View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Heriot Watt Pure

An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis

Eshrag Refaee and Verena Rieser
Interaction Lab, Heriot-Watt University,
EH144AS Edinburgh, Untied Kingdom.
eaar1@hw.ac.uk, v.t.rieser@hw.ac.uk

Abstract

We present a newly collected data set of 8,868 gold-standard annotated Arabic twitter feeds. The corpus is manually labelled for subjectivity and sentiment analysis (SSA) ($\kappa = 0.816$). In addition, the corpus is annotated with a variety of linguistically motivated feature-sets that have previously shown positive impact on classification performance. The paper highlights issues posed by twitter as a genre, such as a mixture of language varieties and topic-shifts. Our next step is to extend the current corpus, using online semi-supervised learning. A first sub-corpus will be released via the ELRA repository as part of this submission.

Keywords: Subjectivity and Sentiment Analysis, Twitter, Arabic

1. Introduction

Given the recent political unrest in the Middle East (2011), there has been an increasing interest in harvesting information written in Arabic language from live online forums, such as twitter. Subjectivity and sentiment analysis (SSA) aims to determine the attitude of the twitter's user with respect to a topic or the overall contextual polarity of an utterance. Compared to other languages, such as English, research on Arabic text for SSA is sparse. A possible reason for this is the complex morphological, structural and grammatical nature of Arabic (Habash, 2010), in addition to the limitation of annotated resources like labelled corpora (Abdul-Mageed et al., 2011). Our overall aim is to create a continuous flow of reliably annotated Arabic twitter data by using semi-supervised online learning, see e.g. (Refaee and Rieser, 2014a). In this paper, we describe our annotation efforts in creating an initial training and test datasets, which we intend to release to the LREC community as part of this submission.

1.1. Challenges of Arabic SSA in Social Networks

Arabic can be classified with respect to its morphology, syntax and lexical combinations into three different categories: Classical Arabic (CA), Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Users on social networks typically use the latter, i.e. varieties of Arabic such as Egyptian Arabic and Gulf Arabic (Al-Sabbagh and Girju, 2012). Dealing with DA creates additional challenges for researchers working on NLP: Being mainly spoken dialects, they lack standardisation, are written in free-text and show significant variation from MSA (Zaidan and Callison-Burch, 2013). People posting text on social networks tend to use informal writing style and bi/multi-lingual users tend to use a mixture of languages, as in example (1) taken from our corpus.

(1) فَالَنتَاين — *Valentine's day* in English, spelled using the Arabic alphabet.

SSA for twitter is not a trivial task due to the complexity and variability of sentiment indicator(s) that a single tweet can contain. Although they are short, twitter messages can be composed of a significant amount of information in a compressed form (Bifet and Frank, 2010). In addition, tweets may also convey sarcasm, mixed and/or unclear polarity content (see examples in Table 1).In contrast to grammar- or lexicon-based approaches to SSA, machine learning techniques are in general robust to a great extent, however, require annotated corpora. While there is a growing interest within the NLP community to build Arabic corpora by harvesting the web, e.g. (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012a; Zaidan and Callison-Burch, 2013), these resources have not been publicly released yet. We therefore present a newly collected corpus of annotated twitter feeds annotated for subjectivity and sentiment analysis.

2. Arabic Twitter Corpus

2.1. Corpus Collection

We use the Twitter Search Application Programming Interface (API)[1] for corpus collection, which allows harvesting a stream of real-time tweets by querying their content. In order to retrieve tweets which are relevant for SA, we create a set of search queries to increase the chance of obtaining tweets that convey opinions, attitudes or emotions towards the specified entities (as in 2, following (Al-Sabbagh and Girju, 2012). Note that for training a classifier, these query terms are replaced by placeholders.

The extracted data was cleaned in a pre-processing step, e.g. by normalising user-names and digits, and eliminating Latin characters (i.e. URLs, emails). In particular, we harvested two datasets at two different time steps:

Development data: This dataset contains 7,503 multi-dialectal Arabic tweets randomly retrieved over the period from 20th of January to 21st of February 2014.

Test data: This dataset contains a total of 1,365 instances retrieved during the period of 6th to 15th of November 2013 .

In related work (Refaee and Rieser, 2014b) we use a similar data set for training and evaluation. Please note, that

[1] https://dev.twitter.com/

Label	Definition	Example	English translation
polar	Positive or negative emotion, evaluation, attitude	السيَاحه في اليمن، جمَال لَا يصدق	*Tourism in Yemen, unbelievable beauty.*
positive	Clear positive indicator	كم انت عظيم يَا بشَار الاَسد	*How great you are, Bashar Al-Asad.*
negative	Clear negative indicator	حنَا لّلاسف نستخدم ايفون	*Unfortunately, we use the iPhone.*
neutral	• Simple factual statements / news • Questions with no emotions indicated	حَالة وفَاة جديده بَاتش ٧ان٩ بَالصين بكم سعر الآيفون ٥ حَالياً؟	*A new reported death case with H7N9 in China.* *What is the price of the iPhone 5 these days?*
mixed	• Mixed positive and negative indicators	نحن نعشق الدبموقرَاطيه و نكره فوضِيّ الاَخوَان المسلمين التي تريد تدمير حريَاتنَا	*We love democracy, but hate the mess that Muslim Brotherhood is making to destroy our freedom*

Table 1: Sentiment labelling criteria for Arabic Twitter Corpus

Products/brands	iPhone, Chanel
Social Issues	divorce, health, education
Public figures	Obama, Mandilla, Khamenei, Erdogan
Sport	Chelsea , Al-Ahli FC
International Committees	United Nations, league of Arab States
Internet	YouTube, Instagram, Google

Table 2: Examples of query-terms used for collecting the Arabic Twitter Corpus

the data set described here is different, since we were not allowed to release the original tweets. For further details on the release format see Section 3.

2.2. Corpus SSA Annotation

We manually annotate a random subset of 8,868 examples of the collected data for subjectivity, i.e. teasing apart subjective/ polar and objective tweets, and sentiment. Following (Wilson et al., 2009), we define a sentiment as a positive or negative emotion, opinion, or attitude. Each data instance (tweet) is marked with only a single tag that denotes the interpretation that is ultimately conveyed by the complete piece of text, taking into account only the writer's perspectives: *neutral*, *mixed*, *positive* and *negative*, where the latter two are both subsumed under the label *polar*, i.e. subjective, see Table 2. The label *mixed* covers the cases where tweets are composed of positive and negative emotions simultaneously (Liu, 2012).

In cases where annotators are not able to decide on one of the aforementioned labels, they can label tweet with *other/uncertain*, see examples (2) and (3) from our dataset. Tweets labelled with *other/uncertain* by at least one of the annotators were excluded from the dataset. The annotators were asked to assign an additional *skip* label to tweets with redundant or advertising content. Data instances in this category were also excluded from the dataset. In addition, we asked annotators to declare their reason for label selection. We used this information to iteratively refine the annotation scheme on a small subset of tweets.

(2) Undeterminable indicator:

المسَاوَاه في قمع الحريَات الشخصية عدل

Equality in suppressing personal freedom is justice

(3) Sarcasm

احيَانَا فهمنَا الَّامور بطريقه خطَا يكون هو الصح

Sometimes, the wrong understanding of things leads to the right thing.

2.2.1. Agreement Study

For annotation, we recruited two native speakers of Arabic. In order to measure the reliability of the sentiment annotations, we conducted an inter-annotator agreement study on the annotated tweets. We use Cohen's Kappa (Cohen, 1960) which measures the degree of agreement among the assigned labels, correcting for agreement by chance. The overall observed agreement is 91.74% and resulting weighted Kappa reached $\kappa = 0.816$, which indicates reliable annotations (Carletta, 1996).

Table 3 shows some example annotations from the corpus. For instance, tweet # 1 represents an agreement among annotators in labelling tweets with a clear negative polarity, while the next example indicates a lower level of agreement: That is, when the annotators agree on the selected label neutral, but disagree on the reason for the annotation. The third example reflects a significant level of disagreement regarding the neutral and positive categories. Furthermore, sarcastic and heterogeneous tweets have created

ID	Original tweet	English translation	Annotator 1	Annotator 2
1	لنري قوتكم يَا ارهَابية بشَار الاسد لنسحقكم هّهّه و نحن لَا نتشرف بلقيَاكم يَا كلَاب النَاتو	*Let us see your power you the terrorists of Bashar Al-Assad to crush you (laugh) and we do not even want to see you, you NATO's dogs.*	Negative	Negative
2	يوجد ايفون بين كل اربعه هوَاتف ذكيه	*There is an iPhone among each of the 4 smart phones.*	Neutral (facts)	Neutral (no-clear positive evaluation)
3	تعتبر السيَاحه مورد هَام لّلَاقتصَاد البحريني، حيث زَارهَا في ٢٠٠٧ حوَالي ٥ مليون شَايح، و يتوقع أن يزدَاد هذَا العدّد بشكل كبير جدَا	*Tourism is considered an important revenue of Bahrain's economy, as the number of tourists in 2007 has reached 5 M and is expected to increase in a (really) big way.*	Positive	Neutral (news).
4	علمتنَا الثورَات العربيه ان بشَار الاسد عنده حق	*The political revolution (Arab Spring) has taught us that Bashar Al-Assad is right.*	Uncertain (unclear sentiment indicator)	Negative

Table 3: Example annotations from the corpus.

a challenge even to human annotators. Tweet # 4 shows a disagreement among annotators, whether it is a sarcastic view of very complicated and tragic circumstances, or just a negative attitude. In the context of SA, sarcasm is difficult to detect because it uses positive indicators to express negative emotions, e.g. *Oh, what a great day!!* – while meaning the opposite (Liu, 2012). In future work, we will address this issue by using contextual features to detect sarcasm.

2.2.2. Topic Change

Table 4 shows the distribution of the gold standard annotations for the test and development set. For the development corpus, the distribution is clearly skewed towards neutral utterances, whereas in the test corpus we can observe more polar, i.e. subjective utterances. We hypothesise that this difference in distributions is due to the time-changing nature of topics in the twitter stream (Bifet and Frank, 2010). To further investigate this hypothesis, we compare the distribution of the top 5 most frequent content word tokens (excluding function words) in the two data sets, see Table 5, which confirms that topics change over time: Only one of the top 5 most frequent content words occurs in both data sets, e.g. سوريَا (Syria). Some of the words, such as الفَالنتَاين (Valentine) confirm "cyclic effects" of topic change throughout the year observed in social media (Eisenstein, 2013).

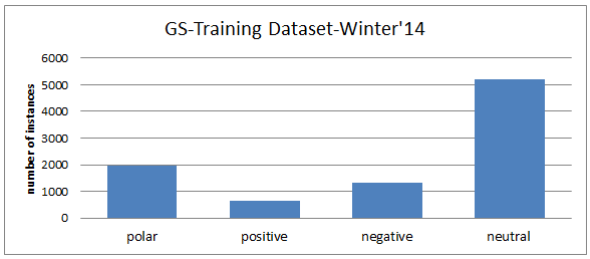
2.3. Automatic Feature Extraction

We annotate the corpus with a rich set of linguistically motivated features, see Table 6, where a subset has been showing an increase in the performance of sentiment analysis on MSA news-wire texts (Abdul-Mageed et al., 2011). We employ morphological features, simple syntactic features, such as n-grams, as well as semantic features. Table 7 provides example annotations for the word token يَحتَرِم (*respect*).

Syntactic Features/ Word Tokens: We experiment with lexical representations of 1st, 2nd and 3rd order of word-based n-grams.

Morphological Features : Considering the morphologically rich nature of Arabic, we annotate the following features: aspect, gender, mood (e.g. indicative), number, person, and voice (e.g. active). We utilise a state-of-art automatic morphological analyser for Arabic text to obtain these features. In particular, we incorporate the current version of MADA+TOKAN (v 3.2) (Habash and Rambow, 2005; Nizar Habash and Roth, 2009) which performs tokenization, diacrization, morphological disambiguation, Part-of-Speech (POS) tagging, stemming and lemmatization for Arabic. It is important to note that MADA is developed for Modern Standard Arabic (MSA) only. Tweets, in contrast, contain dialectal and/or misspelled words where the analyser is incapable of generating morphological interpretations. We therefore include a feature `has_morphological_analysis`. That is, the morphological features for DA words can be expected to be noisy, however, still useful to improve performance for classification. In related work, we evaluate the performance of these (noisy) features by measuring their performance on SSA classification (Refaee and Rieser, 2014b).

Semantic Features: This feature set includes a number of binary features that check the presence of sentiment-bearing words of a polarity lexicon in each given tweet. To obtain this set of features, we exploit an existing manually annotated subjectivity lexicon, namely ArabSenti (Abdul-Mageed et al., 2011). In addition, we make use of a publicly available English subjectivity lexicon, MPQA (Wilson et al., 2009), which we automatically translate using



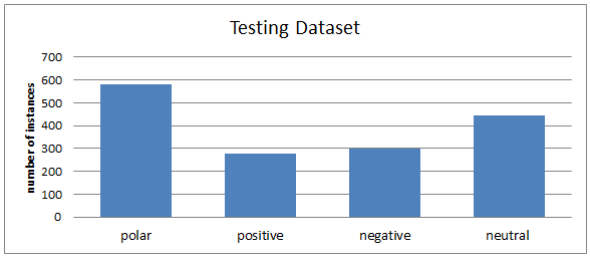


Table 4: Sentiment label distribution of the gold-standard manually annotated training data-sets.

	Development Set			Test Set		
ID	**Arabic**	**English**	**Frequency**	**Arabic**	**English**	**Frequency**
1	سيَاسه	Politics	1652	مصر	Egypt	756
2	سوريَا	Syria	1648	سوريَا	Syria	215
3	بشَار	Bashar	925	جميل	Beautiful	168
4	السيَاحه	Tourism	636	الَاسد	Al-Asad /lion	138
5	الفَالنتَاين	Valentine	426	الّله	Allah/God	111

Table 5: The 5 most frequently used word tokens in the two data sets.

Google Translate, following a similar technique to (Mourad and Darwish, 2013). The translated lexicon is manually corrected by removing translations with neutral or no clear sentiment indicator. For instance, *the day of judgement* is assigned with a negative label while its Arabic translation is neutral, considering the context-independent polarity. This results in 2,627 translated instances after correction. We then construct a third dialectal lexicon of 484 words that we extracted from an independent Twitter development set and manually annotated for sentiment. All lexicons were merged into a combined lexicon of 4,422 annotated sentiment words and phrases (duplicates removed). The dialectal lexicon will be released as part of this submission.

Stylistic Features: This feature-set includes two binary features that check the presence of positive/negative emoticons.

Social Signal Features: This feature-set includes binary features that check the presence of a set of social signals like: consent, dazzle, laugh, regret, sigh.

3. Release Format

The Arabic twitter train and test sets will be released as part of this submission via the ELRA data repository. The released data will contain manual SSA annotations (see Section 2.2.) as well as automatically extracted features (see Section 2.3.), saved in Comma Separated (CSV) and Attribute-Relation File Format (ARFF) file formats. Due to twitter privacy restrictions we replaced the original tweet with its ID. After excluding tweets for which at least one annotator was uncertain about the sentiment label, the total number of instances in the training dataset is 7,503 tweets. The corpus has 96,493 word frequencies and 26,724 unique word tokens. Note that this subset is only the first of a number of planned releases: The overall aim of this work is to create a continuous flow of annotated Arabic twitter data by using semi-supervised online learning, see (Refaee and Rieser, 2014a).

4. Related Work

There is a growing interest within the NLP community to build Arabic corpora by harvesting the web. However, none of these resources are publicly released yet. The corpus described in this paper will be the first Arabic SSA corpus, which is publicy released via the ELRA repository.

The YADAC corpus (Al-Sabbagh and Girju, 2012), for example, is a multi-genre dialectal Arabic corpus, which includes web data from micro-blogs (i.e. twitter), blogs/ forums and online market services. Among their current efforts is the development of reliable POS tagging and phrase chunking tools for DA, based on this corpus.

Work by (Abdul-Mageed and Diab, 2012b) presents a multi-genre corpus of Modern Standard Arabic (MSA) labeled for subjectivity and sentiment analysis.

Furthermore, (Diab et al., 2010) describe new resources and processing tools for dialectal Arabic Blogs. They present a morphological analyser/generator, MAGEAD, which can handle both MSA and DA, but requires some further preprocessing steps, including manual lemmatisation of DA words. In future work, we will investigate using those newly developed tools (once they are released) to replace the features we currently get from MADA, which was developed for MSA only.

5. Conclusions and Future Work

We present a gold-standard annotated corpus to support sensitivity and sentiment analysis (SSA) of Arabic twitter feeds, which is the first publicly released corpus for this task.

We collect a corpus of 8,868 tweets which are manually annotated by two annotators. Our annotations indicate reliable inter-annotator agreement ($\kappa = 0.816$). The corpus comprises a development set (7,503 tweets) and a test set (1,365 tweets), which are harvested at different time slots. An analysis of the top 5 most frequent words shows a

Type	Feature-sets
Morphological	Diacritic, Aspect, Gender, Mood, Person, Part_of_speech, State, Voice, has_morphological_analysis .
Syntactic	n-grams of words and POS, lexemes, including Bag_of_Words (BOW), Bag_of_lexemes.
Semantic	Has_positive_lexicon, Has_negative_lexicon, Has_neutral_lexicon, Has_negator
Stylistic	Has_positive_emoticon, Has_negative_emoticon.
Social Signals	Has_consent, Has_dazzle, Has_laugh, Has_regret, Has_sigh

Table 6: Overview of annotated feature-sets

Type	Feature	Values	Example يَحتَرِم (*respect*)
Morphological	Gender	Masculine (m), feminine (f), not-applicable (na)	`Gen: m`
	Mood	Indicative (i), jussive (j), subjective (s), not-available (na) undefined (u),	`mod: i`
	Number	Singular (s), plural(p), dual(d), undefined (na)	`num: s`
	Person	1st, 2nd, 3rd, not applicable	`per: 3`
	State	Indefinite (i), definite (d), construct (c), not-applicable (na), undefined (u)	`stt: na`
	Voice	Active (a), passive (p),not-applicable (na), undefined (u)	`vox: a`

Table 7: Example of annotated features

change in topic over time, confirming previously observed "cyclic effects" in social media (Eisenstein, 2013).

In related work, we use a similar gold standard corpus (which we are unable to release due to privacy restrictions) in order to learn an automatic classifier for SSA (Refaee and Rieser, 2014b). We find that models trained on the training set do not generalise well to unseen instances in the test set, which is due to the unseen word-based features, i.e. topic shift. We therefore turn to semi-supervised techniques, in particular distant supervision using emoticons as noisy labels (Refaee and Rieser, 2014a). While this approach performs well for subjectivity analysis, we observe significant drop in performance for sentiment analysis. An error analysis shows that this is due to the noise introduced by the emoticon-based labels. In particular, we find that for some cases the direction of facing of emoticons is unclear due to the right-to-left direction of the Arabic alphabet. In future work, we will investigate lexicon-based approaches to distant supervision, following (Zhang et al., 2011), which uses sentiment words from the subjectivity lexicon as automatic labels.

Acknowledgements

The first author would like to thank the Saudi Arabian government for supporting her with a PhD scholarship.

6. References

Abdul-Mageed, M. and Diab, M. (2012a). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Abdul-Mageed, M. and Diab, M. (2012b). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Abdul-Mageed, M., Diab, M. T., and Korayem, M. (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: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal arabic corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Bifet, A. and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. In *Proceedings of the LREC Workshop on Semitic Language Processing)*. European Language Resources Association (ELRA).

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Habash, N. (2010). *Introduction to Arabic Natural Lan-*

guage Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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

Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. *WASSA 2013*, page 55.

Nizar Habash, O. R. and Roth, R. (2009). MADA+TOKAN: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Choukri, K. and Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

Refaee, E. and Rieser, V. (2014a). Can we read emotions from a smiley face? Emoticon-based distant supervision for subjectivity and sentiment analysis of arabic twitter feeds. In *5th International Workshop on Emotion, Social Signal, Sentiment & Linked Open Data*.

Refaee, E. and Rieser, V. (2014b). Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *9th International Conference on Language Resources and Evaluation (LREC'14)*.

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

Zaidan, O. F. and Callison-Burch, C. (2013). Arabic dialect identification. *Computational Linguistics*.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2011). Combining lexiconbased and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89.