SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis

Muhammad Abdul-Mageed, Mona Diab[†]
School of Informatics & Computing and Department of Linguistics
Indiana University, Bloomington, IN
The Department of Computer Science, The George Washington University, Washington, DC[†]
mabdulma@indiana.edu, mtdiab@gwu.edu[†]

Abstract

The computational treatment of subjectivity and sentiment in natural language is usually significantly improved by applying features exploiting lexical resources where entries are tagged with semantic orientation (e.g., *positive, negative* values). In spite of the fair amount of work on Arabic sentiment analysis over the past few years, e.g., (Abbasi et al., 2008; Abdul-Mageed et al., 2014; Abdul-Mageed et al., 2012; Abdul-Mageed and Diab, 2012a; Abdul-Mageed and Diab, 2012b; Abdul-Mageed et al., 2011a; Abdul-Mageed and Diab, 2011), the language remains under-resourced as to these polarity repositories compared to the English language. In this paper, we report efforts to build and present SANA, a large-scale, multi-genre, multi-dialect multi-lingual lexicon for the subjectivity and sentiment analysis of the Arabic language and dialects.

Keywords: Arabic dialects, sentiment analysis, Arabic lexical resource

1. Introduction

We present SANA, a large-scale multi-genre, multi-dialectal multi-lingual lexical resource for subjectivity and sentiment analysis of the Arabic and dialects. Language use varies across genres and SANA caters for that fact by encompassing lexica derived from four main genres: Online newswire, chat turns, Twitter tweets, and YouTube comments. In addition to Modern Standard Arabic (MSA), where most NLP efforts have been focused for the past few years, SANA also covers both Egyptian Dialectal Arabic (EDA) and Levantine Dialectal Arabic (LDA), along with providing English glosses. A significant portion of SANA entries, as will be spelled out below, is also augmented with part of speech (POS) tags, diacritics, gender, number, rationality, and genre class features.

SANA is developed both manually and automatically. For the manual step, we extract and hand-label two extensive word lists from two different genres: 1) *SIFAAT* (Arabic for "adjectives"), which is composed of 3,325 Arabic adjectives extracted from the first four parts of the Penn Arabic Treebank (PATB) (Maamouri et al., 2004; Abdul-Mageed and Diab, 2011), and 2) *HUDA*, a lexicon extracted from an Egyptian Arabic chat data set. The automatic step is articulated using two main methods, i.e., a statistical method based on pointwise mutual information (PMI) (Church and Hanks, 1989), a popular word association measure, and another method based on simple machine translation.

The rest of the paper is organized as follows: In Section 2. we discuss efforts to manually label a number of resources that SANA exploits. Section 3. discusses the automatic methods used to develop SANA. In Section 4., we discuss the different ways of evaluating SANA and its various components. Section 5. is about the scope and attributes of SANA. Section 6. is about related work, and Section 7. is where we conclude and discuss limitations and future work.

2. Manually Labeling Polarity Lexica

We introduce two manually labeled polarity lexica, *SIFAAT* and *HUDA*. While SIFAAT is tagged by two college-educated native speakers of Arabic independently and cases of differences are settled by the two annotators after adjudication, HUDA is labeled by a single college-educated native Egyptian Arabic speaker. In each case, the entries are labeled with one tag from the set {*positive (Pos), negative (Neg), neutral (Obj)*}.

For both SIFAAT and HUDA, annotators were instructed *not* to use their personal experiences when labeling the entries, but rather what they believe would be the orientation of a general use of each term. Annotators were also instructed to assign what they think is the majority orientation of a term, in cases where a term would have both positive and negative use depending on context. The coders were also exposed to a linguistic background related to the concepts of subjectivity and sentiment expression in natural language and the task of annotating data for these phenomena. Although labeling lexica outside context is not necessarily the same as labeling contextual text units like sentences, as is explained in as is explained in e.g., (Abdul-Mageed and Diab, 2011; Abdul-Mageed and Diab, 2012a) (e.g., speech act, good vs. bad news, annotator's background knowledge), we believe that knowledge of the overall background of subjectivity and sentiment annotation in context is still relevant and useful for labeling lexica for prior polarity. Descriptions of both SIFAAT and HUDA are provided below.

2.1. SIFAAT[1]

We introduce *SIFAAT*, a manually created lexicon of 3,325 Arabic adjectives which pertain to the newswire domain and were extracted from the first four parts of the Penn

[1] *SIFAAT* or "صفات" in Arabic script is Arabic for "adjectives."

Arabic Treebank (PATB) (Maamouri et al., 2004; Abdul-Mageed and Diab, 2011). The class distribution in SIFAAT is as follows: 617 Pos, 550 Neg, and 2,158 Obj. Examples of the adjectives labeled as Pos are "بطولي" "heroic," "مرموق" "prestigious," and "مشرق" "splendid." Examples of entries assigned a Neg tag are "مشؤوم" "inauspicious," "خاطف" "kidnapper," and "تحريضي" "provocative." In addition to the polarity labels and part of speech (POS) tags, SIFAAT entries are enriched with English glosses and diacritics.

2.2. HUDA

HUDA is composed of 4,905 entries extracted from an 11 million chat turns corpus collected by us from the Egypt chat room of Yahoo Maktoob!, the Arabic popular chat portal over about 10 months during the year 2008. The entries labeled are the most frequent in the collection, with the most frequent word "هه" "haha" occurring 1,057,670 times and the least frequent "افرح" "be happy" occurring 124 times. The entries were labeled with no prior preprocessing. Since we extracted only the very highly frequent words, we do not envisage this as a big limitation. After all, pre-processing would come at the cost of some losses (especially, given the noisiness of the data). The distribution of the tags in HUDA is as follows: 1,900 Pos, 1,080 Neg, and 1,925 Obj. Examples of the entries labeled with Pos are "نورك" "you're elegant," "حلو" "beautiful," and "حبيبي" "my sweetie." Examples of Neg entries are "معرص" "jerk," "شرموطه" "morally loose woman," and "مالك" "it's none your business."

3. Automatic Augmentation of SANA

3.1. Leveraging English and Multilingual Resources with Machine Translation

Recognizing the need for a wide coverage lexical resource that covers multiple genres of Arabic, we experiment with fast automatic acquisition of an Arabic polarity lexicon using three existing English and multilingual lexica. We introduce each of these lexica in the next subsections. For the first three resources below (i.e., SentiWordNet, Youtube Lexicon, and General Inquirer), we use Google's translation API to acquire the Arabic translation of each entry.

3.1.1. WordNet and SentiWordNet (SWN3)

WordNet (Miller, 1995; Fellbaum, 1998) is a large lexical database of English where nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept. The synsets are interlinked by means of conceptual-semantic and lexical relations. (Baccianella et al., 2010) use a "random-walk" algorithm that exploits WordNet to build *SentiWordNet 3 (SWN3.)*, a lexical database with polarity scores assigned to each synset. SWN3 is composed of more than 117,000 synsets from which we acquire 205,823 unique entries. From these entries, the Google Translation API returns 127,291 Arabic cases (i.e., 61,84%). The rest of entries are either ones for which the API returned the same English entry or an empty string. The Arabic Google-translated SWN3 has 14,543 Pos tokens, 16,285 Neg tokens, and 96,463 Obj cases. Examples of Pos translated entries are "محظوظ" "lucky," "ربحية" "profitableness," and "مسرور" "delighted." Examples of Neg entries in the translated lexicon are "خراب" "ruin," "انقلاب" "coup d'état," and "عار" "shame." Instances of Obj entries are "معبد" "temple," "عداء" "runner," and "قلنسوة" "hood."

3.1.2. Youtube Lexicon (YT)

(Abdul-Mageed et al., 2011b) used Google's Youtube Data API to crawl all comments on 1,000 Youtube videos using the query "obama health care". They refer to the 229,177-comments resulting corpus as *Youtube Health Corpus [YuHC]*. After reducing all repeated letters of frequency > 2 to only 2 (e.g., the word *cooool* is reduced to *cool*), they extracted the top 29,991 words[2] and manually labeled them with semantic orientation tags. Each word was given a label from the set {*Pos, Neg, Obj*}, with a distribution of 3,768 Pos tokens, 6,224 Neg tokens, and 19,999 Obj tokens, which we Google-translated into Arabic. Examples of Pos entries in the YT translated part of SANA include "تغيّر" "change," "مؤمّن عليه" "insured," and "قدامى المحاربين" "veterans;" whereas Neg examples include "اللّعنه" "fuck," "عنصرى" "racist," and "ركود" "recession."

3.1.3. Affect Control Theory Lexicon (ACT)

The Affect Control Theory lexicon (ACT) is composed of 3,448 words and phrases manually rated for polarity by native speakers of English from the U.S. and was developed as part of software accompanying work in social psychology, e.g., (Heise, 2007). ACT is composed of 1,005 verbs (expressing various types of behavior), 1,522 nouns (expressing various types of identities), 540 adjectives (qualifying identities), and 381 nouns of places (describing various place settings). Each entry in ACT is assigned either a positive or a negative score, based on the manual ratings acquired from human beings. The ACT entries were matched against all the SAMA (Buckwalter, 2002; Maamouri et al., 2010; Graff et al., 2009) English glosses and a total of 2,479 matches were found. Since some of the matched English glosses have more than one SAMA Arabic equivalent, the set was expanded by these Arabic accompanying equivalents to 6,965 entries. To illustrate this equivalent-expansion, the English gloss "distress," for instance, gives us access to the Arabic SAMA entries "شجا," "كدر," "شجن," and "ويل." The resulting expanded entries have 3,640 Pos cases and 3,289 Neg cases. The expanded entries also have 36 cases that were assigned both Pos and Neg tags. To keep a fairly dependable quality, we removed these 36 mixed sentiment cases from the final ACT database. The final ACT lexicon thus totals 6,929 entries. Instances of Pos entries in the translated ACT dictionary are "سامر" "entertainer," "يجذب" "attract," and "يشجع" "encourage;" and instances of Neg entries are "صارم" "ruthless," "مستوحش"

[2]Extracted words were of frequency of 3 or more.

"lonesome," and "حسوّد" "envious."

3.1.4. General Inquirer (GI)

The General Inquirer (GI) lexicon (Stone et al., 1966) is part of the GI system, a content analysis program that exploits terms manually classified on a large number of categories. The lexicon contains a total of 11,788 terms, 1,636 of them are labeled as Pos and 2,007 are labeled as Neg, whereas the remaining items, not assigned either Neg or Pos tags, may be considered Obj. The GI lexicon has been widely used in SSA, e.g. (Kamps et al., 2004; Esuli and Sebastiani, 2005; Turney and Littman, 2003). We Googled-translated only the Pos and Neg entries into Arabic. Similar to the procedure followed with ACT, we used a pivoting approach to expand the GI: We exploited the English GI entries to access the several SAMA Arabic equivalents of these English entries. For example, the English gloss "compromise" enables us to access Arabic entries as "تراضي," "تصافي," and "مصالحة." The final GI lexicon totals 4,269 Arabic entries, with 1,859 Pos cases and 2,410 Neg cases. Examples of the Pos entries are "غير مكلف" "inexpensive," "رومانسي" "romantic," and "قابل للتكيّف" "adaptable;" and examples of the Neg entries are "إحراج" "embarrasement," "طريق مسدود" "impasse," and "كبش فداء" "scape goat."

3.1.5. THARWA

The version of THARWA (Diab et al., 2014) we use here (which is yet to be released externally) is a large-scale four way dictionary from Egyptian Arabic to Levantine Arabic, Modern Standard Arabic, and English and is augmented with various elements of linguistic information (e.g., POS tags, gender, number, rationality). THARWA (Diab et al., 2014) was developed both manually and automatically and the version exploited here has 71,000 MSA tokens and 69.000 dialectal entries.

Since the resources we exploit for creating SANA are of two main types with regard to the method of their development (i.e., manual and [semi-]automatic), we wanted to exploit THARWA using each of these two main development types independently. In this way, we can maintain the development distinction as we believe this can be significant for the final quality of SANA. For this reason, we merge with THARWA in two different ways: First, we use the translated SWN3 as the link; second, we use the resources manually labeled for sentiment and then translated into Arabic as the bridge. With each of these two ways, we create a different instance of THARWA. We describe each of these instances below.

THARWA-SWN3. THARWA contains two types of English glosses for each Arabic entry. First, it contains an "English equivalent" gloss as acquired from a number of digitized dictionaries. Second, it contains SAMA glosses. We acquire a dictionary of each of these two types of glosses and merge them together. The dictionary acquired from the 'English equivalent' glosses (henceforth EQ-SAMA) totaled 33,064 entries, and the one comprising SAMA glosses totaled 35,456 entries. Whereas, the merged dictionary resulting from the two is at 48,650 entries. We use this merged dictionary to match against all the entries in the Google-translated SWN3 described above. This process results in a total of 21,436 matching English entries. We are thus able to propagate the sentiment tags from the translated SWN3 to the English entries and hence to the rest of related THARWA fields. We refer to this resulting resource as THARWA-SWN3. THARWA-SWN3 totals 3,189 Pos entries, 3,561 Neg entries, and 14,686 Obj entries.

TAG	ENG	MSA	EGY	LEV
Pos	intelligent	لبيب	ناصح	حربوق
	fragrance	عبير	ريحه	عطر
	contended	راض	مبسوط	مرتضي
Neg	chatterbox	ثرثار	رغاي	حكوجي
	huffy	غاضب	زعلان	معصب
	fault	وزر	غلطة	خطية

Table 1: SANA-THARWA Examples

THARWA-SAGY. We create SAGY, an English gloss dictionary from the merged English entries of the SIFAAT, ACT, GI, and YT resources. SAGY totals 11,858 English entries. We use SAGY to match against the EQ-SAMA dictionary described above to link to THARWA. We are able to find a total of 7,692 matching entries that exist both in THARWA's EQ-SAMA and SAGY. Using this bridging technique, we acquire THARWA-SAGY. THARWA-SAGY has 2,760 Pos, 3,416 Neg, and 1,516 Obj entries.

Table !1 shows examples of Pos and Neg entries from the resulting SANA-THARWA (in this case THARWA-SWN3) across the different languages and dialects.

3.2. Pointwise Mutual Information (PMI)

Pointwise mutual information (PMI) (Church and Hanks, 1989) 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 word "w" and its class "c" (i.e., the Pos vs. the Neg class in our current case) is:

$$PMI(w,c) = log_2 \frac{PMI(w,c)}{P(w)P(c)}$$

For PMI lexicon extraction, we collect two data sets one from the Twitter genre (coming from multiple Arabic dialects) and another from the chat genre (primarily in Egyptian Arabic). We use an approximation where each unit of analysis (i.e., tweets and chat turns) is self-labeled by social media users through employment of emoticons (i.e., emotion icons). This *distance supervision* works such that a unit of analysis with a smiley face ":)" is considered positive and one with a frowny face ":(" is considered negative. While we use this approximation with the Twitter data set and the Pos case for the chat data set, we have found it difficult to acquire enough data points for the Neg case with the chat data. This is the reason because of the extreme noise

in the chat data (which makes it difficult to extract a clean collection with the frowny face) and the meagerness of the use of the frowny face ":(" itself based on our observation of the data. As a solution, we use a seed set of 8 highly profane words that we observed frequently in the data for collecting Neg instances. The set includes words like "قحب" "jerk" and "شرموط" "morally loose." We refer to the Twitter collection of the noisy-labeled data as TAD and that of the Egyptian chat as TED, respectively.

3.2.1. TAD

The TAD collection is comprised of a pool of 971,659 tweets assigned an "ar" tag (for Arabic) by the Twitter API. Observing that this overall collection still includes non-Arabic and short tweets, we wrote a small script to filter out all instances with $< 50\%$ Arabic characters and those whose body text is shorter than two words. In addition, in order to avoid biasing the PMI calculation, we remove all re-tweets. The resulting filtered TAD database comprises 128,760 tweets, with 93,567 positive cases and 35,193 negative cases. For this work, we make no attempt to preprocess (e.g., lemmatize) the collection before we run the PMI code.

We run PMI on TAD with five different word frequency thresholds (5, 10, 15, 20, and 25). For each threshold, tokens with frequency less than the threshold are not considered. With a threshold of 5, PMI-TAD extracts 27,213 Pos cases and 25,152 Neg cases. As we increment the threshold, the number of acquired entries decreases. With the 25 threshold, we retrieve 6,572 Pos cases and 6,157 Neg cases. Examples of Pos entries with top PMI values with the conservative threshold of 25 are "دعوتكم" "inviting you," "والسرور" "and happiness," and "الثقافية" "the cultural;" and instances of Neg entries are "مسكينه" "poor + fem.," "القبور" "the graves," and "انفطر" "[someone's heart] was broken."

3.2.2. TED

As mentioned earlier, we collected an 11 million chat turns corpus from the Egypt room of Yahoo Makttob! From this pool, we extract all instances (n= 687,494) of length > 2 words with at leat one occurrence of the smiley face ":)" and one of 8 profane words as mentioned earlier. The number of positive data points in TED is 456,714 and the number of negative data points is 230,780. We run PMI-TED with the same five thresholds of 5, 10, 15 and 20, and 25 similar to PMI-TAD above. With the 5 word frequency threshold, we acquire a total of 50,389 cases, of which 28,687 are PMI-5[3] Pos and 21,702 are PMI-5 Neg. With the higher 25 threshold, we acquire a total of 18,346 TED-25 polarized tokens, of which 8,468 are Pos and 9,878 are Neg. Examples of Pos entries in TED-25 are "اشتقتلك" "I miss you," "كداب" "liar," and "تعيّيش" "YOU da man!." Examples of Neg entries are "نعلة" "his shoes," "يلكلب" "you bitch!," and "معفنين" "rotten + plural."

[3]Clearly, we refer to the section of TED acquired with the 5 threshold for PMI by simply suffixing the threshold to the name.

4. Evaluation

4.1. Evaluating SIFAAT and HUDA

SIFAAT was manually labeled by two native speakers and hence does not need further manual evaluation. In addition, SIFAAT was successfully used for enhancing Arabic subjectivity and sentiment analysis as reported in (Abdul-Mageed et al., 2011a), with approximately 6% F-measure improvement for subjectivity classification and more than 40% F-measure improvement for sentiment classification. These results show the utility and need for a resource such as SIFAAT for Arabic SSA. Since HUDA was annotated by a single human coder, we randomly evaluated 200 cases of the entries to be able to identify the extent to which the decisions made by the annotator are dependable. We found that our annotator (also a native speaker of Arabic) agrees with the first coder's decisions in 86% of the cases, whereas he disagrees completely (i.e., in cases where he believes the decision should be reversed from Pos to Neg or vice versa) in 4% of the cases. In the remaining 10%, the second annotator also disagreed about Pos or Neg cases that he thinks should be tagged as Obj. This evaluation shows that HUDA is a valuable resource, with a reasonable prior polarity annotation quality.

4.2. Evaluating PMI-TAD and PMI-TED

We manually evaluated a total of 200 randomly sampled cases of the automatically-labeled entries from PMI-TAD-25 (100 Pos and 100 Neg). For the PMI-Pos cases, the human coder assigned a Pos tag to 75%, a Neg tag to 4%, and an Obj tag to the remaining 21%. Whereas, for the PMI-Neg cases, the annotator assigned tags with the distribution of 88% Neg, 2% Pos, and 10% Obj. This human judgment shows that PMI-TAD-25 is at least 75% accurate in assigning polarity tags, which also suggests that thresholds < 25 for PMI extraction of polarity lexica from the data are likely to result in errors $> 25\%$ of the time.

We also hand-labeled a random sample of 200 cases (100 Pos and 100 Neg) of PMI-TED-25. We found that 50% of the PMI-TED-25 Pos cases are human-judged as Pos, 18% as Neg, and 32% as Obj. Regarding the PMI-TED-25 Neg cases, we judged 80% of the cases as Neg, only 2% as Pos, and 18% as Obj. This human evaluation shows that PMI is more dependable as a method for extracting Neg polarity than it is with Pos polarity with regard to the chat genre. While there may be some bias related to the way we chose the data for both the Pos and Neg runs of PMI-TED, we believe that these results are also related to the playfulness of the chat genre where users may like to tease one another, especially give the anonymity of the medium (Herring, 2007; Christopherson, 2007).

4.3. Evaluating ACT, GI, and YT

Although resources that are manually labeled in one language and are automatically translated into another are expected to maintain a high annotation quality, there is no guarantee this will be the case due to potential translation errors. This remains an empirical question, nonetheless, and so we decided to manually evaluate random samples of each of the three resources that were manually labeled in in their original source language (in our case English)

and machine-translated into Arabic (either via Google or through matching against other resources like SAMA as described earlier). For each of the ACT, GI, and YT resources, we manually evaluated 200 instances. In the case of ACT, the human coder agreed with the tag for the translated word 90% of the cases, disagreed for 1% of the cases, and was not sure in 9% of the cases that an entry should be polarized. In these latter 9%, the coder thought these cases should rather be assigned Obj tags.

For the GI lexicon, the human coder agreed 96% of the time with the tags assigned to polarized words (i.e., Pos and Neg tags), and disagreed 4% of the time (including 3% about cases that would reverse the assigned polarity and only 1% about cases where the assigned tag should be Obj, rather than Pos or Neg). With regard to the YT lexicon, the human annotator agreed with the tag assigned to the translated word 88% of the time. In 8% of the cases, the coder disagreed about assigning a polarized tag and preferred to use an Obj tag; whereas in the remaining 4%, the coder decided to shift the assigned polarity. The manual evaluation of the three resources here (i.e., ACT, GI, and YT) shows that translating a manually tagged resource is a very useful measure, given the availability of a good machine-translation service.[4]

4.3.1. YT for English Sentiment Analysis

Although it is more limited in size when compared to a resource like SWN3, YT is manually developed and hence is of a fairly dependable quality. In addition, SWN3 was not developed with social media in mind. As such, SWN3 does not cover some of the modern linguistic variations characteristic of social media communication. To test this claim, we compared YT to SWN3. We found no matches in SWN3 for a total of 13,945 YT entries.The big number is partly due to the fact that the YT lexicon is not processed (e.g., word forms are not lemmatized) and so includes various word forms for each word type (e.g., it has the token "protesters" along with the type "protest"). However, the wider coverage is also due to the absence of certain slang terms (e.g., "bruh," "dammit"), modern named entities (e.g., "Obama," "Mugabe"), and commonly-used Web expressions (e.g., "suck," "lol") in SWN3. YT, however, is far from a comprehensive resource for English sentiment analysis and can best be used as a supporting database along with other resources.

4.4. Evaluating Arabic-SWN3

In order to further evaluate SANA components, we use the English glosses from SIFAAT to evaluate the adjectives of SWN3. We first average the scores assigned to the various senses of each SWN3 adjective. Then, we search SWN3 for each SIFAAT adjective gloss and, if found, we compare the tags assigned to it in each lexicon. Using this procedure we map the SIFAAT adjectives to their counterparts in the SWN3 pool of 29,816 adjectives. The tags assigned agree with a Kappa *(k)* = 0.284. This indicates a fair agreement (Landis and Koch, 1977). In order to gain higher agreement rates, we experiment with two thresholds of scores. Using the thresholds, we retrieve two sub-SWN3 databases, as follows:

- **SentiWN_LEX**: This lexicon is composed of all Pos (N= 1,949) and Neg (N= 2,580) entries with a score > 0.25 plus all Obj entries (N= 5,243). The list thus totals 9,772 entries.
- **SentiWN_Strong_LEX**: This lexicon is composed of all Pos (N= 624) and Neg (N= 1,156) entries with a score > 0.50 as well as all the Obj entries mentioned above (i.e., N= 5,243). The list totals 7,023 entries.

We find that the higher the threshold, the better the agreement we acquire. Thus, with the > 0.25 threshold, we get a higher (although still 'fair') agreement (with a Kappa *(k)* = 0.346). With the higher threshold of > 0.50, a 'moderate' agreement (Landis and Koch, 1977) with a higher Kappa *(k)* = 0.426 is achieved. However, the higher threshold comes at the cost of lower coverage (i.e., 7,023 SWN3 adjectives with the > 0.50 threshold vs. the whole 29,816 SWN3 adjective pool).

4.5. Evaluating SANA-THARWA

We also compare the two instances of SANA-THARWA, THARWA-SWN3 and THARWA-SAGY, introduced above. We match the two dictionaries against each other using the English glosses as keys. We find 6,175 matching entries (28,80% of THARWA-SWN3 and 80,28% of THARWA-SAGY). The sentiment tags assigned to these matching entries agree in 40,62% and disagree for the rest (i.e., 59,48%). The two sets of tag decisions agree with a Kappa *(k)* = 0.277 (a 'fair' agreement). Table 2 shows the tag assignment agreement between the THARWA-SWN3 and THARWA-SAGY. As the Table shows, the distribution of the tags is useful as to the behavior of each of the two instances of THARWA. For example, whereas only 20.10% of the cases in THARWA-SAGY (n=1,238 as shown in row 2 in Table 2) are assigned an Obj class, 55.60% of the tags in THARWA-SWN3 (n=3,452 as shown in column 2 in Table 2) belong to the Obj case. Also, it is clear that each of the Pos and Neg tags is confused with the Obj tag more than it is confused with its polar counterpart. In fact, when the Obj examples are removed from the data points, the agreement increases to a Kappa *(k)* = 0.454. One way of benefiting of the double labeling when using SANA-THARWA can be to prefer THARWA-SAGY to THARWA-SWN3, especially with regard to the Obj tags. Such a preference is motivated by the fact that THARWA-SAGY source entries are manually labeled.

5. Scope and Attributes

As mentioned before, SANA is developed both manually and automatically. The manual part involves hand-labeling two lexica, SIFAAT and HUDA. Both of these

[4] Although we did not evaluate the translation process per se, we observe that the yields acquired from either Google or the matching we performed are of quite good quality.

	Neg	Obj	Pos	Total
Neg	1,008	1310	371	**2,689**
Obj	184	821	233	**1,238**
Pos	247	1321	680	**2,248**
Total	**1,439**	**3,452**	**1,284**	**6,175**

Table 2: THARWA-SWN3 and THARWA-SAGY Agreement

Resource	Pos	Neg	Obj	ALL
ACT	3,640	3,289	NA	**6,929**
PMI-TAD-25	6,572	6,157	NA	**12,729**
PMI-TED-25	8,468	9,878	NA	**18,346**
GI	1,859	2,410	NA	**4,269**
HUDA	1,900	1,080	1,925	**4,905**
SIFAAT	617.00	550.00	2,158	**3,325**
SWN3	14,543	16,285	96,463	**127,291**
THAR-SWN3	3,189	3,561	14,686	21,436
THAR-SAGY	2,760	3,416	1,516	7,692
THAR-ALL	2,417	2,850	11,512	**16,779**
YT	3,768	6,224	19,999	**29,991**
ALL	**43,784**	**48,723**	**132,057**	**224,564**

Table 3: SANA Statistics

lexica were coded by college-educated native speakers of Arabic. For the automatic step, as we mentioned above, we use Google's translation API to render all expressions in the ACT, GI, SWN3. and YT lexica into Arabic. We then expand some of these resources (i.e., ACT, GI, and YT) using SAMA. Our next step was to use the SWN3 Google-translated resource and the expanded SAGY lexica as bridges to propagate the sentiment labels to the matching THARWA database entries. In addition, we use a statistical approach (i.e., PMI) to acquire polarized tokens from both Twitter (TAD) and chat (TED) data. Table 3 shows the number of entries in the various lexica and the total number of entries per category. We do not include the GI-Obj cases, and we only include PMI-extracted words with the conservative threshold of 25 with both TAD and TED. In addition, we do not include the overlap between THARWA-SWN3 and THARWA-SAGY in the overall numbers. As the table shows, SANA comprises a total of 224,564 entries.

In addition to the Arabic form (a lemma in the cases of ACT, GI, SIFAAT, SWN3, and THARWA; and a surface form in the cases of HUDA and YT), SANA is augmented with several other types of features that vary per individual lexicon. These features are part of speech (POS) tag, English gloss, diacritization (Diac), number (Num), Gender (G), and rationality (R), as follows:

RESOURCE	POS	Diac	NUM	G	R
ACT	✓				
GI	✓				
HUDA					
SIFAAT	✓	✓			
SWN3	✓				
THARWA	✓	✓	✓	✓	✓
YT					

Table 4: SANA Augmentation

- **POS and Diacritics.** Certain POS information is useful for SSA. For example, sentences that contain adjectives are more likely to be subjective than objective. With the exception of HUDA, YT, TAD, and TED, SANA components are augmented with POS tags with various degrees of granularity. SANA-THARWA contains parts of speech pertaining *nominals* (e.g., noun, noun_num [number noun], noun_prop, noun_quant, adj) and *verbs* (e.g., verb, pple_act [active participle], pple_pass [passive participle], vbn [verbal noun]). It also includes other tags like adj, modal, interj, pron, prep, etc. ACT lexica only constitute nouns and verbs and so contains tags from the set {adj, verb, noun_person, noun_place}. Arabic-SWN3 and GI also contain tags for major POS. As for diacritics, all SANA components except HUDA, TAD, TED, and SWN3 are enriched with diacritics.

- **Number, Gender, Rationality.** SANA-THARWA also covers includes information pertaining number from the set {singular, plural, dual, mass noun, collective noun}, gender from the set {feminine, masculine and BOTH (for nouns that can be both feminine and masculine)}, and rationality from the set {rational, non-rational and BOTH (for entities that can be considered as both rational and non-rational)}.

- **Genre, Language Variety, and Gloss.** The source genre of each entry is specified at the entry level. Genres from the set {newswire, chat, microblogs, YT, OTHER (for entries coming from general purpose resources like SWN3)}. In addition, for SANA-THARWA, language variety from the set {MSA, DEA, DLA} is provided. For HUDA and TED, although our observation is that most entries are EGY Arabic, an under-specified language variety tag MSA-DEGA is provided for each entry belonging to these components. All machine-translated entries are tagged with an MSA tag, since Google rendered translations exclusively in MSA. With the exception of HUDA, TAD, and TED entries, each SANA entry is accompanied by at least one English gloss. In the case of SANA-THARWA, each entry is accompanied by an English SAMA gloss as well as another 'English equivalent' translation as acquired from various resources and revised by human annotators.

Table 4 provides information pertaining the various SANA features across the individual lexica. Table 5 provides information about the various language varieties SANA carries per entry across the various lexica.

6. Related Work

Learning the semantic orientation of words from texts has received considerable attention in the literature, especially

RESOURCE	Eng. Gloss	MSA	EGY	LEV
ACT	✓	✓		
GI	✓	✓		
HUDA			✓	
SIFAAT	✓	✓		
SWN3	✓	✓		
THARWA	✓	✓	✓	✓
YT	✓	✓		

Table 5: SANA Language and Dialect Coverage

for a language like English. The general idea that has been developed over the years is to expand an initial seed set with a clear prior polarity in a usually unsupervised fashion based on textual and contextual attributes in corpora. The resulting lexicon is then used to enhance subjectivity and sentiment classification at levels beyond that of the word or phrase (e.g., sentence, paragraph, document).

Several approaches have been employed for learning an expression's prior polarity. Some of these approaches depend on expanding the initial seed set using semantic relations housed in lexical databases. For example, (Dave et al., 2003; Kim and Hovy, 2004; Mullen and Collier, 2004) have successfully used WordNet (Miller, 1995; Fellbaum, 1998) to retrieve synonyms and antonyms of expressions with known prior polarity. Another interesting approach exploits the linguistic context where certain types of parts of speech occur as an approximation of learning the polarity values. The classical example is work by (Hatzivassiloglou and McKeown, 1997) who report linguistically-motivated efforts to learn the semantic orientation of adjectives from a corpus. They maintain that *conjunctions* between adjectives are specifically useful, since they impose constraints on the semantic orientation of their arguments. Another observation they make is that conjoined adjectives that have related forms (e.g., "adequate-inadequate") almost always have different semantic orientation. Based on these observations, they use an unsupervised learning algorithm to infer the orientation of adjectives and report an accuracy of 92.37%. One advantage of (Hatzivassiloglou and McKeown, 1997)'s method is that it is unsupervised, and so there is no need to manually annotate data. However, the method assumes the existence of a huge POS-tagged corpus, which is challenging due to the potential unavailability of either huge corpora or high-performance POS taggers, or both, for some languages. Another limitation is that even with the existence of corpora and processing tools, social media data is usually noisy and cause a drop in performance for methods depending on a processing pipeline.

A third important approach introduced by (Turney, 2002) is the unsupervised algorithm for learning the prior polarity of adjectives or adverbs in phrases in a POS-tagged corpus based on Pointwise Mutual Information (PMI) (Church and Hanks, 1989). In this work, (Turney, 2002) calculates the semantic orientation (SO) of a phrase by comparing its similarity to the word "excellent" to its similarity to the word "poor." He reports a classification accuracy of 74.39% across the four domains of his data. (Turney, 2002)'s algorithm is conceptually simple, but is constrained in the sense that it is dependent on results retrieved from a search engine with specific settings (i.e., use of the NEAR operator) that no longer exist.

However, PMI is an attractive method for learning the polarity of words from contexts where the polarity class can be approximated. Especially because PMI employs a frequency threshold, it can be deployed to learn the semantic orientation of expressions from social media data where users self-assign tags or use symbols or words providing clues about the type of polarity of an overall unit of analysis. This is especially the case for Twitter tweets and chat posts with hashtags and/or emoticons functioning as good sentiment clues. This fact has been exploited in the literature by e.g., (Mohammad, 2012).

7. Conclusion

We presented SANA, a large-scale, multi-genre, multi-dialect multi-lingual lexical resource for especially Arabic SSA. We explained the approach we took in developing SANA, over-viewed its various components, and presented an evaluation of it. SANA has the advantage of being genre-nuanced in that it is built exploiting data from several genres like the ATB newswire domain, the Twitter, the YouTube, and the chat genres. The range of dialects SANA covers makes it attractive from a dialect modeling perspective. For these reasons, we hope SANA will be useful for Arabic SSA across the range of Arabic dialects reported here (i.e., MSA, EDA, and LDA). In addition, the English social media component of SANA can be useful for English SSA especially that targeting the health policy domain to which the YT lexicon belongs and, in general, the rest of social media platforms.

In spite its current coverage and scale, SANA is far from complete. For example, although we made attempts to remove duplicates from the SANA entries (e.g., overlap between the THARWA-SWN3 AND THARWA-SAGY components of SANA), we still plan to invest more time on quality control issues like that of deduplication. In addition, we plan to investigate the effect of using pre-proccessed data from the social media domains on the quality of acquired entries with the PMI method. We also plan to expand the dialectal coverage of SANA beyond MSA, DEA, and DLA.

Finally, although some of SANA components have been successfully used for Arabic SSA (e.g., (Abdul-Mageed et al., 2011a) and English subjectivity detection (Abdul-Mageed et al., 2011b), SANA is not yet fully applied to SSA tasks. In our future work, we plan to use SANA for enhancing our SAMAR system (Abdul-Mageed et al., 2014; Abdul-Mageed et al., 2012), *inter alia*.

8. References

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

Abdul-Mageed, M. and Diab, M. (2011). Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annota-*

tion Workshop, pages 110–118, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Abdul-Mageed, M. and Diab, M. (2012b). Toward building a large-scale arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet Conference*, Matsue, Japan, January.

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

Abdul-Mageed, M., Korayem, M., and YoussefAgha, A. (2011b). "yes we can?": Subjectivity annotation and tagging for the health domain. In *Proceedings of RANLP2011*.

Abdul-Mageed, M., Kübler, S., and Diab, M. (2012). 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*, WASSA '12, pages 19–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abdul-Mageed, M., Diab, M., and Kbler, S. (2014). Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech Language*, 28(1):20 – 37.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, page 2010.

Buckwalter, T. (2002). Arabic morphological analyzer version 2.0. Linguistic Data Consortium, 2004.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in internet social interactions:on the internet, nobody knows youre a dog. *Computers in Human Behavior*, 23(6):3038–3056.

Church, K. and Hanks, P. (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.

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

Diab, M., Al-Badrashiny, M., Aminian, M., Attia, M., Dasigi, P., Elfardy, H., Eskander, R., Habash, N., Hawwari, A., Salloum, W., and Yusef, M. (2014). A large scale three-way egyptian arabic dictionary.

Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *ACL-EACL*, pages 174–181.

Heise, D. R. (2007). *Expressive order: Confirming sentiments in social actions*. Springer.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@ internet*, 4(1):1–37.

Kamps, J., Marx, M., Mokken, R., and De Rijke, M. (2004). Using wordnet to measure semantic orientations of adjectives.

Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

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

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. (2010). Ldc standard arabic morphological analyzer (sama) v. 3.1. *LDC Catalog No. LDC2010L01. ISBN*, pages 1–58563.

Miller, G. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mohammad, S. M. (2012). # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP) Conference*, pages 412–418, Barcelona, Spain.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

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