undertaking’) carries a neutral polarity, but is wrongly segmented with the ASMA-SEGS setting into *l+Alqy+Am* where the segment *Alqy* can have negative polarity (Eng. ‘threw’) or neutral polarity (Eng. ‘gave [e.g., a speech]’).
- The word *Alm&Amrp* that (Eng. ‘the conspiracy’) that carries a negative polarity is wrongly segmented with the ASMA-SEGS setting into *Al+m&A+mr+p* where all segments carry neutral polarity.

In addition to the examples above, the polarity of an S-NEG word can theoretically change upon wrong segmentation that results in valid segments. The following is an example:

1. The S-NEG surface form *AlmrwEp* (Eng. 'the horrifying+fem. sing.') which should be rightly segmented into *Al+mrwE+p* (Eng. 'the+horrifying+fem. sing.') can be wrongly segmented into the S-NEG segment *Alm* (Eng. 'pain') and the S-POS segment *rwEp* (Eng. 'terrific').

The polarity of an S-NEG word can also theoretically change upon wrong segmentation that results in one or more valid segments plus one or more invalid segments. The following is an example:

2. The S-NEG surface form *mrwEp* (Eng. 'horrifying+fem. sing.'), which should be rightly segmented into *mrwE+p* (Eng. 'terrifying+fem. sing.'), can be wrongly segmented into invalid segment *m* and the S-POS segment *rwEp* (Eng. 'terrific').

The finding that ASMA-SEGS improves classification over BASE is quite predictable, again due to the reduction of data sparsity that comes with segmenting the surface forms. The OOV rate drops from 51.57% to 28.40% on DEV and from 44.49% to 24.32% on TEST, as shown in Table 7. The situation with the results of ASMA-SEGS as compared to GOLD-SEGS is different, in that there is no clear correspondence between the sparsity rate and the improvement gained (or lack thereof) over DEV and TEST. While the sparsity rates with ASMA-SEGS are higher than their counterparts with GOLD-SEGS (with 4.33% higher OOV on DEV and 4.01% higher OOV on TEST), ASMA-SEGS still outperforms GOLD-SEGS slightly on TEST. This shows that there is one or more other factor(s), other than data sparsity, that interact(s) with classification in ways that result in the performance reported here. One of these factors could be the existence of errors in the gold-segmented ATB data,[12] which affects the GOLD-SEGS setting results but not necessarily the ASMA-SEGS results.

In an attempt to identify such factors further, an analysis of the errors made by GOLD-SEGS but *not* by ASMA-SEGS on the DEV data was performed. It was found that there are 10 instances that are correctly classified with ASMA-SEGS but not with GOLD-SEGS. Two of these instances (i.e., 20% of the cases) are examples where sentiment is expressed subtly with some tokens that carry positive sentiment and others that carry negative sentiment. For such cases, the segmentation errors made by ASMA bias the classifier towards one of the two classes. This happens when the tokens with semantic orientation (whether positive or negative) are the loci of such ASMA segmentation errors. The following is one of the two examples that occurs in the data:

3. *ftsmH lhA tlk Aldwl bAlfwz bbED AlnqAT ky ytsnY lhA <lhA&hA bEDhA bbED wtbqY hy fwq AlSrAE.* (Eng. 'And these countries allow them to achieve a few gains and remain preoccupied with one another such that these countries stay in control of the conflict.')

This sentence has both positive (i.e., *bAlfwz* [Eng. 'gaining']) and negative (i.e., *<lhA&hA* [Eng. 'pacifying/keeping them preoccupied'] and *AlSrAE* [Eng. 'the conflict']) surface word forms and is assigned an overall S-POS tag in the GOLD-SEGS setting. The surface form *bAlfwz* carries a positive semantic orientation and is gold segmented into *b+Al+fwz* (Eng. 'with+the+gain') where the segment *fwaz* that carries positive polarity remains intact. With ASMA-SEGS, however, the word *bAlfwz* is wrongly segmented into *b+Alf+wz,* which causes the loss of the positive semantic

[12] This observation about errors in the ATB segmentation was also brought to my attention by Mona Diab (personal communication, December 1, 2015).

orientation the surface form carries and hence interacts with the classifier decision. With the ASMA-SEGS setting the sentence is rightly assigned an S-NEG tag (now that the segment carrying positive polarity is absent), whereas in the GOLD-SEGS setting it is predicted as S-POS.

The rest of the cases (i.e., 80%) are also accounted for by ASMA segmentation errors, but in these cases not ones that interact explicitly with the semantic orientation of words. Rather, these errors result in creating wrong segments that also happen to exist in the TRAIN set. In the 10 DEV set investigated instances, 13 segments were found to exist with ASMA-SEGS (but not with GOLD-SEGS) that also existed in TRAIN. These 13 segments include, for example, the two segments *AlOjhz* and *p,* which are the result of a wrong segmentation of the word *AlOjhzp* (Eng. 'the forces'[13]). In all, based on this error analysis and knowledge of wider NLP tasks, SSA does not seem as sensitive to segmentation errors as tasks like POS tagging or parsing.

5 Conclusion

In this paper, we raised the question: "How can morphological richness in Arabic be handled in the context of SSA?" For answering this question, we focused on segmentation: We investigated the effect of segmentation (i.e., the process of identifying morphological units within a given surface word form) on Arabic SSA. The results show that segmentation, whether GOLD- or ASMA-based, improves both subjectivity classification and sentiment classification. As explained above, the gains are the result of reduced sparsity and/or reduced ambiguity as the surface word forms are broken down to their component segments. It was observed, however, that the process of segmentation resulted in some ambiguity that was responsible for a share of the errors the classifiers still make. This suggests a potential need for word sense disambiguation on the segmented text. It is also noteworthy that the errors ASMA made in segmenting text negatively impacted neither subjectivity classification nor sentiment classification, based on comparing ASMA-SEGS to GOLD-SEGS on TEST. This finding shows the utility of state-of-the-art, machine-predicted segmentation for SSA.

The findings show that modeling subjectivity and sentiment in lexical space should be nuanced to the level of morphological richness of a given language, and that for a morphologically-rich language like Arabic it is useful to segment the input text and model functional and content segments differently. The results also show that although subjectivity and sentiment are social meaning concepts (i.e., expressed at the levels of semantics and pragmatics), modeling them can benefit from lower linguistics levels in lexical space. One of the limitations of the current study is that the data were primarily labeled by a single human annotator. In spite of the fact that the recuruited annotator has the highest level of training (Ph.D. in linguistics) and motivation we can aspire to, and that our initial statge of double labeling a sample of the data clearly showed that the annotator provided very high quality annotations, it would have been better to double label the full dataset. For future extensions of the current research, we plan to investigate further ways to exploit the lexical space in the context of SSA.

[13] The word is translated into 'forces' since it occurs in a political register to refer to 'security forces,' otherwise it is usually translated into 'sets' or 'equipments' as in the noun phrase 'electrical sets.'

6 References

Abbasi, A. (2007). Affect intensity analysis of dark Web forums. In *Intelligence and Security Informatics* (pp. 282–288). New Brunswick, NJ: IEEE Press. Retrieved March 12, 2010 from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4258712

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 (TOIS)*, *26*(3), 1–34.

Abdul-Mageed, M. M. (2008). Online news sites and journalism 2.0: Reader comments on Al Jazeera Arabic. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, *6*(2), 59-76. Retrieved June 20, 2014 from http://www.triple-c.at/index.php/tripleC/article/download/78/70

Abdul-Mageed, M. (2013a, March 2). *Social Media Arabic*. Paper presented at the 27th Annual Symposium on Arabic Linguistics, Indiana University, Bloomington, IN.

Abdul-Mageed, M. (2013b, October 12). *Social media mining with natural language processing: Challenges and initial solutions*. Paper presented at the ILS Doctoral Student Forum, School of Informatics and Computing, Indiana University, Bloomington, IN.

Abdul-Mageed, M., & Albogmi, H. (2011, March 11). *Taghreed?: What Arabs say on Twitter and how they say it.* Poster session presented at the 11th Georgetown University Round Table on Languages and Linguistics (GURT2011): Language and New Media: Discourse 2.0. Washington, DC: Georgetown University Press.

Abdul-Mageed, M., Brown, C., & Abul-Hija, D. (2013, October 23). *Twitter in the context of Arab Spring.* Paper presented at the 14th annual conference of the Association of Internet Researchers (Internet Research 14.0 – Resistance + Appropriation), Denver, CO.

Abdul-Mageed, M., & Diab, M. (2011a). Linguistically-motivated subjectivity and sentiment annotation and tagging of Modern Standard Arabic. *International Journal on Social Media MMM: Monitoring, Measurement, and Mining, 2*(1-2), 19-38. Retrieved January 10, 2012 from http://www.konvoj.cz/www/en/Publishing/Journals/MMM/Archive/MMM2011/MMM2011-1002.pdf

Abdul-Mageed, M., & Diab, M. T. (2011b, June). Subjectivity and sentiment annotation of Modern Standard Arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop* (pp. 110-118). Stroudsburg, PA: Association for Computational Linguistics

Abdul-Mageed, M., & Diab, M. (2012a, January). Toward building a large-scale Arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet Conference* (pp. 18-22). Matsue, Japan: Global WordNet Association.

Abdul-Mageed, M., & Diab, M. (2012b, May). AWATIF: A multi-genre corpus for Arabic subjectivity and sentiment analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012).* Istanbul, Turkey: European Language Resources Association (ELRA).

Abdul-Mageed, M., & Diab, M. (2014, May). SANA: A large scale, multi-genre, multi-dialect lexicon for Arabic sentiment analysis. *Proceedings of The 9th International Conference on Language Resources and Evaluation (LREC2014)* (pp. 1162-1129). Reykjavik, Iceland: European Language Resources Association (ELRA).

Abdul-Mageed, M., Diab, M. T., & Korayem, M. (2011, June). Subjectivity and sentiment analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 587-591). Stroudsburg, PA: Association for Computational Linguistics.

Abdul-Mageed, M., Diab, M., & Kübler, S. (2013, September). ASMA: A system for automatic segmentation and morpho-syntactic disambiguation of Modern Standard Arabic. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 1-8). Hissar, Bulgaria: INCOMA Ltd. Shoumen.

Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, *28*(1), 20-37.

Abdul-Mageed, M., Kübler, S., & Diab, M. (2012, July). Samar: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 19-28). Stroudsburg, PA: Association for Computational Linguistics.

Al-Kabi, M. N., Abdulla, N. A., & Al-Ayyoub, M. (2013, December). An analytical study of Arabic sentiments: Maktoob case study. In *Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference for* (pp. 89-94). IEEE.

Al-Ayyoub, M., Essa, S. B., & Alsmadi, I. (2015). Lexicon-based sentiment analysis of Arabic tweets. *International Journal of Social Network Mining*, *2*(2), 101-114.

Al Shboul, B., Al-Ayyoub, M., & Jararweh, Y. (2015, April). Multi-way sentiment classification of Arabic reviews. In *Information and Communication Systems (ICICS), 2015 6th International Conference on* (pp. 206-211). IEEE.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation, Malta* (pp. 2200-2204). Valletta, Malta: European Language Resources Association (ELRA). Retrieved July 13, 2012 from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

Badawi, S. M. (1973). *Mustawayat al-'Arabiyya al-mu'asira fi Misr (Levels of Modern Arabic in Egypt).* Cairo: Dar al-Ma'arif.

Banfield, A. (1982). *Unspeakable sentences: Narration and representation in the language of fiction*. London: Routledge.

Bassiouney, R. (2009). *Arabic sociolinguistics*. Edinburgh: Edinburgh University Press.

Bateson, M. (1967). *Arabic language handbook.* Washington, DC: Center for Applied Linguistics.

Chiang, D., Diab, M., Habash, N., Rambow, O., & Shareef, S. (2006, April). Parsing Arabic dialects. In *Proceedings of the European Chapter of ACL (EACL)* (pp.

369-376). Stroudsburg, PA: Association for Computational Linguistics.

Dalal, M. K., & Zaveri, M. A. (2014). Opinion mining from online user reviews using fuzzy linguistic hedges. *Applied Computational Intelligence and Soft Computing*, Volume 2014 (pp. 1-9). New York: Hindawi Publishing Corp.

Darwish, K. (2002). Building a shallow Arabic morphological analyzer in one day. In *Proceedings of the Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of the Data Engineering Workshop, IEEE 24th International Conference on Data Engineering* (pp. 507–512). Washington, DC: IEEE Computer Society.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings of HLT-NAACL 2004: Short Papers on XX* (pp. 149–152). Stroudsburg, PA: Association for Computational Linguistics.

Diab, M., Hacioglu, K., & Jurafsky, D. (2007). Automatic processing of Modern Standard Arabic text. In A. Soudi, G. Neumann, & A. Van den Bosch (Eds.), *Arabic computational morphology: Knowledge-based and empirical methods*. Dordrecht, Netherlands: Springer Verlag.

Esuli, A., & Sebastiani, F. (2005, July). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 617–624). Ney York, NY: Association of Computing Machinery.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* (Vol. 6, pp. 417–422). Genoa, Italy: European Language Resources Association (ELRA).

Esuli, A., & Sebastiani, F. (2007). Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 20th International Conference on Computational Linguistics* (Vol. 7, pp. 442-431). Stroudsburg, PA: Association for Computational Linguistics.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., & Quwaider, M. (2014, August). Cross-lingual short-text document classification for facebook comments. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on* (pp. 573-578). IEEE.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, *8*(4), 1–22.

Grefenstette, G., Qu, Y., Evans, D. A., & Shanahan, J. G. (2006). Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theory and applications* (Vol. 20, pp. 93-107). Dordrecht, The Netherlands: Springer.

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

Habash, N., & Rambow, O. (2005, June). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 573–580). Stroudsburg, PA: Association for Computational Linguistics.

Habash, N., Rambow, O., & Roth, R. (2009, April). Mada+ tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)* (pp. 102-109). Cairo, Egypt: MEDAR.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 174–181). Stroudsburg, PA: Association for Computational Linguistics.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics,* Vol. 1 (pp. 299–305). Stroudsburg, PA: Association for Computational Linguistics.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Washington, DC: Georgetown University Press.

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* (pp. 168–177). New York: Association of Computing Machinery.

Jang, H., & Shin, H. (2010, August). Language-specific sentiment analysis in morphologically rich languages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 498-506). Stroudsburg, PA: Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. New York: Prentice Hall.

Kamps, J., & Marx, M. (2002). Words with attitude. In *Proceedings of the First International Conference on Global WordNet* (pp. 332-341). Mysore, India: Global WordNet Association.

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING2004)* (pp. 1367-1373). Stroudsburg, PA: Association for Computational Linguistics.

Kim, S. M., & Hovy, E. H. (2007). Crystal: Analyzing predictive opinions on the web. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning* (pp. 1056-1064). Stroudsburg, PA: Association for Computational Linguistics.

Lee, Y. S., Papineni, K., Roukos, S., Emam, O., & Hassan, H. (2003). Language model based Arabic word segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Volume 1 (pp. 399–406). Stroudsburg, PA: Association for Computational Linguistics.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.

Lyons, J. (1977). *Semantics.* Cambridge, UK: Cambridge University Press.

Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge, UK: Cambridge University Press.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools* (pp. 102–109). Cairo, Egypt: Magnet Creative Communications.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marton, Y., Habash, N., & Rambow, O. (2010). Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages* (SPMRL'10) (pp. 13–21). Stroudsburg, PA: Association for Computational Linguistics.

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* (pp. 341–349). Edmonton, Canada: Association for Computing Machinery.

Mullen, T., & Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 412-418). Stroudsburg, PA: Association for Computational Linguistics.

Ng, V., Dasgupta, S., & Arifin, S. M. (2006, July). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics* (pp. 611–618). Sydney, Australia: Association for Computing Machinery.

Palva, H. (1982). Patterns of Koineization in modern colloquial Arabic. *Acta orientalia*, *43*, 13–32.

Palva, H. (2006). Dialects: Classification. In K. Versteegh, M. Eid, A. Elgibali, M. A. Woidich, & A. Zaborski (Eds.), *Encyclopedia of Arabic language and linguistics.* (pp. 604–613). Leiden/Boston: Brill.

Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 271–278). Stroudsburg, PA: Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, October). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 40th Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing,* Volume 10 (pp. 79–86). Stroudsburg, PA: Association for Computational Linguistics.

Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning sentiment lexicons in

Spanish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 3077-3081). Istanbul, Turkey: European Language Resources Association (ELRA).

Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., et al. (2005). Parsing Arabic dialects. *Final Report, JHU Summer Workshop*. Retrieved June 22, 2014 from http://idiom.ucsd.edu/~rlevy/papers/rambow-etal-2006-techreport-final.pdf

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge, UK: Cambridge University Press.

Smrž, O. (2007). *Functional Arabic morphology: Formal system and implementation*. Ph.D. thesis, Charles University in Prague. Retrieved December 7, 2012 from: http://ufal.mff.cuni.cz/~smrz.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Terveen, L., Hill, W., Amento, B., McDonald, D., & Creter, J. (1997). PHOAKS: A system for sharing recommendations. *Communications of the ACM*, *40*(3), 59–62.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler , S., Candito, M., Foster, J., Versley, Y., et al. (2010). Statistical parsing of morphologically rich languages (SPMRL): What, how and whither. In *The First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*. Retrieved June 14, 2011 from http://www.aclweb.org/anthology/W10-1401.

Tsur, O., Davidov, D., & Rappoport, A. (2010). ICWSM-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proceeding of the International Conference on Weblogs and Social Media (ICWSM 2010)* (pp. 162–169). Menlo Park, California: AAAI Press.

Versteegh, K. (2001). *The Arabic language*. Edinburgh, Scotland: Edinburgh University Press.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20*(2), 233–287.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, *30*(3), 277-308.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Stroudsburg, PA: Association for Computational Linguistics.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3), 399–433.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing, Volume 10* (pp. 129–136). Stroudsburg, PA: Association for Computational Linguistics.