Not All Segments are Created Equal: Syntactically Motivated Sentiment Analysis in Lexical Space

Muhammad Abdul-Mageed
School of Library, Archival, and Information Studies
University of British Columbia, Vancouver, BC
muhammad.mageed@ubc.ca

Abstract

Although there is by now a considerable amount of research on subjectivity and sentiment analysis on morphologically-rich languages, it is still unclear how lexical information can best be modeled in these languages. To bridge this gap, we build effective models exploiting exclusively gold and machine-segmented lexical input and successfully employ syntactically motivated feature selection to improve classification. Our best models achieve significantly above the baselines, with 67.93% and 69.37% accuracies for subjectivity and sentiment classification respectively.

1 Introduction

The task of *subjectivity* detection refers to identifying aspects of language that are *objective* (i.e., *I have a meeting at 2:00pm.*) vs. those that express opinions, feelings, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and hence are *subjective*. Subjective language is further classified based on its *sentiment* into *positive* (e.g., *The new machines are revolutionary!*), *negative* (e.g., *The Syria war is terrifying!*), *neutral* (e.g., *The new models may be released next week.*), or, sometimes, *mixed* (e.g., *I really like this phone, but it is way too expensive!*). The field of subjectivity and sentiment analysis (SSA) is a very vibrant one and there has been a flurry of research on especially the English language (Wiebe et al., 2004; Liu, 2010; Dave et al., 2003; Pang and Lee, 2008; Chaovalit and Zhou, 2005; Zhuang et al., 2006). By now, there is also a fair amount of work on morphologically rich languages (MRL) (Tsarfaty et al., 2010) like Arabic (Abdul-Mageed and Diab, 2011; Abdul-Mageed et al., 2011; Abdul-Mageed and Diab, 2012; Abdul-Mageed et al., 2014; Aly and Atiya, 2013; Refaee and Rieser, 2014; Nabil et al., 2015; Salameh et al., 2015; Refaee and Rieser, 2016). SSA work on MRLs, however, is still in an early stage as MRLs raise a range of questions on their own. In the current work, we focus on answering the question: "How it is that Arabic can be modeled within lexical space?" More specifically, we investigate the utility of teasing apart lexical input based on grammatical criteria and empirically weigh the contribution of features therein toward SSA. The current work is a follow up on submitted work (Abdul-Mageed, 2017) where we measure both gold and machine-predicted tree-bank style segmentation (Maamouri et al., 2004) on the two tasks of subjectivity and sentiment.

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 Penn Arabic Treebank (ATB) (Maamouri et al., 2004), a segment can be a stem, an inflectional affix, or a clitic. For example, the surface word *wbHsnAthm* (Eng. '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*. In (Abdul-Mageed, 2017), we have shown how reducing a word to its component segments is a desirable measure for SSA since it reduces the number of observed forms and hence alleviates *sparsity*: The system does not see as many forms at test time that have not been seen at training time. Providing all lexical segmented input to a classifier, however, may or may not be an ideal procedure. In English, usually words like 'a,' 'the,' and 'from' are treated as stop words and hence removed before classification. These tokens are viewed as *functional* words that do not usually contribute to classification accuracy. Are there

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 147–156,
Valencia, Spain, April 3, 2017.

ways to break down the lexical space based on relevant, if not comparable, grammatical grounds? That is the question we seek to answer in the current work. Overall, we make the following contributions: (1) We present a new human-labeled ATB dataset for SSA; (2) We introduce a new syntactically motivated feature selection method for SSA on Arabic that can arguably also help classification on other languages of rich morphology; and (3) We present detailed linguistically-motivated (error) analyses of the behavior of the lexical models, against the background of Arabic morphological complexity.

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

2 Dataset and Methods

Data: We gold-label a subset from each of the first three parts of the ATB (i.e., the first 70 documents from ATB1V4.1, the first 50 documents from ATB2V3.1, and the first 58 documents from ATB3V3.2) at the sentence level with tags from the set {*OBJ, subjective-positive (S-POS), subjective-negative (S-NEG), subjective-mixed (S-MIXED)*}. The data belong to the newswire genre and were manually labeled by the Linguistic Data Consortium (LDC) for part-of-speech (POS), morphology, gloss, and syntactic treebank annotation. A single annotator, with a Ph.D. in linguistics and a native Arabic fluency, labeled the data after being provided written guidelines and several sessions of training and discussions with the authors. We followed the guidelines in the literature (Abdul-Mageed and Diab, 2011; Abdul-Mageed and Diab, 2012). To ensure quality, 5% of the data (n=250 sentences) was double labeled by a second annotator. Inter-annotator agreement reached 83% without adjudication, and hence the first annotator's decisions were judged sufficient. Table 1 shows class distribution in our data.

Procedure: We divide each of the three treebank parts into 80% training, 10% development, and 10% test. The training parts from each Treebank are then added up to build TRAIN, the development parts are added up to build DEV, and the test parts are combined to build TEST. For our experiments, results are reported both on DEV and TEST. Importantly, only the DEV set is used for tuning classifier performance and error analyses. TEST is used as a fully blind set. We follow a two-stage classification process where the *first stage* is to tease apart the OBJ and SUBJ classes, and the *second stage* is to distinguish the S-POS and the S-NEG classes. For this work, we do not handle the MIXED class, since it is minimal in our data.

Dataset	OBJ	S-P	S-N	S-M	ALL
ATB1V4.1	582	183	188	39	992
ATB2V3.1	623	151	227	3	1,004
ATB3V3.2	1,472	462	414	6	2,354
ALL	2,677	796	829	48	4,350

Table 1: Data statistics. S-P= *subjective positive* and S-N= *subjective negative*.

Settings: We use two settings based on text preprocessing: *Gold* and *machine-predicted*. For the gold setting, human-annotated segmentation and morphosyntactic disambiguation as labeled by LDC are exploited. For the machine-predicted setting, we use the ASMA tool (Abdul-Mageed et al., 2013), which renders state of the art segmentation and morphosyntactic tagging for MSA. For all the subjectivity and sentiment experiments, we use SVMs with a linear kernel.

3 Results

3.1 Subjectivity with Lexical Filtering

As pointed out earlier, we follow up on previous work (Abdul-Mageed, 2017) where we show the utility of representing lexical input in the form of segments. As such, we cite results from that work with both surface word forms and segmented text and compare the current work to these results. We now set out to answer the question: "Are all segments equally useful to subjectivity and/or sentiment classification?" From a linguistics perspective, segmented lexical input can be viewed as comprised of *content segments* (i.e., those corresponding to verbs or nominals [nouns, adjectives, and adverbs]) and *functional segements* (e.g., definite articles). Content segments are often thought to carry the important semantic content in a sentence, and hence we investigate their utility for SSA. In other words, we employ *lexical filtering*: We filter out functional segments (e.g., clitics and affixes after segmentation) and use content segments exclusively as classifier input. We use the POS tags in Table 2 to identify content segments. Table 3 shows results of subjectiv-

				OBJ			SUBJ		
		Acc	**Avg-F**	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
DEV	surf	62.07	57.4	78.75	29.86	43.3	58.31	92.41	71.5
	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
	gold-cont	61.15	57.34	72.34	32.23	44.59	58.06	88.39	70.09
	asma-cont	62.07	58.47	73.96	33.65	46.25	58.7	88.84	70.69
	gold-cont-M1	**68.28**	**65.71**	**84.76**	**42.18**	**56.33**	**63.03**	**92.86**	**75.09**
	asma-cont-M1	66.9	63.92	83.84	39.34	53.55	61.9	92.86	74.29
TEST	surf	60.58	60.55	90.91	44.22	59.5	46.41	91.61	61.61
	gold-segs	65.03	65	91.02	51.7	65.94	49.65	90.32	64.07
	asma-segs	66.59	66.57	93.37	52.72	67.39	50.88	92.9	65.75
	gold-cont	57.68	57.63	87.14	41.5	56.22	44.34	88.39	59.05
	asma-cont	59.69	59.66	89.51	43.54	58.58	45.75	90.32	60.74
	gold-cont-M1	66.15	66.12	92.26	52.72	67.1	50.53	91.61	65.14
	asma-cont-M1	**67.93**	**67.9**	**94.64**	**54.08**	**68.83**	**51.96**	**94.19**	**66.97**

Table 3: Subjectivity classification with syntactically motivated feature selection. Th prefixes *gold-* and *asma-* refer to Treebank-acquired and ASMA-acquired segments (i.e., *-segs*), content segments (i.e., *-cont*), and select content segments (i.e., *-cont-M*$*$), respectively.

ity classification with both the gold and ASMA syntactically motivated lexical filtering (**gold-cont** and **asma-cont**, respectively) where only content segments are provided as classifier input. For this set of experiments, we use two baselines: 1) performance with surface word forms (surf), and, (in order to compare to performance with both gold- and ASMA-segmented text forms as reported in (Abdul-Mageed, 2017)), 2) **gold-segs** and 3) **asma-segs**, respectively.

As Table 3 shows, for subjectivity classification with gold-cont, no improvement is acquired over the surface word forms (surf). On DEV, gold-cont is 0.92% accuracy below surf. On TEST, gold-cont is at 2.90% accuracy below surf. Similarly, apart from the OBJ class classification on DEV (where 1.29% F_1 gain is acquired), gold-cont loses against surf across all evaluation metrics for both the OBJ and SUBJ classes. On TEST, gold-cont is outperformed by surf with 3.28% accuracy. Comparing the results acquired with gold-cont to those acquired without lexical filtering (i.e., with gold-segs and asma-segs) shows that gold-cont causes classification losses on both DEV and TEST. On DEV, gold-cont causes a classification loss with 7.13% accuracy compared to gold-segs and 3.91% accuracy compared to asma-segs. On TEST, gold-cont is outperformed by gold-segs with 7.35% accuracy and also by asma-segs with 6.90% accuracy. In addition, as Table 3 also shows, asma-cont is outperformed by asma-segs and gold-segs on DEV. On TEST, asma-cont is outperformed by surf (with 0.89% accuracy), asma-segs (6.90% accuracy), and gold-segs (5.34% accuracy).

These results show that removing functional segments is not a useful measure for subjectivity classification, regardless whether the segments kept are gold (gold-cont) or ASMA-predicted (asma-cont). As we show in (Abdul-Mageed, 2017), with regard to results acquired using gold-segs and asma-segs as compared to surf, segmented input text helps reduce data sparsity, which partially accounts for classification improvement with these settings. The situation when we remove functional segments as we do here is different: Even though this type of lexical filtering with both gold-cont and asma-cont does reduce sparsity significantly as compared to surf, as is shown in Table 4, performance with these two settings drops. This shows that data sparsity reduction is not the sole deciding factor as to classifier performance, and that the removed functional segments are important input for the subjectivity task. This conclusion is clearly supported by the fact that lexical filtering settings (i.e., gold-cont and asma-cont) are outperformed by their segmentation counterparts (i.e., gold-segs and asma-segs), even though the differences in sparsity rates between the two are minimal.

VERBS	
VERB	verb
PSEUDO_VERB	pseudo-verb
PV	perfective verb
PV_PASS	perfective passive verb
IV	imperfective verb
IV_PASS	Imperfective passive verb
CV	imperative/command verb
NOMINALS	
NOUN	noun
NOUN_NUM	nominal/cardinal number
NOUN_QUANT	quantifier noun
NOUN.VN	deverbal noun
NOUN_PROP	proper noun
ADJ	adjective
ADJ_COMP	comparative adjective
ADJ_NUM	adjectival/ordinal number
ADJ.VN	deverbal adjective
ADJ_PROP	proper adjective
ADV	adverb
REL_ADV	relative adverb
INTERROG_ADV	interrogative adverb

Table 2: POS tags for content segments

	TRAIN	**DEV**	
	# types	**# types**	**% OOV**
surf	13,201	3,028	44.25%
gold-segs	6,254	2,006	22.88%
asma-segs	7,053	2,159	26.40%
gold-cont	6,124	1,916	23.49%
asma-cont	6,888	2,066	27.11%

Table 4: Type statistics and OOV percentages for gold and ASMA-predicted content segments

To further help interpret the results, we perform an investigation of the distribution of functional segments in the TRAIN set for both gold-segs and asma-segs. To help explain the distribution of functional segments, we introduce the concept of *distributional relative frequency* (RF): RF is the frequency of these segments within a given class divided by the total number of data points making up that specific class. The distribution of functional segments is calculated based on RF to cater for the unbalanced class distribution in the TRAIN data where the number of OBJ cases=$1,259$ and the number of SUBJ cases=$3,840$. Also, RF is calculated based on absolute values (i.e., after reducing the segments of frequency > 1 to 1 [i.e., segments that are repeated multiple times in one sentence are rendered to one occurrence only], to match the presence vs. absence vectors). In acquiring the RF of segments across both the OBJ and SUBJ classes, we use a threshold parameter specifying the number of times a segment occurs in one of the two classes more than the other. This parameter is used with values from the set $\{1, 2, 3, 4\}$.

Functional segments occur with different distributions in the two classes. As extracted from gold-segs TRAIN, there is a total of 160 functional segments out of which 99 occur in gold-segs DEV set. On TRAIN, 60% (n=96) of functional segments occur at least two times in one of the two classes more than the other class, 48.75% (n=78) of them occur at least three times, and 0.05% (n=8) of them occur at least four times. On DEV, 57.57% (n=57) of functional segments occur at least two times in one of the two classes, 49.49% (n=49) of them occur at least three times, and 45.45% (n=45) of them occur at least four times. For ASMA, there is similarly a discrepancy of distribution between the functional segments occurring in the two classes: Within the asma-segs training set, a total of 149 functional segments occur. For a considerable percentage of these segments (%=37.58, n=56), each segment is found to occur with a relative frequency that is four times or more in one of the two classes than its occurrence in the other. When the relative frequency threshold is lowered to three times or more, it is found that 41.61% (n=62) of these functional segments satisfy this lowered threshold of class distribution. When the relative frequency threshold is lowered to two times or more, 57.05% (n=85) of segments satisfy that threshold.

The different distribution of functional segments across the OBJ and SUBJ classes is linguistically motivated, as these segments are related to a host of linguistic phenomena that interact with expression of subjectivity. The following is a number of such phenomena that we find to be used more frequently with the SUBJ class:

Negation: Negation is used in natural language for various purposes, including those related to the 'etiquette' of involving in a conversation or politeness (Brown and Levinson, 1987) in discourse. For example, it is usually considered more polite to say that something is 'not good' (and hence employ negation) rather than saying it is 'bad.' Negation is used in newswire discourse in various con-

texts. For example, politicians use negation when they 'denounce' an action or 'deny' a stance. Often times, contexts where negation is employed would be more SUBJ than OBJ, based on the distribution of negation particles in the TRAIN and DEV data. Examples of negation particles that occur more frequently in the SUBJ class are *lA* (which negates imperfective verbs) *lm* (which negates things in the past) and *ln* (which negates things in the future).

Restriction: In situations where it is necessarily to be precise, restrict a statement, stress a position, etc., employment of restriction particles is useful. Restriction particles like *<lA* (Eng., 'except') and *<nmA* (Eng. 'but for') are used more with the SUBJ class in the data. An example sentence where *<lA*is used in a SUBJ-POS context is *lA ysE Al <nsAn <lA <n yqdrhA* (Eng. 'One can only appreciate it').

Interactional resources: Writers use a number of linguistic resources, often referred to as interactional resources (Hyland and Tse, 2004), to engage readers in the argument in ways that interact with expression of subjectivity. Interactional resources include self-mentions via first person pronouns, engagement markers like second person pronouns, epistemic modality markers that serve to convey how confident people are about the ideational material they convey (Palmer, 1986) (whether these are hedges like *rbmA* [Eng. 'perhaps'] or boosters like *mwkd* [Eng. 'it is certain']), and attitude markers like *llOsf* (Eng. 'unfortunately'). Self-mentions and engagement markers are, more often than not, expressed via functional segments in Arabic, namely first and second person pronouns. Epistemic modality and attitude markers are either expressed adverbially or as phrases involving functional segments like prepositions. For example, the phrase *mn Almdh$* On (Eng. 'it is surprising that') acts as an attitude marker. Filtering out functional segments removes these interactional resources which are useful devices for expression of subjectivity, leaving the phrase as *mdh$* (Eng. 'surprising'), which carries less intense polarity. This, in turn, adds to the classification drop.

Conditionals: In situations when a writer/speaker needs to describe a hypothetical scenario or condition on occurrence on another, etc., conditionals are used. These, as such, are not as much associated with facts as they are with what the writer/speaker believes is possible, likely, probable, etc. In this way, they can be associated with hedges when they are employed to restrict a claim. Conditionals like *<*A* (Eng. 'if') and *lw* (Eng. 'if') are thus distributed more frequently with the subjective class.

3.2 Subjectivity with Syntactically Motivated Feature Selection

In order to further investigate the utility of functional segments for subjectivity classification, we perform a set of experiments based on pointwise mutual information (PMI) (Church and Hanks, 1989; Church and Hanks, 1990) feature selection focused at these segments. PMI is a statistical measure of the co-occurrence of two events that captures the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions. The PMI between a functional segment 'fs' and its class 'c' (e.g., the OBJ vs. the SUBJ class) is:

$$PMI(fs|c) = \log_2 \frac{p(fs|c)}{p(fs)p(c)} \qquad (1)$$

PMI is a filter feature selection method that is used to keep only features important to the classification process and filter out the rest. In the current case, only functional segments that occur in one of the two OBJ and SUBJ classes more than the other with a relative frequency (RF) that is $\geq$ a certain threshold are kept for classification (in addition to the content segments) while the rest are filtered out. Experiments with PMI are run with RF different threshold parameters from the set $\{1, 2, 3, 4\}$. It is found that with an RF threshold ≥ 1, PMI feature selection results in improvements over all previous settings whether gold (i.e., gold-segs and gold-cont) or ASMA-predicted (i.e., asma-segs and asma-cont). Table 3 shows related results. The experiments with only certain functional segments filtered out with gold segmentation are referred to as gold-cont-M1 and those with ASMA are referred to as asma-cont-M1, with the suffix 'M1' standing for 'modified' with a threshold of is ≥ 1 in both cases.

As Table 3 shows, modified lexical filtering helps improve classification across the board over surf, over comparable lexical filtering (i.e., gold-cont and asma-cont), and over segmentation settings (i.e., gold-segs and asma-segs). On DEV,

gold-cont-M1 achieves identical accuracy scores (i.e., 68.28%) as gold-segs. On TEST, gold-cont-M1 improves 1.12% accuracy on gold-segs. For ASMA-predicted settings, asma-cont-M1 improves over asma-segs (with 0.92% accuracy on DEV and 1.34% accuracy on TEST). Results of modified lexical filtering show that some, but not all, functional segments are important for subjectivity classification. In addition to the linguistic analysis provided earlier in this section about the importance of some of the functional segments, it is worth mentioning that filtering out all functional segments also deprives the classifier of any access to multiword expressions (MWE). Although the sentences in the experiments reported above are not represented beyond unigrams and hence MWEs are not explicitly provided to the classifier, there is still a possibility for the classifier to benefit from these expressions when all segments in a sentence are accessible. The following is an example of an MWE carrying SUBJ content and explanations of accompanying filtered out functional segments: In the phrase *wqf fy wjh* (Eng. 'he stood against') removal of functional segment preposition *fy* (Eng. 'face') results in the string *wqf+wjh* (Eng. 'he stood+face'). Again, these resulting CONT segments do not carry SUBJ content themselves.

3.3 Sentiment with Lexical Filtering

Table 5 shows results of sentiment classification with both gold lexical filtering (gold-cont) and ASMA lexical filtering (asma-cont). For comparison, earlier results with the segmented unfiltered setting (gold-segs and gold-segs) and results with syntactically motivated feature selection (which we refer to as *gold-cont-M1* and *asma-cont-M1*), as is explained below, are also provided in Table 5. As the Table shows, on both DEV and TEST, syntactically motivated lexical filtering (whether gold-cont or asma-cont) improves classification over surf. On DEV, gold-cont improves 1.98% accuracy and asma-segs improves 3.29% accuracy over surf. On TEST, gold-cont improves 4.51% accuracy and asma-segs improves 2.71% accuracy over surf. Comparing gold-cont to segmented text, however, shows a trend similar to that of subjectivity classification where segmentation without lexical filtering is still competitive: On DEV, gold-cont and asma-cont are both outperformed by segmented text (i.e., both gold-segs and asma-segs). On TEST, gold-cont outperforms gold-segs. These results show that although lexical filtering is able to outperform surf, it is not consistently useful compared to segmented text. A consideration of the data sparsity in DEV and TEST indicates that there is no consistent correlation between the percentage of OOV segments and performance. For example, although gold-cont has less OOV percentage than gold-segs and asma-segs on DEV, it is still outperformed by these two settings.

Similar to subjectivity classification with lexical filtering, we performed an analysis of the relative frequency (RF) of functional segments as occurring in segmented text across the S-POS and S-NEG classes as is reported in Table 5. The analysis shows a similar trend to that of subjectivity classification where functional segments have different RF distribution across the two classes across both the gold-segs and asma-segs settings.

In order to further investigate the utility of functional segments for sentiment classification, again, we perform a set of experiments based on PMI (Church and Hanks 1989; 1990) feature selection focused at these segments. Similar to subjectivity classification above, all functional segments that occurred in one or another of the two S-POS and S-NEG classes with a PMI value more than that of the other (i.e., with a relative frequency of $\geq$ a threshold from the set $\{1, 2, 3, 4\}$) are kept unfiltered for this set of experiments. The best results are achieved with the RF ≥ 1. As is reported in Table 5, with this modified lexical filtering (gold-cont-M1 and asma-cont-M1), an improvement is gained on DEV as compared to surf, segmented text settings (gold-segs and asma-segs), and lexical filtering (gold-cont and asma-cont). The case is different, however, on TEST where no such improvement is possible with the modified lexical filtering settings. Comparing the performance of modified lexical filtering in the case of subjectivity classification as presented earlier to the current performance of modified lexical filtering on sentiment classification shows a discrepancy of the utility of modified/partial lexical filtering. This is the case since the two classification tasks are different, as expression of sentiment itself is different from that of subjectivity. Functional segments are associated with subjective content in general regardless of the specific type of sentiment expressed, for which case these segments do not play as much of a role in distinguishing the S-POS from the S-

				S-POS			S-NEG		
		Acc	**Avg- F**	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
DEV	surf	57.89	57.71	65.33	56.32	60.49	50.65	60	54.93
	gold-segs	65.13	64.45	69.77	68.97	69.36	59.09	60	59.54
	asma-segs	61.84	61.52	68.35	62.07	65.06	54.79	61.54	57.97
	gold-cont	59.87	59.86	69.12	54.02	60.65	52.38	67.69	59.06
	asma-cont	61.18	61.19	71.21	54.02	61.44	53.49	70.77	60.93
	gold-cont-M1	**66.45**	**66.11**	**72.5**	**66.67**	**69.46**	**59.72**	**66.15**	**62.77**
	asma-cont-M1	61.84	61.81	71.01	56.32	62.82	54.22	69.23	60.81
TEST	surf	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**
	gold-cont	69.37	69.17	73.91	60.71	66.67	66.15	78.18	71.67
	asma-cont	67.57	67.35	71.74	58.93	64.71	64.62	76.36	70
	gold-cont-M1	64.86	64.85	64.91	66.07	65.49	64.81	63.64	64.22
	asma-cont-M1	63.96	63.97	64.29	64.29	64.29	63.64	63.64	63.64

Table 5: Sentiment classification with syntactically motivated feature selection.

NEG classes as they do in distinguishing the OBJ and SUBJ classes. What supports this analysis is the fact that although the distribution of functional segments differs from one class to another, this distribution is not as pronounced with the RF values 2, 3, and 4 as in the case of subjectivity classification. For example, while on TRAIN and DEV combined 58.785% of gold-segs occur with an RF=2 in either the SUBJ or the OBJ classes, only 43.155% of these with the same RF=2 occur in either the S-POS or S-NEG classes (also as derived from TRAIN and DEV combined). The situation is also similar with ASMA functional segments over TRAIN+DEV, where 56.075% occur with an RF=2 in one or the other of subjectivity classes whereas only 42.475% of them occur with the same RF threshold in one or the other of the two sentiment classes.

In order to better understand why it is that full gold lexical filtering yields lower performance than gold segmented data, we perform an error analysis of the gold-cont cases (n=26) that are correctly classified by the gold-segs classifier on the DEV set. The error analysis shows that there is a host of linguistic phenomena that interact with expression of sentiment as follows:

Negation: Negation particles belong to functional segments and can cause a change of the polarity of content segments. When negation particles are absent, and hence not accessible to the classifier unlike content segments that potentially carry the opposite of the label of a sentence, classification errors occur.

Interactional resources: Only one category of interactional resources (Hyland Tse, 2004) is found to be most important to sentiment expression: *First person pronouns*. First person singular, but more frequently, plural pronouns are found to be used with higher distribution with the S-POS class. This is specially the case since the data involve discourse where politicians do their best trying to draw a positive image of themselves and/or the political entities they represent and hence cite self. This is an example from the error analysis data: *qAl Alr}ys AlsngAly: 'lA xyAr OmAmnA swy AltjmE'*. (Eng. 'The Senegalese president said: "we have no other choice but uniting".')

The example is human-annotated with S-POS and involves first person plural pronouns (e.g., the possessive pronoun *nA* [Eng. 'our'], the imperfective verb prefixal *n-* [Eng. 'we']) that is filtered out with the lexical filtering setting (both gold-cont and asma-cont) and hence the classifiers do not have access to these as signals of positive sentiment, which results in the erroneous prediction.

The finding that full lexical filtering improves over segmented text on TEST but not on DEV is one that also calls for further investigation. An error analysis of the examples (n=18) gold-cont correctly identifies but gold-segs fails to predict on DEV was performed to better interpret this finding. Among these 18 examples (%=77.77) are human-labeled as S-NEG and hence the gold-cont classifier performs better on the S-NEG class.

This indicates that expression of negative content is more likely to be carried by content segments rather than a combination of both functional and content segments. The addition of certain functional segments (i.e., those that occur with higher RF with the S-POS class) is responsible for misclassification errors. For example, since the first personal plural pronouns mentioned above (i.e., the possessive pronoun *nA* [Eng. 'our'], the imperfective verb prefixal *n-* [Eng. 'we']) are more frequently occurring with the S-POS class, they contribute to causing the gold-segs classifier assigning an S-POS tag to the following S-NEG sentence that was rightly predicted with the gold-cont setting: *mqr bOnh 'ElynA On nEtrf bOn bED jwAnb AlmEAhd p lyst wADHp'*. (Eng. 'Attesting that "we must admit that some aspects of the treaty are not clear".') *-nA* and *n-* have higer RF with the POS class.

A close investigation of the examples wrongly classified by gold-segs also shows that they all carry (very) strong sentiment. Although sentiment intensity is not manually labeled in the data and the current work does not involve predicting degrees of intensity in data, it is worth discussing this aspect as it relates to the current error analysis. Sentiment intensity in Arabic is primarily expressed via content segments and hence these content segments, rather than functional segments, are the important signals in (very) strongly polarized examples. The following example that carries a strong negative sentiment illustrates this point: *wtSwrhA bEbE yEml ElY IhAnp AlnAs wAl-IsA'p IlY Alm$trkyn.* (Eng. 'And portrays it as a monstrous ghost working to humiliate people and wronging participants.'). In this last sentence, the strong sentiment is carried by the content segments *bEbE* (Eng. 'monstrous ghost') and *IsA'* (Eng. 'wronging-related') rather than by any functional segments. This utility of content segments in expressing strong sentiment makes them more crucial for the task and adding functional segments may be 'distractive' to the classifier especially in S-NEG examples as the comparison between the performance of gold-cont and gold-segs shows.

4 Related Work

Sentiment analysis has been a popular NLP task, with a lot of work focused at mining movie and product reviews (Dave et al., 2003; Hu and Liu, 2004; Turney, 2002). Recently, there has been a number of SemEval tasks devoted to sentiment (Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016).

For Arabic, early work includes (Abbasi et al., 2008) who detect hostility in Arabic and English web fora and (Abdul-Mageed et al., 2011) who use gold-labeled morphological features and a polarity lexicon from the news domain. This current work differs in that we use automatically predicted morphological features in addition to gold features. (Abdul-Mageed et al., 2014) is also related to our work in that the we also investigate ways to best represent lexical information, yet on newswire data rather than social media. A number of studies have reported models using n-gram features after preprocessing input data (Abdulla et al., 2013; Aly and Atiya, 2013; ElSahar and El-Beltagy, 2015; Mourad and Darwish, 2013; Saleh et al., 2011). The focus of our work is different in that we seek to break the space of lexical input based on syntactic criteria and introduce a method to weigh the informativity of the resulting spaces via feature selection. We also have shown how linguistic phenomena interact with sentiment expression via detailed error analyses of model output.

5 Conclusion

In this work, we introduced a new human-labeled ATB dataset for SSA and investigated ways to model subjectivity and sentiment in lexical space in Arabic, a language of rich morphology. We demonstrated how each of these tasks can be performed with both gold and machine-predicted segments under different grammar-based conditions. Our results show that not all lexical input is relevant to the tasks and that some syntactically-defined segments are more relevant to a given task than another, thus motivating our syntactically motivated feature selection method. We found functional segments to be more vehicles for carrying subjective content than devices for communicating positive and negative content. Our detailed error analyses helped uncover a host of linguistic phenomena that interact in intricate ways with both subjectivity and sentiment expression in the Arabic newswire genre. Our 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 knowledge at lower linguistics levels in lexical space.

References

A. Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26:1–34.

Muhammad Abdul-Mageed and Mona Diab. 2011. Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 110–118, Portland, Oregon, USA, June. Association for Computational Linguistics.

Muhammad Abdul-Mageed and Mona Diab. 2012. *AWATIF*: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of LREC*, volume 12.

Muhammad Abdul-Mageed, Mona Diab, and Mohamed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Mona T Diab, and Sandra Kübler. 2013. Asma: A system for automatic segmentation and morpho-syntactic disambiguation of modern standard arabic. In *RANLP*, pages 1–8.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Muhammad Abdul-Mageed. 2017. Modeling subjectivity and sentiment in lexical space. In *Submitted*.

Nawaf Abdulla, N Mahyoub, M Shehab, and M Al-Ayyoub. 2013. Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

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

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

P. Brown and S.C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge Univ Pr.

Pimwadee Chaovalit and Lina Zhou. 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE.

K.W. Church and P. Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the ACL*, pages 76–83. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

K. Dave, S. Lawrence, and D.M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Ken Hyland and Polly Tse. 2004. Metadiscourse in academic writing: A reappraisal. *Applied linguistics*, 25(2):156–177.

B. Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 978–1420085921.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval*, pages 1–18.

F. Palmer. 1986. *Mood and Modality. 1986*. Cambridge: Cambridge University Press.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Eshrag Refaee and Verena Rieser. 2014. An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Eshrag Refaee and Verena Rieser. 2016. ilab-edinburgh at semeval-2016 task 7: A hybrid approach for determining sentiment intensity of arabic twitter phrases. *Proceedings of SemEval*, pages 474–480.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 73–80. Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777.

M Rushdi Saleh, Maria Teresa Martín-Valdivia, Arturo Montejo-Ráez, and LA Ureña-López. 2011. Experiments with svm to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804.

R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kuebler, Y. Versley, M. Candito, J. Foster, I. Rehbein, and L. Tounsi. 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.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

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

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM.