Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs

Ahmed Mourad and Kareem Darwish
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
{amourad, kdarwish}@qf.org.qa

Abstract

Though much research has been conducted on Subjectivity and Sentiment Analysis (SSA) during the last decade, little work has focused on Arabic. In this work, we focus on SSA for both Modern Standard Arabic (MSA) news articles and dialectal Arabic microblogs from Twitter. We showcase some of the challenges associated with SSA on microblogs. We adopted a random graph walk approach to extend the Arabic SSA lexicon using Arabic-English phrase tables, leading to improvements for SSA on Arabic microblogs. We used different features for both subjectivity and sentiment classification including stemming, part-of-speech tagging, as well as tweet specific features. Our classification features yield results that surpass Arabic SSA results in the literature.

1 Introduction

Subjectivity and Sentiment Analysis has gained considerable attention in the last few years. SSA has many applications ranging from identifying consumer sentiment towards products to voters' reaction to political adverts. A significant amount of work has focused on analyzing English text with measurable success on news articles and product reviews. There has been recent efforts pertaining to expanding SSA to languages other than English and to analyzing social text such as tweets. To enable effective SSA for new languages and genres, two main requirements are necessary: (a) subjectivity lexicons that broadly cover sentiment carrying words in the genre or language; and (b) tagged corpora to train subjectivity and sentiment classifiers. These two are often scarce or nonexistent when expanding to new languages or genres. In this paper we focus on performing SSA on Arabic news articles and microblogs. There has been some recent work on Arabic SSA. However, the available resources continue to lag in the following ways:
(1) The size of existing subjectivity lexicons is small, with low coverage in practical application.
(2) The available tagged corpora are limited to the news domain, with no publicly available tagged corpora for tweets.
To address the issue of limited lexicons, we applied two methods to build large coverage lexicons. In the first, we used Machine Translation (MT) to translate an existing English subjectivity lexicon. In the second, we employed a random graph walk method to automatically expand a manually curated Arabic lexicon. For the later method, we used Arabic-English MT phrase tables that include both Modern Standard Arabic (MSA) as well as dialectal Arabic. As for tagged corpora, we annotated a new corpus that includes 2,300 Arabic tweets. We describe in detail the process of collecting tweets and some of the major attributes of tweets.

The contribution of this paper is as follows:
- We introduce strong baselines that employ Arabic specific processing including stemming, POS tagging, and tweets normalization. The baseline outperforms state-of-the-art subjectivity classification for the news domain.
- We provide a new annotated dataset for Arabic tweet SSA.
- We employ a random graph walk algorithm to ex-

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 55–64,
Atlanta, Georgia, 14 June 2013.

pand SSA lexicons, leading to improvements for SSA for Arabic tweets.

The remainder of this paper is organized as follows: Section 2 surveys related work; section 3 introduces some of the challenges associated with Arabic SSA; section 4 describes the lexicons we used; section 5 presents the experimental setup and results; and section 6 concludes the paper and discusses future work.

2 Related Work

There has been a fair amount work on SSA. Liu (2010) offers a thorough survey of SSA research. He defines the problem of sentiment analysis including associated SSA terms such as object, opinion, opinion holder, emotions, sentence subjectivity, etc. He also discusses the more popular two stage sentiment and subjectivity classification approach at different granularities (document and sentence levels) using different machine learning approaches (supervised and unsupervised) along with different ways to construct the required data resources (corpora and lexicon). In our work, we classify subjectivity and sentiment in a cascaded fashion following Wilson et al. (2005).

2.1 Subjectivity Analysis

One of most prominent features for subjectivity analysis is the existence of words in a subjectivity lexicon. Mihalcea et al. (2007) translated an existing English subjectivity lexicon from Wiebe and Riloff (2005) using a bilingual dictionary. They also used a subjectivity classifier to automatically annotate the English side of an English-Romanian parallel corpus and then project the annotations to the Romanian side. The projected annotations were used to train a subjectivity classifier. In follow on work, Banea et al. (2010) used MT to exploit annotated SSA English corpora for other languages, including Arabic. They also integrated features from multiple languages to train a combined classifier. In Banea et al. (2008), they compared the automatic annotation of non-English text that was machine translated into English to automatically or manually translating annotated English text to train a classifier in the target language. In all these cases, they concluded that translation can help avail the need for building language specific resources. In performing both subjectivity and sentiment classification, researchers have used word, phrase, sentence, and topic level features. Wilson et al. (2005) report on such features in detail, and we use some of their features in our baseline runs. For Arabic subjectivity classification, Abdul-Mageed et al. (2011) performed sentence-level binary classification. They used a manually curated subjectivity lexicon and corpus that was drawn from news articles (from Penn Arabic tree bank). They used features that are akin to those developed by Wilson et al. (2005). In later work, Abdul-Mageed et al. (2012) extended their work to social content including chat sessions, tweets, Wikipedia discussion pages, and online forums. Unfortunately, their tweets corpus is not publicly available. They added social media features such as author information (person vs. organization and gender). They also explored Arabic specific features that include stemming, POS tagging, and dialect vs. MSA. Their most notable conclusions are: (a) POS tagging helps and (b) Most dialectal Arabic tweets are subjective. Concerning work on subjectivity classification on English tweets, Pak and Paroubek (2010) created a corpus of tweets for SSA. They made a few fundamental assumptions that do not generalize to Arabic well, namely:

- They assumed that smiley and sad emoticons imply positive and negative sentiment respectively. Due to the right-to-left orientation of Arabic text, smiley and sad emoticons can be easily interchanged by mistake in Arabic.
- They also assumed that news tweets posted by newspapers Twitter accounts are neutral. This assumption is not valid for Arabic news articles because many Arabic newspapers are overly critical or biased in their reporting of news. Thus, the majority of news site tweets have sentiment. Consider the following headline:

اللّجنة الدينية تهاجم استمرار تحكم أمن الدولة في تعيين أئمة المساجد

meaning: Religious Council critical of State Security over interference in hiring of clerics.

- They constructed their tweet sets to be uniformly distributed between subjective and objective classes. However, our random sample of Arabic tweets showed that 70% of Arabic tweets are subjective. So this kind of training is misleading especially for a Naïve Bayesian classifier that utilizes the prior probability of classes.

2.2 Sentiment Analysis

Abbasi et al. (2008) focused on conducting sentiment classification at document level. They used

syntactic, stylistic, and morphological (for Arabic) features to perform classification. Abdul-Mageed et al. (2011) performed sentence-level sentiment classification for MSA. They concluded that the appearance of a positive or negative adjective, based on their lexicon, is the most important feature. In later work, Abdul-Mageed et al. (2012) extended their work to social text. They concluded that: (a) POS tags are not as effective in sentiment classification as in the subjectivity classification, and (b) most dialectal Arabic tweets are negative. Lastly, they projected that extending/adapting polarity lexicon to new domains; e.g. social media; would result in higher gains. Kok and Brockett (2010) introduced a random-walk-base approach to generate paraphrases from parallel corpora. They proved to be more effective in generating more paraphrases by traversing paths of lengths longer than 2. El-Kahky et al. (2011) applied graph reinforcement on transliteration mining problem to infer mappings that were unseen in training. We used this graph reinforcement method in our work.

3 Challenges for SSA of Arabic

Arabic SSA faces many challenges due to the poorness of language resources and to Arabic-specific linguistic features.

Lexicon: Lexicons containing words with prior polarity are crucial feature for SSA. The most common English lexicon that has been used in literature is the Multi-Perspective Question Answering (MPQA) lexicon, which contains 8,000 words. Some relied on the use of MT to translate English lexicons to languages that lack SSA resources (Mihalcea et al., 2007). A lexicon that is translated into Arabic may have poor coverage due to the morphological and orthographic complexities of Arabic. Arabic nouns and verbs are typically derived from a set of 10,000 roots that are cast into stems using templates that may add infixes, double letters, or remove letters. Stems can accept the attachment of prefixes or suffixes, such as prepositions, determiners, pronouns, etc. The number of possible Arabic surface forms is in the order of billions. In this work, we employed stemming and graph reinforcement to improve the converge of lexicons.

Negation: Negation in dialects can be expressed in many ways. In MSA, the word ليس (meaning "not") is typically used to negate adjectives. Dialects use many words to negate adjectives including:ماهو, مش, مو, ما, منو, etc. These words can have other meanings also. For example, ماهو also means "what is". As for verbs, some dialects like Egyptian and Levantine use a negation construct akin to the "ne ... pas" construct in French. All these make detecting negation hard. We use word n-gram features to overcome this problem.

Emoticons: Another challenge has to do with the limited usefulness of emoticons, because Arabic's smileys and sad emoticons are often mistakenly interchanged. Thus, many tweets have words and emoticons that are contradictory in sentiment. For example:

بسم الله عليك من الألم :(

meaning: with the help of God over your pain (positive) : followed by a sad face

أنا عندي أخت حسبي الله عليها :)

meaning: I have a sister from which I seek the protection of Allah (negative) : followed by a smilie

Use of dialects: Though most Arabic speakers can read and understand MSA, they generally use different Arabic dialects in their daily interactions including online social interaction [1]. There are 6 dominant dialects, namely Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni. Dialects introduce many new words into the language, particularly stopwords (ex. ماحد and شنو mean "no one" and "what" respectively). Dialects lack spelling standards (ex. ماعرفتش and معرفتش are varying spellings of "I did not know" in Egyptian). Different dialects make different lexical choices for concepts (ex. باهي and صافي mean "good" in Morrocan and Libyan respectively). Due to morphological divergence of dialectal text from MSA, word prefixes and suffixes could be different. For example, Egyptian and Levantine tend to insert the letter ب ("ba") before verbs in present tense. Building lexicons that cover multiple dialects is cumbersome. Further, using MT to build SSA lexicons would be suboptimal because most MT systems perform poorly on dialects of Ara-

[1] http://en.wikipedia.org/wiki/Varieties_of_Arabic

bic.

Tweet specific phenomena: Tweets may contain transliterated words ("LOL" → لول) and non-Arabic words, particularly hashtags such as #syria. Tweets are often characterized by the informality of language and the presence of name mentions (@user_mention), hashtags, and URL's. Further, tweets often contain a significant percentage of misspelled words.

Contradictory language: Often words with negative sentiment are used to express positive sentiment:
تتظاهر الأنثي بالبرود وعدم الاهتمام بك، وتبدء بقول

ألفاظ قد توءلمك .. اعلم أنها كانت تعشقك الي حد الألم
meaning: a female pretends to be cold and uninterested and may even use hurtful words. Know that she painfully loves you.

Other observations: We also observed the following:
- Users tend to express their feelings through extensive use of Quranic verses, Prophetic sayings, proverbs, and poetry.
- Of the annotated tweets in our corpus, nearly 13.5% were sarcastic.
- People primarily use tweets to share their thoughts and feelings and to report facts to a lesser extent. In the set we annotated, 70% of the tweets were subjective and 30% were objective. Of the subjective tweets (positive and negative only), the percentage of positive tweets was 66% compared to 34% for negative tweets.

4 SSA Lexicon

We employed two lexicons that were available to us, namely:
- The MPQA lexicon, which contains 8,000 English words that were manually annotated as strong subjective (subjective in most contexts) or weak subjective (subjective in some contexts) and with their prior polarity (positive, negative, neutral, or both). We used the Bing online MT system [2] to translate the MPQA lexicon into Arabic.
- The ArabSenti lexicon (Abdul-Mageed et al., 2011) containing 3,982 adjectives that were extracted from news data and labeled as positive, negative, or neutral. We optionally used graph reinforcement to expand the ArabSenti lexicon using MT phrase tables, which were modeled as a bipartite graph (El-Kahky et al., 2011). As shown in Figure 1, given a seed lexicon, graph reinforcement is then used to enrich the lexicon by inferring additional mappings. Specifically, given the word with the dotted outline, it may map to the words "unfair" and "unjust" in English that in turn map to other Arabic words, which are potentially synonymous to the original word. We applied a single graph reinforcement iteration over two phrase tables that were generated using Moses (Koehn et al., 2007). The two phrase tables were:

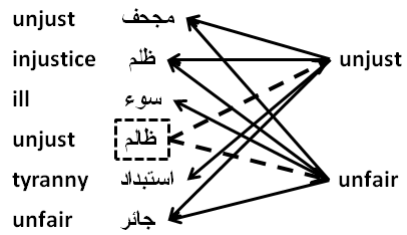

Figure 1: Example mappings seen in phrase table

- an **English-MSA** phrase table, which was trained on a set of 3.69 million parallel sentences containing 123.4 million English tokens. The sentences were drawn from the UN parallel data along with a variety of parallel news data from LDC and the GALE project. The Arabic side was stemmed (by removing just prefixes) using the Stanford word segmenter (Green and DeNero, 2012).
- an **English-Dialect** phrase table, which was trained on 176K short parallel sentences containing 1.8M Egyptian, Levantine, and Gulf dialectal words and 2.1M English words (Zbib et al., 2012). The Arabic side was also stemmed using the Stanford word segmenter.

More formally, Arabic seed words and their English translations were represented using a bipartite graph G = (S, T, M), where S was the set of Arabic words, T was the set of English words, and M was the set of mappings (links or edges) between S and T. First, we found all possible English translations $T' \subseteq T$ for each Arabic word $s_i \subseteq S$ in the seed lexicon. Then, we found all possible Arabic translations $S' \subseteq S$ of the English translations T'. The mapping score $m(s_j \subseteq S'|s_i)$ would be computed

[2] http://www.bing.com/translator/

as:

$$1 - \prod_{\forall s_j, s_i \in S, t \in T'} (1 - \frac{p(t|s_i)}{\sum_t p(s_i|t)} \frac{p(s_j|t)}{\sum_{s_j} p(t|s_j)}) \quad (1)$$

where the terms in the denominator are normalization factors and the product computes the probability that a mapping is not correct given all the paths from which it was produced. Hence, the score of an inferred mapping would be boosted if it was obtained from multiple paths, because the product would have a lower value.

5 Experimental Setup

5.1 Corpus, Classification, and Processing

For subjectivity and sentiment classification experiments on Arabic MSA news, we used the translated MPQA dataset and the ArabSenti dataset respectively. As for SSA on Arabic tweets, to the best of our knowledge, there is no publicly available dataset. Thus, we built our own. We crawled Twitter using the Twitter4j API (Yanamoto, 2011) using the query "lang:ar" to restrict tweets to Arabic ones only. In all, we collected 65 million unique Arabic tweets in the time period starting from January to December 2012; we made sure that duplicate tweets were ignored during crawling. Then we randomly sampled 2300 tweets (nearly 30k words) from the collected set and we gave them to two native Arabic speakers to manually annotate. If the two annotators disagreed on the annotation of a tweet, they discussed it to resolve the disagreement. If they couldn't resolve the disagreement, then the tweet was discarded, which would somewhat affect the SSA effectiveness numbers. They applied one of five possible labels to the tweets, namely: neutral, positive, negative, both, or sarcastic. For subjectivity analysis, all classes other than neutral were considered subjective. As for sentiment analysis, we only considered positive and negative tweets. For both subjectivity and sentiment classification experiments, we used 10-fold cross validation with 90/10 training/test splits. We used the NLTK (Bird, 2006) implementation of the Naïve Bayesian classifier for all our experiments. In offline experiments, the Bayesian classifier performed slightly better than an SVM classifier. The classifier assigned a sentence or tweet the class $c \in C$ that maximizes:

$$\underset{c \in C}{argmax} P(c) \prod_{i=1}^{n} P(f_i|c) \quad (2)$$

where f is the feature vector and C is the set of pre-defined classes. As for stemming and POS Tagging, we used an in-house reimplementation of AMIRA (Diab, 2009). We report accuracy as well as precision, recall and F-measure for each class.

5.2 Baseline: SSA for MSA

5.2.1 Subjectivity Classification

As mentioned in section 2, we employed some of the SSA features that were shown to be successful in the literature (Wiebe and Riloff, 2005; Wilson et al., 2005; Yu and Hatzivassiloglou, 2003) to construct our baseline objective-subjective classifier. We used the automatically translated MPQA and the ArabSenti lexicons. We tokenized and stemmed all words in the dataset and the lexicon. Part of the tokenization involved performing letter normalization where the variant forms of alef (آ, أ, and إ) were normalized to the bare alef (ا), different forms of hamza (ؤ and ىء) were normalized to hamza (ء), ta marbouta (ة) was normalized to ha (ه), and alef maqsoura (ى) was normalized to ya (ي). We used the following features:

Stem-level features:

- *Stem* is a binary features that indicates the presence of the stem in the sentence.
- *Stem prior polarity* as indicated in the translated MPQA and ArabSenti lexicons (positive, negative, both or neutral). Stems and their prior polarity were reportedly the most important features in Wilson et al. (2005).
- *Stem POS*, which has been shown to be effective in the work done by (Wiebe and Riloff, 2005; Yu and Hatzivassiloglou, 2003). Although Abdul-Mageed et al. (2011) used a feature to indicate if a stem is an adjective or not, other tags, such as adverbs, nouns, and verbs, may be good indicators of sentiment. Thus, we used a feature that indicates the POS tag of a stem as being: adjective, adverb, noun, IV, PV, or other, concatenated with the stem. For example, the stem "play" may be assigned "play-noun" if it appears as a noun in a sentence. We chose this reduced POS set based on the frequency distribution

	Acc	Prec		Rec		F-Meas	
		Obj	Subj	Obj	Subj	Obj	Subj
Banea et al. (2010)	72.2	72.6	72.0	60.8	81.5	66.2	76.4
Baseline-MPQA	77.2	83.4	74.2	61.4	90.0	70.7	81.4
Baseline-ArabSenti	76.7	82.4	73.9	60.9	89.5	70.0	80.9
Expanded-ArabSenti-MSA	76.7	83.2	73.6	60.0	90.2	69.7	81.0
Expanded-ArabSenti-MSA+Dialect	76.7	82.9	73.7	60.4	89.9	69.9	81.0

Table 1: Baseline Results for MSA Subjectivity Classifier.

	Acc	Prec		Rec		F-Meas	
		Pos	Neg	Pos	Neg	Pos	Neg
Baseline-MPQA	80.6	75.4	84.0	78.0	82.5	76.5	83.2
Baseline-ArabSenti	80.5	75.4	84.6	78.6	81.5	76.8	82.9
Expanded-ArabSenti-MSA	80.0	74.9	83.9	77.8	81.4	76.2	82.6
Expanded-ArabSenti-Dialect	79.2	73.7	82.8	76.0	81.2	74.6	81.9

Table 2: Baseline Results for MSA Polarity Classifier.

of POS tags and subjectivity classes in the training data.

- *Stem context* as the stem bi-gram containing the stem along with the previous stem. We experimented with higher order stem n-grams, but bigrams yielded the best results.

Sentence features: These features have been shown to be effective by Wiebe and Riloff (2005). They include:

- *Counts of stems belonging to so-called reliability classes* (Wiebe and Riloff, 2005), which are basically either strong-subjective and weak-subjective tokens (as indicated in the SSA lexicon).
- *Counts of POS tags* where we used the counts of the POS tags that used for stem features (adjective, adverb, noun, IV, and PV).

We compared our baseline results with the results reported by Banea et al. (2010) for Arabic subjectivity classification. We used their Arabic MPQA corpus that has been automatically translated from English and then projected subjectivity labels with the same training/test splits. The 9,700 sentences in this corpus are nearly balanced with a 55/45 subjective/objective ratio. Table 1 shows the results for MSA subjectivity classification compared to the results of Banea et al. (2010). Our baseline system improved upon the results of Banea et al. (2010) by 5% (absolute) in accuracy with significant gains in both precision and recall. Using MPQA or ArabSenti lexicons yielded comparable results with MPQA yielding marginally better results. We think that much of improvement that we achieve over the results of Banea et al. (2010) could be attributed to stemming and POS tagging.

5.2.2 Polarity Classification

For polarity classification experiments, we used the positive and negative sentences from the ArabSenti dataset (Abdul-Mageed and Diab, 2011). From the 2,855 sentences in ArabSenti, 45% were objective, 17.2% were positive, 24.1% were negative and the rest were both. We employed the following features:

Stem-level features:

- *Stem*, *Stem prior polarity*, and *Stem POS tag* as in subjectivity classification
- *Stem context* where we considered a stem and the two preceding stems. In offline experiments, we tried looking at more and less context and using the two previous stems yielded the best results. The intuition to use stem context is to compensate for the difficulties associated with 'negation' in Arabic (as mentioned earlier section 3).

Sentence-level features: We used only one binary feature that checks for the occurrence of positive adjectives in the sentence. We experimented with other features that aggregate other POS tags with their prior polarity including negative adjectives and all led to worse classification results.

Table 2 reports on the baseline results of doing sentiment classification. The results of using either MPQA or ArabSenti lexicons were comparable.

	Acc	Prec		Rec		F-Meas	
		Obj	Subj	Obj	Subj	Obj	Subj
Baseline-Majority-Class	70.0	0.0	70.0	0.0	100.0	0.0	83.0
Baseline-MSA	55.1	53.8	56.4	54.5	55.8	54.1	56.1
Baseline-MPQA	64.8	44.9	81.4	66.5	64.0	53.5	71.5
Baseline-ArabSenti	63.9	43.8	80.8	65.9	62.9	52.5	70.7
Expanded-ArabSenti-MSA	64.1	44.2	81.1	66.3	63.3	52.8	71.0
Expanded-ArabSenti-Dialect	63.1	43.2	80.3	65.5	62.1	51.9	70.0

Table 3: Baseline Results for Arabic Tweets Subjectivity Classifier.

	Acc	Prec		Rec		F-Meas	
		Pos	Neg	Pos	Neg	Pos	Neg
Baseline-MSA	54.8	63.2	45.7	55.5	53.8	59.1	49.4
Baseline-MPQA	72.2	85.9	57.0	69.0	77.8	76.3	65.5
Baseline-Arabsenti	71.1	83.9	55.9	69.2	74.8	75.8	63.8
Expanded-ArabSenti-MSA	72.5	86.1	57.7	69.1	79.3	76.5	66.4
Expanded-ArabSenti-Dialect	71.3	85.5	56.3	68.0	77.8	75.6	65.1

Table 4: Baseline Results for Arabic Tweets Polarity Classifier.

5.3 Baseline: SSA of Arabic Microblogs

5.3.1 Subjectivity Classification

We have four baselines for subjectivity classification of Arabic tweets, namely:

Baseline-Majority-Class for which we considered all the tweets to be subjective, where "subjective" was the majority class.

Baseline-MSA for which we used the aforementioned MSA subjectivity classifier using the MPQA lexicon (section 5.2).

Baseline-MPQA and **Baseline-ArabSenti** for which we used microblog specific features and the MPQA and ArabSenti lexicons respectively. We used the following features:

Stem-level features:

- *Stems*, where we normalized words using the scheme described by Darwish et al. (2012). Their work extended the basic Arabic normalization to handle non-Arabic characters that were borrowed from Farsi and Urdu for decoration decorate and words elongation and shortening. After normalization, words were stemmed.
- *MSA or dialect*, which is a binary feature that indicates whether the stem appears in a large MSA stem list (containing 82,380 stems) which was extracted from a large Arabic news corpus from Aljazeera.net.
- *Stem prior polarity* and *Stem POS* as those for MSA subjectivity classification.

Tweets-specific features: Following Barbosa and Feng (2010) and Kothari et al. (2013), we took advantage of tweet specific features, namely:

- Presence of hashtag (#tag).
- Presence of user mention (@some_user) and position in the tweet (start, end and middle).
- Presence of URL and position in the tweet (start, end and middle).
- Presence of retweet symbol "RT" and position in the tweet (start, end and middle).

"RT" and URL's usually appear in the beginning and end of tweets respectively, particularly when retweeting news articles. A change in their position may indicate that the person retweeting added text to the tweet, often containing opinions or sentiment.

Language-independent features: These are binary features that look for non-lexical markers that may indicate sentiment. They are:

- Usage of decorating characters. e.g. گ instead of ك.
- Elongation (detecting both repeated uni-gram & bi-gram character patterns. e.g. لووووول (looool), هاهاها (hahaha).
- Punctuation; exclamation and question marks.
- Elongated punctuation marks (e.g. ???, !!!!!)
- Emoticons (e.g. :), :(, :P ... etc.).

Sentence-level features: We used the counts of so-called reliability classes, which count the number of strong-subjective and weak-subjective words.

Table 3 shows the results for subjectivity analysis on tweets. Baseline-Majority-Class was the best given that most Arabic tweets were subjec-

tive. Tweet-specific features were not discriminative enough to outperform Baseline-Majority-Class. Thus, assuming that all tweets are subjective seems to be the most effective option. However, it is worth noting that using a classifier that was trained on dialectal tweets yielded better results than using a classifier that was trained on news in MSA. Again using either lexicon made little difference.

5.3.2 Polarity Classification

Our work on MSA showed that *stem* and *stem prior polarity* are the most important features for this task. We used these two features, and we added a third binary feature that indicates the presence of positive emoticons. Negative emoticons appeared infrequently in both training and test sets. Hence using a feature that indicates the presence of negative emoticons would be unreliable. Again we used the MPQA or ArabSenti lexicons, both of which were constructed from news domain (**Baseline-MPQA** and **Baseline-ArabSenti** respectively). For reference, we used the sentiment classifier trained on the MSA news set as a reference (**Baseline-MSA**). Table 4 shows the results for sentiment classification on tweets. Training a classifier with in-domain data (tweets) enhanced classification effectiveness significantly with a gain of 17.4% (absolute) in accuracy and 17.2% and 16.1% (absolute) improvement in F-measure for positive and negative classes respectively. We saw that MPQA led to slightly better results than ArabSenti.

5.4 Lexicon Expansion

We chose to expand the ArabSenti lexicon using graph reinforcement instead of the MPQA lexicon because the ArabSenti was curated manually. The MPQA lexicon had many translation errors and automatic expansion would have likely magnified the errors. We repeated all our **Baseline-ArabSenti** experiments using the expanded ArabSenti lexicon. We expanded using the English-MSA (**Expanded-ArabSenti-MSA**) and the English-Dialect (**Expanded-ArabSenti-Dialect**) phrase tables.

Table 1 reports on the expansion results for MSA news subjectivity classification. The expanded lexicon marginally lowered classification effectiveness. This is surprising given that the number of tokens that matched the lexicon increased more than five fold compared to the baseline (105k matches for the baseline and 567k and 550k matches for the English-MSA and English-Dialect phrase tables respectively). As shown in Table 2, we observed a similar outcome for the expanded lexicon results, compared to baseline results, for MSA sentiment classification. Though expansion had little effect on classification, we believe that the expanded lexicon can help generalize the lexicon to new out-of-domain data.

Tables 3 and 4 report subjectivity and sentiment classification of Arabic tweets respectively. Lexicon expansion had some positive impact on subjectivity classification with improvements in both accuracy, precision, and recall. Lexicon expansion had a larger effect on sentiment classification for tweets with improvement accuracy, precision, and recall with improvements ranging between 1-3% (absolute). The coverage of the lexicon increased nearly 4-folds compared to the baseline (19k matches for baseline compared to 75k matches with expansion for subjectivity, and 7k matches for baseline compared to 28k matches with expansion for sentiment classification). For both subjectivity and sentiment classification, using the English-MSA phrase table was better than using the English-Dialect phrase table. This is not surprising given the large difference in size between the two phrase tables.

6 Conclusion and Future Work

In this paper we presented a strong baseline system for performing SSA for Arabic news and tweets. In our baseline, we employed stemming and POS tagging, leading to results that surpass state-of-the-art results for MSA news subjectivity classification. We also introduced a new tweet corpus for SSA, which we plan to release publicly. We also employed tweet specific language processing to improve classification. Beyond our baseline, we employed graph reinforcement based on random graph walks to expand the SSA lexicon. The expanded lexicon had much broader coverage than the original lexicon. This led to improvements in both subjectivity and sentiment classification for Arabic tweets.

For future work, we plan to explore other features that may be more discriminative. We would like to

investigate automatic methods to increase the size of SSA training data. This can be achieved by either utilizing bootstrapping methods or applying MT on large English tweets corpora. Another problem that deserves thorough inspection is the identification of polarity modifiers such as negation.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

Muhammad Abdul-Mageed, Mona Diab, and Mohammed 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, 2011.

Muhammad Abdul-Mageed and Mona T. Diab. 2011. Subjectivity and sentiment annotation of modern standard arabic newswire. *ACL HLT 2011*, page 110, 2011.

Muhammad Abdul-Mageed, Sandra Kübler, and Mona Diab. 2012. Samar: A system for subjectivity and sentiment analysis of arabic social media. *WASSA 2012*, page 19, 2012.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 28–36. Association for Computational Linguistics, 2010.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2427–2430. ACM, 2012.

Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In *2nd International Conference on Arabic Language Resources and Tools*, 2009.

Ali El-Kahky, Kareem Darwish, Ahmed Saad Aldein, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed Ammar. 2011. Improved transliteration mining using graph reinforcement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1384–1393. Association for Computational Linguistics, 2011.

Spence Green and John DeNero. 2012. A Class-Based Agreement Model for Generating Accurately Inflected Translations. *ACL* 2012.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, and others. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time *HLT-NAACL-2010*, pages 145–153.

Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, and Ahmed Taei. 2013. Detecting Comments on News Articles in Microblogs *ICWSM*, pages 145–153.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, pages 627–666, 2010.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *ACL-2007*, volume 45, page 976, 2007.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010, 2010.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing 2005*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

Yusuke Yanamoto. 2011. Twitter4j: A java library for the twitter api, 2011.

Hong Yu and Vasileios Hatzivassiloglou 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences In *EMNLP-2003*, pages 129–136. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John

Makhoul, Omar F. Zaidan and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of NAACL*.